THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

PROGRAM IN LINGUISTICS

A THEORETICAL AND COMPUTATIONAL ANALYSIS OF SARCASM

JOSEPHINE SODDANO
SPRING 2021

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Mathematics and Linguistics
with honors in Linguistics

Reviewed and approved* by the following:

Deborah Morton
Assistant Teaching Professor of Linguistics
Thesis Supervisor

John Lipski
Edwin Erle Sparks Professor of Spanish and Linguistics
Honors Adviser

*Signatures are on file in the Schreyer Honors College.

Abstract

Sarcasm is perhaps one of the most complex and confusing features of human language. In this thesis, I aim to present a formal theory of sarcasm through the lens of the Rational Speech Acts (RSA) framework, which provides a probabilistic platform for modelling pragmatic and social reasoning in conversation. Many past formal models account for classical instances of sarcasm but fail to account for non-classical instances. This thesis describes one particular model that is capable of handling both classical and non-classical instances of sarcasm. An experimental validation showed that predictions made by this particular model closely correlated to human judgement predictions about sarcasm.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I must first thank my thesis advisor, Deborah Morton for her help and support along the way. Taking introduction to linguistics with Dr. Morton was one of the best decisions of my life, and I would not have discovered this wonderful field without her.

I would additionally like to thank Becky Passonneau for introducing me to natural language processing research at Penn State. I am very grateful for her mentorship.

Thank you to Chris Potts for all the help with this project and for teaching me a whole bunch of stuff about pragmatics, Mike Frank for all of the helpful advice about graduate school, Reuben for helping me understand sarcasm and being incredibly patient with all of my questions, Sebastian for helping me with my Mechanical Turk questions, and Erica for teaching me how to teach a computer to be nice. Finally, thank you to the Stanford CSLI for giving me my first taste of computational linguistics research.

I would also like to thank my lab mates: Hanson, Ben, and Andrew for answering all of my coding questions, Junior for recovering that one Jupyter Notebook for me, and Isabel for our banters about coding. Thanks to Sam and Ross for the adventures and meditation sessions. Finally, thank you to all of my family and friends that have supported me along the way.

Chapter 1

Introduction

1.1 Overview

Language is baf ing. When someone says ”It's raining cats and dogs outside”, they do not literally mean that cats and dogs are falling from the sky. When someone says: ”My boss is a shark”, they do not literally mean that their boss is a vicious sea creature. Oftentimes, we use very creative expressions to communicate our thoughts. But why? Instead of saying: ”It's raining cats and dogs outside”, why couldn't one just say: ”It's raining really hard outside”. Instead of saying ”My boss is a shark”, why couldn't one just say: ”My boss is a deceitful man”? Every non-literal expression has a literal equivalent, so why is non-literal language even necessary?

Humans yearn to be creative. We want to express ourselves using a full range of emotions. Creativity stimulates and encourages us to connect and interact with our surroundings. A world without non-literal language would be dull. We would still be able to communicate information and accomplish the same goals, but communication would be very different. We value our individuality and need an outlet to express that. We strive to entertain and wish to be entertained in return [1]. We use language not only to express information but also to communicate our feelings and emotions. Language allows us to connect with others.

Formalizing language has long posed a daunting task to researchers. Researchers have typically avoided the ”annoying” parts of language in formal and quantitative approaches. These ”annoying” parts of language tend to be variable, ambiguous, and dif cult to model accurately due to external contextual factors. In 1950, linguist Martin Joos wrote:

> ”We can allow other people to speak artistically, imprecisely, about language. But as linguists we lay upon ourselves the condition that we must speak precisely about language or not at all. ...We must make our ”linguistics” a kind of mathematics, within which inconsistency is by de nition impossible” [2]

Joos suggests limiting the eld of linguistics to instances of language that can be precisely de ned. In 1971, philosopher Yehoshua Bar-Hillel coined pragmatics– the study of language meaning in social contexts– as a the ”wastebasket” of linguistics for the reasons that Joos highlights above [3]. As a result, formalizing pragmatics has posed a great challenge to the linguistics community [4].

The Rational Speech Acts Framework (RSA), rst proposed by Mike Frank and Noah Goodman in 2012, offers a solution to formalizing complex pragmatic phenomena in conversation [5]. In this thesis, I aim to present (1) an overview of the RSA framework (2) show how sarcasm can be implemented into the RSA framework (3) show how RSA can be computationally modelled, and (4) present an experimental validation of an RSA model of sarcasm.

In chapter 2 of this thesis, I offer a linguistic analysis of sarcasm and provide an overview of previous attempts to formally model sarcasm. Chapter 3 offers an overview of the Rational Speech Acts Framework. Chapter 4 will discuss previous work by Kao et al. (2015) and Cohn-Gordon and Bergen (2019) on implementing sarcasm into the RSA framework [6] [7]. Chapters 5 and 6 provide an experimental validation of the RSA model of sarcasm proposed by Cohn-Gordon and Bergen (2019). Chapter 5 describes an experiment that was launched to collect human judgement data on sarcasm and another experiment to collect data to train the sarcasm RSA model. Chapter 6 shows (1) how the RSA can be modelled computationally through the probabilistic programming language, WebPPL and (2) a comparison of human judgement of sarcasm versus computational predictions. Chapter 7 offers suggestions for future work and implications.

1.2 Meaning is Use

In Section 43 of *Philosophical Investigations*, influential 20th century philosopher Ludwig Wittgenstein writes the following:

> "For a large class of cases—though not for all—in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language."[8]

In perhaps the most important quote from *Philosophical Investigations*, Wittgenstein expresses the idea that it isn't words that give rise to meaning but rather the way in which the words are said and the context in which they are uttered that gives them their meaning. Wittgenstein additionally proposes the notion of the language game, which he defines as "consisting of language and the actions into which it is woven" [8]. The key idea is that the meaning of language can vary depending on the context in which it is uttered. Wittgenstein offers the following example to illustrate this idea: "If a lion could talk, we should not be able to understand him" [8]. We should not be able to understand lions because they engage in very different language games than us since non-language related contextual factors inhibit humans from understanding lion speech. Additionally, the same lexical item can communicate many different meanings. Consider the following four scenarios:

1. You are in the desert and see water for the first time in two days
2. A king tells his servant to fetch him a glass of water
3. You are at a restaurant and the waiter asks what you would like to drink
4. Your professor asks you what the chemical formula H_2O stands for

In the above four scenarios, each speaker utters "Water!". However, this utterance takes on a different meaning in each scenario. (1) is an exclamation, (2) is a command, (3) is a request, and (4) is an answer to a question. "Water!" might also be a code word in a certain situation by a group that has adopted another meaning for the word water. Depending on the situation in which the word is uttered, it takes on a completely different meaning.

Wittgenstein also introduces the famous thought experiment of a beetle in a box:

> Suppose everyone had a box with something in it: we call it a "beetle". No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at his beetle. [8]

Imagine a situation in which we have to talk about our beetles. There is no way in which we can peer into another person's box, so our understanding of what is in someone else's box can come solely from the person to whom the box belongs to. While it is assumed that everyone has the same beetle, it could also be the case that everyone has different beetles or even nothing at all. This analogy corresponds to the idea that we all only have access to our own individual and private thoughts. There is no way to understand what is going on in someone else's mind. Though we cannot transfer our own individual thoughts to other people, we can attempt to communicate to others the beetle that we have in mind to the best of our abilities. Now, consider the following two scenarios:

1. The skies are clear and there is plenty of sunshine

2. It is hailing, sleeting, and freezing cold

Imagine that in these two scenarios, your friend says: ”What beautiful weather”! The meaning of this utterance is different for each scenario. In the rst scenario, your friend is most likely communicating her approval of the weather. In the second scenario, your friend is most likely communicating her disapproval of the weather. What accounts for this difference in meaning? Scenarios 1 and 2 represent two different language games where the speaker has different communicative goals. We can infer that the utterance in scenario 1 is meant to be taken at face value while the utterance in scenario 2 is meant to be taken sarcastically. Though it is impossible to know for certain what the speaker is truly meaning to communicate (we cannot peer into the speaker's box to look at her beetle), contextual cues can provide insight into one's communicative goals.

Why would a speaker be sarcastic in one situation but not another? There must be some motivating factor for using sarcasm over literal language. What does sarcasm communicate that literal language does not? These issues will be addressed in the following chapter.

Chapter 2

A Linguistic Analysis of Irony

2.1 Introduction

Sarcasm is a complex phenomenon—because of its complexity, it is often misunderstood. For example, consider a situation where a man is at a bar and wants to order a drink. The man and the bartender share the following conversation:

> Man: Could I have a drink?
>
> Bartender: No.

This scenario is rather ambiguous. Is the bartender serious or is he not? In order to better judge this situation, we need more information about the relationship that these two gentlemen have. Consider three different scenarios that describe the relationship that these two men share and how their relationship in uences how the bartender's utterance is interpreted:

1. In the case that this man is a frequent of the bar and the gentlemen share a fond relationship, the bartender's utterance is more likely to be interpreted as sarcastic.

2. Perhaps the man at the bar is a travelling businessman who has never been to this bar before. The two gentlemen do not share any previous encounters. Then, it is likely that the utterance is likely to be interpreted as not sarcastic.

3. Suppose the man visits the bar from time to time but de nitely would not be considered a regular. The bartender is acquainted with the man, but their relationship doesn't go far beyond that. Then, it is ambiguous whether this utterance was meant to be interpreted as sarcastic.

From the three interpretations listed above, we can see that it is dif cult to discern the bartender's intention behind his utterance without knowing more information. This is part of the reason why sarcasm is so dif cult to understand– because its interpretation is so context-driven.

One thing to note before delving further into this section is that sarcasm and irony will be used interchangeably at times as sarcasm is considered to be a subset of irony [9]. All theories pertaining to irony also apply to sarcasm.

2.2 The Classical View

The Classical View of Irony was proposed by the Roman educator Quintilianus nearly 2,000 years ago. It states the following:

> Irony is speech in which we understand something which is the opposite of what is actually said [10]

Consider for instance that an utterance is intended to communicate some proposition P. Under the Classical View of Irony, a sarcastic utterance is intended to communicate a meaning that is the opposite of P, denoted opposite(P). To explain this in more concrete terms, consider the previous scenario describing the interaction between a bartender and a man who wants a drink. Suppose the man asks the bartender if he can have a drink. Upon hearing this utterance, the bartender decides

that he would like to respond to the man sarcastically. The bartender intends to communicate P: Yes, you can have a drink. However, the bartender utters opposite to P: No. The bartender utters this in hopes that he will effectively be able to use irony to communicate his intended P.

