Towards functional, agent-based models of *dogwhistle* communication

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

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

Abstract

Henderson and McCready 2017, 2018, 2019 build a novel theory of so-called 'dogwhistle' communication by extending the *social meaning games* of Burnett 2017. This work reports on an ongoing project to build systems to model the evolution of dogwhistle communication in a population based on probability monads (Erwig and Kollmansberger, 2006; Kidd, 2007). The ultimate results will be useful not just for dogwhistles, but modeling the diffusion and evolution of social meaning in populations in general. The initial results presented here is a computational implementation of Henderson and McCready 2018, which will serve as the basis for models with multiple speakers and repeated interactions.

1 Introduction

It is the 2016 US presidential election and Jill Stein is in a predicament. She is doing a Reddit AMA[1] and has just been asked about vaccines. We assume she believes her base is uniformly anti-corporate, but also contains a passionate anti-vax minority that hold a position others in her party don't like. She knows that her anti-corporate bona fides are solid, but the question wouldn't be coming up unless there was some uncertainty about her stance on vaccines. This is the perfect occasion for a dogwhistle. She says:

> By the same token, being "tested" and "reviewed" by agencies tied to **big pharma** and the **chemical industry** is also problematic.

Phrases like 'big pharma' and 'chemical industry' could be read as generic anti-corporate speak, but people familiar with anti-vax discourse know that these phrases are a staple of that genre. By using phrases like this, the general population of listeners, who are unfamiliar with anti-vax discourse, might assume that Stein is being anti-corporate, while anti-vaxers, who are obviously familiar with this discourse, might assume that Stein is one of them because she speaks like one of them. That is, Stein can use a dogwhistle to signal allegiance to ingroup members, while signaling a different, more palatable allegiance to naive outgroup members. Crucially, this signal is plausibly deniable. When outgroup members who were savvy about anti-vax discourse called out Stein for using dogwhistles, she could fall back on statements she had made asserting the efficacy of vaccines.

This is an isolated example, but dogwhistles have been well studied in the political science (Albertson, 2015; Hurwitz and Peffley, 2005; Mendelberg, 2001; White, 2007) and advertising literature (Kanner, 2000; Palmer, 2000). In particular, Albertson 2015 shows that religious dogwhistles are in fact effective in signaling religious affiliation to ingroup prospective voters in ways that non-religious voters who would othwerise disapprove of religious appeals in politics are unable to detect. The linguistic work on dogwhistles is sparse, but Henderson and McCready 2017, 2018, 2019 build a novel theory by extending the *social meaning games* of Burnett 2017, which itself builds of of work in game-theoretic pragmatics, in particular, Bayesian Rational Speech Act theory (e.g., Goodman and Frank 2016; Franke and Jäger 2016; Franke and Degen 2016).

Having an account of dogwhistles, especially one that connects with social meaning and pragmatic reasoning more broadly, is an advance, but the proposal in Henderson and McCready 2017, 2018, 2019 makes no attempt to study the dynamics of dogwhistles in a population. As we have seen though, even the rather schematic Stein ex-

[1]An AMA ('Ask Me Anything') is an online forum for free discussion hosted by Reddit.

ample, the structure of the population is critical. The Stein example imagines a scenario with three groups—ingroup, naive outgroup, and savvy outgroup. Dogwhistles should evolve under the following conditions: (i) ingroup members, in virtue of speaking with each other, should develop linguistic variants that occur at a lower rate than outgroup members, (ii) most outgroup members (naive) should be unaware of these linguistic variants that signal ingroup members, though some savvy outgroup members may be away of ingroup language, and (iii) group membership is punished by outgroup members, but rewarded by ingroup members. In this scenario, and in a specific communication event, speakers could choose one of these linguistic variants if the structure of the audience (proportion of ingroup / savvy outgroup / naive outgroup) is such that it will lead to a positive payoff.

