Social Meaning in Repeated Interactions

Robert Henderson
Department of Linguistics
University of Arizona
rhenderson@arizona.edu

Elin McCready
Department of English
Aoyama Gakuin University
mccready@cl.aoyama.ac.jp

Abstract

Judgements about communicative agents evolve over the course of interactions both in how individuals are judged for testimonial reliability and for (ideological) trustworthiness. This paper combines a theory of social meaning and persona with a theory of reliability within a game-theoretic view of communication, giving a formal model involving interactional histories, repeated game models and ways of evaluating social meaning and trustworthiness.

1 Overview

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. To this end, there has been significant research in the computational linguistics and artificial intelligence communities in this domain. Within semantics and pragmatics, the motivation has been to identify and understand the kinds of meanings carried by expressions with socially significant content, and to find ways of formally modeling their effects on discourse, norms of behavior, and non-linguistic structures.

One active area of research has been political and hate speech. In many cases, though of course not all (for example slurs: see e.g. Camp 2013; Davis and McCready 2018 for work on this topic), it is difficult to determine what counts as hate speech, or what aspects of speech have political overtones. Prominent in this area is the phenomenon of *dogwhistles*, expressions which have the dual function of 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, while appearing to those not in the know as carrying only more innocuous meanings. Work in this area is the starting point for the present paper.

Henderson and McCready (2018, 2017) present a theory of dogwhistles set in an extension of the game-theoretic framework proposed by Burnett (2018). The theory will be detailed in §2, but in essence involves a game in which utilities depend on recognition of the persona the speaker means to express, where 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, which they take to contrast with the notion of reliability in McCready 2015, which takes testimonial reliability to be determined by communicational histories and initial judgements about the likelihood that a source is reliable; this theory is outlined in §3. The basic idea of Henderson and McCready 2019 is 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 regardless of the degree to which one finds them reliable in the sense of truth-tracking in communicative behavior.

The main goal of this paper is to 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; this extension is presented in §4, after

which the paper concludes with some future directions in §5.

2 Social Meaning and Dogwhistles

This section briefly describes the theory of dogwhistles given by (Henderson and McCready, 2018). Dogwhistles are prevalent in political speech, and also of course used elsewhere; they serve to 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. Further, the meanings they convey are not obviously part of any of the traditional categories of semantic and pragmatic meaning: at-issue content, presupposition, conversational implicature and so on. Henderson and McCready (2018) pursue an analysis which ties dogwhistles directly to the expression of social meaning, and claim they fall into a new kind of category of meaning.

Within sociolinguistics, the category of *indexical meanings* has been used for decades (e.g. Eckert 2008; Silverstein 2003). Such meanings are tied to (for example) phonological or stylistic features and express aspects of the speaker's identity; as such, their efficacy is contingent on recognition by the interpreter of the kinds of identity associated with the feature. Burnett (2018) provides a game-theoretic model for such features using a modified version of standard signaling games involving *personas*, roughly definable as social presentations, which are quite various and cover traits such as social features such as friendliness/professionalism and political ideologies. In her model, 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 (2018) extend this model to provide an analysis of dogwhistles. The basic idea is that 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), which combines the value of the 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, with 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 in §4 with the view of McCready (2015) on reliability, which we turn to next.

3 Reliability

McCready (2015) presents a model of how epistemic agents can make judgements about the reliability of an individual's testimony. Reliability here refers exclusively to the degree to which the individual's utterances can be expected to accurately convey information about the world, so reliability corresponds to the probability with which the individual's testimony conveys the truth. According to this work, such judgements come from two sources: initial impressions of an individual's reliability based on experience and world knowledge, and learning about reliability from interactions with that individual.

The first aspect comes into play when making initial judgements about an agent's reliability. Many have observed that such judgements are conditioned on aspects of presentation – e.g. clothing, grooming, context, and various properties like age, race, gender, and physical form which, when used as bases for judging reliability, often lead to pernicious results (Fricker, 2007) – together with stereotypical judgements about how such properties correlate with truth-telling and reliability (see McCready and Winterstein 2019). In the present paper, we are more concerned with the second aspect: the way in which agent interaction

influences subsequent judgements about reliability.

