Dogwhistles as Identity-based Interpretative Variation

Quentin Dénigot and **Heather Burnett**
Laboratoire de Linguistique Formelle
5, rue Thomas Mann
F-75205 Paris Cedex 13
qdenigot@linguist.univ-paris-diderot.fr
heather.susan.burnett@gmail.com

Abstract

The following paper presents a formal model for the description of *dogwhistles*. Dogwhistles are a class of expressions often used in political discourse that aim at being interpreted in different ways by listeners of different communities. The model presented here describes this phenomenon using a variation on the Social Meaning Games framework that uses probability distributions over possible interpretation functions.

1 Introduction

Pragmatics has underlined the importance of context in determining the meaning of utterances, and Gricean pragmatics in particular has established a normative framework for the successful transmission of a message between two cooperating agents (Grice, 1975). The insights into human communication that are Grice's conversational maxims have led to formal implementations since Lewis 1969. Most notably, the maxim of quality is the basis for the emergence of scalar implicatures in the Rational Speech Act (RSA) framework (Frank and Goodman, 2012). Grice's maxims, much like Lewis' signalling games, only seek to describe situations where language is used for the sole goal of transmitting accurate information from one speaker to a listener. This is in part what is meant by "cooperation": both sides share the same goal of having the information properly transmitted (whether it be by choosing the right message for the speaker, or choosing the right interpretation for the listener).

This vision, however, only describes a subset of language. It is reasonable, for example, to think that the information content of a linguistic utterance is not limited to the content of the message itself, but that the way in which the message is articulated, either in terms of pronunciation or choice of words, can convey information about the speaker themself. This is what sociolinguists call *social meaning* (Eckert, 2008, 2012): the part of a linguistic signal that conveys information about the person producing the signal rather than the content of the signal itself. It has been shown that intuitions about the social meaning contained in certain accents, for example, has an influence on the reception of a message by the listeners, leading to systematic interpretations of signals that could be at odds with the message conveyed by the content of the message (Acton, 2020). The traditional approach is limited in its scope in the sense that it fails to account for the existence of at least two sources of what could be called "information" in any given linguistic utterance: message content and social meaning.

Works on Social Meaning Games (SMG) (Burnett, 2017, 2019) fill this gap by offering a framework based on game theory (like Lewis' works and like many formal approaches to pragmatics, including RSA) which treats socially significant linguistic variation as another source of meaning. This leads to a variation on signaling games in which the *personae* signaled by the speaker and retrieved by the listener have to match in order to maximize both players' utilities. Crucially, we are talking of *personae*, not *social identity*, because we have to account for cases where the speaker is trying to convey a specific set of traits about themself to the listener for a given goal; they are trying to communicate how they want to be seen in this situation. Here, we are reaching a point where the maxim of quality is, to some extent, flouted, or at least not as relevant.

Examples used for the illustration of SMGs in Burnett 2019 are often political in nature. Political discourse (whether in debates or speeches) is a great field of inquiry for these phenomena because they involve speakers that are publicly known and for whom we can usually access several discourses, including discourses in many different contexts. SMGs can give us an intuitive

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 17–25

view of how social meaning is conveyed, which is key in understanding political discourse, but they fail to account for situations described as *dogwhistle politics*. The term *dogwhistle* refers to a class of expressions often used in political discourse; the goal in using them is to convey two different messages to two different communities. SMGs do not take into account the fact that the audience of a political discourse might be ideologically heterogeneous, leading to differing interpretations of a given message according to prior beliefs and social background. The goal of this work is to define what form situations of dogwhistling might take and to give a formal model describing the contexts in which they are more likely to be used.

2 *Dogwhistles* and *dogwhistle politics*

Dogwhistle politics is generally defined as sending a message to an audience in such a way that a subset of the audience will understand the message differently from the rest of the audience. In more political terms, it is a "way of sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive" (Goodin and Saward, 2005).