Understanding how dogwhistles arise, are used in particular speech situations, and then fall out of use clearly calls for some kind of agent-based modeling. The fact that Henderson and McCready do not do so is due to lack of tooling. There are currently no existing, off-the-shelf resources for computationally modeling populations of agents playing social meaning games, or even the simpler games discussed in the RSA literature. This paper will present ongoing efforts to develop agent-based models of dogwhistle communication in a population based on probability monads (Erwig and Kollmansberger, 2006; Kidd, 2007), especially the implementation of simple RSA models in Bumford and Charlow 2018.

2 Dogwhistles in social meaning games with probability monads

The Haskell[2] type system provides a clean way to lay out social meaning models, that is, the models in which expressions have their social meaning.[3] One of the core ideas of so-called *Third Wave* variationist sociolinguistics (see Eckert 2012 for a review), is that sociolinguistic practice is deeply creative, with speakers, though their linguistic choices constantly creating a place for themselves in social space. Under this view, linguistic variation is the ferment from which speakers construct, entrench, and mutate social identities though their stylistic practices. Eckert (2008) calls this ferment the indexical field, which is made up of opposing features. Speakers, through selection of language variants creatively construct persona which are sets of these features.

In our Stein example, we treat features as types which can be grouped into an indexical field, or list of indexes of opposing types.

(1)
```
data Feature =
  AntiVax | ProVax | ProCorp |
      AntiCorp
type Index = [Feature]
indices =
  [[AntiVax,ProVax], [AntiCorp,
      ProCorp]]
```

A persona is a maximally consistent list of features drawn from the indexical field. The set of all personas is what Burnett (2017) calls the Eckert-Montague Field. We can generate all possible personas, or the EMField, from selecting one Feature by each Index in every way possible.[4]

(2)
```
personae :: [Index] -> EMField
personae p = sequence p
```

We can now introduce messages with social meaning. We assume that messages have their normal truth conditional meaning, but when we turn to social meaning they are not interpreted in worlds, but instead denote sets of Features. In the Stein example, we assume that expressions like ‘Big Pharma’ is both anti-vax and anti-corporate language, while Stein could have selected some other variant, like ‘Corporate Scientists’, which would be anti-corporate, but in virtue of invoking science, could be interpreted as pro-vax.

(3)
```
data Message = BigPharma | CorpSci
type Denotation = Message -> [
    Feature]
deno :: Denotation
deno BigPharma = [AntiVax, AntiCorp]
deno CorpSci = [ProVax, AntiCorp]
```

The effect of uttering one of these variants is for the listener to rule out assigning the speaker any persona that is inconsistent with that variant. That is, ‘Big Pharma’ tells the listener the speaker is definitely not both pro-vax and pro-corporate. Eval implements this logic, which takes a message and a context (some field of possible personas),

[2] We direct the intererested reader who is not familiar with Haskell to Hackage, https://hackage.haskell.org/, which provides documentation of built-in functions used here.

[3] One can see the complete code discussed here with a working example at https://github.com/bkeej/SocialMeaningExp/blob/master/src/RSAsoc.hs

[4] Not all these personas may be active in a community. Following a reviewer’s suggestion, we could set the prior that an agent bears such a persona to 0 in a community to model this.

and returns just those personas that overlap with the denotation of the message.

(4)
```
type Lexicon = Message -> EMField ->
  [Persona]
eval :: Lexicon
eval m f = nub $ [i | i <- f,
                     p <- i,
                     p `elem` (deno m)]
```

This completes the implementation of the model theoretic aspects of social meaning in Henderson and McCready 2017, 2018, 2019. The real action takes place as speakers use these expressions and listeners infer their personas in a probabilistic setting. Before defining this, though, note that there are actually different kinds of listeners, and these are meant to react differently to dogwhistles. We have ingroup listeners, as well as two kinds of outgroup listeners, those savvy to ingroup language and those who are naive.

(5)
```
data Group = Ingroup | Naive | Savvy
```

The listener's priors for the speaker's persona, as well as how speakers of different personas tend to speak, will now now be conditioned what group they belong to. The probability monad toolkit as described in Kidd 2007; Erwig and Kollmansberger 2006 and implemented in Bumford and Charlow 2018 is built on a set of monad transformers that enrich monads with probabilistic notions that can be computed in the background (e.g., weights, Bayes' theorem, etc.), separating them from code describing the structure at hand.