Though the classical view can explain many accounts of irony, a flaw with the classical view is that it does not account for all possible scenarios. The following sections of this chapter will discuss shortcomings in the classical view and alternative theories to formalizing sarcastic language.

2.3 Shortcomings of the Classical View

As mentioned previously, the classical view can explain many accounts of irony, but not all accounts of irony.

(1) Interrogatives – Consider the following scenario where Bill and Julie have just finished watching a movie. They both agree that it was a really terrible movie. Bill utters: "Wasn't that movie amazing?". Bill's intention here was most likely to communicate that the movie was in fact not amazing. However, this interpretation fails under the classical view as the opposite of an interrogative is ill-defined.

(2) Ill-Defined Opposite – Consider another scenario where Bill is discussing potato salad that he had just tried but did not enjoy. Bill utters: "Those raisins made that potato salad fantastic!". Bill intends to communicate here that raisins ruined the potato salad that he just ate. The Classical View fails to capture this intended interpretation because it isn't the case that the meaning of the utterance is reversed. Rather, the irony is linked to a presupposition that the potato salad was bad.

(3) Ironic Understatements – Consider the following meme:

Figure 2.1: "This is Fine" Dog

The interpretation of the dog's utterance also fails under the classical view of irony. The dog in the image is under emotional distress yet produces an utterance that does not match the intensity of the emotion that he is feeling; he is understating his emotion in response to his given situation.

Note also that the classical view does not account for why we even use irony in the first place. For example, in the scenario where Bill and Julie saw a terrible movie, why couldn't Bill just say that the movie was terrible? The classical view does not offer an explanation of why Bill couldn't just communicate his thoughts directly.

2.4 A Gricean Explanation of Irony

Many modern philosophers have attempted to improve and expand upon the Classical model. Perhaps one of the best known cases comes from 20th century British philosopher, Paul Grice. In his 1975 publication *Logic and Conversation*, Grice introduces the cooperative principle, which describes how people can achieve rational, clear, and effective communication in common social situations. The cooperative principle is comprised of four maxims, each of which explain the connection between what is said in an utterance and what is meant by an utterance. These four maxims are: quantity, manner, relevance, and quality.

The four maxims are summarized as follows:

- Quality- don't lie or say things that aren't supported by adequate evidence
- Quantity- strive to be as informative as possible, but do not share too much information
- Manner- be clear, brief, and orderly
- Relevance- be relevant

Grice also introduces the notion of conversational implicatures. Implicatures refers to things that are implied by a speaker in the production of an utterance (that aren't literally communicated in her speech). In his discussion of conversational implicatures, Grice notes that irony is a type of conversational implicature.

In this section, I will focus primarily on discussing the maxim of quality as it relates speci cally to irony. During a conversation between two interlocutors, both interlocutors generally assume that the other person is telling the truth (or what the speaker deems to be considered the truth). Consider the situation where a mother asks her son if he has nished his homework yet. Knowing very well that he has been playing video games the entire evening and hasn't even taken his homework out from his backpack, the son responds with “Yes, mom”. This violates the maxim of quality.

Consider an alternative scenario where it is hailing and sleeting outside. You sarcastically utter the following: ”The weather is absolutely fantastic today”. The speaker in this instance is not being truthful as the weather outside is actually terrible. Does this violate the maxim of quality? No. Though the speaker is being untruthful, she is also giving enough context to convey that she is intending to be sarcastic. In other words, the speaker is not deliberately trying to mislead her conversational partner, but she is deliberately not observing Grice's maxims. This is known as outing a maxim. The outing of a maxim is a type of conversational implicature. When a speaker outs the maxim of quality to express irony, the speaker assumes that she is providing enough context to effectively communicate irony. Note that the difference in violating and outing a maxim is that a speaker who outs a maxim disobeys a maxim without trying to intentionally deceive the listener. On the other hand, one who violates a maxim has the intention to deceive.

However, the Gricean account of irony fails for similar reasons to why the classical view fails. His theory does not explain the motivations of using irony in certain situations over others. In fact, Grice acknowledges the shortcomings of his model himself. It is still unclear why a speaker would out a maxim when she could express the same idea using literal language.

2.5 Echoic and Pretense Theory

The echoic and pretense theories of irony completely reject the Classical and Gricean accounts of irony, that an ironic utterance communicates the opposite of the meaning intended by the speaker. Instead, the echoic and pretense theories propose that ironic utterances are produced in order to draw a distinction between a hypothetical world that the speaker is trying to communicate and the actual state of the world. The echoic and pretense theories of irony are sometimes seen as indistinguishable and many hybrid models incorporate elements from both accounts. However, I will discuss these theories separately in this section.

(1) Echoic Theory: The echoic theory states that irony is used "to dissociate the speaker from an attributed thought or utterance which she wants to suggest is more or less obviously false, irrelevant or under-informative" [11]. Consider the following conversation:

Bill: I just met the Pope!

Julie: You met the Pope? Wow, me too!

Julie echoes Bill's previous utterance not to remind him of what he said but rather to show that she considered his utterance and wanted to convey her attitude toward his utterance. She displays mockery toward Bill's utterance to show that she nds it false and absurd.

(2) Pretense Theory Pretense theory offers an alternative to the echoic theory of irony. In the echoic account, a speaker communicates irony by taking on a pretend perspective with the expectation that the listener will "see through the pretense and recognize the mocking or critical attitude behind it". Consider the same conversation again:

Bill: I just met the Pope!

Julie: You met the Pope? Wow, me too!

Julie is pretending to adopt the perspective that she believes that Bill met the Pope in hopes that Bill will recognize that she is actually mocking his utterance.

Despite being improvements to the Gricean account of irony, there are downfalls to both of these models. One criticism of the echoic theory is that the utterance that is echoed may have not been explicitly mentioned previously. A criticism of pretense theory is that it is assumed that the listener will adopt the speaker's pretend perspective, but this is not guaranteed

2.6 Irony as Countersignaling

In an attempt to improve upon all of the aforementioned theories of irony, Cohn-Gordon and Bergen (2019) [7] propose that irony is a form of linguistic countersignaling, where conversational agents engage in reasoning about pretense in order to communicate about the common ground. Common ground refers to the set of knowledge that is shared between conversational agents. The strength of this model is that it offers an explanation of why people are motivated to use irony in conversation. A sketch of the countersignaling model given by Cohn-Gordon and Bergen (2019) is given as follows:

When a listener hears an utterance u that communicates an unlikely state of the world w, she infers that the speaker said u in order to pretend about w being real. The speaker utters u assuming that the actual state of the world w is already in the common ground. The speaker reasons that it is more desirable to produce an utterance that communicates w than an utterance that communicates w' for reasons that will be discussed in chapter 4.4. The speaker chooses to use irony in order to communicate what she believes the common ground to be between her and the listener. The main idea is that speakers produce ironic utterances in order to communicate about the common ground.

To describe a concrete example, consider the scenario described earlier in this chapter:

Man: Could I have a drink?

Bartender: No

If the bartender's goal was to be sarcastic, then he must have assumed that it was already in the common ground that the man could have a drink.

The countersignaling theory of irony proposed by Cohn-Gordon and Bergen (2019) will be the motivating theory for the remainder of this thesis. Chapter 4 will discuss how the countersignaling model can be implemented into the Rational Speech Acts Framework. Additionally, Chapters 5 and 6 will offer an experimental validation of this particular model.

Chapter 3

The Rational Speech Acts Framework (RSA)

3.1 Introduction to the Rational Speech Acts (RSA) Framework

The Rational Speech Acts Framework (RSA) was first introduced by Noah Goodman and Michael Frank in 2012 [5]. The RSA framework describes a game-theoretic probabilistic approach to pragmatic and social reasoning. The framework, which formalizes Grice's principles that govern clear and effective communication, models how people extract pragmatic meanings of utterances during conversations through recursive social reasoning [1]. Just as classical logic serves as a means to model semantics, the RSA framework serves as a means to model pragmatics. The foundation of RSA is based on the fact that probability can be used to represent uncertainty in an agent's (in this case, either a speaker or a listener) internal representation of the world within defined hypotheses spaces. These models primarily build upon Bayesian models of social reasoning [12].

3.1.1 Brief Overview of Bayesian Inference

Bayesian inference is a method of statistical inference about the state of the world or some event given relevant prior knowledge. A Bayesian model assumes that an agent updates her information state B by conditioning on A, shown as follows:

$$P(b|a) = \frac{P(a|b) \cdot P(b)}{P(a)} = \frac{P(b \cap a)}{P(b)} \tag{3.1}$$

where:

- $a \in A$ and $b \in B$
- for $\forall a \in A; P(A) \in [0, 1]$
- for $\forall b \in B; P(B) \in [0, 1]$
- $P(A) = 1$ and $P(B) = 1$
- $P(A)$ represents the likelihood of A being true and is usually referred to as the prior
- $P(B)$ represents the likelihood of B being true and is usually referred to as the marginalization
- $P(A|B)$ is the probability of A occurring given that B is true. $P(A|B)$ is referred to as the posterior probability distribution.
- $P(B|A)$ is the probability of B occurring given that A is true

Note that $P(A)$ and $P(B)$ represent independent probabilities of A and B such that $P(A \cup B) = P(A) + P(B)$. In the case that $B = \{b_1, b_2, ...\}$ is a partition of B (every element of B is included in exactly one subset), then

$$P(b_i|a) = \frac{P(a|b_i) \cdot P(b_i)}{\sum_{j=1}^{|B|} P(a|b_j) \cdot P(b_j)} \tag{3.2}$$

where every $b_i \in B$ represents the set of hypotheses spaces used to infer how a was generated. Equation 3.2 is often expressed more simply as:

$$P(b_i|a) \propto P(a|b_i)\ P(b_i) \tag{3.3}$$

Equation 3.3 states that the probability that b_i is true given the observation of a is proportional to (1) the probability of observing a given a is true multiplied by (2) the probability that was assigned to b_i being true prior to observing a. The denominator in Equation 3.3 is constant regardless of b_i, so it is only necessary to ensure that the resulting distribution sums to 1.