To what extent are such practices indeed noticed and what effects do they *actually* have on public opinion? There is a compelling literature on the subject, showing notably that phrases like "*inner cities*" can be responsible for the fact that discussions of nonracial policies can be biased by racial thinking in White voters (Hurwitz and Peffley, 2005) while having a different effect on Black voters (White, 2007). Likewise, it has been shown that the use of religious discourse can also have a significant impact on both opinions and voting intentions for Evangelical voters (Calfano and Djupe, 2009; Albertson, 2015). The effects of dogwhistle speech are backed by empirical evidence and these effects are congruent with the effects that are intuitively attached to the practice: dogwhistle speech reinforces the support of core supporters while being ignored by moderates, in situations where explicit reference to religion or race has negative effects on moderates.

As far as the intentional use of such terms is concerned, we can mention Kuo 2006, who clearly acknowledges it:

> "We threw in a few obscure turns of phrase known clearly to any evangelical, yet unlikely to be noticed by anyone else [. . .]"

The topic, however, has barely been discussed in the linguistics literature. Several theories exist regarding how and why dogwhistles actually work and only recently (starting with Stanley 2015) have these efforts focused on analyzing the language *per se* and trying to give a linguistically consistent description of the phenomenon.

A first approach consists in saying that dogwhistle words have an *explicit* meaning and an *implicit* meaning. This is the approach favored by Mendelberg 2001; Stanley 2015; Henderson and McCready 2019b and Saul 2018. One way of thinking about this (Mendelberg, 2001) is *ambiguity*, each word would have several meanings, for example one racial and the other nonracial, and the use of that term would trigger (or not) one or both of the interpretations in the audience. This makes intuitive sense, but it has important problems, one of them being that the ambiguity that takes place here does not appear to be symmetrical. Khoo 2017 uses the counterexample of the ambiguous word "*funny*" in English, which can either mean "*humorous*" or "*strange*", and remarks that (1) poses no problem.

(1) Smith is a funny man who is not humorous.

Compare with a sentence like (2), which sounds very uncanny.

(2) #Smith is an inner-city pastor who is from, works and lives, in the suburbs.

If the word "*inner-city*" was indeed ambiguous between a racial and a nonracial meaning, one should be able to cancel out one of the two meanings, but it appears that the nonracial meaning is not cancellable, whereas (3) does not seem to cause any weirdness in terms of interpretation.

(3) Smith is an inner-city pastor who is not African American.

If the word "*inner-city*" was properly ambiguous, one would call upon either one of its meanings while disregarding, or even cancelling the other, and this does not appear to be the case.

Stanley 2015 proposes an approach relying on the concepts of *at-issue* and *not-at-issue* contents. The idea here is that dogwhistle words would not be ambiguous *per se*, but that through con-

ventional use, they have acquired a secondary, *not-at-issue* meaning. The problem with this approach, however, is underlined in Henderson and McCready 2019b and Khoo 2017: conventional meanings are generally thought to be non-cancellable, which makes the crucial deniability part of dogwhistles impossible. Compare, for example, with slurs, where the added conventional meaning that gives the listener information about the speaker's attitude towards the community they are referring to is not cancellable (examples from (Henderson and McCready, 2019b)), compare (4) with (5), where "welfare" is thought to dogwhistle a negative attitude towards social programs:

(4) A: Angela Merkel is a kraut!
B: What do you have against Germans?
A: #I don't have anything against Germans. Why do you think I might?

(5) A: Donald is on welfare.
B: What do you have against social programs?
A: I don't have anything against social programs. Why do you think I might?

That deniability is a key point of dogwhistles that differentiates them from slurs or other lexical items imbued with added conventional meaning.

