Sociolinguistic Interaction and Identity Construction: The View from Game-Theoretic Pragmatics*

Heather Burnett
Laboratoire de Linguistique Formelle
CNRS-Université de Paris-Diderot

Abstract

Understanding the dynamics that characterize interaction between conversational participants is a fundamental goal of most theories of socially conditioned language use and identity construction through language. In this paper, I outline a class of formal tools that, I suggest, can be helpful in making progress towards this goal. More precisely, this paper explores how *Bayesian signalling game* models can be used to formalize key aspects of current sociolinguistic theories, and, in doing so, contribute to our knowledge of how speakers use their linguistic resources to communicate information and carve out their place in the social world. The Bayesian framework has become increasingly popular for the analysis of pragmatic phenomena of many different types, and, more generally, these models have become a dominant paradigm for the explanation of non-linguistic cognitive processes. As such, I argue that this approach has the potential to yield a formalized theory of personal and social identity construction and to situate the study of sociolinguistic interaction within a broader theory of rationalistic cognition.

1 Introduction

Understanding the dynamics that characterize interaction between conversational participants is a fundamental goal of most theories of socially conditioned language use and/or

*This research has been partially supported by the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL), and a fellowship from the Center for the Study of Language and Information at Stanford University. I thank Leon Bergen, Olivier Bonami, Adrian Brasoveanu, Judith Degen, Fabio del Prete, Bernard Fradin, Chantal Gratton, Erez Levon, Eric McCready, Jessica Rett, Devyani Sharma, Elizabeth Smith, Sali Tagliamonte, Meredith Tamminga, audiences at UCL, Institut Jean Nicod, LLF Paris-Diderot, Stanford, UCLA, UCSC and *NWAV45*, and especially Penny Eckert and Dan Lassiter for very helpful comments and discussions. All errors are my own.

identity construction through language (see Goffman, 1961, 1967; Gumperz, 1982a,b; Bell, 1984, 1997; Giles et al., 1991; Ochs, 1993; Eckert and McConnell-Ginet, 1995; Bucholtz and Hall, 2005, 2008, among many others). In this paper, I outline a class of formal tools that, I suggest, can be helpful in making progress towards this goal. More precisely, this paper explores how *epistemic game theory* and, more specifically, *Bayesian signalling game* models can be used to formalize key aspects of current sociolinguistic frameworks, in doing so, contribute to our knowledge of how speakers use their linguistic resources to communicate information and carve out their place in the social world.

Linguistic communication and identity construction through language are extremely complex cognitive and social phenomena, and a lot of open issues in the study of language, variation and identity are very subtle. Formalization can be a very powerful tool for helping us carefully distinguish between different aspects of theoretical proposals and for precisely identifying empirical predictions made by competing analyses. This being said, in order for a mathematical approach to sociolinguistic interaction and identity construction to be helpful, we need to use a formalism that is appropriate for the type of data that we want to model, and it turns out that this is not a trivial matter. In fact, many mathematical approaches to the study of meaning allow contextual factors to play only a restricted role[1] and tend to study only the behaviour of the listener, not the speaker. As such, many formal frameworks are ill-equipped to capture the context-dependent interplay between conversational participants that lies at the heart of studies of interaction in sociolinguistics and linguistic anthropology.

Developing appropriate, mathematically precise frameworks for capturing the relation between language, meaning and use is a longstanding problem in linguistics. Already in her 1985[2011] paper *Feminism in Linguistics*, Sally McConnell-Ginet reflects on the supposed 'trade-off' between formal rigor and interactivity as follows (McConnell-Ginet, 2011, 64):

> Many critics would say that rigor in linguistics has been achieved at the price of rigor mortis. The radical operation required to 'isolate' the language system has killed it: formal rules and representations provide no insight into language as a human activity. The defense against this malpractice charge, of course, is to develop an account of the relation between abstract linguistic systems and the mental states and processes, social actions and cultural values, that infuse them with life.

In this paper, I propose that game theory gives us a way to answer McConnell-Ginet's challenge of providing 'vibrant' formal theories of linguistic communication. Indeed, the idea that language can be conceptualized as a game dates back at least to Wittgenstein (1953), and the proposal that concepts and mechanisms from game theory could be useful for analyzing language use has previously been explored by Goffman (1970); Bourdieu (1977); Myers-Scotton and Bolonyai (2001); Dror et al. (2013, 2014); Clark (2014), among others. In their seminal article on the potential of game theory to illuminate questions in

[1] See the discussions in Recanati (2004, 2010).

variationist sociolinguistics, (Dror et al., 2013, 562) say,

> There is a precedent for considering a link between linguistic practices and 'game theory' (Benz et al., 2005). A fair amount of work has been published over recent years considering linguistic diffusion from an evolutionary game theory perspective, including work on game theory and typology (Jäger, 2007), on the semantics of numbers (Jäger, 2012) and the pragmatics of 'epistemically lifted game' phrasing or number choice (Franke, 2009) [...], and on Gricean theories of 'pragmatics' (Jäger, 2008b, 2012).

This paper picks up where Dror et al. leave off, detailing how (what they call) *epistemically lifted game theory* has been used in the field of formal pragmatics to formalize Gricean theories of linguistic interaction, and explores how these recent developments in pragmatics might be fruitfully applied to the modelling of sociolinguistic interaction. More technically: I focus on **signalling game architecture** (Lewis, 1969) paired with a **probabilistic/Bayesian approach** to speaker/listener reasoning (see Oaksford and Chater, 2007, for an overview). This framework has become increasingly popular for the analysis of pragmatic phenomena of many different types, and Bayesian game-theoretic models (more generally) have become a dominant paradigm for the explanation of non-linguistic cognitive processes (to be discussed below). As such, I argue that such models have the potential to yield a formalized theory of identity construction through language and to situate the study of sociolinguistic interaction within a broader theory of rationalistic cognition[2].

In this article, I will mostly concentrate on building a model that can relate insights from formal semantics/pragmatics (and Bayesian cognitive science more generally) to the quantitative patterns of linguistic variation that are the main object of study of **quantitative/variationist** sociolinguistics (Labov, 1966, et seq.). This particular focus is motivated both by my own personal research interests and by the desire to help bridge the substantial theoretical and empirical gap that has historically existed between these areas[3]. This being said, it is my hope that sociolinguists working in more qualitative approaches might also find the topics discussed here to be useful, possibly as a way to provide a precise language into which to state theories concerning the relationship between language and the interactive process. Indeed, there has been at least a small amount of interest in formal modelling in the study of interaction in the traditions of Goffman and Gumperz: in his 1970 work *Strategic Interaction*, Goffman even develops his own class of game-theoretic models[4], called *expression games* (see Goffman, 1970, 11-46, for definitions), for captur-

[2]The focus of this work will be the construction of a mathematical model of identity construction and its place within the cognitive sciences. Readers interested in the relationship between social meaning and other kinds of pragmatic meaning within the game-theoretic perspective are referred to Burnett (2017).

[3]Fortunately, this gap is starting to close with recent works such as Acton and Potts (2014); Acton (2014); Beltrama (2016); Burnett (2017), and this work aims to contribute to this research programme in 'socio-semantics'.

[4]*Strategic interaction* was written after Goffman spent a sabbatical year at Harvard working with the Nobel prize winning game theorist Thomas Schelling (Manning, 1992). Thus, in addition to a contribution to the study of the dynamics of deception, Goffman's expression games also make a contribution to the

ing linguistic (and other) behaviour associated with deception and deception detection. Gumperz (1982a) also acknowledges the potential usefulness of (appropriate) formal modelling in developing more goal-oriented interactive theories of linguistic variation. Possibly evoking evolutionary game theory (Maynard Smith and Price, 1973), he says (p.29),

> There is a need for a sociolinguistic theory which accounts for the communicative functions of linguistic variability and for its relation to speakers' goals without reference to untestable functionalist assumptions about conformity or nonconformance to closed systems of norms. Since speaking is interacting, such a theory must ultimately draw its basic postulates from what we know about interaction. [...] Empirical methods must be found to determine the extent to which underlying knowledge is shared - perhaps through models of social aggregates patterned on modern theories of ecosystems, which specify constraints on interpretation and behavior but do not seek to predict what is actually used and how it is evaluated.

Thus, I suggest that Bayesian game-theoretic modelling is one of the kinds of empirical methods that can help us discover the constraints on interpretation and behaviour that characterize situations of communicative interaction, and therefore can be of interest to researchers working in a wide range of sociolinguistic traditions.

