**Modeling "Non-literal" Social Meaning with Bayesian Pragmatics**

Truth-conditional and socially indexical meanings have traditionally been studied in separate subfields. However, recent years have seen promising attempts to integrate the semantics and pragmatics of social meaning (e.g. Smith, Hall, and Munson 2010, Acton and Potts 2014). In particular, Burnett (2017, in press) introduces a formalization of social meaning in terms of the Rational Speech Act (RSA) paradigm (Goodman and Frank 2016). Building on this work, we address a central observation of contemporary sociolinguistics, that a linguistic variant may be used to convey only *some aspects* of the social meaning with which it is conventionally associated (Eckert 2008). For instance, an adult can use childlike language features to convey not that they are a child (first order indexicality), but that they have certain traits associated with children, like cuteness or innocence (second order indexicality). Similarly, Eckert (2008) notes that some suburban Detroit teenagers use phonetic and syntactic forms conventionally associated with urban Detroit (such as vowel backing and negative concord), and hypothesizes that this is not to signal urbanity (i.e., *I am from urban Detroit*) per se, but rather to affiliate with certain perceived aspects of urbanity, such as being "autonomous, tough, and street-smart". The idea is that a speaker can exploit the stereotypical correlation between being urban and being tough to signal that they are the latter, even if the former does not apply to them. Consequently, a listener can infer that only toughness is intended to be communicated, if they hear a form also associated with being urban. We address this with the mechanism of projection functions (Kao, Bergen, and Goodman 2014) to allow for utterances which are informative only along particular dimensions of meaning.

**Applying RSA to Social Meaning** RSA is a paradigm for formalizing pragmatics which has been applied to a wide variety of pragmatic inferences. It involves a correspondence between a set of possible worlds, $W$, and a set of possible utterances, $U$. A listener $L_n$ hears an utterance and infers what world a hypothetical speaker $S_n$ would have been in to have produced that utterance. $S_n$, given a world $w$, chooses the utterance that would cause a hypothetical $L_{n-1}$ to be in $w$. This recursive process grounds out in a semantics, which states which utterances and worlds are compatible. In the case of social indexical meaning, the crucial difference is that elements of $W$ are no longer states of the world as a whole, but rather representations of the speaker's identity. Elements of $U$ are linguistic variants which may be truth-conditionally equivalent, but index different social identities.[1]

Generalizing Burnett's formulation, we propose to model social identities (i.e. elements $w \in W$) as lists of variables. For instance, Burnett presents an example in which an identity (or in her terms, a persona) can be formulated as consisting of two binary variables, *friendliness* and *competence*, both of which can be either valued true or false. This formulation is flexible with regard to our theoretical position: we can choose discrete or continuous variables which measure macrosocial categories such as gender, class and race, as well as microsocial categories, which are of increasing concern for sociolinguists (Eckert 2012).

**Modeling Second Order Indexicality** With the example of a suburban Detroit teenager using negative concord (e.g. *I ain't never seen him*), we show that the mechanism employed by Kao, Bergen, and Goodman (2014) to model metaphorical language can be used to model

[1]For present purposes, we do not consider the interaction of truth-conditional and social meaning, but note that this framework is able to accommodate such interaction.

second order indexicality. Let $u$ be a binary variable indicating use of negative concord, and $w$ a tuple of two binary variables, *urbanity* and *toughness*. In the setting of social meaning, a semantics represents the conventional common knowledge of which variants are compatible with which social identities in a given speech community. As a simple example, we stipulate that the *neg-concord* feature is compatible with urban speakers, i.e., $[\![u]\!](w)$ is defined as follows.[2]

| $u$ \ $w$ | $\langle \mathbf{urban}, \mathbf{tough} \rangle$ | $\langle \mathbf{urban}, \neg\mathbf{tough} \rangle$ | $\langle \neg\mathbf{urban}, \mathbf{tough} \rangle$ | $\langle \neg\mathbf{urban}, \neg\mathbf{tough} \rangle$ |
|---|---|---|---|---|
| *no-concord* | 0 | 0 | 1 | 1 |
| *neg-concord* | 1 | 1 | 0 | 0 |