3 Formal model

There have been very few attempts at sketching out a formal representation of dogwhistles and their use, and we argue that any attempt at doing so should present a solution that satisfies the following properties of the phenomenon: dogwhistles are cases of INTERPRETATIVE VARIABILITY, where different listeners should assign different interpretations to a speaker's single utterance. Dogwhistles are most common in situations of POLITICAL CONFLICT between conversational participants (Goodin and Saward, 2005; Saul, 2018; Stanley, 2015). Furthermore, interpretative variability is IDENTITY-BASED: listeners who attribute a religious identity, or *persona* (Eckert, 2008; Agha, 2003), similar to theirs to the speaker will be more likely to interpret the dogwhistle in the religious way than those who believe the speaker holds no specific religious beliefs (Albertson, 2015). Since racist interpretations are conditioned on but not determined by identity, use of a dogwhistle often provides some PLAUSIBLE DENIABILITY to the speaker, which can be useful to them to satisfy the political requirements of a diverse audience. This deniability is important because a SAVVY OPPONENT, someone who does not share the speaker's political ideology but who understands the racist way the dogwhistle can be used, can call the speaker out for this use (Stanley, 2015; Saul, 2018). Finally, as observed by Khoo 2017, whether or not an expression will display identity-based interpretative variability depends on its SPECIFIC FORM: expressions that are truth-conditionally equivalent to *inner city*, such as *city center*, are not semantically variable in the same way.

3.1 Previous approaches

Because of the strategic aspect of dogwhistling, authors such as Henderson and McCready 2019b,a and Asher and Paul 2018a have found game-theoretic pragmatics to be useful for solving the puzzles described above. Henderson & McCready propose a framework in which a speaker, S, sends a dogwhistle message m_D to a listener, L. L has particular beliefs about the persona of the speaker, and they update their beliefs about the world by taking into account L's hypothesized persona and the m_D's literal meaning. To account for SPECIFIC FORM, Henderson & McCready (2019b:6) use axiom schemata (6) which are triggered by the form of the dogwhistle. In the case of *inner city*, S's use of this message (and this message only) triggers the proposition "All neighborhoods at the center of the city are urban African American" in the mind of L, which then allows L to infer that *African American neighborhood* is S's intended meaning. They say (p.8), "The following axiom (6) states that, given that a speaker S with a persona π uses the dogwhistle *inner city*, and given that the hearer believes that inner city neighborhoods are all African American, then normally the speaker intends the inference from his phrasing to this enriched meaning to be made".

(6) $$\begin{aligned} & Use(S, \pi, [inner\_city]) && \wedge \\ & Bel(L, \forall x(inner\_city(x) && \rightarrow \\ & urban\_AA\_neighborhood(x)) && > \\ & Intend(S, Bel(L, urban\_AA\_neighborhood(x))) \end{aligned}$$

Although this innovative framework provides a game-theoretic foundation for identity based interpretative variability, we argue it could be improved. For one thing, axioms such as (6) are required for each dogwhistle, even though pat-

terns of speaker/listener behavior are exactly what game-theoretic systems aim to derive. The account for SAVVY OPPONENT is also not clear: according to (Henderson and McCready, 2019b), the listener beliefs mentioned in (6) are necessary for the dogwhistled content to be inferred; however, politically informed non-racist listeners can detect dogwhistles without them.

3.2 Dogwhistle games

We therefore propose to modify the system presented in Henderson and McCready 2019b to arrive at one which can account for SAVVY OPPONENTS and in which we can prove statements similar to (6) as theorems. Our proposal also builds on Asher and Paul 2018b,a, who highlight the importance of POLITICAL CONFLICT in dogwhistles and use a special *Jury* player to determine conversational success. Rather than invoking an abstract *Jury*, we will have dogwhistles arise from political conflict between the conversational participants themselves.

A *dogwhistle game* G_{DW} is a tuple:
$G_{DW} = \langle \{S, L^i, L^j\}, W, M, \text{PERS}, \text{INT}, \text{I-LEX}, Pr_\pi(\cdot), Pr_w(\cdot), \mu_S, U_S \rangle$
where

- S, L^i, L^j are the speaker and two listeners.
- W is a set of worlds w.
- M is a set of messages m.
- PERS is a set of personae π.
- INT is a set of interpretation functions $[\![\cdot]\!]$.
- I-LEX : PERS $\rightarrow \Delta$INT is a function from personae to probability distributions over interpretation functions.
- $Pr_\pi(\cdot)$ is a probability distribution over personae.
- $Pr_w(\cdot)$ is a probability distribution over worlds.
- μ_S is S's preference function from worlds to N of the form $\mu_S(w)$ where w stands for a message interpretation.
- U_S, a utility function from $M \times W$ to R of standard RSA form.