The paper is laid out as follows: In section 2, I give a general overview of game-theoretic models and a more detailed overview of the particular class of models that are commonly employed in formal pragmatics: **signalling games**. Then I outline some recent work in game-theoretic pragmatics that formalizes Gricean reasoning, focusing on *Iterated Best Response* (Franke, 2009) and *Rational Speech Act* (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013) models. I show how this framework has been applied to modelling the calculation of **scalar implicatures** and argue that, in these approaches, scalar reasoning shares many of the properties that have been proposed to characterize identity construction through language. These similarities, I propose, motivate their application to sociolinguistic phenomena. In section 3, I explore such an application through giving a formalization of the *Third Wave* approach to the meaning of variation (Eckert, 2000, 2008, 2012) in terms of Bayesian Signalling Game models, and I illustrate how this framework works through modelling two empirical studies: Gratton (2016)'s production study of non-binary people's use of (ING) (i.e. *work**ing*** vs *work**in'***) and Levon (2014)'s perception study of the relationship between listener gender stereotypes and the interpretation of high/low pitch in the speech of British men. Section 4 concludes.

field of game theory proper.

2 Game Theory and Bayesian Reasoning

Probably the most compelling argument in favour of using game theory to analyze the interactive and strategic aspects of sociolinguistic variation is simply that game theory is, in its essence, a mathematical formalism for describing situations of strategic interaction. In a nutshell, a **game** (as game-theorists define it) is composed of two basic parts[5]. The first component is the **architecture** of the situation of interaction itself: for a situation to be a game (in this technical sense), there must be at least two players. The players must interact and this interaction must result in a particular outcome which must depend on the choice of strategy of each player; that is, players' actions have to play some role in determining what happens. Finally, each player must have a preferential ordering over outcomes: they must prefer some things to happen more than other things. The value that players assign to particular outcomes and actions is called their **utility**. The second crucial component to a game is the **solution concept**: a rule or algorithm that determines how the game is played, i.e. what actions the particular players take.

As stated, this definition of a *game* is very abstract and broad. This reflects the fact that game-theoretic models have been used to analyze very many instances of interaction (both human and non-human) across the economic, biological, social and cognitive sciences[6]. However, when it comes to dealing with linguistic communication it makes sense to start looking at a much more narrow class of games: **signalling games** (Lewis, 1969). Informally speaking[7], in a signalling game there are two players: the **speaker** (S) and the **listener** (L). S knows a piece of (truth conditional) information that they want to communicate to L. L wants to learn the information that S is trying to communicate to them, and in order to help them to transmit their piece of information, S has a set of messages that they can choose to send to L. In formal pragmatics, we usually assume that messages are particular linguistic forms paired with semantic meanings. S's action is to pick a message to send to L, i.e to say something. Then L's action is to assign an interpretation to the message, i.e. to understand it in some way. The game thus has two outcomes:

1. **L interprets the message in the way that S intended**, and so they learn the information that S wanted to tell them. This outcome is good for S because they managed to communicate their information, and it's good for L because they got to learn the information that they were looking for.

2. **L doesn't interpret the message in the way that S intended**, and so they do

[5]For a more technical introduction, see Osborne and Rubinstein (1994). For linguistically oriented introductions, see Benz et al. (2005); Jäger (2011) (game theory+pragmatics) or Dror et al. (2013, 2014) (game theory+sociolinguistics).

[6]In addition to the cognitive processes discussed in this paper, game-theoretic analyses have been applied to an enormous range of topics from human and animal population dynamics and reproduction, to voting strategy, auctions and pricing, and economic and political bargaining

[7]Readers who are more formally inclined may enjoy the linguistic and philosophical introduction to signalling games in Franke (2009).

not learn the information. This outcome is bad for S since they didn't communicate their information, and furthermore it's bad for L since they didn't learn the fact that S was trying to tell them.

Since outcome 1 is preferred by both S and L (i.e. they both 'win') and outcome 2 is dispreferred by both S and L (i.e they both 'lose'), the signalling game is a game of **cooperation** (Schelling, 1960).

With this architecture in mind, we now turn to the solution concept: what determines which message the speaker will pick to try to communicate their desired piece of information and which meaning the listener will assign to S's message?

When it comes to linguistic communication, be it propositional communication or identity construction, a natural idea is that both the speaker and listener's actions will be largely determined by properties of their **beliefs** about their conversational partner and their **reasoning** about how their partner will act (see the discussion in Franke, 2009). A very influential recent idea in formal pragmatics is that the approach that we adopt to analyze human reasoning is the one that is found in the **Bayesian/probabilistic** approach to cognitive science (see Tenenbaum et al., 2011; Zeevat and Schmitz, 2015; Franke and Jäger, 2016, for recent overviews of Bayesian pragmatics). More specifically, as discussed in (Tenenbaum et al., 2011, 1279), the Bayesian approach to cognitive science can be summarized as a set of answers to the following questions concerning the nature of knowledge and cognition:

(1) a. How does abstract knowledge guide learning and inference from sparse data?
b. What forms does abstract knowledge take, across different domains and tasks?
c. How is abstract knowledge itself acquired?

The Bayesian answer to questions (1-a) and (1-c) is that learning and acquisition are products of **statistical inference**. More specifically, Bayesians propose that the fundamental rule of human reasoning is **Bayesian Inference**[8]: humans draw a conclusion B after having observed event A (we write this as $P(B|A)$, read as *the probability of B given A*) through combining two things:

1. How likely they think A is to indicate B (written $P(A|B)$, read *'the likelihood of A given B'*).
2. How likely they thought B was to begin with (written $Pr(B)$, read *'their prior belief that B is the case'*).

The Bayesian answer to (1-b) is that knowledge takes the form of rich, structured represen-

[8] This inference rule is laid out in its gory detail in (i).

(i) $$P(B_i|A) = \frac{Pr(B_i) \times P(A|B_i)}{\sum_{j=1}^{|B|} Pr(B_j) \times P(A|B_j)}$$

tations, such as phonological or syntactic tree structures (Chomsky, 1957; Chomsky and Halle, 1968, among many others), structured semantic representations (Link, 1983; Bach, 1986, among others), or even, as I will propose below, indexical fields (Eckert, 2008). This approach is therefore innovative because it allows for a **synthesis** of symbolic approaches to language (common in formal linguistics) with statistical, frequency-based approaches (common in more functionally oriented linguistics). As Tenenbaum et al. (2011) say (p.1279),

> Until recently, cognitive modelers were forced to choose between two alternatives (Pinker 1997): powerful statistical learning operating over the simplest, unstructured forms of knowledge [...] or richly structured symbolic knowledge equipped with only the simplest, non-statistical forms of learning, checks for logical inconsistency between hypotheses and observed data, as in nativist accounts of language acquisition (Niyogi, 2006). It appeared necessary to accept either that people's abstract knowledge is not learned or induced in a nontrivial sense from experience (hence essentially innate) or that human knowledge is not nearly as abstract or structured (as "knowledge-like") as it seems (hence simply associations).

By virtue of their generality, Bayesian models have found wide applications across the cognitive sciences, being used to model phenomena related to vision (Kersten and Yuille, 2003; Yuille and Kersten, 2006, among many others), memory (Shiffrin and Steyvers, 1997; Steyvers et al., 2006), sensorimotor systems (Körding and Wolpert, 2006), and, of course, language (see Chater and Manning, 2006; Lassiter and Goodman, 2013; Zeevat and Schmitz, 2015; Franke and Jäger, 2016, for overviews). As such, when we propose to use Bayesian game-theoretic models to analyze identity construction, we are making a particular proposal concerning the formalization of sociolinguistic theories and, at the same time, integrating the study of the identity construction process into the broader field of Bayesian cognitive science.

In order to understand how these models work, in the next section, we will see how they can be used to formalize a particular theory of linguistic interaction, **Gricean pragmatics** (Grice, 1975), and how they can be used to model a context-sensitive interactive phenomenon: *scalar implicature calculation*.

2.1 Gricean Reasoning in Bayesian GT Pragmatics

One of the principal pragmatic phenomena that has been treated in Bayesian game-theoretic pragmatics is **implicature calculation**: 'extra' inferences drawn by the listener that are triggered by the speaker's use of one linguistic form over another. For example, in many situations, if we hear a speaker say an utterance with *some*, such as (2-a), we will conclude the negation of the corresponding utterance with *all* (2-b).

(2) a. Mary ate **some** of the cookies.

b. $\rightsquigarrow$ Mary did not eat **all** of the cookies.

Although the inference (2-b) may seem automatic, there are reasons to think that this **implicature** is not directly encoded into the meaning of *some*. For instance, implicature calculation is restricted to certain linguistic environments, something that would be unexpected if the *but not all* inference was entailed by the literal meaning of *some*. In particular, if *some* is embedded within the antecedent of a conditional (3-a) or within a question (3-b), the *but not all* implicature is not drawn, i.e. it is not possible to answer 'yes' to (3-b) if you have eaten all the cookies.

(3) a. If you eat **some** of the cookies, I'll be angry.
$\not\rightsquigarrow$ If you eat some but not all of the cookies, I'll be angry.
b. Did you eat **some** of the cookies?
$\not\rightsquigarrow$ Did you eat some but not all of the cookies?