The RSA model captures the recursive social reasoning that both a listener and a speaker perform during a conversation. In this model, a listener reasons about a speaker. A speaker, in turn, reasons about the listener. The most basic vanilla RSA model consists of three agents: (1) a naive literal listener L_0, (2) a speaker S_1, and (3) a (sophisticated) pragmatic listener L_1. Figure 3.1 illustrates recursive reasoning in the RSA framework.

Figure 3.1: Recursive Reasoning in RSA

The model starts with a S_1 speaker who reasons about a literal L_0 speaker in order to choose the best utterance u in which to communicate some state of the world w to L_0. Note that L_0 will take any utterance heard by S_1 at face value. Even if S_1 told L_0 a complete lie, L_0 would update her knowledge of the state of the world based solely on the semantics of the utterance that she heard from S_1. L_1 is a more sophisticated listener than L_0 that is able to reason about why S_1 chose to utter u. L_1 reasons about what state of the world led S_1 to choose u. Lastly, the pragmatically-aware sophisticated S_2 speaker reasons about L_1 in order to produce an utterance u.

Figure 3.2 shows the recursive reasoning within the vanilla (most basic) RSA framework. A pragmatic S_2 speaker reasons about a pragmatic L_1 listener, who reasons about the S_1 speaker, who is in turn reasoning about the L_0 to produce an utterance.

Figure 3.2: Recursive Reasoning in the Vanilla RSA Model

3.2 The ”Vanilla” RSA Model

3.2.1 Overview

In the following equations, suppose that U represents the set of all possible utterances. $u \in U$ represents a specific utterance spoken by a speaker. Now, consider another set W that represents the set of all possible states of the world. Let $w \in W$ represent one specific state of the world.

3.2.2 The Literal Listener: L_0

L_0 makes an inference about the state of the world w based upon hearing an utterance u from a speaker. The following model describes how L_0 assigns probabilities to $w \in W$:

$$L_0(w|u) = \frac{[\![u]\!](w) \cdot P_L(w)}{\sum_{w' \in W} [\![u]\!](w') \cdot P_L(w')} \tag{3.4}$$

where the semantic interpretation function $[\![u]\!](w)$ is defined as follows. This function takes in u and w as inputs and maps each utterance-world pair to a truth value.

$$[\![u]\!](w) = \begin{cases} 0 & w \notin [\![u]\!] \\ 1 & w \in [\![u]\!] \end{cases}$$

where $[\![u]\!]$ is the intended meaning of u. P_L represents the listener's prior beliefs about a state of the world w. In other words, the literal listener considers all possible world states that are semantically compatible with the utterance that she heard.

Consider the following concrete example that is illustrated in Figure 3.3. Suppose that a speaker is describing one of his friends to a listener. Suppose that in this extremely simplified world that the speaker can only possibly be talking about two people: either someone with glasses and brown hair or someone with brown hair and no glasses. The speaker says: ”My friend has brown hair”.

(a) Friend with brown hair and glasses (b) Friend with brown hair and no glasses

Figure 3.3: Two Possible States of the World

In this formal model,

- Let $w_1 \in W$ refer to the friend with glasses and brown hair
- Let $w_2 \in W$ refer to the friend with brown hair and no glasses
- Let $u_1 \in U$ be the utterance, "My friend has brown hair"
- Let $u_2 \in U$ be the utterance, "My friend has brown hair and glasses"
- Let $P_L(W) = \{w_1 : 0.5, w_2 : 0.5\}$

A pragmatic listener would assume that since the speaker did not mention that his friend wears glasses that the friend he most likely is referring to is the friend with brown hair and no glasses (w_2). However, the literal listener considers all possible world states that are compatible with the utterance that she hears. In this case, both w_1 and w_2 are compatible with u_1. Therefore, L_0 assigns equal probability to both w_1 and w_2. More formally, $L_0(w_1|u_1) = \frac{1}{2}$ and $L_0(w_2|u_1) = \frac{1}{2}$. Figure 3.4 shows the probability mappings for L_0.

Figure 3.4: L_0 Probability Mappings

3.2.3 The Informative Speaker: S_1

The informative speaker produces an utterance with the ultimate goal of maximizing informativity. Therefore, S_1 tries to produce an utterance in hopes that L_0 will place the most probability weight on w after hearing her utterance. S_1 is defined as follows:

$$S_1(u|w) = \frac{e^{U(L_0(w|u))}}{\sum_{u'} L_0(w|u')} \tag{3.5}$$

where

$$U(u; w) = \ln(L_0(w|u)) \tag{3.6}$$

$U(u; w)$ is a utility function that is defined as L_0's probability of inferring w based on hearing u.

$S_1(u|w)$ represents S_1's decision function, which is soft-max. Soft-max choice rules make the speaker's choices approximately rational [13]. This function has a parameter $\alpha > 0$. An $\alpha = 1$ models very rational behavior (typically used in game- theoretic models) while smaller values of α model less rational behavior.

Consider again the set of world states and utterances mentioned in Section 3.2.2. Given that the speaker wants to communicate the world state that his friend has brown hair but no glasses, he should always choose the utterance 'my friend has brown hair' in order to maximize informativity. Given that the speaker wants to communicate the world state that his friend has both brown hair and glasses, the speaker would not be as informative as he possibly could have been if he produced the utterance 'my friend has brown hair'. If S_1 strives to communicate the fact that his friend has both glasses and brown hair, then he should produce the utterance 'my friend has brown hair and glasses' in order to be as informative as possible. RSA predicts $P(u_1|w_2) = 1$ and $P(u_2|w_1) > P(u_2|w_2)$. Figure 3.5 below shows the probability mappings for S_1.

Figure 3.5: S_1 Probability Mappings

3.2.4 The Pragmatic Listener: L_1

L_1 is a more sophisticated version of L_0, a listener that is able to reason about pragmatic indicators in conversation. Recall that RSA is a recursive framework. At the L_1 level, the pragmatic listener is reasoning about how S_1 is reasoning about L_0. This is illustrated in Figure 3.6:

Figure 3.6: L_1 Recursive Reasoning

Note that it is highly unlikely to ever encounter a naive L_0 speaker in real life. Nor is it it likely to ever encounter such a non-maximal S_1 speaker in real life either. L_0 and S_1 exist in the RSA framework simply because they are inherently nested in a pragmatic listener's psychology. The literal listener L_0 provides a foundation for compositional semantics to input conventionalized semantic information into the pragmatic reasoning process [14]. Compositional semantics is a vehicle that allows languages to construct complex meanings from combinations of simpler elements [15]. In this case, a naive L_0 facilitates the construction of a recursive pragmatic chain as shown in Figure 3.6. One last remark about this recursive reasoning chain is that it does not depict back-and-forth interaction between people. Rather, it describes a process in which a single person (a listener) reasons about his own beliefs about the world. This person's thought process

is in uenced by another person (a speaker) who is also working on her beliefs about the world. Philosopher David Lewis writes:

> In our subsequent reasoning we are windowless monads doing our best to mirror each other, mirroring each other mirroring each other, and so on [16].

This process is capable of existing ad in nitum. However, it has been estimated that humans reason about 1.61 recursive levels on average [17]. It is only necessary to utilize deeper depths of recursion i.e. L_2 and S_3 when modelling more complex phenomena. The pragmatic listener is modelled as follows:

$$L_1(w|u) = \frac{S_1(u|w) \quad P_L(w)}{\sum_{w'} S_1(w'|u) \quad P_L(w')} \tag{3.7}$$

In the L_1, the listener is not merely thinking about semantic compatibility of an utterance is reasoning about the world state that S_1 must have been thinking of in order to produce . We begin by de ning the following set of utterances and world states:

- Let $w_1 \in W$ be the friend with glasses and brown hair
- Let $w_2 \in W$ be the friend with brown hair and no glasses
- Let $u_1 \in U$ be the utterance, ”My friend has brown hair”
- Let $u_2 \in U$ be the utterance, ”My friend has brown hair and glasses”
- Let $P_L(W) = \{w_1 : 0.25, w_2 : 0.75\}$

An issue that we encountered earlier with was that she places equal probability weight on inferring both world states having heard the utterance: 'My friend has brown hair'. However, since L_1 is now reasoning about what world state is in, she now assigns a higher probability weight to w_2 rather than w_1. Therefore, $P_{L_1}(w_2|u_1) > P_{L_1}(w_1|u_1)$. L_1 probability mappings are shown in Figure 3.7:

Figure 3.7: L_1 Probability Mappings

3.3 The Pragmatic Speaker: S_2

The pragmatic speaker S_2 is a higher-order speaker who evaluates the world state that L_1 has in to produce her utterance and also reasons that S_1 attends to communicate the most informative utterance to L_0. I will not go into very much detail about this section because many of the concepts in this model have been introduced in previous sections. The S_2 model is given as follows:

$$S_2(u|w) = \frac{e^{U(L_1(w|u))}}{\sum_{u'} L_1(w|u')} \tag{3.8}$$

where

$$U(u;w) = \ln(L_1(w|u)) \tag{3.9}$$

$U(u;w)$ is a utility function that is defined as L_1's probability of inferring w based on hearing u.

3.4 Modelling More Complex Phenomena with the RSA Framework

As mentioned earlier, the vanilla RSA framework works fine for modelling many pragmatic phenomena. In many cases, adding additional levels of recursion beyond the first level would essentially be redundant. However, research has been done to model many pragmatic phenomena into the RSA framework including: metaphor [1], hyperbole [18], vagueness [14], politeness [19], questions under discussion (QUD) [20], etc. Another pragmatic phenomenon that can be implemented into the RSA framework is sarcasm. This will be demonstrated in Chapter 4.

Chapter 4

Previous Work on Modelling Sarcasm

4.1 Kao et. al (2015)

Kao et al. (2015) proposes a model of irony within the RSA framework that considers that listeners might be uncertain what question under discussion (topic of the conversation) a speaker aims to address when formulating an utterance [6]. Kao et al. (2015) refers to this model as the qRSA model. In the qRSA model, the speaker not only reasons about world state but also about two dimensions of affect: valence or arousal. Emotional valence refers to the extent to which the speaker feels negative or positive valence toward some world state. Emotional arousal corresponds to the intensity of emotion that is felt by the speaker. For example, glee is a positive valence, high arousal emotion and contentment is a neutral/positive valence, low arousal emotion.