As is the case in Henderson and McCready 2019b, we have a set of words $w \in W$, a set of messages $m \in M$, and a set of personae $\pi \in$ PERS. We differ from their model in that we have a **set** of interpretation functions: $[\![\cdot]\!] \in$ INT and a lexical interpretation function, I-LEX, mapping personae to probability distributions over INT. The idea is that a speaker's persona will be informative for their form-meaning associations. Given that listeners are rarely certain about the state of the world or even S's political identity, we will represent this uncertainty as prior probability distributions over worlds ($Pr_w(\cdot)$) and personae ($Pr_\pi(\cdot)$).

Following Asher and Paul 2018a, we will allow considerations other than communication to influence S's actions. As in standard SMGs (Burnett, 2019), we have a preference function μ; in our case however, it is applied to preferred worlds for the speaker. The idea behind this is that in dogwhistling situations, we can assume the speaker might not respect the maxim of quality. The goal of the speaker is not to communicate the "*actual state of the world*" but to ensure the support of the audience (by conveying a state of the world that satisfies their beliefs). It is important to point out that the goal of the speaker is, to some extent, to deceive the audience: we are *not* in a cooperative situation, although our listeners will largely assume that we are. In other words, we are in an asymmetrical context.

We conceptualize the interaction situation as parallel to a signaling game for the listeners: they are trying to figure out S's message. Correspondingly, L^i and L^j's interpretation process will closely follow the *Rational Speech Act* model (Frank and Goodman, 2012). In our situation, however, the speaker is duplicitous, and we will represent this duplicity by the use of a preference function μ^*, that takes as input ordered pairs of worlds corresponding to each listener's possible interpretation. Similarly, this duplicitous speaker has their own U^*_S utility function that also takes ordered pairs of worlds as input.

3.3 Determining the listeners' interpretations

The listeners in this model are almost identical to standard RSA listeners, in that they also infer their interpretations from what a speaker faced with a literal listener would say. In standard RSA, speakers and listeners can reason about each other's rea-

soning, leading to an interpretation of messages that relies on their *literal meaning*, but is not necessarily determined by it.

There is one thing in our model that is added to the listeners: it is assumed that they have priors over the possible personae of the speaker and that they can derive different interpretation functions from these priors. A key point of the model presented here is that there exist different possible interpretations for a given message. This will be illustrated with an example in section 4.

From these two priors, using the I-LEX function, listeners can associate a probability distribution over interpretation functions dependent on the priors over personae that they have. The intuition behind this is that listeners assume that a speaker displays a certain persona, and they assume that people belonging to the group that the speaker appears to belong to speak in a certain way. The strength of these assumptions depends on the speaker.

The probability $P(\llbracket\cdot\rrbracket)$ that a certain interpretation function will be used is computed as follows :

(7) For all $\llbracket\cdot\rrbracket, P(\llbracket\cdot\rrbracket) = \sum_{\pi \in \text{PERS}} Pr(\pi) * \text{I-LEX}(\pi, \llbracket\cdot\rrbracket)$

Then each message can be interpreted using one interpretation function or another:

(8) $Pr(w|\llbracket m\rrbracket) = \frac{Pr(\{w\} \cap \llbracket m\rrbracket)}{Pr(\llbracket m\rrbracket)}$

And finally, the meaning of any given message, taking into account all the ways it could have been meant, uses both values, giving us the *literal listener*:

(9) $P_{L_0}(w|m) = \sum_{\llbracket\cdot\rrbracket \in \text{INT}} P(\llbracket\cdot\rrbracket) * Pr(w|\llbracket m\rrbracket)$

The subsequent steps are similar to standard RSA in that the speaker computes the utility of each message, and using this utility score for a given message, we can have probability distribution over the different messages the speaker can send, given what they want to convey. Where we differ from standard RSA is in the use of the μ function giving the preference of the speaker over *possible interpretations*. What we assume is that listeners have a bias towards thinking that speakers think like them and therefore have similar preferences.