Instead of being an aspect of the literal meaning of *some*, meanings such as (2-b) are commonly proposed to arise through the combination of the speaker's action (choosing to use *some*, rather than *all*) and the listener's particular interpretation of that choice in the discourse context (Strawson, 1950; Grice, 1975; Levinson, 1983; Horn, 1989, among very many others). Thus, at a basic level, scalar implicature have been analyzed in pragmatics along the same lines as many researchers in linguistics, anthropology and philosophy have analyzed personal and social identity: not as a "a stable structure located primarily in the individual psyche or in fixed social categories", but as "a relational and socio-cultural phenomenon that emerges and circulates in local discourse contexts of interaction" (Bucholtz and Hall, 2005, 586).

The clearest way to see how this framework works is through an example: Suppose we have two agents: the speaker (S) and the listener (L). S and L baked three cookies, and then, while L was out, Mary stopped by and possibly ate some of them. Suppose that L calls the house and wants to know how many of the cookies Mary ate. What should S say and how should L understand what S says to them?

In this example, there are four possibilities (shown in Table 1): the situation (or *world*, in formal semantics terminology) in which Mary didn't eat any of the cookies (call this w_0); the situation in which she ate one cookie (w_1); the situation in which she ate two cookies (w_2); and the situation in which she ate all three of the cookies (w_3).

Suppose that S sees that Mary actually ate two of the cookies; therefore, S's wants to communicate that we are in w_2 in this example. In order to try to communicate this fact to L, S needs to pick a message. For simplicity, we will assume that S can choose from the three messages shown in Table 2:

For illustration, we will limit the messages in the model to those three. Of course, we could have also included the longer *Mary ate some but not all of the cookies* and the non-partitive

World	Description
w_0	Mary ate 0 cookies
w_1	Mary ate 1 cookie
w_2	Mary ate 2 cookies
w_3	Mary ate 3 cookies

Table 1 – Possible worlds

Short name	message	⟦message⟧
NONE	Mary ate none of the cookies	$\{w_0\}$
SOME	Mary ate some of the cookies	$\{w_1, w_2, w_3\}$
ALL	Mary ate all of the cookies	$\{w_3\}$

Table 2 – Messages in cookie example

Mary ate some cookies into the message set of the game. Indeed, we would include them in a comprehensive analysis of the use and interpretation of scalar quantifiers in English, but we will keep things simple here.

As shown in Table 2, messages are associated with semantic meanings, which we will take to be the sets of possible worlds/situations in which they are true. For example, *Mary ate none of the cookies* (NONE) is true only in one world: the one in which she eats zero cookies. On the other hand, *Mary ate some of the cookies* (SOME) is true in three worlds: every one except w_0.

The speaker's first step in choosing what to say is to make a hypothesis about their interlocutor's belief state concerning which cookies may (or may not) have been eaten. Suppose that S thinks that L has no prior expectations about how many cookies Mary ate. We can represent the listener's uncertain belief state through having their prior beliefs, Pr, be **uniform** over the set of possible worlds, as shown in Table 3.

w_0	w_1	w_2	w_3
0.25	0.25	0.25	0.25

Table 3 – L has uniform prior beliefs ($\Pr(w)$).

With L's prior beliefs in mind, S picks a message to say. Following Franke (2009) and Frank and Goodman (2012), when the listener hears a message m, the first thing that they do is restrict their attention to the worlds in which m is true. More technically, L conditions their beliefs on the meaning of the message, which is equivalent to intersection followed by renormalization of the measure. In other words, after hearing a message m, the listener 'zooms in' on the worlds in which the m is true, discards the ones in which m is false as impossible, and then re-adjusts their beliefs. Assuming L's prior beliefs are uniform (Table 3), L's beliefs immediately after hearing a message are shown in Table 4: after hearing NONE, L is certain that Mary ate zero cookies; after hearing ALL, L is certain that she ate

the three cookies; and after hearing SOME, L is certain that Mary did not eat zero cookies, but is equally uncertain about how many she ate.

Message	w_0	w_1	w_2	w_3
NONE	1	0	0	0
ALL	0	0	0	1
SOME	0	$0.33\dot{3}$	$0.33\dot{3}$	$0.33\dot{3}$

Table 4 – L's beliefs immediately after hearing m ($\Pr(w|m)$).

Following Grice, we assume that speakers aim to make the most informative statement possible, and informativity supplies a point for S and L to coordinate on (Lewis, 1969)[9]. These models formalize Grice's maxims of quantity[10] by building informativity into the speaker's *utility function* (U_S). S's utility function is a measure of how useful a message would be for S to communicate their desired piece of information to L. Following Frank and Goodman (2012), who follow Shannon (1948), the informativity of a message is measured as the natural log of the listener's beliefs conditioned on the meaning of the message, as shown in (4).

(4) Utility of a message m to communicate w ($U_S(m, w)$):
$U_S(m, w) = ln(Pr(w|m))$

In signalling games, speaker utility functions often also encode information associated with **costs** for different messages. For example, if we were also considering messages like *Mary ate some but not all the cookies*, we might want to assign it a penalty to reflect the fact that it is much longer than *Mary ate some of the cookies*. Likewise, if we are comparing *Mary ate some **of** the cookies* with *Mary ate some cookies*, we might want to penalize the non-partitive sentence, since listeners have been shown to be less likely to draw implicatures in the partitive construction than in the non-partitive (Degen, 2015). Since, in our small example, there is no major length or other grammatical difference between the messages under consideration, we will not incorporate message costs into the speaker utility function in this paper; however, message costs are an important way in which linguistic conditioning factors can be captured in game-theoretic pragmatics.

Thus, to generate measures of utility for each message, we plug the numbers in Table 4 into the equation in (4), which generates Table 5[11]. Messages whose denotation does not contain a world are assigned the utility $-\infty$ for communicating that world because $ln(0) = -\infty$. So, by virtue of the fact that the conditionalized probability of interpreting

[9]Note that, under this view, technically speaking, the listener does not have to be positively disposed to the speaker and actively wish to engage in some meaning making process with them; they just try to extract the most information from S's utterance as possible.

[10](Grice, 1975, 45): 1. Make your contribution as informative as is required (for the current purposes of the exchange). 2. Do not make your contribution more informative than is required.

[11]The values in Table 5 have been rounded to three decimal places where appropriate.

w_0 after immediately hearing SOME is 0 (see Table 4), the speaker's utility of using SOME to communicate w_0 is as low as it possibly can be: $-\infty$. As shown in Table 5, the only useful message for communicating w_0 is NONE (since $ln(1) = 0$); the only useful message for communicating w_1 and w_2 is SOME (since $ln(0.33\dot{3}) \approx -1.099$); and the most useful message for communicating that Mary ate the three cookies is ALL (because $0 > -1.099$).

Message	w_0	w_1	w_2	w_3
NONE	0	$-\infty$	$-\infty$	$-\infty$
ALL	$-\infty$	$-\infty$	$-\infty$	0
SOME	$-\infty$	-1.099	-1.099	-1.099

Table 5 – S's utility for m for communicating w ($U_S(m, w)$).

One of the great benefits of Bayesian game-theoretic models is that they can be used to make gradient quantitative predictions concerning linguistic production and interpretation. This arises under the hypothesis that speakers are **approximately rational**: they are trying to make the choice that will have the best chance of accomplishing their goals (whatever they may be) (see Anderson, 1991, among others), but they may not always pick the optimal action. That is to say, if human speakers were fully rational, we would always expect them to perform the action that has the highest utility. So, if S wishes to communicate that we are in w_1 or w_2, we would expect them to pick SOME 100% of the time; whereas, if they wish to communicate w_3, then we expect them to pick ALL 100% of the time (never SOME). However, we know that human mental computation can be impeded by a variety of time/resource constraints (fatigue, working memory etc.), and, as discussed above, scalar implicature generation and calculated has been observed to be variable. So to account for variability in action selection, we assume that the speaker chooses which message to say non-deterministically: using the *Soft-Max* choice rule (Luce, 1959; Sutton and Barto, 1998)[12]. With this rule, the probability of making a choice increases as the utility of the choice increases, and the amount of variability is governed by the values of a parameter α, called the *temperature*. A parameter in a model is a particular number that is chosen for a particular dataset, which contributes to determining the model's predictions for that dataset. In this context, we use α to encode how much inherent variability we think there is in the system that we are studying. When α is set to ∞ (infinity), the speaker picks the message with the highest utility 100% of the time (so no inherent variability); whereas, anything lower than ∞ will predict some variability, even if it is just a tiny amount. The lower the value of α is set, the more variable the speaker's choice will be. The basic structure of the model therefore predicts a **range** of probabilities for the use of a variant; whereas, exact probabilities will depend on which number for α is selected. For

[12]The Soft-Max choice rule is written out formally as follows: For a world w, a message m and a real number α, called the *temperature*,

$$P_S(m|w) = \frac{exp(\alpha \times U_S(w, m))}{\sum_{m' \in M} exp(\alpha \times U_S(w, m'))}$$

example, the predicted probabilities of using NONE, SOME vs ALL to communicate that we are in w_3 for values of α less than 12 are shown in Figure 1. Note that once α gets higher than around 6, choice becomes almost deterministic in this model. In particular, as α gets larger than zero, the probability of using ALL sharply rises and the probability of using SOME sharply decreases. NONE is never predicted to be used.