For instance, the PerhapsT monad transformer attaches probabilities to each computation in the list monad, while the MaybeT monad transformer allows us to throw out branches of of the computation that fail, which permits an implementation of Bayes' theorem via normalizing probabilities of non-failed branches.

(6)
```
type BBDist = MaybeT DDist
```

We start by setting priors for personas via calls to weighted, which constructs a weighed distribution from a list of weights and values.

(7)
```
personaPrior :: Dist m =>
  Group -> m Persona
personaPrior g =
  weighted [Mass 5 [ProVax,ProCorp
      ]...
```

In principle, priors for the speaker's persona can vary by listener type, but for this example, we assume that all listeners are fairly certain Stein is not ProVax,ProCorp=5%, most likely not AntiVax,ProCorp=%15, but think it is equally likely that she is AntiCorp,AntiVax=40% or AntiCorp,AntiVax=40%. This uncertainty is what makes using a dogwhistle a potentially profitable strategy.

Listeners also have beliefs about the probability that they will hear certain messages. The fact that these beliefs can vary by listener type is what will make a particular linguistic expression a dogwhistle. That is, an Ingroup member on knowing a speaker is AntiVax might expect them to use BigPharma because they are familiar with anti-vax rhetoric (the same for Savvy outgroup members). In contrast, a Naive outgroup member would assign a lower probability, maybe placing more probability on purely AntiCorp speakers using the phrase. We see this in the definition of messagePrior:

(8)
```
messagePrior :: Dist m =>
  Group -> Persona -> m Message
messagePrior Ingroup [AntiVax,
    AntiCorp] =
  weighted [Mass 80 BigPharma,...
messagePrior Naive [AntiVax,AntiCorp
    ] =
  weighted [Mass 15 BigPharma,...
```

Finally, we can define the recursive RSA-style reasoning, following the example in Bumford and Charlow 2018, where the literal speaker produces messages based on their persona and priors on how speakers with that persona speak, while listeners guess the speaker's persona based on their priors and a model of what the literal speaker will do. By providing higher integers we get a tower of back-and-forth, probabilistic reasoning between speakers and listeners. Note the guard condition in the literal speaker. The computation will fail for messages whose denotation is not consistent with the given persona. This triggers a reapportioning of probability mass over the surviving branching by the monad transformer BBDist. The result is that Bayesian reasoning happens in the background, while we preserve a clean presentation in code of the structure of these games, exactly as promised by the probability monads.

(9)
```
speaker :: Int -> Group -> Persona
    ->
Lexicon -> BDDist Message
speaker n g p sem = bayes $ do
  m <- messagePrior g p
  scaleProb m $
  if n <= 0 --lit. speaker
   then guard (p `elem` sem m field)
  else do --lit. listener
   p' <- listener n g m sem
   guard (p' == p)
  return m
```

(10)
```
listener :: Int -> Group -> Message
    ->
 Lexicon -> BDDist Persona
listener n g m sem = bayes $ do
  p  <- personaPrior g
  m' <- speaker (n-1) g p sem
  guard (m' == m)
  return p
```

With this recursive reasoning, we can already observe the dogwhistle effect. For instance, assuming the message priors above, on hearing Stein say 'Big Pharma', an Ingroup or Savvy outgroup member assigns a 60% probability that Stein is AntiVax, up from 40%, while the Naive outgroup member only assign a 42% chance, just slightly up from the prior of 40%.

Starting from the speaker's perspective (i.e., literal-speaker vs. literal-listener) makes sense in these sociolinguistic games. Actually, already, before worrying about issues of audience design, we have implemented a probabilistic model so-called 'First Wave' sociolinguistics. That is, speakers are assigned a persona and mechanistically produce variants at a rate given by that speech community—i.e., by messagePrior. We have seen that we can produce the dogwhistle effect even in this First Wave model. As discussed above, Third Wave sociolinguistics is much richer, assuming that speakers (along with their listeners) are constantly collaboratively choosing variants to construct a persona.