We can then define the literal listener $L_0(w|u) \propto \Pr(w) \cdot [\![u]\!](w)$, which starts with a prior belief about the speaker, and updates it based on the semantics of the heard utterance. We choose a prior $\Pr(w)$ which encodes the listener's belief that the speaker is unlikely to be urban (probability 0.2), and the ideology that urban people are more likely to be tough than not ($0.15 > 0.05$), whereas suburban people are less likely to be tough ($0.35 < 0.45$).

| $w$ | $\langle \mathbf{urban}, \mathbf{tough} \rangle$ | $\langle \mathbf{urban}, \neg\mathbf{tough} \rangle$ | $\langle \neg\mathbf{urban}, \mathbf{tough} \rangle$ | $\langle \neg\mathbf{urban}, \neg\mathbf{tough} \rangle$ |
|---|---|---|---|---|
| $\Pr(w)$ | 0.15 | 0.05 | 0.35 | 0.45 |
| $L_0(w \mid \textit{neg-concord})$ | <u>0.75</u> | 0.25 | 0 | 0 |
| $L_0(w \mid \textit{no-concord})$ | 0 | 0 | <u>0.438</u> | 0.562 |

If $L_0$ hears the negative concord feature, they will infer that the speaker must be urban. Therefore if a speaker who is not urban wants to convey every dimension of their identity, they will never use negative concord. However, if the speaker only cared about conveying that they were tough, they would prefer to use negative concord because it better signals toughness ($0.75 > 0.438$). This motivates us to adopt Kao et al.'s speaker model $S_1$, which takes the speaker's interest into account. Formally, $S_1(u|w, q)$ is conditioned on an identity $w$ and a *projection* q, which is a function that maps $w$ to the subset of $W$ containing all $w'$ which agree with $w$ on one variable. For instance, $q_{\text{toughness}}(\langle \mathbf{urban}, \mathbf{tough} \rangle) = \{\langle \mathbf{urban}, \mathbf{tough} \rangle, \langle \neg\mathbf{urban}, \mathbf{tough} \rangle\}$. $S_1$ is trying to communicate $w$ along $q$: they prefer utterances with higher likelihood of making $L_0$ agrees with them on this dimension of $w$:

(1) $S_1(u|q, w) \propto \Pr(u) \cdot \sum_{w':\, q(w')=q(w)} L_0(w'|u)$

Because of the correlation between urbanity and toughness in the $L_0$ prior, $S_1$ will be motivated to use *neg-concord* if they want to communicate that they are tough. Assuming a uniform prior over utterances, $S_1(\textit{neg-concord} \mid \langle \mathbf{urban}, \mathbf{tough} \rangle, q_{\text{toughness}}) = 0.63$.

With a uniform prior over projections, the pragmatic listener $L_1$ then jointly infers a speaker's identity and projection (2). On hearing the suburban speaker use negative concord, $L_1$ infers that this speaker is most likely tough but not urban, and is trying to communicate toughness. $L_1(q_{\text{toughness}}, \langle \neg\mathbf{urban}, \mathbf{tough} \rangle \mid \textit{neg-concord}) = 0.33$. Thus, we model second order indexicality as deriving from pragmatic reasoning about the identity of the speaker.

(2) $L_1(q, w|u) \propto \Pr(q) \cdot \Pr(w) \cdot S_1(u|q, w)$

**Selected references** ∘ Burnett, Heather. (in press). Signalling games, sociolinguistic variation and the construction of style. *Linguistics & Philosophy* ∘ Burnett, Heather. (2017). Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Journal of Sociolinguistics*

[2]This can be relaxed by replacing $[\![u]\!](w)$ with a probabilistic speaker $S_0(u \mid w)$, who is more likely to use negative concord if they are urban (see Henderson & McCready's (2017) analysis of dogwhistles).