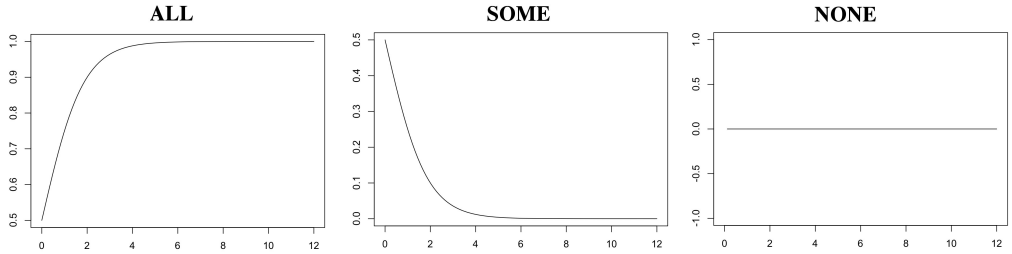

Figure 1 – Predicted probabilities of using ALL vs SOME vs NONE to communicate w_3, by α

When we are modelling actual quantitative studies, the value for α that best fits the observed data can be estimated (as in Goodman and Stuhlmüller, 2013; Franke and Degen, 2016); indeed, in the next section, we will choose values for α that best fit the quantitative sociolinguistic variation data that we will try to capture. However, in order to exemplify how the scalar implicature model works, I will pick a value for α that will allow for some variation: $\alpha = 4$. Plugging the utility values in Table 5 and $\alpha = 4$ into the Soft-max choice rule generates the probability distributions over messages shown in Table 6. If S wants to communicate w_0, then the model predicts that they will categorically say *Mary ate none of the cookies*. If S wants to communicate w_1 or w_2, then the model predicts that they will categorically say *Mary ate some of the cookies*. However, if the speaker wants to communicate w_3, the model predicts variable language use: S will say *Mary ate all of the cookies* 99% of the time, but *Mary ate some of the cookies* a highly disfavoured 1% of the time.

Message	w_0	w_1	w_2	w_3
NONE	1	0	0	0
ALL	0	0	0	0.99
SOME	0	1	1	0.01
Prediction	Cat. NONE	Cat. SOME	Cat. SOME	Favored ALL

Table 6 – S's predicted use of m, given w with $\alpha = 4$ ($P_S(m|w)$).

Finally, listeners interpret messages using their hypotheses that speakers are (approximately) rational and motivated by informativity, combined with their prior beliefs. In other words, as discussed in the previous section, we treat linguistic interpretation as Bayesian inference, and, from the values in Table 6 (the **probability that S would use** m **given** w) and the values in Table 3 (L's **prior beliefs concerning** w), we derive a probabil-

ity distribution over interpretations of messages, as shown in Table 7[13]. Crucially, Table 7 should be read from the listener's perspective, inverted from Table 6 which is from the speaker's perspective. For example, Table 6 indicates that the probability of S using ALL to communicate w_3 is 0.99; whereas, the same cell in Table 7 indicates that L has a probability of 1 (100%) of interpreting w_3 if they hear ALL.

Message	w_0	w_1	w_2	w_3	PREDICTION
NONE	1	0	0	0	Categorical w_0
ALL	0	0	0	1	Categorical w_3
SOME	0	0.498	0.498	0.005	Favoured w_1, w_2

Table 7 – L's predicted interpretation of w, given m ($P_L(w|m)$).

In particular, the model predicts that if L hears NONE, then they will categorically understand that Mary ate zero cookies. However, if L hears SOME, then we predict that they will largely hesitate between w_1 and w_2, with a very tiny probability assigned to w_3. In other words, the model predicts that if the speaker says *Mary ate some of the cookies*, the listener will almost certainly understand that Mary ate some but not all of the cookies. Hence, the model predicts the (variable) *but not all* implicature.

Note that the exact probability distributions shown above are dependent on parameters of the model. In particular, for exposition, we made the assumption that L started off with no prior expectation concerning how many cookies Mary would eat. However, suppose that L knows that Mary usually likes to have two cookies for her dessert. So, before hearing what S has to say, they are expecting Mary to have eaten two cookies. This belief can be represented through changing L's prior probability distribution to one that is heavily weighted on w_2, as shown in Table 8.

w_0	w_1	w_2	w_3
0.1	0.1	0.7	0.1

Table 8 – L's priors heavily weighted on w_2.

In this case, L's interpretation probabilities change, and L's probability of interpreting w_2 after hearing SOME jumps up to **0.87** from 0.498. This is because, as discussed above,

[13]The numbers in Table 7 have been rounded to two decimal points. Calculating these probability distributions by hand can be tedious. Fortunately, to facilitate calculations and prediction testing, a number of computational implementations of the Rational Speech Act model have been developed that the interested reader might find helpful:

1. Potts' implementation in python: `https://github.com/cgpotts/pypragmods`.
2. Goodman and Tenenbaum's implementation in Church: `https://probmods.org/`.
3. Goodman and Stuhlmüller's implementation in WebPPL: `http://dippl.org/examples/pragmatics.html`.

Bayesian interpretation takes L's prior beliefs into account. Since L's prior probability distribution assigns w_2 a probability of 0.7 and w_1 only a probability of 0.1, L will think that it is much more likely that S is trying to tell them that we are in w_2 than in w_1 if they hear SOME. In other words, in these models, aspects of listener's prior beliefs and expectations can dramatically influence how they interpret the linguistic expressions that S offers to them.

So far in this paper, we have seen that Bayesian game-theoretic models provide a framework for capturing instances of interactive co-construction of propositional meaning, as exemplified by scalar implicature calculation. Observe that, in this framework, the *not all* inference is not part of the semantic meaning of SOME. Rather, this implicature arises as a product of coordination between the speaker and listener based on reasoning about a set of **scalar alternatives** (NONE vs SOME vs ALL). In the rest of the article, we will explore how Bayesian game-theoretic models can be applied to capturing the linguistic co-construction of personal and social identity as a product of speaker-listener coordination based on reasoning about other kinds of alternatives: **sociolinguistic variants**.

3 Identity Construction in GT Pragmatics

This section presents a new formal model of identity co-construction based on the Bayesian signalling game models commonly used in game-theoretic pragmatics defined above. I define the models and then show how they can be applied to the analysis of two empirical studies of sociolinguistic variation and interpretation. But first, I identify a number of empirical generalizations taken from the literature on sociolinguistic perception/interpretation that, I argue, a formal model of socially conditioned variation and identity construction through language should capture.

3.1 Social Meaning, Variation and Identity Construction

One of the most common ways that social meaning and sociolinguistic interpretation has been investigated is through the use of a particular experimental paradigm called the *Matched Guise Technique* (MGT) (Lambert et al., 1960). This is an experimental method widely used in social psychology and more recently in variationist sociolinguistics to assess listeners' implicit attitudes towards speakers of different linguistic varieties. In this paradigm, participants listen to samples of recorded speech that have been designed to differ in specific and controlled ways. Each participant hears one of two recordings of the same speaker (called *guises*) which differ only in the alternation studied. In this way, we know that any significant differences between guises that we may observe are due to the linguistic forms under study, and not to something else (content, other aspects of voices etc.). After hearing a recording, participants' attitudes towards the recorded speaker are assessed, usually via focus groups and/or questionnaires.

In a series of papers on the social interpretation of variable (ING) (5), Campbell-Kibler (2006, 2007, 2008) reports on a MGT study that she performed with 124 participants using stimuli formed from the speech of 8 different speakers.

(5) a. I'm work**ing** on my paper. *-ing*
b. I'm work**in'** on my paper. *-in'*

Given its size, this study yielded a variety of complex patterns. Among them was that all speakers were rated as significantly more *educated* and more *articulate* in their *-ing* guises than in their *-in'* guises. Likewise, speakers were significantly more likely to be described as a *redneck* in their *in'* guises than in their *-ing* guises. Finally, one male speaker, who she calls *Jason*, is significantly more likely to be described as *gay* when he says *-ing* than when he says *in'*.

Similarly, Podesva et al. (2015) performed a MGT study with 70 participants investigating the interpretation of /t/ release (6) using stimuli formed from political speeches of 6 American politicians (Barak Obama, John Edwards, Nancy Pelosi, George W. Bush, Hillary Clinton, and Condoleezza Rice).

