Modeling "non-literal" social meaning[1]

Reuben COHN-GORDON — *Stanford University*
Ciyang QING — *Stanford University*

Abstract

Truth-conditional and socially indexical meanings have traditionally been studied in separate subfields. However, recent years have seen promising attempts to unify the semantics and pragmatics of the two (e.g. Smith et al., 2010; Acton and Potts, 2014). In particular, Burnett (2017, 2019) 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 index only *some aspects* of a speaker's identity. 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".

We model such uses of indexes with the mechanism of *projection functions* (Kao et al., 2014), to allow for utterances which are informative only along particular dimensions of meaning.

Keywords: Bayesian models, pragmatics, sociolinguistics, metaphor, higher order indexes.

1. Introduction

While truth-conditional semantics is concerned with the relation between linguistic utterances and the state of the world, a central concern of sociolinguistics is the relation between often truth conditionally equivalent *variants* and the identity of the *speaker* that these variants *index*.

Early sociolinguistics studies, such as Labov (1968), study the relation between membership in particular social categories (e.g. class, gender) and truth-conditionally equivalent phonetic variants (e.g. presence or absence of th-stopping). In more contemporary work, often referred to as *third wave sociolinguistics* (see (Eckert, 2012)), focus has turned to finer-grained notions of social identity, indexed by the use of correlated linguistic features, which make up *styles*. For instance, Pratt and D'Onofrio (2017) studies how the use of phonetic features such as creaky voice, in tandem with paralinguistic choices like jaw-setting are used to index a socially constructed valley girl *persona*, or social identity. As D'Onofrio (2016) puts it,

"This approach treats variation as a semiotic system (Eckert, 2008), in which linguistic features are viewed not as passive markers of speaker age, gender, or region, but as resources that speakers use toward any number of social or interactional ends."

[1]We would like to thank Judith Degen for her very helpful guidance in the early stages of this project, as well as the rest of the Stanford linguistics department.

Proceedings of Sinn und Bedeutung 23, vol. 1, pp. 301–309. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès).

Probabilistic Perspective Following Burnett (2017, 2019), we argue both here and in (Qing and Cohn-Gordon, in press) for the appropriateness of Bayesian inference as a tool to model aspects of sociolinguistic theory, particularly from the third wave perspective.

Concretely, let $\mathbf{u} \in U$ be a style, represented as an n-tuple $(u_1 ... u_n)$ where each u_i is a linguistic feature in a speaker's production. Similarly, let an identity $\mathbf{w} \in W$ be an n-tuple of socially relevant variables $(w_1 ... w_n)$. For instance u_1 might represent the phonetic articulation of a particular vowel, and w_1 could be the age of the speaker.

Then, the conventional associations between identities and styles in a particular community of practice can be modeled as a conditional probability distribution[2] $S_0(\mathbf{u}|\mathbf{w})$, with the job of a listener L to *infer* (using Bayes' rule) the social identity $\mathbf{w}$ of their interlocutor based on their style $\mathbf{u}$. In this respect, social meaning aligns with use-conditional meaning, a connection explored closely in (Qing and Cohn-Gordon, in press).

To say that linguistic features $(u_1 ... u_n)$ have social meaning in concert as a style, rather than separately, is to say that $S_0(u_1 ... u_n|w) \neq S_0(u_1|w) * ... S_0(u_n|w)$, i.e. that $u_1 ... u_n$ are not independent. Similarly, the aspects composing the interlocutor's social identity $(w_1 ... w_n)$ are not generally independent under the listener's prior or posterior.

To say that the meaning of a variant (or style) depends on context is to say that it depends on the conventions dictated by the S_0, as well as the prior beliefs about the speaker held by the listener, so that a particular feature u_1 may signal entirely different things depending on the context and the style it appear in.

Higher Order Indexes One particular way in which speakers signal their social identity is by using variants conventionally associated with a macrosocial category they do not belong to, in order to convey attributes associated with that category. For example, Eckert (2012) notes the use of urban-led sound changes (such as vowel backing and negative concord) by suburban teenagers who want to index perceived urban qualities:

> "...just as women are not making direct gender claims when they use female-led changes, burnouts are not making direct urban claims when they use urban-led changes...The urban kids that burnouts identified with were white kids who knew how to cope in the dangerous urban environment – kids they saw as autonomous, tough, and street-smart. Presumably in adopting urban forms, suburban kids were affiliating with those qualities, not claiming to be urban.".

We refer to this sort of use of a variant as *higher order indexicality*[3]. It allows for a speaker to exploit interspeaker variation in their communities of practice (for example, the fact that urban Detroit speakers use more negative concord) in order to communicate information other than just membership of a given macrosocial category.