The motivation of incorporating these dimensions is that forms of verbal irony involve expressing negative meanings with positive utterances and vice versa. For example, in order to understand an utterance, a listener may need to reason about both the QUD and what is communicated via a speaker's utterance. Suppose that a speaker gets stuck in the middle of a snow storm and utters the following: ”This weather is just fantastic!”. In this utterance, the intention of the speaker is not to remark on the state of the weather. Instead, the speaker is communicating about his emotion regarding the state of the weather. In his utterance, the speaker is literally communicating a positive valence, high arousal emotion. Because his utterance is meant to be interpreted ironically, his utterance actually conveys a negative valence, high arousal emotion since he is upset that he is stuck in a snowstorm. Since a pragmatic listener will most likely infer that the speaker is upset about the weather, the best explanation as to why the speaker chose this utterance was to communicate the intensity of his emotion toward the weather.

Figure 4.1: Boy Stuck in a Snowstorm

One issue with this model is that it only considers the ironic utterances that align with the classical view of irony. Recall that the classical view defines an ironic utterance as one that communicates a proposition that is the opposite of P. The model presented in Kao et al. (2015) is unable to model non-classical forms of irony such as ironic understatement, interrogatives, and cases in which the opposite of P is undefined. However, Cohn-Gordon and Bergen (2019) propose an RSA model that is able to handle non-classical examples of irony, which will be discussed later in this section.

4.1.1 Joint Reasoning in the RSA Framework

Before delving into the irony RSA models, it is important to update our vanilla RSA with a new parameter: joint reasoning. In the vanilla model, a listener makes a judgment about w upon hearing u. However, when the listener has uncertainty about the speaker (say that the listener is unsure about what speaker model is most appropriate), then we must incorporate uncertainty into the model). This is a simple change. In a joint reasoning model, the listener reasons about both w and s, some factor that might influence the speaker's behavior, upon hearing an utterance. This is formalized as follows:

$$P_L(w; s|u) = \frac{P_S(u|w; s)P(s)\ P_L(w)}{\sum_{w'} P_S(w|u')\ P_L(w')} \tag{4.1}$$

4.2 Modelling Questions Under Discussion (QUD) with qRSA

In the qRSA model, S_1 is modelled as follows:

$$U(u; w; a; q) \propto \ln\Big(\sum_{s';a'} \mathbb{1}_{q(w;a)=q(w';a')} L_0(w', a'|u)\Big) \tag{4.2}$$

where a listener infers two dimensions this time: (1) the state of the world W and (2) the speakers' affect. q represents the projection of the listener's inferred meaning into the relevant QUD dimensions. The indicator function is defined as follows:

$$\mathbb{1}_{q(w;a)=q(w';a')} = \begin{cases} 1 & q(w; a) = q(w', a') \\ 0 & \text{else} \end{cases}$$

The inclusion of this indicator function allows the speaker to only be informative along the specific QUD dimension. As a result, a speaker is able to bypass the Gricean maxim of quality. A model of S_1 is given as the following softmax function:

$$S_1(u|w; a; q) \propto e^{U(u;w;a;q)} \tag{4.3}$$

and L_1 is modelled as follows:

$$L_1(w; a|u) \propto P(w)P(a|w) \sum_{q} P(q)W(u|w; a; q) \tag{4.4}$$

The pragmatic listener in qRSA considers prior knowledge and his own internal model of the speaker in order to (1) make a prediction about w and (2) to determine the speaker's affect [6].

4.3 Shortcomings of the qRSA framework

In the qRSA framework, L_1 uses her prior knowledge of the state of the world in order to assume that the literal meaning of an utterance meant to be ironic is unlikely. Upon finding an appropriate projection of an inferred meaning to a relevant QUD dimension, L_1 is able to come to the conclusion that the communicated true world is of the opposite emotional valence but the same

intensity [7]. However, similar to why the Classical View of irony fails, the qRSA model fails when it comes to modelling instances of ironic understatement, where opposite emotional valence is communicated but not at the same intensity.

For example, recall the following meme from chapter 2:

Figure 4.2: ”This is Fine” Dog

Figure 4.2 is a clear illustration of ironic understatement. The dog in the image is under emotional distress yet produces an utterance that does not match the intensity of the emotion that he is feeling. The qRSA model would not be able to correctly interpret the true intended meaning of this ironic utterance since opposite emotional valence is conveyed, but not at the same intensity (ne anddistressedare not words of opposite intensity). Section 4.4 will introduce an alternative model of how to incorporate irony into the RSA framework, which is able to handle instances of irony beyond the Classical View.

4.4 Cohn-Gordon and Bergen (2019)

Cohn-Gordon and Bergen (2019) propose a model of irony that is a form of linguistic countersignaling where the speaker and the listener reason about pretense in order to communicate about the common ground [7]. In the RSA framework, a speaker and a listener recursively reason about each other's mental states in order to communicate effectively. Recall that the Vanilla RSA framework models up to the rst level of recursion (S_2 ! L_1 ! S_1 ! L_0). However, the irony model proposed by Cohn-Gordon and Bergen (2019), which will henceforth be referred to as IronyRSA, includes agents that reason about each other at a second level of recursion. In other words: (S_3 ! L_2 ! S_2 ! L_1^P ! S_1 ! L_0). Note that L_1 is now referred to as L_1^P in this model. Figure 4.3 illustrates the recursive reasoning in the IronyRSA model.

Figure 4.3: Recursive Reasoning in the IronyRSA Model

L_0 and S_1 are modelled almost exactly the same way as they were modelled in the vanilla RSA framework. However, one new variable, c, which represents the common ground between the listener and the speaker, is introduced into the model. Recall that common ground is generally the shared knowledge between the listener and the speaker. Common ground will be discussed in more detail in Section 4.4.2.
A model for L_0 is given as follows:

$$L_0(w|u;c) = \frac{[\![u]\!](w) \cdot P_L(w|c)}{\sum_{w' \in W} [\![u]\!](w') \cdot P_L(w')} \tag{4.5}$$

A model for S_1 is given as follows:

$$S_1(u|w;c) = \frac{e^{U(u;w;c)}}{\sum_{u'} L_0(w|u'))} \tag{4.6}$$

where

$$U(u;w;c) = \ln(L_0(w|u;c)) \tag{4.7}$$

where $U(u;w;c)$ is a utility function that is defined as L_0's probability of inferring w based on hearing u and reasoning about c.

4.4.1 Incorporating Pretense into the RSA Model

There are significant additions to the model starting at the L_1 level. Note that in the IronyRSA framework, L_1 is now referred to as L_1^P, where the pragmatic listener is now capable of reasoning about pretense. According to pretense theory, a speaker does not directly utter a false utterance. Instead, the speaker communicates about a world such that u is true. Though u may be false in the actual world, u is compatible with the choice of w that the speaker communicates. A fundamental claim of this model is that every act of pretending has a corresponding form of irony. It is not the case, however, that all pretenses are ironic (irony is a subset of pretense). In order for irony to be

communicated properly, it is necessary for a listener to understand that the speaker is pretending. The listener must adopt a different perspective in order to process pretense.

Because of the direct correspondence between pretense and irony, Cohn-Gordon and Bergen (2019) introduce a Maxim of Pretense in the Gricean tradition:

Maxim of Pretense. Make it clear that you are pretending. If a pretend perspective could be taken by the listener to be the real one, don't pretend to have this perspective [7].

With this Maxim of Pretense in mind, it is now appropriate to introduce the L_1^P and S_1 levels of the IronyRSA model.

L_1^P: A listener who can detect pretense

It is extremely imperative to note here that L_1^P is not able to reason about sarcasm at this level. L_1^P is able to reason about pretense, but he is not able to detect sarcasm yet (this will come in the L_2 level). The purpose of L_1^P is to introduce the notion of pretense into the IronyRSA model. The difference between L_1 in the Vanilla RSA model and L_1^P is that L_1^P is now able to consider scenarios in which the speaker is not communicating about the actual world w but instead some non-actual world w'. There are two things that are introduced in the L_1^P model: (1) $P_{pretense}$: the probability of pretense and (2) $P_{channel}(w'|w; c; P_L)$: the pretense channel.

1. $P_{pretense}$ represents the listener's belief that the speaker was engaging in pretense before hearing the utterance (the probability that the listener had that the speaker would communicate a pretend world)

2. $P_{channel}(w'|w; c; P_L)$ represents the listener's beliefs about how the speaker chose the world that she intended to communicate.

Define $P_{channel}(w'|w; c; P_L)$:

$$P_{channel}(w'|w; c; P_L) \propto P_L(w'|c) \quad _{(w \neq w')} \tag{4.8}$$

where

$$_{(w \neq w')} = \begin{cases} 1; & w \neq w' \\ 0; & \text{else} \end{cases}$$

Finally, L_1^P is defined as:

$$L_1^P(w|u; c) \propto [P_{pretense}(True) \quad P(w|c) \sum_{w_{prior}} P_{channel}(w_{prior}|w) \quad S_1(u|w_{prior}; c)]$$
$$+ [P_{pretense}(False) \quad P(w|c) \quad S_1(u|w; c)] \tag{4.9}$$

To present a concrete example, suppose that you have a friend that despises cats and has mentioned this fact to you several times in the past. As you two are walking down the street, you spot a cat. You point at the cat and ask your friend, "Isn't that cat adorable?".

In this model, there are two possible worlds. Let w_1 represent the world where you think that your friend thinks the cat is adorable and w_2 represent the world where you don't think that your friend thinks that the cat is adorable.

We can define the parameters for our L_1^P model as such:

- $W = \{w_1, w_2\}$
- $U = \{$"Isn't that cat ugly?", "Isn't that cat adorable?"$\}$
- $Prior_{world} = \{w_1 = 0.01, w_2 = 0.99\}$
- $P_{pretense} = \{$pretense $= w_1: 0.95$, pretense $= w_2: 0.05\}$

In this situation, your friend would assume you are more likely to be communicating about w_2 than w_1. The listener's prior belief is that the speaker believes w_2 is true. Additionally, the listener has belief that the speaker is pretending that w_1 is true. Note again that L_1^P does not model sarcasm. Rather, it provides a foundation for modelling sarcasm by modelling pretense.