The utility of the speaker is computed as follows, where C is a cost function:

(10) $U_S(m, w) = log(P_{L_0}(w|m)) - C(m)$

The probability distribution over their possible messages is computed as follows, where α is a temperature parameter governing how much variability the system allows:

(11) $P_S(m|w) \propto exp(\alpha U_S(m, w)) * exp(\alpha' \mu(w))$

Adapting from (Burnett, 2019), we can infer from the μ function and the P_S values a probability distribution over the possible messages to have an idea of the speaker's behavior as it is envisioned by all speakers using the following two formulae:

(12) $P_W(w; \mu) = \frac{exp(\alpha' * \mu(w))}{\sum_{w' \in W} exp(\alpha' \mu(w'))}$

(13) $\mathcal{P}_S(m) = \sum_w P_W(w; \mu) * P_S(m|w)$

Finally, given those assumptions about the speaker, pragmatic listeners L_1 will try to interpret the meaning of what was just said by S using:

(14) $P_{L_1}(w|m) \propto Pr(w) * P_S(m|w)$

3.4 The duplicitous speaker

This would be a classic RSA model involving one speaker and one listener. But the case of dogwhistles, it is assumed that there are several listeners in the crowd. More specifically, there are at least two different listeners with different opinions, and the goal of the speaker is to satisfy them both.

We will distinguish several instances of the speaker in this model. Some of these instances are the speakers we described in the previous section. We will call them the *honest speakers*; they are speakers that standard listeners assume they are talking to. But the context in which dogwhistles appear call for another kind of speaker, that we will call *duplicitous*. The duplicitous speaker differs from the honest speakers in one major way: they always consider pairs of worlds when computing any of the aforementioned probabilities. Therefore, instead of a utility function $U_S : M \times W \rightarrow R$, they have a utility function U_S^* defined as follows, where the indices i and j serve to differentiate the two different listeners:

(15) $U_S^*(m, \langle w, w' \rangle) = log(P_{L_0}^i(w|m)) + log(P_{L_0}^j(w'|m)) - C(m)$

Similarly, the preference function μ becomes μ^*

and takes ordered pairs of worlds as inputs. It is important that the pairs of worlds are ordered, because the duplicitous speaker seeks to treat the two different listeners differently.

3.5 The savvy listener

We mentioned previously that a satisfactory model for dogwhistles should take into account the possibility of a *savvy listener*, i.e. a listener that can see through what the speaker is saying and infer that there are in fact several messages being communicated here.

In this model, the savvy listener is equivalent to a listener that would act like the duplicitous speaker does; in other words, the savvy listener also takes into account the fact that there are various listeners with various beliefs and preferences, and assumes that the speaker is going to try and take advantage of that fact. Therefore, they assume that the speaker uses a function μ^* and a utility function U_S^* and computes the intended meaning using this supplementary data, using (16):

(16) $P^{Savvy}(\langle w, w' \rangle | m) \propto P_W(\langle w, w' \rangle; \mu^*) * P_{S_{Dup}}(m | \langle w, w' \rangle)$

4 '*Inner cities*' example

By way of illustration, we will focus on the often-used example of *"inner cities"*. Following (Henderson and McCready, 2019a,b), let S be Representative Paul Ryan trying to discuss issues in urban areas while also trying to gather the support of his more right-wing voters. Let L^i be a stand-in for a more conservative voter and L^j a less conservative voter. More specifically, L^j refuses to speak openly of race in that context and Ryan would lose their support if he did, whereas L^i would be more likely to offer their support to Ryan if he did take into account race in his discourse.