3 Adding audiences in the Third Wave

One way to think of the system in its current guise is that it purely models information transfer in the social meaning domain. To get a Third Wave theory, one that can handle richer aspects of dogwhistles in agent-based models, we need to endow speakers and listeners with preferences for personas.[5] This will allow speakers, not just to report their persona, but also to pick messages that allow them to have a persona they like (and the audience likes) in a particular situation.

Once again, we take the speaker's perspective and model the social utility of message and person for a speaker given a listener.

(11)
```
vL :: Group -> Persona -> Float
vS :: Persona -> Float

uSoc :: Message -> Persona ->
 Group -> Lexicon -> Float
uSoc m p g l =
 pr + (vL g p * pr) + (vS p * pr)
   where Sum pr = sum $ [x | Mass x (
       Just y)
    <- runMassT (runMaybeT
       (RSAsoc.listener 1 g m eval)),
            y == p]
```

Speakers now pick a message based on its efficacy in informing a listener about a persona (i.e., pr in uSoc) modified by how listeners and the speaker themself will react to listeners assigning them that persona (i.e., vS and vL in uSoc), weighted by the probability the lister will assign that persona.

Treating an audience as a list of listener types, the utility of a message for a speaker is just the sum of the utility calculation for each listener.

(12)
```
Type Audience = [Group]
uSSoc :: Audience -> Message ->
 Persona -> Lexicon -> Float
uSSoc a m p l = sum $
 map (\g -> uSoc m p g l) a
```

We now have a 'Third Wave'-style model of social meaning for dogwhistles, and make good on the promise made in Section 1. That is, if Stein looks out at her audience and thinks there are a large number of Naive outgroup members, it will increase her social utility to use an anti-vax dogwhistles. The reason is the low probability of detection by Naive outgroup members will meaning high negative affective value for the AntiVax persona will be weighted downward. In contrast, if the audience is mostly Savvy outgroup members, it will not be safe to do so.

4 Conclusions

This paper provides a computational implementation of Henderson and McCready 2018 using probability monads as implemented in Bumford and Charlow 2018. To make the implementation we extend the RSA-style games in that work with model-theoretic logic for social meaning, uncertainty for messages based on persona type, different listener types, and a social utility function implementing social costs for audiences with one or more listeners. In this way, we get formal verification of the work in Henderson and McCready 2017, 2018, 2019, as well as firm foundation for future work developing functional, agent-based models of *dogwhistle* communication.

[5] Note, there are aspects of Third Wave theory that we do not model like bricolage—agents convey parts of multiple personas at once, or the fact that the indexical field itself is dynamic, i.e., "fluidity".

References

Bethany L Albertson. 2015. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37(1):3–26.

Dylan Bumford and Simon Charlow. 2018. Prob-tools.

Heather Burnett. 2017. Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Linguistics and Philosophy*.

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

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

Martin Erwig and Steve Kollmansberger. 2006. Functional pearls: Probabilistic functional programming in Haskell. *Journal of Functional Programming*, 16(1):21–34.

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

Michael Franke and Gerhard Jäger. 2016. Probabilistic pragmatics, or why Bayes??? rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, 35(1):3–44.

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

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. *Signaling without Saying: The Semantics and Pragmatics of Dogwhistles*. To appear from Oxford University Press.

Jon Hurwitz and Mark Peffley. 2005. Playing the race card in the post—Willie Horton era the impact of racialized code words on support for punitive crime policy. *Public Opinion Quarterly*, 69(1):99–112.

Bernice Kanner. 2000. Hide in plain sight. *Working Woman*, 25(3):14.

Eric Kidd. 2007. Build your own probability monads. *Draft paper for Hac 07 in Freiburg*, 7.

Tali Mendelberg. 2001. *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton University Press.

Kimberly Palmer. 2000. Gay consumers in the driver's seat: Subaru's new ad campaign is among those signaling to homosexual buyers. *The Washington Post*.

Ismail White. 2007. When race matters and when it doesn't: Racial group differences in response to racial cues. *American Political Science Review*, 101(2):339–354.