S_2: A speaker who knows the listener can detect pretense

S_1 abides by the Maxim of Quality, meaning that she will always want to produce utterances that are as informative as possible. S_1 will never produce false utterances. S_2, however, is able to reason about L_1^P, and therefore produces her utterance knowing that L_1^P is able to detect pretense. As a result, this equips S_2 with the ability to produce an utterance that is literally false with respect to the actual world but true with respect to the world that she is communicating about. S_2 The model is given as follows:

$$S_2(u|w; c) \propto e^{U(u;w;c)} \tag{4.10}$$

where

$$U(u; w; c) = \ln(L_1^P(w|u; c)) \tag{4.11}$$

One important thing to note here is that even though S_2 is capable of producing ironic utterances, she still prefers to produce utterances which are compatible with the actual world. However S_2 has less of a need to be informative when she believes that the common ground is high or, in other words, that she knows L_1^P has a high prior probability of knowing the true world. This corresponds to the Maxim of Pretense: "if a pretend perspective could be taken by the listener to be the real one, don't pretend to have this perspective". Because S_2 is merely aware of pretense but isn't capable of reasoning about it, she is not motivated to deviate from the truth.

4.4.2 The Role of Common Ground in Conversation

Common ground is usually referred to as the shared knowledge between interlocutors. In a conversation between a speaker and a listener, common ground is the set of propositions that: (1) both interlocutors assume, (2) both interlocutors assume each other assumes, and (3) higher order beliefs, and so on.

Before getting into the second level of social recursion, i.e. L_2 and S_3, this section will elaborate more on the notion of common ground. Common ground can be interpreted in two different ways:

Reasoning 1: Suppose a listener hears an utterance u that conveys the proposition P, where P is not taken to be prior knowledge of the listener. Then, the listener learns: (1) P and (2) that P wasn't already in the common ground.

Now, consider the following example. I need to introduce myself to a group of people that I have never met before. I utter the following: 'Hello, my name is Josephine.' This utterance is meant to be taken at face value. The people that I introduce myself to learn two things: (1) that my name is Josephine and (2) that this information wasn't already established in our common ground.

Here is also another way in which common ground can be interpreted:

Reasoning 2: Suppose a listener hears an ironic utterance that communicates the proposition P. Through this utterance, the listener learns (1) opposite of P and (2) P was already in the common ground.

Now, consider the following example. Your friend is baking a cake and you see her add 5 cups of sugar into the cake batter. You say: "Wow, I think you should add more sugar." Your friend learns two things here: (1) you think that she should not add more sugar to the cake batter and (2) the fact that the cake does not need more sugar is already encoded in the common ground.

Common ground in the traditional sense refers to the shared knowledge between interlocutors [21]. However, a listener reasons about what the common ground must have been in order for the speaker to produce her utterance. The flaw in viewing common ground in the traditional sense is treating it as fixed value. It is quite possible that higher-order listeners (like L_2) may have uncertainty over c. This fact will be implemented into the L_2 model.

L_2: A listener who jointly reasons about the state of the world and pretense

Recall that a listener may possess uncertainty about what the common ground might have been in order for a speaker to produce an utterance. Our higher-order listener is now making predictions about both (1) the state of the world that the speaker is communicating and (2) the common ground. In order to model the uncertainty over c, $P_{hyperprior}$ is introduced into the model. $P_{hyperprior}$ is a distribution over distributions that represents the listener's beliefs about what prior over w that he believed the speaker to have before producing her utterance. The L_2 model is given as:

$$L_2(w; c|u) \propto P_{hyperprior}(c)\ P(w|c)\ S_2(u|w; c) \tag{4.12}$$

Upon hearing utterance u, L_2 reasons about how likely it was that the speaker intended to communicate world state w given that the speaker assumed the common ground between the interlocutors to be c.

4.4.3 Sarcasm as Countersignaling

Consider the following situations:

1. Suppose you are strolling down a street late at night with a friend. You notice that all of the stores are closed. You say to your friend: "Looks like everything is open".

2. Suppose that students in a class just learned that their professor assigned 200 pages of required reading due tomorrow. A student turns to her friend and says: "I'm definitely looking forward to doing those readings!".

3. You and a friend see the same movie. You really loved it, but you're not quite sure what your friend thought about it. You say to your friend: ”Gosh, that movie sucked!”

The utterance in situation 1 risks being misinterpreted. The common ground is too certain regarding the proposition that all the stores are closed. This utterance would most likely cause confusion because there is simply nothing left to communicate about.

The utterance in situation 3 also risks miscommunication. In this scenario, the common ground is not certain enough regarding the proposition that the movie was not enjoyable. If you enjoyed the movie, but your friend did not, then there is a high risk that your friend will take this utterance at face-value.

The utterance in situation 2 is an effective way of communicating irony because there is an appropriate amount of that is communicated in the utterance. Upon hearing this utterance, the student's friend will not only learn that the speaker is not looking forward to doing the readings but also that this information was already in the common ground.

Therefore, the speaker has to make a very strategic choice in communicating the common ground with her conversational partner. The goal of the speaker is to produce an utterance that will not only effectively communicate a state of the world but also the state of the common ground between the two interlocutors. This is known as linguistic countersignaling. In communicating irony, the speaker presupposes that the false state of the world that she communicates was already encoded in the common ground.

S_3: A speaker who produces an utterance with the goal of communicating the common ground

S_3 is a model of linguistic countersignaling. The higher-order speaker produces an utterance that will optimally communicate her intended w and c. The model for S_3 is given as follows:

$$S_3(u|w;c) \propto e^{U(u;w;c)} \tag{4.13}$$

where

$$U(u;w;c) = \ln(L_2(w;c|u)) \tag{4.14}$$

The key to S_3 communicating irony effectively is making sure that she communicates an appropriate c, as shown previously. Otherwise, she will run the risk of miscommunication. Hence, we have finally arrived at a speaker who is capable of communicating ironic utterances.

We have now theoretically arrived at a model that is capable of handling all instances of irony both in classical and the non-classical sense. In order to test if this model is able to predict sarcasm in the same capacity as humans, Chapters 5 and 6 will present an experimental validation of the IronyRSA model. Chapter 5 will describe experiments that were performed to collect human judgment data on sarcasm and to collect data to train the IronyRSA model. Chapter 6 will show that predictions made by the IronyRSA model were very similar to predictions made by humans.

Chapter 5

Experimental Validation Part I: Data Collection

5.1 Overview

Chapter 4 offered an overview of the theoretical IronyRSA model presented in Cohn-Gordon and Bergen (2019). Chapters 5 and 6 offer an experimental validation of the IronyRSA model in order to con rm that it truly predicts irony in conversation. Chapter 5 describes two studies that were conducted to collect human judgment data: (1) a sarcasm study that collected data on how people interpret sarcasm and (2) a norming study that collected data on common ground. Then, in Chapter 6, data from the norming study was fed into a computational implementation of the IronyRSA model. Results from the sarcasm study and the IronyRSA computational model predictions were compared to see how well the model's predictions align with human judgment of sarcasm.

5.2 Sarcasm Study

A sarcasm study was conducted in order to collect data on how humans judge sarcasm in a given situation. Participants were asked to read about a hypothetical scenario where you and and your friend, Mary are packing for a weekend camping trip. You realize that you would like to bring some apples, but you don't have any at home. Your friend John kindly offers to buy some apples from the store that you and Mary can bring on your trip. John returns back with x apples, where x is a varied value such that $x = \{2, 3, 4, 5, 10, 20\}$. The x value that each participant read about was randomly generated. After John leaves, Mary utters the following: ”We could use more apples”. Note that the classical model of irony would have trouble in making predictions about this scenario as Mary's utterance is sarcastically ambiguous based on the context.

The purpose of this experiment was to investigate how participants judged sarcasm based on the context of the number of apples John brought back. In other words, Mary's utterance could be taken either at face-value or sarcastically depending on the number of apples that John brought back for the camping trip. The survey consisted of two parts: Part I asked participants about their judgements of sarcasm in an implicit way and Part II asked participants about their judgments of sarcasm explicitly.

Figure 5.1 shows Part I of the sarcasm survey. Participants were rst asked to interpret Mary's utterance: ”We could use more apples”. If participants chose the rst option: Mary thinks that John bought too many apples”, then it is likely that they thought that Mary's utterance was sarcastic. If participants chose the second option: Mary thinks that John bought too few apples, then it is likely that they took Mary's utterance at face value. Participants were then asked to rate their con dence in their answer on a likert scale from (1-7), with 1 meaning that they felt very uncertain in their response to the previous question and 7 meaning that they felt very certain in their response to the previous question.

Figure 5.1: Survey Part I

Figure 5.2 shows Part II of the sarcasm survey. Participants were rst asked if Mary's utterance was sarcastic or not. Then, they were again asked to rate their con dence in their answer on a likert scale from (1-7), with 1 meaning that they felt very uncertain in their response to the previous question and 7 meaning that they felt very certain in their response to the previous question. Note that participants were disabled from returning back to Part I of the survey. This was done in order to ensure that participants would not be able to change their answers from Part I after viewing Part II. The goal was to investigate if participants' judgments of sarcasm changed from Part I once they were primed to consider sarcasm in Part II.

Figure 5.2: Sarcasm Survey: Part II

5.2.1 Materials and Methods

This study was launched on Amazon Mechanical Turk[1]. 143 participants were recruited to take Part In the survey. Participants were screened by the following criteria:

- Must be a United States citizen
- Must have at least a 95% approval rating on Mechanical Turk

All participants were compensated at the posted rate for their participation in the survey. This survey was approved by the Stanford University Institutional Review Board (IRB) before being launched.

[1]The following link will direct you to the sarcasm survey: https://jsoddano.github.io/irony-experiments/experiments/template_survey/index.html

5.2.2 Results

Table 5.1 shows the distribution of participants assigned to each context (number of apples):

Context	Number of Participants
2	22
3	19
4	19
5	23
10	35
20	25

Table 5.1: Participants per Context

Table 5.2 shows the percentage of participants who thought Mary's utterance: "We could use more apples" was sarcastic based on the given context. The second column labeled "% Mean: too many" corresponds to the percentage of participants who chose the option of Mary thought that John bought too many apples in Part I (sarcastic interpretation). The third column labeled "% Mean: sarcastic" corresponds to the percentage of participants who chose the option that Mary's utterance was sarcastic in Part II. Column 2 encodes data for implicit sarcasm judgments while Column 3 encodes data for explicit sarcasm judgments.