This is the perfect situation for a dogwhistle. Here is the original statement by Ryan:

(17) We have got this tailspin of culture, in our *inner cities* in particular, of men not working and just generations of men not even thinking about working or learning the value and the culture of work.

Let M be the set of expressions available to S. To simplify, we will distinguish between three possible messages: {"inner cities", "city centers", "African-American neighborhoods"}. These differ by being more or less open references to race, with "city centers" ignoring race completely and "African-American neighborhoods" putting it front and center. "Inner cities" is our dogwhistle term. Because of those properties, we will label these messages as follows:

$M = \{m_D, m_{\neg R}, m_R\}$

To this we add the set W of worlds, which distinguishes between worlds where the issue is about race and worlds where it is not:

$W = \{w_R, w_{\neg R}\}$

Let PERS contain the possible personae of "racist conservative" and "non-racist conservative" :

PERS $= \{\pi_i, \pi_j\}$

Finally, we also set the following :

- INT $= \{[\![\cdot]\!]_i, [\![\cdot]\!]_j\}$

  There are two interpretation functions, one corresponding to each of the personae we are considering here.

- I-LEX$(\pi_\rho, [\![\cdot]\!]_\rho) = 1$

  For simplicity, we assume that for any persona π_ρ, our listeners will associate it invariably with the corresponding interpretation function, meaning that they assume that people displaying persona π_ρ will always mean $[\![m]\!]_\rho$.

- $Pr_w^{L_{1/2}}(w_{R/\neg R}) = 0.5$

  The probability distribution over worlds is uniform, meaning that people have no priors regarding what is going to be said.

- $Pr_\pi^{L^i}(\pi_i) = 0.6 = Pr_\pi^{L^j}(\pi_j)$
  $Pr_\pi^{L^i}(\pi_j) = 0.4 = Pr_\pi^{L^j}(\pi_i)$

  Each listener assumes that the speaker is a bit more likely to display one persona over the other. In this scenario, we can interpret this as L^i recognizing themselves more in persona π_i and therefore having a bias towards thinking that S is more likely to be displaying that same persona, and symmetrically for L_j.

- For simplicity of computation, we assume that messages are *costless* and we set temperature parameters α are set to 1.

See Table 1 for the result of applying each interpretation function $[\![\cdot]\!]$ to each message m.

$Pr(w\|[\![\cdot]\!]_i)$	w_R	$w_{\neg R}$
m_R	1	0
$m_{\neg R}$	0	1
m_D	1	0
$Pr(w\|[\![\cdot]\!]_j)$	w_R	$w_{\neg R}$
m_R	1	0
$m_{\neg R}$	0	1
m_D	0	1

Table 1: Interpretation of each message according to different interpretation functions.

The important value we want to have here are the probabilities ascribed to the interpretation of each message $P_{L_{i/j}}(w|m)$ for each listener, as well as the $P_S(m|w)$ score for each message of what the speaker is expected to say given that they wish to communicate w. After the relevant computations, applying the μ functions in table 2, we have the numbers in table 3 and table 4.

	w_R	$w_{\neg R}$
μ^i	2	1
μ^j	0	2

Table 2: μ functions as envisioned by honest listeners. L^j assumes that the speaker does not want to convey race-based interpretations (because they themselves despise them).

$P_{L_1^i}(w\|m)$	w_R	$w_{\neg R}$
m_R	1	0
$m_{\neg R}$	0	1
m_D	≈ 0.567	≈ 0.432
$P_{L_1^j}(w\|m)$	w_R	$w_{\neg R}$
m_R	1	0
$m_{\neg R}$	0	1
m_D	≈ 0.254	≈ 0.745

Table 3: Interpretations for honest pragmatic listeners of each message depending on their priors.