[2] A distribution $P(A)$ over a set A is the pair (A, f), where f is a function $A \to \mathscr{R}$, assigning each element of A a real-valued weight between 0 and 1, such that $\sum_{a \in A} f(a) = 1$. A conditional distribution $P(A|B)$ is a function $B \to Dist(A)$, where $Dist(A)$ is the set of all possible distributions on A. In other words, a conditional distribution takes (i.e. is conditioned on) $b \in B$ and returns a distribution over A.

[3] This term is coined by (Silverstein, 2003), although our precise usage may differ.

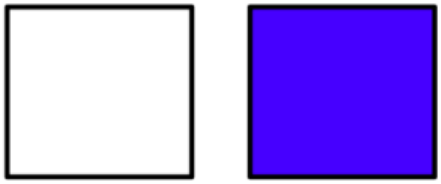

Figure 1: $U : \{square, blue\ square\}, W : \{R_1, R_2\}$

Our goal in this paper is to suggest a model of higher order indexicality, closely related to a Bayesian model of metaphor proposed by (Kao et al., 2014).

2. Bayesian Models of Semantics and Pragmatics

A body of recent work (often collectively referred to as the Rational Speech Acts, or RSA, framework) uses nested Bayesian models of speakers and listeners to formalize a number of the pragmatic phenomena envisioned by Grice (1975) to derive from inter-agent reasoning, such as scalar implicature (Frank and Goodman, 2012), manner implicature (Bergen et al., 2016), metaphor (Kao et al., 2014), hyperbole (Kao et al., 2014), and presupposition accommodation (Qing et al., 2016).

These models involves a correspondence between a set of possible worlds, W, and a set of possible utterances, U. A listener L_n hears an utterance $u \in U$ and infers what world $w \in W$ a hypothetical speaker S_n would have been in to have produced u. S_n, given a world w, chooses the utterance u that would cause a hypothetical L_{n-1} to infer w. This recursive process grounds out in a semantics, which states which utterances and worlds are compatible.

A simple two agent communicative task known as a *reference game* provides an example of linguistic communication that serves as a useful demonstration case for modeling semantics and pragmatics generally. We begin by summarizing the application of a simple RSA model to reference games, and show how this serves as a basis for richer models of linguistic communication.

In a reference game, a speaker and listener see a set W of referents. The speaker is assigned one of these referents as their target, and aims to communicate which referent this is to the listener. The speaker does so by choosing an utterance from a set U of possible utterances. Figure 1 provides a concrete example.

Assuming that both the speaker and listener share a semantics, so that *blue square* can only refer to R_2 while *square* can refer to either, the most informative utterance for a speaker whose target is R_2 is *blue square*.

A listener who assumes that the speaker acts informatively in this way can draw an inference on hearing *square*. They can infer that R_1 is the referent, since had R_2 been the referent, the speaker would have said *blue square*.

It is clear that a model of semantics is not sufficient to derive either the behavior of the informative speaker or the listener who reasons about this speaker. A speaker who only cares about

producing true utterances will be equally inclined to say *blue square* or *square* when referring to R_2, since both are true. Likewise, a listener who only attends to semantics will be agnostic as to whether the intended referent is R_1 or R_2 on hearing *square*, since that utterance is compatible with either.

As such, what is needed is a model which formalizes the process of reasoning about one's interlocutor. We first consider a model of a listener who only reasons about a semantics, L_0:

$$(1) \quad L_0(w|u) = \frac{[\![u]\!](w)p(w)}{\sum_{w' \in W} [\![u]\!](w')p(w')}$$

This model, on hearing *blue square*, is certain that the referent is R_2
(since $L_0(R_2|\textit{blue square}) = 1.0$), but on hearing *square*, does not draw any inference that the reference is R_1 (since $L_0(R_1|\textit{square}) = L_0(R_2|\textit{square}) = 0.5$).