Context	Part I Sarcastic (%)	Part II Sarcastic (%)
2	0.00	0.91
3	5.26	10.53
4	0.00	10.53
5	8.70	5.35
10	17.14	15.29
20	25.00	52.00

Table 5.2: Part I and II Sarcasm Judgments

Table 5.2 shows that in both Parts I and II, sarcasm detection increased as number of apples increased. This was to be expected because bringing a very large number of apples on a camping trip is rather absurd. Comparing Part I and II results, sarcasm was detected more frequently for each given context compared to Part I. This may have been due to the fact that participants were primed to think about sarcasm in Part II. For example, In the 20 apples context, 24% of participants judged Mary's utterance as sarcastic in Part I while 52% of participants judged Mary's utterance as sarcastic in Part II.

Figure 5.3 shows Part I implicit sarcasm judgment data. The x axis represents the number of apples in context and the y axis represents the number of participants who thought that Mary's utterance was sarcastic or not. The data shows that sarcasm is increasingly detected as the number of apples in the situation increases.

Figure 5.3: Part I Implcit Sarcasm Judgments

Figure 5.4 shows Part II explicit sarcasm judgment data. Again, the x axis represents the number of apples in context and the y axis represents the number of participants who thought that Mary's utterance was sarcastic or not. Similar to Part I data, sarcasm is increasingly detected as the number of apples in the situation increases. However, we see that sarcasm is detected in higher proportions in Part II.

Figure 5.4: Part II Explicit Sarcasm Judgments

An explanation for this pattern is that participants may have been somewhat confused about their task. This was something that was voiced in the survey comments. Participants were not aware that they were being asked to reason about sarcasm until Part II.

Mary uttered: ”We could use more apples”. From a literal standpoint, the correct option to describe the situation would be that ”Mary thinks that John bought too few apples”. This is the

option that most participants chose. However, if the participants were also taking sarcasm into account, then they may have chosen the other option. Being made explicitly aware that the survey was asking about sarcasm, participants more often detected sarcasm in Part II (especially for the 20 apples context). Though overall, it could be the case that sarcasm isn't very easily detected over the internet and caused some confusion amongst participants. Additionally, the experiment was designed such that Mary's utterance was sarcastically ambiguous based on the context of number of apples given. It could have been that participants found the situation hard to interpret.

Table 5.3 shows participants' average con dence ratings (out of 7) for Part I. The second column "Sarcastic Con dence" gives the average con dence score of those who chose the option: Mary thinks that John bought too many apples (the sarcastic option). The third column "Not Sarcastic" Con dence gives the average con dence score of those who chose the option: Mary thinks that John bought too few apples (the non-sarcastic option). The fourth column gives the overall average con dence of all participants per context.

Context	Sarcastic Con dence	Not Sarcastic Con dence	Overall Con dence
2	—	6.86	6.86
3	7.00	6.89	6.89
4	—	6.63	6.63
5	5.00	6.52	6.30
10	5.17	6.69	6.43
20	5.33	5.95	5.80

Table 5.3: Part I Con dence Ratings (by context)

Con dence ratings generally decreased as number of apples in the given situation increased regardless if participants thought Mary's utterance was sarcastic or not. As number of apples increased, more participants deviated from the literal interpretation of Mary's utterance and were also increasingly uncertain about their interpretation. The decreasing con dence trend is present for those who interpreted Mary's utterance as non-sarcastic too. This might hint that participants who chose this option were considering the fact that Mary's utterance could be taken non-literally despite the fact that they chose the literal option.

The following table shows participants' average con dence ratings (out of 7) for Part II. The second column "Sarcastic" Con dence gives the average con dence score of those who believed that Mary's utterance was sarcastic. The third column "Not Sarcastic" Con dence gives the average con dence score of those who believed Mary's utterance was not sarcastic. The fourth column gives the overall average con dence of all participants per context.

Similar to Table 5.3, Table 5.4 shows the same decreasing trend in con dence as the number of apples in context increases. However, con dence scores on average are lower in Part II compared to Part I. Perhaps mentioning sarcasm in Part II caused participants to re-evaluate their interpretation of the given utterance.

Context	”Sarcastic” Con dence	”Not Sarcastic” Con dence	Overall Con dence
2	6.00	5.85	5.86
3	5.00	6.47	6.32
4	5.50	5.88	5.74
5	6.00	5.00	5.04
10	5.80	5.40	5.31
20	5.31	5.50	5.92

Table 5.4: Part II Con dence Ratings (by context)

Figures 5.5 and 5.6 show that there was a lot more variability in con dence ratings from Part II compared to Part I. Again, this could have been caused by the fact that explicitly mentioning sarcasm in Part II caused some participants to re-evaluate their responses from Part I. In both g-ures, the x axis represents the number of apples in context and the y axis represents the average con dence rating of participants on a scale from (1-7).

Figure 5.5: Part I Con dence Ratings (by Context)

Figure 5.6: Part II Con dence Ratings (by Context)

Table 5.5 shows con dence ratings from Parts I and II based on whether the participant interpreted Mary's utterance as sarcastic or not.

Those who thought Mary's utterance was sarcastic had the same amount of con dence in both Parts I and II. However, con dence decreased from Part I and Part II for those who interpreted Mary's utterance as non-sarcastic.

Sarcastic	Part I Con dence	Part II Con dence
Yes	5.20	5.20
No	6.60	5.53

Table 5.5: Con dence Ratings (grouped by sarcasm judgment)

5.3 Norming Study

Common ground refers to the knowledge that interlocutors share in a conversation, which is derived from each conversational partner recursively reasoning about the other's beliefs. Recall from section that common ground plays a pivotal role in how people reason about communication. In the sarcasm survey described in section 5.2, a participant's interpretation of Mary's utterance depends on their understanding of the common ground between you and Mary. These patterns of reasoning about the common ground are illustrated as follows:

1. Suppose that you interpret Mary's utterance at face-value. Then, you learn (1) that Mary believes that you need more apples for the camping trip and (2) this information wasn't already in the common ground.

2. Suppose that you interpret Mary's utterance sarcastically. Then, you learn (1) that Mary does*not* believe that you need more apples for the camping trip and (2) this information was already in the common ground.

The goal of the norming study was to collect information on common ground. In this study, participants were asked to read about a very similar situation as the one presented in the sarcasm study. In order to extract information about common ground, simply asking participants how many apples is too many apples to bring on a camping trip would not be suf cient. Instead, participants were placed in the listener's shoes. They were explicitly made aware that they are taking this survey along with many other people. Their goal was to select the option that the majority of survey takers will also choose. Participants were then asked to rate their con dence about their response to the previous question. This task yielded information about how the interlocutors in this particular situation reason about the common ground.

This study was inspired by work published in Von Ahn and Dabbish (2008) on how to design games with a purpose. This task was designed in such a way that ”facts are collected as a side effect of playing” [22]. It is important that tasks are not only enjoyable for the participant but also allow the researcher to collect necessary information. Figure 5.7 shows the norming experiment that was presented to participants.

Figure 5.7: Norming Study

The common ground data collected from this study was used to train a computational implementation of the IronyRSA model. The predictions from the computational model were compared to the human judgment data discussed in Section 5.2.

5.3.1 Materials and Methods

This study was launched on the crowdsourcing platform, Amazon Mechanical Turk[2]. 315 participants were recruited to take Part In the survey. Participants were screened by the following criteria:

- Must be a United States citizen

[2]The following link will direct you to the norming study: https://jsoddano.github.io/irony-experiments/experiments/0_apple-ratings/index.html

- Must have at least a 95% approval rating on Mechanical Turk

All participants were compensated at the posted rate for their participation in the survey. This survey was approved by the Stanford University Institutional Review Board (IRB) before being launched.

5.3.2 Results

The distribution of participants assigned to each context (number of apples) is given in the following table:

Context	Number of Participants
2	49
3	55
4	56
5	57
10	47
20	51

Table 5.6: Participants per Context

Table 5.7 shows the results from participants' responses to the first question. The second column labeled "% too many" corresponds to the percentage of participants who believed that the number of apples listed in the given scenario was too many apples to bring on a camping trip.

Context	% too many
2	10.20
3	12.73
4	30.36
5	70.21
10	78.72
20	96.08

Table 5.7: Apples Judgment

Figure 5.8 shows that the proportion of participants who thought that number of apples was too many apples to bring on a camping trip increases as increases. This was to be expected. The x axis represents the number of apples in context and the y axis represents the number of participants who thought that apples was either too many or too few.

Figure 5.8: Norming Experiment: Apples Judgment

Table 5.8 shows participants' average con dence ratings (out of 7) for their responses to the previous question. The second column "Too Many" Con dence gives the average con dence scores of those who thought the number of apples in the given context was too many to bring on a weekend camping trip. The third column "Too Few" Con dencegives the average con dence scores of those who thought the number of apples in the given context was too few to bring on a weekend camping trip. The fourth column gives the overall con dence scores per context assigned.

Note that con dence ratings for all respondents increases as the number of apples in a context increases.

Context	"Too Many" Con dence	"Too Few" Con dence	Overall Con dence
2	5.80	6.00	5.88
3	5.43	5.52	5.38
4	5.65	5.44	5.20
5	5.15	5.46	5.28
10	5.89	5.10	5.72
20	6.10	6.50	6.12

Table 5.8: Part I Con dence Ratings (by context)

Table 5.9 shows con dence ratings grouped by if participants thought that the number of apples was too many or too few. There is no discernible difference in the con dence ratings between these groups.

Apples	Average Con dence
Too Many	5.55
Too Few	5.60

Table 5.9: Con dence Ratings (grouped by Many/Few Judgment)

The surveys launched in this section equip us with the necessary data to perform a comparative analysis between how the IronyRSA predicts sarcasm compared to humans. Note that the sarcasm experiment was designed such that Mary's utterance was sarcastically ambiguous. As a result, there was a lot of uncertainty and variability in the data. In order to prove that the IronyRSA model is effective at modelling sarcastic phenomena, it needs to be able to capture the same nuances and variability present in the human evaluation. An evaluation of the IronyRSA model is presented in Chapter 6.

Chapter 6

Experimental Validation Part II: A Computational Implementation of RSA

6.1 Introduction to WebPPL

Data from the norming study described in Chapter 5 was used to train a computational implementation of the IronyRSA model. This computational model was constructed with the probabilistic programming language, WebPPL. Once predictions were obtained from the IronyRSA model, they were compared with human judgments of sarcasm. Section 6.2 will show how basic RSA is implemented in WebPPL, Section 6.3 will show how IronyRSA can be implemented in WebPPL, and Section 5.4 will show a comparison of results between human judgment and computational model predictions.

