Social Meaning in Repeated Interactions

Elin McCready & Robert Henderson

Aoyama Gakuin University & University of Arizona

mccready@cl.aoyama.ac.jp / rhenderson@arizona.edu

Overview and introduction

Social meaning has been a topic of much recent attention in computational linguistics and in semantics and pragmatics (Yoon et al 2016, Burnett 2018, McCready 2019).

- One reason for this has been the need to address and identify bad actors in online speech through automatic means.
- Semantics and pragmatics: attempt to identify and understand the kinds of meanings carried by expressions with socially significant content, and to find ways of formally modeling their effects.

One active area of research has been political and hate speech.

- Often difficult to determine what counts as hate speech, or what aspects of speech have political overtones.
- **Dogwhistles**: expressions with a dual function:
  1. signaling a speaker's (usually objectionable or controversial) political stance to a set of savvy interpreters with the requisite background to catch the coded message,
  2. and appearing to those not in the know as carrying only more innocuous (truth-conditional) meanings.

$\Longrightarrow$ Starting point for the present paper.

Henderson and McCready (2017,2018,2020) present a theory of dogwhistles set in an extension of the game-theoretic framework proposed by Burnett (2018).

- Game in which utilities depend on recognition of the persona the speaker means to express.
- Key: a persona is understood as a kind of social role or stance on certain socially relevant issues (cf. Jaffe 2009).
- Henderson and McCready (2019) extend this work to an account of trust in communication.
  - Basic idea to ground a notion of trust on social meaning: since social meanings and personas can signal shared values and goals, it is sensible to trust someone on that basis.
- Contrast with McCready (2015), which takes testimonial reliability to be determined by communicational histories and initial judgements about the likelihood that a source is reliable.

Goal: combine these two views into one coherent one.

- Judgements about communicative agents evolve over the course of interactions both in how individuals are judged for testimonial reliability and for (ideological) trustworthiness.
- A formal model of this necessitates combining the insights of McCready (2015) on histories and repeated game models and those of Henderson and McCready (2019) on ideology and trust.
- This paper proposes an extension of McCready (2015) which takes social meaning into account, and how social presentation can change over time.

Social Meaning and Dogwhistles

Dogwhistles show the ideologies and social or political stances and views of the speaker in a way which is both deniable and accessible only to those aware of the coded language they utilize.

- Not obviously part of any of the traditional categories of semantic and pragmatic meaning: at-issue content, presupposition, conversational implicature etc.
- Henderson and McCready tie dogwhistles directly to the expression of social meaning, and claim they fall into a new kind of category of meaning.

Indexical meanings, used for decades in sociolinguistics: phonological or stylistic features and express aspects of the speaker's identity.

- Their efficacy is contingent on recognition by the interpreter of the kinds of identity associated with the feature.
- Burnett (2018): game-theoretic model for such features using a modified version of standard signaling games involving *personas*, covering traits such as social features and political ideologies.
- Utilities depend on hearer recovery of the speaker's presented persona and the way in which hearers assign value, positive or negative, to that persona.

Henderson and McCready extend this model to provide an analysis of dogwhistles.

- Basic idea: the coded message which savvy listeners retrieve from dogwhistles is available as a result of recognizing the speaker's ideological presentation as modeled in the form of a persona.
- Thus Burnett's model must be extended to allow interpreters to vary in the degree to which they associate particular messages with personas.

Utilities are then calculated according to (2).

- Combines two values:
  1. social meaning of the message (1), which depends on the affective values of the range of personas consistent with the message and likelihood of recovering each persona from the message
  2. the value assigned to its truth-conditional content, positive only in case the hearer arrives at the true state of affairs on the basis of the message.
- The two aspects of meaning are weighted with values δ and γ which reflect the relative importance assigned to social and truth-conditional meaning respectively.

(1) $U_S^{Soc}(m, L) = \sum_{p \in [m]} ln(Pr(p|m))+$
$\nu_S(p)Pr(p|m) + \nu_L(p)Pr(p|m)$

Speaker strategies σ are functions from pairs of states and personas to messages; listener strategies ρ are functions from messages to such pairs. Let $\rho(\sigma(p, t)) = (p', t')$. Then

(2) $US(m, L) = US_{Soc}(m, L) + EU(m, L)$, where $EU(m, L) = \sum_{t \in T} \mathbf{Pr}'(t) \times U(t, m, L)$, where $U(t, m, L) > 0$ if $t \in \rho(m)$ and else $= 0$ (cf. van Rooij 2008).

This view will be combined subsequently with the view of McCready(2015) on reliability, which we turn to next.

Reliability

McCready (2015) presents a model of how epistemic agents can make judgements about the reliability of an individual's testimony.

- Reliability here corresponds to the probability with which the individual's testimony conveys the truth.
- Reliability judgements come from two sources:
  1. initial impressions of an individual's reliability based on experience, and
  2. world knowledge, and learning about reliability from interactions with that individual.

Testimonial interaction with an agent produces a *history* consisting of a record of that agent's utterances and the way in which they track truth, modeled in terms of records of their actions in a repeated game.