(6) /t/ release
a. wa[t^h]er released /t/
b. wa[ɾ]er flapped /t/

As in Campbell-Kibler's study, the /t/ release study yielded a number of results concerning associations with released vs flapped/unreleased /t/: for example, John Edwards and Condoleezza Rice were rated as significantly more *articulate* in their released /t/ guises than in their flapped guise. On the other hand, Nancy Pelosi was rated as significantly less *friendly* and less *sincere* when she used released /t/. Thus, the speaker's choice of linguistic form can often make a large difference in how they are perceived by the listener.

This being said, which exact property attributions a particular variant will trigger often depends on which other properties the listener believes hold of the speaker. This can already be seen in Campbell-Kibler and Podesva et al.'s MGT studies: Podesva et al. found significant relationships between articulateness and released /t/ with Edwards and Rice, but only with these two speakers. Likewise, flapping/not releasing made only Nancy Pelosi sound more friendly and sincere. Finally, in Campbell-Kibler (2007), only Jason was heard as gay when he said *-ing*; (ING) made no differences to interpretations of sexual orientation with the other speakers. Thus, listener prior beliefs concerning individual speakers appear to constrain sociolinguistic interpretation.

Furthermore, we have evidence that listener prior beliefs concerning groups of speakers, i.e. *stereotypes*, also influence interpretation. This can be clearly seen in Levon (2014) which reports on a MGT study of the interpretation of styles involving higher ([+raised]) vs

lower ([-raised]) pitch in the speech of British men, among other variables. In addition to the standard MGT response questionnaire, Levon had participants fill out the *Male Role Attitudes Survey* (MRAS (Pleck et al., 1994)). This instrument measures agreement with a series of statements corresponding to 'traditional' male gender norms. In other studies, higher scores on the MRAS have been shown to correlate with higher rates of homophobia, church attendance, promiscuity and other behaviours. Again, Levon (2014)'s study generated a lot of different empirical results; however, the one which is pertinent for this paper is that participants separated into two groups[14]: listeners scoring high on the MRAS attributed more **incompetent** and less **masculine** personae to speakers using [+raised] pitch than [-raised] pitch, and listeners scoring low on the MRAS attributed more incompetent personae to speaker using [+raised] pitch, but there was no difference in masculinity. Thus, listeners' prior ideological beliefs concerning the relationship between masculinity and (in)competence play a role in determining what kinds of personae/identities they attribute to speakers based on their linguistic performances.

In summary, we see from the literature on sociolinguistic perception that hearers make judgments about the properties that characterize speakers based on the linguistic forms that they use; however, social interpretation is crucially constrained by listener prior beliefs. This being said, interpretation is only one side of the coin when it comes to sociolinguistic interaction, and we have reason to believe that speakers strategically exploit listeners' interpretation processes to communicate properties about themselves and, in doing so, construct their identities.

A clear illustration of this phenomenon comes from Kiesling (1998)'s study of the use of (ING) by nine college fraternity members. Among other contexts, Kiesling made recordings of the boys socializing and participating in an organizational meeting. He found that many of the fraternity members displayed a higher rate of *-in'* in the informal socializing context (75% *-in'* on average) than in the more formal meeting context (47% *-in'*, (Kiesling, 1998, 76)). For instance, one of the boys, Mack, uses the *-in'* form **73%** of the time in socialization; however, in the meeting his rate of *-in'* drops to **13%**. This being said, some fraternity members do not decrease their use of *-in'* in the meeting context like Mack does: another boy, Speed, moves only from **95%** *-in'* to **82%**, a non-significant difference. According to Kiesling, both Mack's change across contexts and Mack and Speed's divergences are due to the fact that the different fraternity boys are aiming to construct different identities in different contexts. While both boys adopt a casual persona in socialization, Speed and Mack are constructing different kinds of powerful masculine personae at the meeting. Following Ochs (1992), among others, Kiesling proposes that the *-in'* form *indexes* (i.e. is related to) properties such as *casualness, physical masculinity*[15] and an *anti-establishment* stance; whereas, *-ing* indexes *formality, non-physical masculinity/non-masculinity* and a *pro-establishment* stance. Based on analysis of Speed's speech, Kiesling proposes that Speed's identity at the meeting is "a rebel creating a powerful stance vis-a-vis

[14]This is the formulation of the result in Burnett and Levon (2016).

[15]See Connell and Connell (2005).

the ability power hierarchy in the fraternity that rewards hard work, rather than structural power for its own sake" (Kiesling, 1998, 86), which explains why Speed favours the *-in'* form in this context. Mack, on the other hand, aims to construct the persona of the "leader who knows what is good for his flock" (Kiesling, 1998, 92). Kiesling argues that there is "strong evidence that Mack, who used less [*-in'*] in the meeting than while socializing, is indeed displaying a meeting identity by indexing structurally powerful alignment roles, roles that are also associated with the use of the [*-ing*] variant" (Kiesling, 1998, 92).

Finally, we also have reasons to believe that speakers are sensitive to what their interlocutors think about them, and that these expectations can influence which linguistic forms they use. This can be seen clearly in Gratton (2016)'s recent study of the link between (ING) and gender presentation across contexts. Gratton conducted group interviews with non-binary individuals, that is, individuals whose gender identity does not respect the male/female binary. She focused on two people: Flynn, who was assigned female at birth, and Casey, who was assigned male at birth. Gratton conducted two sets of interviews with these consultants: the first one was in a queer-friendly environment (their home and a queer coffee shop, respectively) and the second one took place in a public coffee shop. Gratton found that neither Flynn nor Casey show a significant difference in their use of *-in'* vs *-ing* in the queer-friendly environments: **44%** and **58%** *-ing* respectively. However, in the public coffee shop, Flynn, who was assigned female at birth, uses significantly more *-in'* (**80%**), while Casey, who was assigned male at birth, uses significantly more *-ing* (**89%**). Based on ethnographic analysis of her interviews, Gratton argues that these patterns are created by variation in how her speakers evaluate their listeners' beliefs. She says (p.6):

> The individuals in this community believe that in queer environments, they can be read as non-normative quite easily, which means, according to them, that they do not need to consciously worry about their gender presentation. However, the same cannot be said for non-queer public contexts. They believe that individuals whom they encounter in non-queer public spaces will pre-suppose a binary gender based mainly on their physiological characteristics. In order to present a non-binary or non-normative gender identity, they must distance themselves from the gender which is presumed–their gender-assigned-at-birth– by utilizing resources which resist the gender norms associated with their respective gender-assigned-at-birth.

In summary, in this section, I outlined a number of properties that, based on research in sociolinguistics, a formal model of identity construction should capture. I argued that we want a framework that can predict variable, quantitative patterns of variation and interpretation, and that can model interaction between conversational participants such that the speaker (tries to) choose the variant that has the best chance to construct their desired persona. Furthermore, we want a model in which the listener's prior beliefs and ideologies constrain interpretation. I highlight that these are many of the same properties that we saw characterize scalar implicature calculation. Therefore, I propose that it is reasonable to extend the game-theoretic models outlined in the previous section to model

the strategic aspect of sociolinguistic variation.

3.2 Formalization of the Third Wave

In the construction of our formal model, we will build on the large amount of previous work theorizing about identity construction through language in sociolinguistics, linguistic anthropology and sociocultural linguistics more generally (see Bucholtz and Hall, 2005, 2008). In particular, we will adopt the *indexicality* theory of social meaning in which abstract mental representations mediate the relationship between language and identity categories (Ochs, 1992, 1993; Silverstein, 1976, 1979, 2003; Eckert, 2008, among others), and, more specifically, how indexicality is developed and related to patterns of language use within *Third Wave* approach to variation (see Eckert, 2012, for an overview).

In the model, as in a classic signalling game, there are two players: the speaker (S) and the listener (L). Third Wave variation theory focuses on how variants combine together to form styles, which construct particular identities or personae (see Podesva, 2004; Eckert, 2008; Zhang, 2008, among many others). In this paper, we will take personae to be particular collections of properties that 'go together'. For our empirical illustration, we will construct personae from the set of properties in (7), where we understand *masculine/feminine* to broadly regroup various kinds of masculinities/non-femininities and non-masculinities/femininities respectively (see Cameron and Kulick, 2003; Eckert and McConnell-Ginet, 2013, among others).

(7) {competent, incompetent, casual, delicate, masculine, feminine}

We also suppose that there are some ideological relations[16] between these properties: for example, suppose that one cannot be both *competent* and *incompetent* at the same time; nor can one be *casual* and *delicate* at the same time, or *masculine* and *feminine* at the same time. So, given this setup, the set of personae generated from (7) is shown in (8).

(8) Set of personae
1.{competent, casual, masculine}
2.{competent, casual, feminine}
3.{competent, delicate, masculine}
4.{competent, delicate, feminine}
5.{incompetent, casual, masculine}
6.{incompetent, casual, feminine}
7.{incompetent, delicate, masculine}
8.{incompetent, delicate, feminine}

[16]See Burnett (2017) for more formal definitions.