6.2 Integrating the Rational Speech Acts Framework into a Computational Model

A simple example will be used to show how WebPPL can be used to model predictions within the RSA framework. The example that will be demonstrated in this section is the scenario that is presented in Frank and Goodman (2012) [5]. In this scenario, a speaker and a listener are able to talk about three objects, a blue square, a blue circle, and a green square, as shown in Figure 6.1. The code discussed in this section was taken directly from Probabilistic language understanding: An introduction to the Rational Speech Act framework and is freely accessible online[1]

Figure 6.1: Reference Game Presented in Frank and Goodman (2012)

In this reference game, a speaker wants to refer to one of these three objects, but is only allowed to choose a single word utterance to do so. As a result, the speaker is able to communicate about either the shape of the given object or the color of the object. The set W of world states and the set U of possible utterances is given as follows:

- W = {blue square, blue circle, green square}
- U = {"blue", "green", "square", "circle"}

[1] Code discussed in this section is freely accessible here: https://michael-franke.github.io/probLang/chapters/01-introduction.html

6.2.1 Modelling L_0 in WebPPL

Recall that L_0 is modelled as follows, where $w \in W$ and $u \in U$:

$$L_0(w|u) = \frac{[\![u]\!](w) \; P_L(w)}{\sum_{w' \in W} [\![u]\!](w') \; P_L(w')}$$

where the semantic interpretation function $[\![u]\!](w)$ is defined as follows:

$$[\![u]\!](w) = \begin{cases} 0 & w \notin [\![u]\!] \\ 1 & w \in [\![u]\!] \end{cases}$$

The following code from Scontras and Tessler [23] demonstrates how the literal listener can be modelled:

```
// set of world states
var world_states =
    [{color: "blue", shape: "square", string: "blue square"},
     {color: "blue", shape: "circle", string: "blue circle"},
     {color: "green", shape: "square", string: "green square"}]
// set of utterances
var utterances = ["blue", "green", "square", "circle"]
// prior over world states
var objectPrior = function(){
  var obj = uniformDraw(world_states)
  return obj.string
}
// meaning function to interpret the utterances
var meaning = function(utterance, obj){
  _.includes(obj, utterance)
}
// literal listener model
var literalListener = function(utterance){
  Infer({model: function(){
    var obj = objectPrior();
    var uttTruthVal = meaning(utterance, obj);
    condition(uttTruthVal == true)
    return obj
  }})
}
viz.table(literalListener("blue"))
viz.table(literalListener("square"))
```

This code produces the following outputs:

w	Probability
Blue Circle	0.5
Blue Square	0.5

(a) L_0 interpretation of hearing the utterance "blue"

w	Probability
Green Square	0.5
Blue Square	0.5

(b) L_0 interpretation of hearing the utterance "square"

The literal listener simply assigns equal probability weights to both the blue circle and blue square upon hearing "blue" since both objects are blue. The literal listener also assigns equal probability weights to both the green square and blue square upon hearing "square" since both objects are square.

6.2.2 Modelling S_1 in WebPPL

Recall that in the vanilla RSA framework, the informative (pragmatic) speaker S_1 can be modelled as such:

$$S_1(u|w) = \frac{e^{\alpha U(L_0(w|u))}}{\sum_{u'} L_0(w|u')}$$

where α is an optimality parameter and

$$U(u; w) = \ln(L_0(w|u))$$

is a utility function that is defined as L_0's probability of inferring w based on hearing u. S_1 can be modelled in WebPPL with the following code[2]

```
// set speaker optimality parameter
var alpha = 1

// pragmatic speaker model
var speaker = function(obj){
  Infer({model: function(){
    var utterance = uniformDraw(utterances)
    factor(alpha * literalListener(utterance).score(obj))
    return utterance
  }})
}
```

[2] Note that $\alpha = 1$ models rational behavior as predicted in game theory while values closer to zero model less rational behavior.

6.2.3 Modelling L_1 in WebPPL

Recall that in the vanilla RSA framework, the pragmatic listener can be modelled as such:

$$L_1(w|u) = \frac{S_1(u|w)\ P_L(w)}{\sum_{w'} S_1(w'|u)\ P_L(w')}$$

L_1 can be modelled in WebPPL with the following code:

```
nnpragmatic listener model
var pragmaticListener = function(utterance){
  Infer({model: function(){
    var obj = objectPrior()
    observe(speaker(obj), utterance)
    return obj
  }})
}
```

Output of the code:

w	Probability
Blue Square	0.6
Blue Circle	0.4

(a) L_1 interpretation of hearing the utterance "blue"

w	Probability
Blue Square	0.6
Green Square	0.4

(b) L_1 interpretation of hearing the utterance "square"

Upon hearing the utterance "blue", L_1 now assigns a higher probability weight to blue square rather than the blue circle. L_1 reasons that if the speaker wanted to communicate about the blue circle, then she would have been more likely to say "circle" rather than "blue". Upon hearing the utterance "square", L_1 assigns higher probability weight to the blue square rather than the green square. L_1 reasons that if the speaker wanted to communicate about the green square, then she would have been more likely to say "green" rather than "square".

Changing the optimality parameter for S_1 from $= 1$ to $= 5$ produces the following output:

w	Probability
Blue Square	0.943
Blue Circle	0.057

(a) L_1 interpretation of hearing the utterance "blue"

w	Probability
Blue Square	0.943
Green Square	0.057

(b) L_1 interpretation of hearing the utterance "square"

By increasing the optimality parameter, L_1 becomes more rational in his judgments of the speaker's utterance.

6.3 Modelling Sarcasm in WebPPL

In Section 6.2, it was demonstrated how the basic RSA framework can be computationally modelled using WebPPL. This section will demonstrate how sarcasm can be modelled into the RSA framework using WebPPL. The following code was adapted from the model presented in [7]. Consider the situation in the experiment described in section 5.2 where John brings back x number of apples for you and Mary's weekend camping trip and Mary says: "We could use more apples". Suppose that in this world, there exists two possible choices for the state of the world and two possible utterances. These are defined as follows:

- $W = \{wTooMany, wTooFew\}$
- $U = \{uTooMany, uTooFew\}$

wTooMany refers to the state where there are too many apples and wTooFew refers to the state where there are too few apples. uTooMany corresponds to the utterance: "We could use more apples" and uTooFew corresponds to the utterance: "We could use less apples".

6.3.1 IronyRSA L_0 in WebPPL

Recall that in the IronyRSA framework, L_0 is modelled as follows:

$$L_0(w|u;c) = \frac{[\![u]\!]^K(w) \cdot P_L(w|c)}{\sum_{w' \in W} [\![u]\!]^K(w') \cdot P_L(w')}$$

One new parameter is introduced in the IronyRSA framework, c, which represents the common ground between the listener and the speaker. The following code will model L_0

```
//set of world states and utterances
var worlds = ["wTooMany","wTooFew"]
var utterances = ["uTooMany", "uTooFew"]
```

```
//models the common ground
var constructPrior = function(paramsf)
 return Infer({ method: 'enumerateg, function() f
   var v = bernoulli({ p: paramsg)
   return v?worlds[0]:worlds[1]
 g)g

//prior over world states
var worldPrior = function()f return Infer({ method: 'enumerateg,
    function() f
   var w = uniformDraw(worlds)
   return wg)g

//prior over utterances
var utterancePrior = function()f return Infer({ method: 'enumerateg,
    function() f
   var u = uniformDraw(utterances)
   return ug)g

//semantic interpretation function
var semantics = function(u,w)f
 return (u[4]==w[4])
g

//literal listener model
var L0 = function(u,priorParams)f
 return Infer({ method: 'enumerateg, function() f
   var b = constructPrior(priorParams)
   var w = sample(b)
   condition(semantics(u,w))
   return w
 g)g
```

6.3.2 IronyRSA S_1 in WebPPL

Recall that in the IronyRSA framework S_1 is modelled as follows:

$$S_1(u|w;c) = \frac{e^{U(u;w;c)}}{\sum_{u'} L_0(w|u')}$$

where

$$U(u;w;c) = \ln(L_0(w|u;c))$$

where $U(u;w;c)$ is a utility function that is de ned as L_0's probability of inferring w based on hearing u and reasoning about c. S_1 produces an utterance based on her knowledge of the world

and the common ground between her and L_0. The following code defines S_1 in WebPPL. Note that α has been set to 1.

```
// informative speaker
var S1 = function(w, priorParams){
 return Infer({method: 'enumerate'}, function() {
   var u = sample(utterancePrior())
   factor(L0(u,priorParams).score(w))
   return u
 })}
```

6.3.3 Sarcastic L_1^P in WebPPL

Two new things are introduced in the L_1^P model: (1) $P_{pretense}$: the probability of pretense and (2) $P_{channel}(w'|w; c; P_L)$: a pretense channel. Please refer to equation 4.8 for the equation of the pretense channel and equation 4.9 for the equation of L_1^P. The WebPPL implementation is given as follows:

```
//define pretense channel
var pretensechannel = function(w, dist){
 return Infer({method: 'enumerate'}, function() {
   var pretenseW = sample(dist)
   condition(pretenseW!=w)
   return pretenseW
 })}
 // L1P model
var L1P = function(u, priorParams){
 return Infer({method: 'enumerate'}, function() {
   var pretense = flip(0.1)
   var b = constructPrior(priorParams)
   var w = sample(b)
   var pretenseW = sample(pretensechannel(w,b))
   var S1Dist = pretense?S1(pretenseW, priorParams):S1(w, priorParams)
   factor(S1Dist.score(u))
   return w
 })}
```

Note that in this implementation, the probability of pretense is set quite low (0.1) since this particular model assumes that the speaker is not pretending in her speech for the most part. This number can be adjusted if needed.

6.3.4 Sarcastic S_2 in WebPPL

S_2 produces an utterance based on her knowledge of the world and the common ground between her and L_1^P.