- Each action a performed by agent i in each iteration of a game g is entered into the record as $a_i = \langle \varphi, \tau \rangle$, where φ is the content of the utterance and τ indicates its (un)truth.
- Records then have the form $Hist_g = \langle a_1, \ldots, a_n \rangle$, for a game g with n repetitions.

The degree of reliability assigned to an agent R_a is defined as, where $t_a = \Sigma_{i \in 1,\ldots,n} val(2(a_i)) = T$ (for '2' is a projection function picking out the second element of the tuple) and $f_a = \Sigma_{i \in 1,\ldots,n} val(2(a_i)) = F$,

$$R_a =_{def} \frac{t_a}{t_a + f_a}.$$

Trust in repeated action

But social meaning is important for decisions about trust too.

- Henderson and McCready (2019) combine the ideas of H&M(2017,2018) and McCready (2015) to help understand how communicative agents who are obviously unreliable in a truth-conditional sense can still be trusted.
- Donald Trump is the obvious example here.
- According to their proposal, trust is not strictly dependent on truth, but rather can involve ideology.

Neither Burnett 2018 nor H&M discuss how hearer values are assigned to personas.

- One way to valuate personas is to compare them to your own: the more similar, the higher the value assigned.
- H&M2019 motivate this view via ideological personas:
  - the closer an ideology is to one's own, the more one likes it, since it expresses a similar political stance.
- Then, if an agent has a similar enough persona to oneself, they can be *trusted*, without precisely being *believed* given a sufficiently low reliability index.

But this idea is not fully formalized, because the only model of discourse-level reliability available is that of McCready 2015, which only covers truth-tracking.

- H&M 2019 observe this point but do not modify the model so that it is capable of handling the full range of facts.
- The goal of this section is to extend that model to account for a notion of trust.

Additional tool: a model of personas set within vector spaces of the same sort used to ground formal models of cognitive lexical semantics (Burnett 2020).

- On this view, ideological structures have the form $\langle D, sim, \text{PERS}, \mu \rangle$, where $\langle D, sim \rangle$is a $|D|$-dimensional vector space and sim a similarity function on points in such spaces.
- PERS is a set of points which correspond to personas in this ideological space.
- μ is a function partitioning personas into positively and negatively valued ones.

In this model, trust amounts to just a similarity comparison.

- Once the persona expressed by the signaler is extracted by the interpreter, sim is used to compare the personas of signaler and interpreter, yielding a value in the real-numbered interval $[0, 1]$.
- Given a sufficiently high degree of similarity, the interpreter will be justified (in terms of closeness of interests) in trusting the signaler.

To extend this model to discourse-level phenomena and thereby make the actions of agents across the lifespan of testimonial interaction genuinely dependent on both social meaning and reliability, we now integrate this view with the histories of McCready 2015.

- Game iterations are now of the form $\langle \varphi, \tau, \pi \rangle$, where φ and τ are as before and $\pi \in$ PERS.
- Now (3) indicates the degree of trust assigned by the interpreter to the signaler a in the initial state.
  - This is just the degree of similarity between the persona π_1 expressed by a in their first interaction, ie. the first game iteration.
- (4) indicates how trust is assigned as the interaction continues: by averaging the trust assigned before the current iteration with the similarity of the interpreter's and the agent's currently expressed personas.

(3) $trust_a^1 = sim(\pi_1, P)$

(4) $trust_a^{i+1} = \frac{sim(\pi_i, P) + trust_a^i}{2}$

This system is extremely simple and gives a high degree of importance to the latest interaction of the two agents.

- This is easy to modify, but we find it intuitive to let the latest interaction of agents be highly determinative of how they judge trustworthiness via social aspects of persona and ideological communication.

Acknowledgements

Thanks to Nicholas Asher, Daisuke Bekki and Heather Burnett for discussion.

References

Heather Burnett. 2018. Signalling games, sociolinguis- tic variation and the construction of style. Linguistics and Philosophy. Heather Burnett. 2020. A persona-based semantics for slurs. To appear in Grazer Philosophische Studien. Robert Henderson and Elin McCready. 2017. How dogwhistles work. In Proceedings of LENLS 14. JSAI. Robert Henderson and Elin McCready. 2018. Dog- whistles and the at-issue/not-at-issue distinction. In Daniel Gutzmann and Katherine Turgay, editors, Secondary Content, pages 191?210. Brill. Robert Henderson and Elin McCready. 2019. Dogwhistles, trust and ideology. In Proceedings of the 22nd Amsterdam Colloquium, pages 152?160. ILLC. Alexandra Jaffe. 2009. Introduction: the sociolinguis- tics of stance. In Alexandra Jaffe, editor, Stance: Sociolinguistic Perspectives, pages 3?28. Oxford Uni- versity Press. Elin McCready. 2015. Reliability in Pragmatics. Oxford University Press. Elin McCready. 2019. The Semantics and Pragmatics of Honorification: Register and Social Meaning. Oxford University Press. Robert van Rooij. 2008. Game theory and quantity implicatures. Journal of Economic Methodology, pages 261?274. Erica J Yoon, Michael Henry Tessler, Noah D Goodman, and Michael C Frank. 2016. Talking with tact: Polite language as a balance between kindness and informativity. In Proceedings of the 38th Annual Conference of the Cognitive Science Society. Cognitive Science Society.