3.2.1 (ING) and gender identity construction

As in signalling games, we will have a set of messages that the speaker can pick from to try to construct their desired personae. As an illustration, we will first consider Gratton (2016)'s study of (ING) and gender identity construction across contexts. The messages for this game are shown in (9).

(9) Messages = {*-ing*, *-in'*}

In Third Wave variation theory, individual variants have meaning that goes beyond their truth conditional meaning. In particular, variants index sets of properties/stances, called their *indexical field* (Eckert, 2008). Following (simplified) Eckert (2008), we propose the indexical fields for (ING) shown in Table 9.

Variant	Eckert field
-ing	{competent, delicate}
-in'	{incompetent, casual}

Table 9 – Eckert indexical fields for (ING)

The representations in Table 9 show the most standard approach to the representation of indexical fields. This being said, the 'Eckert' fields shown will not be exactly the right kinds of objects to be integrated into the larger model. Instead, in the spirit of Montague (1973), we will adopt an **equivalent** characterization indexical fields as the set of personae that they have the potential to construct. The relation between more traditional Eckert fields and equivalent Eckert-Montague fields is shown in Table 10. Observe that, although both variants have the potential to construct competent and casual personae (for example), only *-ing* can construct competent and delicate personae, while only *-in'* can construct incompetent and casual personae.

Variant	Eckert field	Eckert-Montague field
-ing	{competent, delicate}	{comp., delicate, masc.}, {comp., delicate, fem.}, {comp., casual, masc.}, {comp., casual, fem.}, {incomp., delicate, masc.}, {incomp., delicate, fem.}
-in'	{incompetent, casual}	{incomp., casual, masc.}, {incomp., casual, fem.}, {comp., casual, masc.}, {comp., casual, fem}, {incomp., delicate, masc.}, {incomp. delicate, fem}

Table 10 – Messages and Indexical Fields

As in the scalar implicature example, the speaker makes a hypothesis about the listener's prior beliefs concerning which persona(e) they instantiate[17]. Thus, we represent the listener's prior beliefs as a probability distribution over personae (Pr). Pr can encode specific

[17]This can be seen as a way of encoding of *audience design* (Bell, 1984) into the model.

beliefs about the particular speaker (*Mary has X properties*) or more general stereotypes (*Canadian women have X properties*). For example, suppose listeners think that *delicateness* and *femininity* are ideologically linked (Ochs, 1992), as are *casualness* and *masculinity*. This can be represented in the listener's prior beliefs as there being much higher probability mass on personae that have the combination {casual, masculine} and {delicate, feminine} than on {casual, feminine} and {delicate, masculine}. A prior distribution that encodes these ideological relations is shown in Table 11: the persona {competent, casual, masculine} has a higher weight than the persona {competent, delicate, masculine} (0.1 vs 0.05), and the same holds for {incompetent, casual, masculine} compared to {incompetent, delicate, masculine}.

In her study of Flynn (the non-binary individual who was assigned female at birth), Gratton (2016) argues that, in the public coffee shop, they are worried that their interlocutor will incorrectly attribute them a feminine persona. We also represent this specific belief in the model through putting more probability mass on feminine personae than on masculine personae in the listener's prior beliefs, as shown in Table 11. In other words, if we add up the probabilities on feminine personae in Table 11, we see that Flynn thinks that their coffee shop interlocutor assigns them 0.7 probability of being some feminine persona.

persona	Pr(persona)
{competent, casual, masculine}	0.1
{competent, casual, feminine}	0.1
{competent, delicate, masculine}	0.05
{competent, delicate, feminine}	0.25
{incompetent, casual, masculine}	0.1
{incompetent, casual, feminine}	0.1
{incompetent, delicate, masculine}	0.05
{incompetent, delicate, feminine}	0.25

Table 11 – L's prior beliefs (Flynn/coffee shop)

As in the models for implicature calculation presented in section 2, we assume that the first thing that the listener does when they hear a variant is to restrict their attention to the personae that appear in the variants' Eckert-Montague field, and adjust their prior beliefs accordingly. For example, if they hear the variant *-ing*, they discard the possibility that the speaker is incompetent and casual (i.e. they assign 0 probability to {incompetent, casual, masculine} and {incompetent, casual, feminine} personae). Likewise, if they hear *-in'*, they assign 0 probability to {competent, delicate, feminine} and {competent, delicate, masculine} personae.

In parallel with scalar implicature model, we assume that speaker utility is guided by informativity, and, for simplicity, we will assume that there are no cost differences between variants[18]. So, transposed into the identity construction model, speaker utility (U_S) for a

[18]Again we could also incorporate message costs into the model as a way of capturing both social and

variant m to construct a persona P is given by (10): the utility for the speaker to use a variant m, given that they wish to construct persona P, is the natural log of the probability of P conditioned on (i.e. taking into account) the indexical fields of m.

(10) $\quad U_S(P, m) = ln(Pr(P|m))$.

Following Gratton, we assume that Flynn wants to construct a non-feminine persona (for example, a {competent, casual, masculine} persona). So in exactly the same way that we calculated utilities for SOME, NONE and ALL in the previous section, we plug the values in Table 11 into the speaker utility equation in (10), which assigns a utility of -1.946 for Flynn to use *-in'* and a utility of -2.079 to use *-ing*. We can then take these utilities and plug them into the Soft-Max choice rule described in the previous section. If we do so, and we set the temperature $\alpha = 10$ for this dataset, we predict that they will use *-in'* **79**% of the time, which is similar to what Gratton found[19].

Thus, given a particular proposal concerning a speaker's analysis of their communication situation (who they want to be and who others think they are), we predict which variant the speaker will be most likely to use. Note crucially that the framework does not assume that all or even most aspects of message/interpretation selection or utility calculation are conscious or intentional. Indeed, as discussed in section 2, one of the core proposals of this paper is that identity construction arises from the same principles and mechanisms as other kinds of cognitive activities. As mentioned above, Bayesian models have been shown to be useful in the analysis of cognitive processes like vision, perception, memory and motion planning, which are all activities that feature very little conscious awareness. So it by no means follows that because it is appropriate to model aspects of cognition as instances of rationalistic decision making, all such decision making must be operating above the level of consciousness (seeDennett (1993); Graziano (2013) for general discussion and Burnett (2017) for more discussion of this point in the context of sociolinguistic variation and identity construction).

For Casey, who was assigned male at birth, the listener prior beliefs look different: Casey is worried about being attributed a masculine persona in the coffee shop. So, although the general ideological relationships between masculinity, casualness, femininity and delicateness stay the same in Table 12, Casey's representation of their listener's prior beliefs about them is highly weighted on masculine personae.

linguistic conditioning factors. For example, we know that (ING) is conditioned by grammatical category and other abstract properties of morphological structure (Labov, 1966; Houston, 1985; Tamminga, 2014). However, in order to do this properly, we would need more complicated message representations. So the unification of social and linguistic factors within Bayesian game-theoretic pragmatics is left to future research.

[19] These calculations can, again, be tedious to do by hand. Therefore, a computational implementation in python of the model used in this paper (done in collaboration with D. Lassiter) is available for interested readers at `https://github.com/hsburnett/smg`.

persona	Pr(persona)
{competent, casual, masculine}	0.25
{competent, casual, feminine}	0.05
{competent, delicate, masculine}	0.1
{competent, delicate, feminine}	0.1
{incompetent, casual, masculine}	0.25
{incompetent, casual, feminine}	0.05
{incompetent, delicate, masculine}	0.1
{incompetent, delicate, feminine}	0.1

Table 12 – L's prior beliefs (Casey/coffee shop)

Based on these priors, if we suppose that Casey wishes to construct a feminine persona (for example {competent, delicate, feminine}), the model predicts that they will categorically use *-ing*[20]. So given different listener prior beliefs and different desired identities, we predict drastically different patterns of variation.

Of course, speakers' assessments of how they are perceived can change in different situations. Suppose, as described by Gratton, they are fairly certain that they will be attributed their desired persona ({competent, casual, masculine} for Flynn) when they are in a queer friendly environment (Table 13). In this case, particular choices of variants will have less of an effect, and the model predicts that Flynn will use *-in'* **50%** of the time (rather than **79%** in the public setting), with $\alpha = 10$.

persona	Pr(persona)
{competent, casual, masculine}	0.81
{competent, casual, feminine}	0.01
{competent, delicate, masculine}	0.01
{competent, delicate, feminine}	0.05
{incompetent, casual, masculine}	0.05
{incompetent, casual, feminine}	0.01
{incompetent, delicate, masculine}	0.01
{incompetent, delicate, feminine}	0.05

Table 13 – Listener's prior beliefs (Flynn/queer friendly)

More generally, the predicted probabilities for Flynn's language use in the public vs queer-friendly contexts are depicted in Figure 2. This figure shows the model's predicted probabilities with all values of the parameter α less than 50. Recall that lower values for α encode more inherent variability in the system, while higher values make the speaker's choice more deterministic.