The S_2 model is given as follows:

$$S_2(u|w;c) \propto e^{U(u;w;c)}$$

where

$$U(u;w;c) = \ln(L_1^P(w|u;c))$$

The WebPPL code is given:

```
// S2 model
var S2 = function(w, priorParams){
return Infer({method: 'enumerate'}, function() {
 var u = sample(utterancePrior())
 factor(L1P(u, priorParams).score(w))
 return u
})}
```

6.3.5 Sarcasm L_2 in WebPPL

Recall that L_2 in the Sarcasm RSA model is now reasoning about the state of the world that the speaker is communicating and the common ground between the listener and the speaker. S_2 samples possible values from a hyperprior distribution, which is a distribution over distributions that represents the listener's beliefs about what prior over w that he believed the speaker to have before she produced her utterance. Common ground data from the norming study was used to construct the hyperprior distribution. The hyperprior distribution is constructed with data from two variables: priorSupport and priorProbs. priorProbs corresponds to sarcasm judgment ratings while priorSupport corresponds to the con dence that is associated with sarcasm judgment. Sarcasm judgment ratings and corresponding con dence scores were rescaled on the interval [0, 1] before constructing the hyperprior. Recall that the equation for L_2 in the IronyRSA framework is given as:

$$L_2(w;c|u) \propto P_{hyperprior}(c) \; P(w|c) \; S_2(u|w;c)$$

The following code can be implemented in WebPPL to model L_2:

```
//common ground data
var priorSupport = [confidence scores]
var priorProbs = [sarcasm judgment]

//construct hyperprior
var hyperPrior = function(){return Infer({method: 'enumerate'},
function() {
  var p = categorical({ps: priorProbs, vs: priorSupport})
   return p})}

//define L2 model
var L2 = function(u){
 return Infer({method: 'enumerate', samples: 10000}, function() {
```

```
    var priorParams = sample(hyperPrior())
    var w = sample(constructPrior(priorParams))
    var S2Dist = S2(w, priorParams)
    factor(S2Dist.score(u))
    return [priorParams,w]
  })
```

6.3.6 Sarcasm: S_3 in WebPPL

Lastly, S_3 engages in countersignaling in order to produce an utterance based on her knowledge of the world and the common ground. S_3 is modelled as follows:

$$S_3(u|w;c) \propto e^{U(u;w;c)}$$

where

$$U(u;w;c) = \ln(L_2(w;c|u))$$

Here is the corresponding WebPPL code:

```
nnS3 model
var S3 = function(w, priorParams){
return Infer({method: 'enumerate'}, function() {
 var u = sample(utterancePrior())
 factor(L2(u, priorParams).score([priorParams,w]))
 return u
})}
```

6.4 Comparison of Results

Figure 6.4 shows sarcasm predictions from the IronyRSA model:

Context	% sarcastic
2	0.33
3	1.62
4	0.23
5	4.41
10	14.35
20	22.48

Table 6.4: IronyRSA Model Predictions

Figure 6.2 below shows a comparison of results from Part I of the sarcasm study discussed in Section 5.2 and WebPPL predictions. The x axis represents the number of apples in context and the y axis represents percentage of sarcasm detected. Sarcasm detection generally increases as context increases.

Figure 6.2: Human Judgment vs Computational Predictions

For both human judgment and computational predictions, as context increases, sarcasm detection increases. This shows that the IronyRSA model was effective at detecting sarcasm in a similar fashion compared to human judgment. The WebPPL model detected sarcasm at lower frequencies in lower number contexts compared to human judgment data. However, there is a big jump in sarcasm detection from context 4 to 5. This shows that this is the context in which sarcasm is beginning to be most clearly detected. A similar trend appears in the human judgment data- there is a noticeable jump in sarcasm detection from context 4 to 5. However, this data showed almost no sarcasm detected at all at context 2. Then, sarcasm detection increases approximately linearly from context 3 to context 20. Meanwhile, there was some variability in the WebPPL predictions at lower contexts. A little sarcasm was detected at context 2, then went down at 3, then back up at 4. The general trend, however, is that WebPPL predictions generally corresponded to human judgment predictions.

From this data, it has been shown that the IronyRSA was reasonably successful at modelling sarcasm in one particular scenario. It must be mentioned that the classical view would not have been able to handle this scenario due to its sarcastic ambiguity. This shows that the IronyRSA model was effective at making predictions about a scenario that would have otherwise failed under the classical account. However, it must be noted that these computational predictions were generated with a very small set of data. Much more future work is necessary in proving that the IronyRSA model is effective in predicting sarcasm in the same way as humans.

Chapter 7

Conclusion

7.1 Overview

In this thesis, I have demonstrated (1) how the Rational Speech Acts framework can be used to model sarcasm in language and (2) the predictions made by a speci c model of sarcasm show close similarity to human predictions about sarcasm. Chapter 2 provided a linguistic overview of sarcasm. Many past attempts at modelling irony were all shown to be ineffective at fully explaining the role of sarcasm in communication. However, Cohn-Gordon and Bergen (2019) proposed a theory of irony as a form of linguistic countersignaling that is capable of capturing irony in both the classical and non-classical sense. The main idea of the countersignaling model is that a speaker is motivated to use irony in conversation in order to communicate about the common ground.

In Chapter 3, it was shown how the Rational Speech Acts framework is able to capture social and pragmatic reasoning in conversation through probabilistic inference. Chapter 4 demonstrated how sarcasm can be introduced into the Rational Speech Acts Framework. It was shown how the irony framework proposed by Kao et al. (2015) was able to model some sarcastic phenomena but fails to model sarcastic phenomena that do not fall under the Classical View of Irony. Later in this chapter, it is shown how the countersignaling model introduced by Cohn-Gordon and Bergen (2019) is able to handle non-classical accounts of sarcasm. Chapters 5 and 6 offer an experimental validation of the countersignaling model for sarcasm detection.

Chapter 5 describes two surveys that were launched to collect data for the experimental validation. The rst survey that was launched collected data on human judgment of sarcasm. This survey was designed in such a way that the scenario was sarcastically ambiguous and could not be modelled under the Classical View of Irony. The second survey that was launched collected common ground data that was used to train a computational implementation of the IronyRSA model. Chapter 6 provides an overview of how RSA can be modelled using the probabilistic programming language, WebPPL. It was later shown how IronyRSA can be implemented computationally using WebPPL. The last section of Chapter 6 compared sarcasm judgments by human subjects to the computational predictions made by the WebPPL IronyRSA model. A comparison of results showed that the predictions made by the computational model were very similar to human judgment predictions.

7.2 Future Work

Though the results presented in chapter 6 are promising, this is just the rst step of many in proving that the IronyRSA model is capable of modelling the very complex and confusing facet of human language that is sarcasm. Future work will need to show that this model can predict non-classical accounts of sarcasm such as ironic understatement and ironic questions. In order to show this, more experimental validation will need to be performed. One concern in the future is whether computational models will be able to handle more complex scenarios. Most of the scenarios presented in this thesis consist of sets of utterances and worlds that contain no more than three elements per set. These simple scenarios are fairly simple to implement. However, most scenarios are not nearly as simplistic. The computational complexity that is required to model the human cognitive process could be very large, complicating the use of this model.

7.3 Broader Implications

We have come a long way from simply regarding pragmatics as the ”wastebasket” of linguistics. With recent developments in computational tools like the Rational Speech Acts Framework that are able to model very complex pragmatic and social phenomena, we are able to formalize elds that have only been approached qualitatively in the past. These developments may have a substantial impact on the eld of arti cial intelligence (AI), a sub eld of computer science that is concerned with programming computers to think like humans and perform tasks like humans. One of the goals of AI is to program machines to understand human language. Natural language understanding, which refers to the computer comprehension and understanding of human language, is regarded as an AI-complete problem, meaning that solving it would solve a central AI problem. This would result in an AI that is as intelligent as a human being. Recent work has been done to combine the Rational Speech Acts Framework with machine learning models in order to yield more human-like language predictions [24]. Combining these two disciplines can be very promising for improving computational models of natural language in the future.

Bibliography

[1] Justine T Kao. Modeling Creative and Social Uses of Language. Stanford University, 2016.

[2] Martin Joos. Description of language design. The Journal of the Acoustical Society of America, 22(6):701–707, 1950.

[3] Yehoshua Bar-Hillel. Out of the pragmatic wastebasket. Linguistic inquiry, 2(3):401–407, 1971.

[4] Christopher Potts. Formal pragmatics. The Routledge Encyclopedia of Pragmatics, pages 167–170, 2009.

[5] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. Science, 336(6084):998–998, 2012.

[6] Justine T Kao and Noah D Goodman. Let's talk (ironically) about the weather: Modeling verbal irony. In CogSci, 2015.

[7] Reuben Cohn-Gordon and Leon Bergen. Verbal irony, pretense, and the common ground, 2019.

[8] Ludwig Wittgenstein. Philosophical investigations. John Wiley & Sons, 2010.

[9] Christopher J Lee and Albert N Katz. The differential role of ridicule in sarcasm and irony. Metaphor and symbol, 13(1):1–15, 1998.

[10] Marcus Fabius Quintilian. Institutio oratoria, trans. HE Butler. Cambridge, 1920.

[11] Deirdre Wilson. The pragmatics of verbal irony: Echo or pretence? Lingua, 116(10):1722–1743, 2006.

[12] Michael C Frank. Rational speech act models of pragmatic reasoning in reference games. 2016.

[13] R Duncan Luce. Individual choice behavior: A theoretical analysis. Courier Corporation, 2012.

[14] Daniel Lassiter and Noah D Goodman. Adjectival vagueness in a bayesian model of interpretation. Synthese, 194(10):3801–3836, 2017.

[15] Zhiyuan Liu, Yankai Lin, and Maosong Sun. Compositional semantics. In Representation Learning for Natural Language Processing, pages 43–57. Springer, 2020.

[16] LEWIS David. Convention: a philosophical study, 1969.

[17] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. The Quarterly Journal of Economics, 119(3):861–898, 2004.

[18] Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. Nonliteral understanding of number words. Proceedings of the National Academy of Sciences, 111(33):12002–12007, 2014.

[19] Erica J Yoon, Michael Henry Tessler, Noah D Goodman, and Michael C Frank. " i won't lie, it wasn't amazing": Modeling polite indirect speech. In CogSci, 2017.

[20] Robert XD Hawkins, Andreas Stuhlmüller, Judith Degen, and Noah D Goodman. Why do you ask? good questions provoke informative answers. In CogSci. Citeseer, 2015.

[21] Robert J Aumann. Agreeing to disagree. The annals of statistics, pages 1236–1239, 1976.

[22] Luis Von Ahn and Laura Dabbish. Designing games with a purpose. Communications of the ACM, 51(8):58–67, 2008.

[23] G Scontras, Michael Henry Tessler, and M Franke. Probabilistic language understanding: An introduction to the rational speech act framework, 2017.

[24] Will Monroe. Learning in the Rational Speech Acts Model. PhD thesis, Stanford University, 2018.