A speaker S_1 can then be defined which given a referent w, prefers utterances u which convey w to L_0, so that $S_1(\textit{blue square}|R_2) > S_1(\textit{square}|R_2)$:

$$(2) \quad S_1(u|w) = \frac{L_0(w|u)p(u)}{\sum_{u' \in U} L_0(w|u')p(u')}$$

This puts us in a position to define a new listener, L_1, capable of deriving the desired implicature by reasoning about what referent S_1 must have had in order to have produced the heard utterance:

$$(3) \quad L_1(w|u) = \frac{S_1(u|w)p(w)}{\sum_{w' \in W} S_1(u|w')p(w')}$$

To make concrete predictions from our model, we must specify concrete values for U, W, $P(u)$, $P(w)$, and the semantics $[\![\cdot]\!]$, for instance as follows:

- W : $\{R_1, R_2\}$
- U : $\{\textit{square}, \textit{blue square}\}$
- $P(w)$: $\{R_1 : 0.5, R_2 : 0.5\ \}$
- $P(u)$: $\{\textit{square} : 0.5, \textit{blue square} : 0.5\}$
- $[\![\cdot]\!]$: $\{((\textit{blue square}, R_1), 0), ((\textit{blue square}, R_2), 1), ((\textit{square}, R_1), 1), ((\textit{square}, R_2), 1)\}$

Under this interpretation, L_1 prefers R_1 on hearing *square*, although R_2 is still a possibility: $L_1(R_1|\textit{square}) > L_1(R_2|\textit{square})$. This corresponds to the calculation of an implicature.

Projection Functions One important enrichment of the model in equations (1-3) is the L_1^{Q} model of metaphor. In the cases discussed so far, there has been a parameter left implicit that dictates which *aspects* of the world a speaker cares about conveying. For instance, the listener who hears "I ate some of the cookies." is modeled as drawing inferences about the number of cookies eaten, but not about whether it is raining in Timbuktu. We can make this dependence on an aspect of the world explicit by replacing S_1 with S_1^{Q}:

$$(4) \quad S_1^{\mathrm{Q}}(u|w,q) \propto \sum_{w'} \delta_{q(w)=q(w')} * L_0(w'|u)p(u)$$

Here, q is a function of type $W \to \mathscr{P}(W)$, which maps a world w to an equivalence class of worlds. For instance, suppose that q maps a world to all the worlds in which John ate exactly the same number of cookies. Then the goal of S_1^{Q} is to be informative, but only up to the goal of conveying the number of cookies. The S_1^{Q} may mislead the listener with respect to the weather in Timbuktu, in the course of carrying out their goal.

Since q is an explicit variable on which S_1^{Q} depends, one can create a listener L_1^{Q} which *jointly reasons* about the world w and the aspect of the world q which the speaker wishes to communicate.

(5) $L_1^{\mathrm{Q}}(w,q|u) \propto S_1^{\mathrm{Q}}(u|q,w) * P(w) * P(q)$

This model was first proposed in (Kao et al., 2014) to model metaphor. The idea is that, in a metaphorical utterance like "John is a shark.", a listener reasons jointly about what John is like, and what *aspect* of John the listener was attempting to convey.

L_1^{Q} qualitatively differs from L_1 in the following way. It can hear an utterance u and infer a world w which is not compatible with u in the semantics. In particular, it can hear a metaphorical utterance like "John is a shark." and infer that the speaker meant to convey only some aspect of John, perhaps that he is vicious, and not another, like that he is able to breathe underwater (notated below as the property *aquatic*). The following interpretation of L_1^{Q} yields this behavior:

- $P(w)$: $\{(vicious, aquatic) : 0.2, (vicious, \neg aquatic) : 0.2,$
  $(\neg vicious, aquatic) : 0.3, (\neg vicious, \neg aquatic) : 0.3\}$
- $P(U)$: $\{$*shark*: $\frac{1}{3}$, *vicious*: $\frac{1}{3}$, *aquatic*: $\frac{1}{3}\}$
- $P(q)$:
  - $q_{vicious}(\lambda(x,y) : x) : 0.5$
  - $q_{aquatic}(\lambda(x,y) : y) : 0.5$
- The semantics:
  - $[\![shark]\!](w) \mapsto 1$ if $w = (vicious, aquatic)$ else 0
  - $[\![vicious]\!](w) \mapsto 1$ if $w = (vicious, aquatic) \vee w = (vicious, \neg aquatic)$ else 0
  - $[\![aquatic]\!](w) \mapsto 1$ if $w = (vicious, aquatic) \vee w = (\neg vicious, aquatic)$ else 0

On hearing *shark* predicated of John, L_1's favored interpretation is that John is vicious but doesn't breathe underwater, and that the speaker is being informative about the viciousness dimension ($L_1^{\mathrm{Q}}((vicious, \neg aquatic), q_{vicious}|shark) = 0.32$).