[20]In order to capture variation in Casey's use of (ING) in the coffee shop, we could attribute Casey's *-in'* to linguistic/grammatical conditioning and/or adopt a more complicated model which allows some variation in persona selection. See Burnett (2017) for such a model.

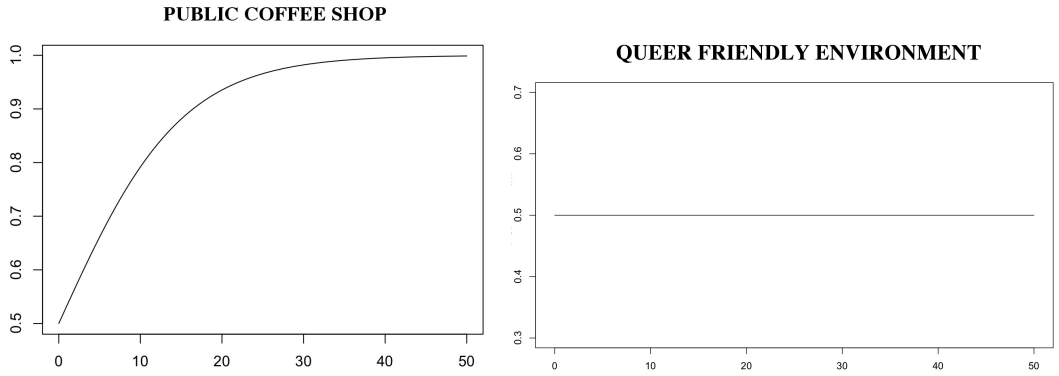

Figure 2 – Predicted probabilities for Flynn's use of *-in'* across contexts for $\alpha \leq 50$

Thus, the models capture the observation that speakers' different representations of listeners' beliefs can have an large effect on their language use.

Up till now, I have been arguing that Bayesian game-theoretic models of the type we have been discussing can formally express the link between observed patterns of language use and certain proposals within sociolinguistics and linguistic anthropology (such as those of Ochs, Eckert, Campbell-Kibler and Gratton), which previously had only been stated at an intuitive level. However, in addition to helping us 'pin down' some of the insights of existing proposals, I argue that game-theoretic modelling can also be a tool for **testing** hypotheses concerning difficult theoretical questions like 1) which indexical fields to assign to which variants, and 2) what speaker ideologies look like. For example, we might wonder whether there are other analyses of the indexical fields of (ING) besides the one proposed in Table 9 that are consistent with Gratton's results. For instance, perhaps the proposal in Table 9 is unnecessarily complex and we could do just as well with the simpler analysis in Table 14 where the more 'standard' form *-ing* indexes *competence* and the 'vernacular' form *-in'* indexes *incompentence*.

Variant	Eckert field
-ing	{competent}
-in'	{incompetent}

Table 14 – Eckert indexical fields for (ING) (competing analysis)

In this case, if we keep the same set up as above (i.e. we suppose that Flynn's desired persona is {competent, casual, masculine} and that their assumed listener priors are as in Table 11), we make the prediction that Flynn will use *-ing* **100%** of the time under all values of α, which is clearly very different from what Gratton found. Thus, Bayesian game-theoretic models provide a way of constructing direct **empirical** arguments in favour of one analysis of the indexical field over another.

3.2.2 Pitch and gender identity construction

Different listener prior beliefs can also produce different sociolinguistic interpretations in the model. To see this, we can consider the Levon (2014) example discussed in section 3.1. Recall that Levon found a distinction between two groups of listeners: *progressives*, who had no particular (diagnosable) stereotypes concerning male behaviour, and *conservatives*, whose beliefs conformed to particular 'traditional' male gender norms. Furthermore, Levon found that differences in male stereotypes translated into differences in interpretation of pitch: conservatives assigned incompetent and non-masculine personae to speakers using raised pitch; whereas, progressives assigned incompetent and both masculine and non-masculine personae to raised pitch (Burnett and Levon, 2016).

This pattern is predicted by the model: in this game, we have two messages (11), and I propose that [+pitch raised] indexes incompetence, while [-pitch raised] indexes competence (Table 15).

(11) Messages = {[+pitch raised], [-pitch raised]}

Variant	Eckert field	Eckert-Montague field
[+pitch raised]	{incompetent}	{incomp., delicate, masc.}, {incomp., delicate, fem.}, {incomp., casual, masc.}, {incomp., casual, fem.}
[-pitch raised]	{competent}	{comp., casual, masc.}, {comp., casual, fem.}, {comp., delicate, masc.}, {comp. delicate, fem}

Table 15 – Messages and Indexical Fields (Levon 2014)

Given Levon's result associated with the MRAS, it is reasonable to suppose that a main difference between progressives and conservatives lies in which kinds of personae they expect to encounter in the world: progressives in Levon's study have no beliefs about how incompetence and non-masculinity/femininity cluster together; therefore, we represent their beliefs as a uniform probability distribution over personae. On the other hand, conservatives in Levon's study think that incompetence is emasculating. We represent this belief in the model as one that it is highly unlikely that they will encounter an incompetent, masculine person (see Table 16).

Based on these prior beliefs (with α again set at 10), the model correctly predicts that speakers using [+pitch raised] will be more likely to be attributed non-masculine personae by conservative listeners; whereas, it is predicted that there will be no such gender-based asymmetry for progressives, as shown in Table 17.

Thus, these formals models can also be useful for capturing the relationship between ideological structure and sociolinguistic interpretation.

persona	Pr(persona)
{competent, casual, masculine}	0.15
{competent, casual, feminine}	0.15
{competent, delicate, masculine}	0.15
{competent, delicate, feminine}	0.15
{incompetent, casual, masculine}	0.05
{incompetent, casual, feminine}	0.15
{incompetent, delicate, masculine}	0.05
{incompetent, delicate, feminine}	0.15

Table 16 – Conservatives' prior beliefs (Pr(persona)) in Levon (2014)

persona	progressives	conservatives
{competent, casual, masculine}	0	0
{competent, casual, feminine}	0	0
{competent, delicate, masculine}	0	0
{competent, delicate, feminine}	0	0
{incompetent, casual, masculine}	0.25	0.13
{incompetent, casual, feminine}	0.25	0.37
{incompetent, delicate, masculine}	0.25	0.13
{incompetent, delicate, feminine}	0.25	0.37

Table 17 – Probability distribution over interpretations [+ pitch raised]

4 Conclusion

In this paper, I gave an exploration of how formal tools commonly used in the growing field of game-theoretic pragmatics could be applied to the analysis of socially conditioned variable linguistic phenomena and identity construction through language. Using a Bayesian game-theoretic framework inspired by the *Iterated Best Response/Rational Speech Act* models, I showed how we can formalize sociolinguistic theories, in this case, *Third Wave Variation* theory, in order to make both qualitative and quantitative predictions about variable language use and interpretation across different contexts and different kinds of speakers. I argued that, unlike many existing mathematical semantic/pragmatic frameworks, these models can capture the interactive co-construction of meaning that forms the basis of how we establish our place in the social world. In particular, inferences associated with properties of the speaker, such as (12-b), are not proposed to be directly encoded into the message, but rather arise as the product of (un)conscious coordination between the speaker and the listener.

(12) a. I have been work**in'** on my paper.
b. $\rightsquigarrow$ The speaker is casual.

Furthermore, I argued that the proposed models can capture the contribution that speaker/listener specific prior beliefs and ideologies make to social interpretation and patterns of context-sensitive linguistic variation. I therefore propose that these models constitute yet another item in the sociolinguist's 'toolkit' for analyzing patterns of variation and interaction.

Importantly, I stress that game-theoretic modelling **complements** existing methodologies in sociolinguistics rather than replacing any aspect of current practice: logistic regression/Goldvarb analyses are crucial for telling us which linguistic and social factors are operative in the dataset that we are studying. In other words, variationist analyses help us identify what the empirical generalizations that our analyses should capture are; whereas, game-theoretic analyses allow us to state different (possibly competing) analyses in such a way that allows us to automatically determine whether or not they do in fact capture these generalizations. More specifically, the framework that I have developed provides a way to relate analyses consisting of indexical fields and ideological structure to probability distributions over variants and interpretations. However, this architecture remains an empty shell in the absence of concrete detailed proposals concerning which properties are in the indexical fields of which variants, and which ideological relations exist between properties. Qualitative ethnographic analysis allows us to identify (or at least to make reasonable hypotheses concerning) the ideological structure assumed by the speakers in the communities under study and to formulate analyses which can then be evaluated using the models. In this way, I suggest that game-theoretic modelling can serve as a **link** between proposals concerning social meaning and ideological structure, which have been greatly studied by researchers in interactional sociolinguistics and conversational analysis, and the fine-grained patterns of grammatical variation studied in more quantitative approaches.