Here, the basic model is frequentist. 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; simplifying slightly, 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 truth or lack thereof; so τ is selected from $\{T, F, ?\}$, for 'true', 'false', and 'indeterminate/unknown' respectively. The value ? is selected when the content either cannot be verified to be true or false at the present time or if it is unclear whether it has a truth-value at all, as in utterances containing only nontruthconditional content or more controversial cases such as sentences expressing subjective judgements ('Life is beautiful.'). Records then have the form $Hist_g = \langle a_1, \ldots, a_n \rangle$, for a game g with n repetitions.

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

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

This simple treatment can be made more sophisticated in various ways (e.g. by weighting more recent interactions over older elements in the history, by introducing awareness, or by introducing other ways to deal with ?-valued elements), but it is sufficient for our purposes to note that all such modifications will still be restricted to judgements about truth-tracking and leave out social meaning entirely.

4 Trust and Reputation

But social meaning is important for decisions about trust. Henderson and McCready (2019) combine the ideas of Henderson and McCready (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.

A lacuna in the proposals of Burnett (2018) and Henderson and McCready (2018) is the way in which 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. Henderson and McCready (2019) 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. It then becomes possible to judge an individual unreliable in the sense of §3 – in that their statements don't consistently track the truth – but still trust them, in the sense that one takes them to have similar goals and thus judges them to act in a way consistent with one's interests. The idea then is that if an agent has a similar enough persona to oneself they can be *trusted*, without precisely being *believed*.

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. Henderson and McCready (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.

Burnett (2020) provides a model of personas set within vector spaces of the same sort used to ground formal models of cognitive lexical semantics. 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, it is easy to see how to incorporate a notion of trust: 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, in the same way as which reliability was handled by McCready (2015).

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, simply by averaging the trust assigned before the current iteration with the similarity of the interpreter's and the agent's currently expressed personas.

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

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

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.

5 Conclusions and Directions

This paper has integrated the model of testimonial reliability of McCready (2015) with the model of trust of Henderson and McCready (2019) via a notion of persona similarity in vector spaces. This integration is successful and brings together notions of reliability in terms of truth-telling and reliability in terms of common interests and ideological similarity, on the assumption that the latter is to be understood in terms of personas. In future work, we intend to incorporate the valuation function μ and thereby rethink the notion of persona. We think that it is likely that agents judge others not just on the basis of the persona they communicate but also in terms of how they evaluate such personas, ie. their general ideological stance. This requires incorporating valuations into the notion of persona in general, an extension of the model of Burnett (2020). Doing so is the next step in the current project.

Acknowledgments

Thanks to Daisuke Bekki and Heather Burnett for discussion.

References

Heather Burnett. 2018. Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*.

Heather Burnett. 2020. A persona-based semantics for slurs. To appear in *Grazer Philosophische Studien*.

Elisabeth Camp. 2013. Slurring perspectives. *Analytic Philosophy*, 54(3):330–349.

Christopher Davis and Elin McCready. 2018. The instability of slurs. To appear in *Grazer Philosophische Studien*.

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

Miranda Fricker. 2007. *Epistemic Injustice*. Oxford University Press.

Robert Henderson and Elin McCready. 2017. How dogwhistles work. In *Proceedings of LENLS 14*. JSAI.

Robert Henderson and Elin McCready. 2018. Dogwhistles 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 sociolinguistics of stance. In Alexandra Jaffe, editor, *Stance: Sociolinguistic Perspectives*, pages 3–28. Oxford University 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.

Elin McCready and Grégoire Winterstein. 2019. Testing epistemic injustice. *Investigationes Linguisticae*, 41:86–104.

Michael Silverstein. 2003. Indexical order and the dialectics of social life. *Language and Communication*, 23:193–229.

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.