In this case, the model's behavior is very simple: the prior knowledge that John is unable to breathe underwater makes clear that $q_{aquatic}$ cannot be the value of q. More interesting dynamics arise when U includes a wider range of utterances, and W, a wider range of properties. The general behavior of L_1^{Q}, on receiving an utterance u is to try to find the pair (w,q) such that $q(w)$ is plausible under the prior but also such that no other utterance u' would better convey $q(w)$. For instance, "John is a shark" is unlikely to be interpreted to mean that John can swim well,

even if it is plausible that he can, if there is an alternative utterance that would have conveyed this property better, like “swimmer”.

3. Bayesian Models of Sociolinguistic Phenomena

3.1. A State Space of Speaker Identities

In the examples of RSA models in section (2), we interpreted $w \in W$ as the state of the world. The crucial move in applying RSA-style models to sociolinguistic phenomena is to interpret w as the social identity of the speaker, and $u \in U$ as the choices that a speaker makes (i.e. variants), which jointly carry information about their social identity.

More concretely, we model social identities (i.e. elements $w \in W$) as n-tuples of variables (following Burnett, 2017, 2019). Likewise, $u \in U$ are n-tuples of features. In the truth-conditional setting discussed above, a semantics (in the form of a compatibility relation) connects u to w, laying the basis for pragmatic enrichments from inter-speaker reasoning. This semantics is conventional knowledge in a given community of practice.

To model social meaning, we replace a relational semantics with a conditional probability distribution $S_0(u|w)$, which represents the *conventional stereotypes* about which types of people produce which types of language. Importantly, S_0, given a speaker with social identity w, is not a model of any speaker's actual language use but rather represents what all agents *treat* as such a speaker's model of language use.

The motivation for using an S_0 is discussed in more detail in (Qing and Cohn-Gordon, in press), as is the method by which it can be integrated into a model which also attends to truth-conditional meaning. For now, we simply observe that S_0 can be understood as representing the *conventional association between social identity and language* that is common ground in a given community of practice.

Thus a listener, on hearing their interlocutor speak, infers a joint distribution over the variables which describe the interlocutor's identity, by reasoning about S_0. For instance, children's speech differs along a number of features from that of adults, ranging from the realization of phones such as /r/, to the absence of complex syntactic constructions. This information is represented in S_0, which predicts that children will produce a child-like style of language. On hearing language with child-like style, L_0 can reason about S_0 to infer that the speaker is a child.

3.2. Modeling Higher Order Indexicality

The key feature of higher order indexicality is that a listener may hear a style u associated with some macrosocial category, and based on their prior belief about their interlocutor, draw an inference that they only possess certain attributes of this macrosocial category. For instance, hearing a child-like voice would not lead to the inference that the speaker is a child if they are known to be an adult. Instead, it might be taken to signal properties associated with children.

The L_0 model is insufficient to capture this sort of inference, since it has no mechanism for deciding what parts of a speaker's signaled identity are relevant. For instance, suppose personae consisted of just two Boolean variables: *youth* and *innocence*. Further suppose that the speaker

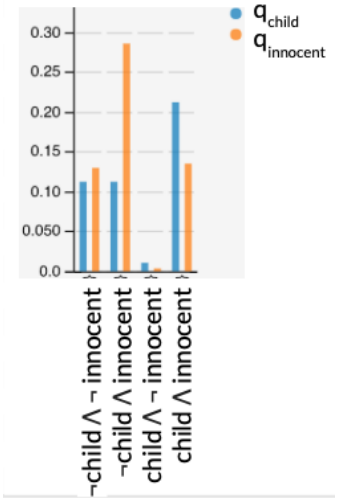

Figure 2: The L_1 posterior, after hearing their interlocutor produce language in a child-like style. The preferred interpretation is that the speaker is not a child, but is attempting to convey innocence. In this example, we do not strictly eliminate the prior possibility that the interlocutor might be a child, so the interpretation that this is the case also receives quite high probability.

is believed to be an adult. We encode this in the model through the listener's prior over W, as shown in (6). This prior prefers states w where *young* is false. It also encodes the correlation between being young and being innocent: when *young* is true, *innocent* is more likely to be true than when *young* is false.

(6) P(w): {(young=true, innocent=true): 0.15, (young=true, innocent=false): 0.05, (young=false, innocent=true): 0.35, (young=false, innocent=false): 0.45,}

(7) p(u): {child-like style: 0.5, adult style: 0.5}

Because L_0 assumes that S_0 is trying to communicate both dimensions (youth and innocence) of their own identity, L_0 will place high probability on both being true after hearing child-like language. By contrast, what we want is a model of a listener that can infer that the speaker only means to communicate about one of these dimensions, and is simply leveraging an index which has other associations.