Furthermore, the shape of this link has its source in the principles underlying Bayesian cognitive science, which have independently been argued to characterize numerous aspects of human cognitive activities. Consequently, we might also expect sociolinguistic studies to have important role to play in the future development of this exciting new field. I therefore conclude that, working in tandem with detailed quantitative and ethnographic studies of linguistic interaction, Bayesian game-theoretic models have useful applications to treat patterns of sociolinguistic variation/interpretation and the potential to situate the study of sociolinguistic interaction "beyond talk", i.e. within a broader theory of human behaviour and cognition.

References

Acton, E. K. (2014). *Pragmatics and the social meaning of determiners*. PhD thesis, Stanford University.

Acton, E. K. and Potts, C. (2014). That straight talk: Sarah palin and the sociolinguistics of demonstratives. *Journal of Sociolinguistics*, 18(1):3–31.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14(03):471–485.

Bach, E. (1986). The algebra of events. *Linguistics and philosophy*, 9(1):5–16.

Bell, A. (1984). Language style as audience design. *Language in society*, 13(02):145–204.

Bell, A. (1997). Language style as audience design. In *Sociolinguistics*, pages 240–250. Springer.

Beltrama, A. (2016). *Bridging the gap. Intensifiers between semantic and social meaning*. PhD thesis, University of Chicago.

Benz, A., Jäger, G., Van Rooij, R., and Van Rooij, R. (2005). *Game theory and pragmatics*. Springer.

Bourdieu, P. (1977). The economics of linguistic exchanges. *Social science information*, 16(6):645–668.

Bucholtz, M. and Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

Bucholtz, M. and Hall, K. (2008). All of the above: New coalitions in sociocultural linguistics1. *Journal of Sociolinguistics*, 12(4):401–431.

Burnett, H. (2017). Signalling games, sociolinguistic variation and the construction of style. *accepted in Linguistics & Philosophy*.

Burnett, H. and Levon, E. (2016). A (more or less) compositional semantics of style. In *Meaning, Optimization and Interaction (MOI) workshop*, Université Paris-Diderot.

Cameron, D. and Kulick, D. (2003). *Language and sexuality*. Cambridge University Press.

Campbell-Kibler, K. (2006). *Listener perceptions of sociolinguistic variables: The case of (ING)*. PhD thesis, Stanford University.

Campbell-Kibler, K. (2007). Accent,(ing), and the social logic of listener perceptions. *American speech*, 82(1):32–64.

Campbell-Kibler, K. (2008). I'll be the judge of that: Diversity in social perceptions of (ing). *Language in Society*, 37(05):637–659.

Chater, N. and Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive science*, pages 335–344.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Clark, R. L. (2014). *Meaningful games: Exploring language with game theory*. MIT Press.

Connell, R. W. and Connell, R. (2005). *Masculinities*. Univ of California Press.

Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8(11):1–55.

Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.

Dror, M., Granot, D., and Yaeger-Dror, M. (2013). Speech variation, utility, and game theory. *Language and Linguistics Compass*, 7(11):561–579.

Dror, M., Granot, D., and Yaeger-Dror, M. (2014). Teaching & learning guide for speech variation, utility, and game theory. *Language and Linguistics Compass*, 8(6):230–242.

Eckert, P. (2000). *Language variation as social practice: The linguistic construction of identity in Belten High*. Wiley-Blackwell.

Eckert, P. (2008). Variation and the indexical field. *Journal of sociolinguistics*, 12(4):453–476.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Eckert, P. and McConnell-Ginet, S. (1995). Constructing meaning, constructing selves. *Gender articulated: Language and the socially constructed self*, pages 469–507.

Eckert, P. and McConnell-Ginet, S. (2013). *Language and gender*. Cambridge University Press.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Franke, M. (2009). *Signal to act: Game theory in pragmatics*. PhD thesis, Institute for Logic, Language and Computation.

Franke, M. and Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, 11(5):e0154854.

Franke, M. and Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift fr̈ Sprachwissenschaft*, 35:3–44.

Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: Communication and context. *Contexts of accommodation: Developments in applied sociolinguistics*, 1.

Goffman, E. (1961). *Encounters: Two studies in the sociology of interaction.* Bobbs-Merrill.

Goffman, E. (1967). *Interaction ritual: essays on face-to-face interaction.* Aldine.

Goffman, E. (1970). *Strategic interaction*, volume 1. University of Pennsylvania Press.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Gratton, C. (2016). Resisting the gender binary: The use of (ING) in the construction of non-binary transgender identities. *Penn working papers in linguistics*, 22.

Graziano, M. S. (2013). *Consciousness and the social brain*. Oxford University Press.

Grice, P. (1975). Logic and conversation. *Syntax and Semantics*, 3:41–58.

Gumperz, J. J. (1982a). *Discourse strategies*, volume 1. Cambridge University Press.

Gumperz, J. J. (1982b). *Language and social identity*, volume 2. Cambridge University Press.

Horn, L. (1989). *A natural history of negation*. University of Chicago Press.

Houston, A. (1985). *Continuity and change in English morphology: The variable (ING)*. PhD thesis, University of Pennsylvania.

Jäger, G. (2011). Game-theoretical pragmatics. In van Benthem, J. and ter Meulen, A., editors, *Handbook of Logic and Language*, pages 467–491. Elsevier, Amsterdam.

Kersten, D. and Yuille, A. (2003). Bayesian models of object perception. *Current opinion in neurobiology*, 13(2):150–158.

Kiesling, S. F. (1998). Men?s identities and sociolinguistic variation: The case of fraternity men. *Journal of Sociolinguistics*, 2(1):69–99.

Körding, K. P. and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences*, 10(7):319–326.

Labov, W. (1966). *The social stratification of English in New York city*. Center for Applied Linguistics.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., and Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, 60(1):44.

Lassiter, D. and Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory*, volume 23, pages 587–610.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Levon, E. (2014). Categories, stereotypes, and the linguistic perception of sexuality. *Language in Society*, 43(05):539–566.

Lewis, D. (1969). *Convention*. Harvard UP, Cambridge.

Link, G. (1983). The logical analysis of plurals and mass terms. In Baeuerle, R., editor, *Meaning, Use and Interpretation of Language*, pages 303–329. DeGruyter.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, 66(2):81.

Manning, P. (1992). *Erving Goffman and modern sociology*. John Wiley & Sons.

Maynard Smith, J. and Price, G. (1973). The logic of animal conflict. *Nature*, 246:15.

McConnell-Ginet, S. (2011). *Gender, Sexuality and Meaning: Linguistic Practice and Politics*. Oxford University Press, Oxford.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pages 221–242. Springer.

Myers-Scotton, C. and Bolonyai, A. (2001). Calculating speakers: Codeswitching in a rational choice model. *Language in Society*, 30(01):1–28.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. MIT press Cambridge, MA:.

Oaksford, M. and Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Ochs, E. (1992). Indexing gender. *Rethinking context: Language as an interactive phenomenon*, 11:335.

Ochs, E. (1993). Constructing social identity: a language socialization perspective. *Research on Language and Social Interaction*, 26:287–306.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

Pleck, J. H., Sonenstein, F. L., and Ku, L. C. (1994). Attitudes toward male roles among adolescent males: A discriminant validity analysis. *Sex roles*, 30(7-8):481–501.

Podesva, R. (2004). On constructing social meaning with stop release bursts. In *Sociolinguistics Symposium*, volume 15.

Podesva, R. J., Reynolds, J., Callier, P., and Baptiste, J. (2015). Constraints on the social meaning of released/t: A production and perception study of us politicians. *Language Variation and Change*, 27(01):59–87.

Recanati, F. (2004). *Literal meaning*. Cambridge University Press.

Recanati, F. (2010). *Truth-conditional pragmatics*. Clarendon Press Oxford.

Schelling, T. C. (1960). *The strategy of conflict*. Harvard university press.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: Rem-retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4(2):145–166.

Silverstein, M. (1976). Shifters, linguistic categories, and cultural description. *Meaning in anthropology*, 1:1–55.

Silverstein, M. (1979). Language structure and linguistic ideology. *The elements: A parasession on linguistic units and levels*, pages 193–247.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, 23(3):193–229.

Steyvers, M., Griffiths, T., and Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in cognitive science*, page 327?334.

Strawson, P. F. (1950). On referring. *Mind*, 59(235):320–344.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Tamminga, M. (2014). *Persistence in the production of linguistic variation*. PhD thesis, University of Pennsylvania.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Wittgenstein, L. (1953). *Philosophical investigations. Philosophische Untersuchungen*. Macmillan.

Yuille, A. and Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive science*, page 301?308.

Zeevat, H. and Schmitz, H.-C. (2015). *Bayesian natural language semantics and pragmatics*, volume 2. Springer.

Zhang, Q. (2008). Rhotacization and the 'beijing smooth operator': the social meaning of a linguistic variable. *Journal of Sociolinguistics*, 12(2):201–222.