This requires precisely the S_1^{Q} and L_1^{Q} models of (4) and (5). This is because S_1^{Q} allows for the production of utterances which are only true with respect to some aspect q of w, and L_1^{Q} is able to infer what this aspect is. In this setting, two projections are possible: $q_{young} = \lambda(x,y).x$ and $q_{innocence} = \lambda(x,y).y$. We set the distribution over these projections $p(Q)$ to uniform.

Because being young and being innocent are correlated in the L_0's prior, S_1^{Q} is able to use child-like language to communicate that they are innocent. This also has the effect of misleading L_0 with respect to whether the speaker is young, but if S_1^{Q}'s projection is $q_{innocence}$ (i.e. they only care about being informative along the *innocence* dimension of meaning), this does not discourage S_1^{Q} from using the child-like style.

Thus, L_1^Q, because of its prior belief that the speaker is not a child, is able to rationalize the child-like language it hears from an adult speaker as a means of communicating the correlated attribute of innocence. As shown in figure 2, L_1^Q puts the most weight on $(w = (young : False, innocent : True), q = q_{innocence})$, after hearing a child-like style from an adult. This bears a close relation to the inference drawn when hearing a metaphor, such as "The party was a riot.", where only *some* aspects of *riot* are pertinent to the party, though the present model employs a distribution S_0 in place of a truth-conditional semantics.

While the example of child-like language given here is very simple, the core dynamic of the model can be extended to much richer cases of higher-order indexicality. For instance, female gender is associated with many different attributes and stances, any of which can be indexed, in an appropriate context, by female led sound changes.

4. Conclusion

The central idea of this paper, applying the L_1^Q model to higher-order indexicality, is part of a larger bridge between truth-conditional semantics and Gricean pragmatics on the one hand, and use-conditional and social meaning on the other. We argue that nested Bayesian models of speakers and listeners are the right tool to understand the similarities (e.g. the existence of a conventional association between form and meaning) and differences (whether this association is a compatibility relation or a conditional distribution). Our perspective connects indexicality with a larger system of non-linguistic semiotics, such as the use of fashion or body language to communicate social identity.

We envision developing this picture in future work, in several ways. One is to investigate how stylistic features are connected by a hierarchical latent variable. For instance, many aspects of stereotypical Californian style are related to displays of "low energy" , such as creaky voice, jaw setting, lengthened vowels, slow speech rate and so on. A hierarchical model, in which a latent "low energy" variable (possibly in concert with other variables) indexed the valley girl persona would be able to connect this variety of displays of low energy in a coherent way.

Another aim is to further develop the notion of social identities. Are the fundamental dimensions which make up identities socially constructed attributes, like gender, or seemingly objective ones, like geographic location, age, etc? A related question is diachronic: what is the process by which the commonly known conventions linking identity and linguistic style change over time and through use? As Coupland (2001) notes, "It is in relation to group norms that stylistic variation becomes meaningful; it is through individual stylistic choices that group norms are produced and reproduced" (2002:198).". From the perspective of the Bayesian models considered here, the natural modeling assumption corresponding to such change is to have uncertainty over the S_0 itself, so that through repeated interactions, agents gradually update their beliefs. Building models of this kind and applying them to real sociolinguistic data constitutes a promising avenue for further work.

References

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

Bergen, L., R. Levy, and N. D. Goodman (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics 9*(20).

Burnett, H. (2017). Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Journal of Sociolinguistics 21*(2), 238–271.

Burnett, H. (2019). Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, 1–32.

Coupland, N. (2001). *Language, situation, and the relational self: Theorizing dialect-style in sociolinguistics*. na.

D'Onofrio, A. (2016). Social meaning in linguistic perception.

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.

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

Goodman, N. D. and M. C. Frank (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences 20*(11), 818–829.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Kao, J. T., L. Bergen, and N. D. Goodman (2014, July). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, Wheat Ridge, CO, pp. 719–724. Cognitive Science Society.

Kao, J. T., J. Y. Wu, L. Bergen, and N. D. Goodman (2014, August). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences 111*(33), 12002–12007.

Labov, W. (1968). *The social stratification of English in New York city*. Cambridge University Press.

Pratt, T. and A. D'Onofrio (2017). Jaw setting and the california vowel shift in parodic performance. *Language in Society 46*(3), 283–312.

Qing, C., N. D. Goodman, and D. Lassiter (2016). A rational speech-act model of projective content. In *Proceedings of the thirty-eighth annual conference of the Cognitive Science Society*.

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

Smith, E. A., K. C. Hall, and B. Munson (2010). Bringing semantics to sociophonetics: Social variables and secondary entailments. *Laboratory Phonology 1*(1), 121–155.