What we can see with these numbers is that honest listeners believe that honest speakers would be more likely to use m_R if they wish to convey w_R, but that using m_D is not seen as impossible. Similarly for $m_{\neg R}$ and $w_{\neg R}$. What it also shows, however, is that if they hear m_D, they are more likely to interpret it according to the interpretation function that they deemed more probable, given their prior beliefs about the speaker.

$P_S^i(m\|w)$	w_R	$w_{\neg R}$
m_R	0.625	0
$m_{\neg R}$	0	≈ 0.714
m_D	0.375	≈ 0.286
$P_S^j(m\|w)$	w_R	$w_{\neg R}$
m_R	≈ 0.714	0
$m_{\neg R}$	0	≈ 0.770
m_D	≈ 0.286	≈ 0.230

Table 4: Speaker probabilities for an honest speaker for each listener.

We can use (12) and (13) to have a better idea of the behavior of the speaker. In this case, we obtain that $\mathcal{P}_S^i(m_D) \approx 0.351$ and $\mathcal{P}_S^j \approx 0.271$, so the use of m_D will be somewhat unexpected, but still more or less in keeping with the idea of such a speaker.

We now consider the duplicitous speaker S_{Dup}, who uses the literal listener values along with the μ^* function presented in table 6. The duplicitous speaker uses these in conjunction with (15), giving us the values in table 5. Using (12) and (13) again to have a better picture of how such a speaker could be predicted to act, we find that $\mathcal{P}_{S_{Dup}}(m_D) \approx 0.752$. Such a speaker is much more likely to use a dogwhistle!

$P_{S_{Dup}}(m\|\langle w, w'\rangle)$	m_R	$m_{\neg R}$	m_D
$\langle w_R, w_R\rangle$	$\approx$0.806	0	$\approx$0.194
$\langle w_R, w_{\neg R}\rangle$	0	0	1
$\langle w_{R\neg R}, w_R\rangle$	0	0	1
$\langle w_{R\neg R}, w_{\neg R}\rangle$	0	$\approx$0.806	$\approx$0.194

Table 5: Speaker probabilities for a duplicitous speaker following the μ^* function in table 6

		w^j	
μ^*		w_R	$w_{\neg R}$
w^i	w_R	0	2
	$w_{\neg R}$	0	1

Table 6: $\mu^*(\langle w^i, w^j\rangle)$ function used by the duplicitous speaker, their main objective is not to be seen as racist by L^j.

A savvy listener L^{Savvy} in this framework is simply a listener who assumes the duplicity of the speaker and bases their interpretation of the speaker message using $P_{S_{Dup}}$ instead of P_S. To compute the intention of the speaker, L^{Savvy} uses the P_W values used at the previous step by the duplicitous listener, following (16) leading to the numbers in table 7. The savvy listener interprets that when the dogwhistle is used there is a very high chance that the speaker is trying to appeal to audiences with opposing points of view!

$P_{L^{Savvy}}(< w, w' > \mid m)$	m_R	$m_{\neg R}$	m_D
$\langle w_R, w_R \rangle$	1	0	$\approx$0.021
$\langle w_R, w_{\neg R} \rangle$	0	0	$\approx$0.811
$\langle w_{R\neg R}, w_R \rangle$	0	0	$\approx$0.109
$\langle w_{R\neg R}, w_{\neg R} \rangle$	0	1	$\approx$0.058

Table 7: Interpretations for savvy listener of each message according to their priors about the speaker.

We argue that our model accounts for the main properties of dogwhistles in the following ways:

- INTERPRETATIVE VARIABILITY: the listeners do not assign the exact same interpretations to dogwhistles. In our example, L^i thinks it is just a bit more likely that m_D conveys a racial meaning rather than no racial meaning at all, and L^j has the opposite view.

- POLITICAL CONFLICT: the use of m_D only presents interest if there is a situation of political conflict, reflected in the duplicitous speaker preferences μ^*. In our example, L^j understanding w_R is what the speaker desires least; whereas, they want for L^i to understand w_R. In fact it can be shown that if we do not have political conflict in this sense, then the game reduces to a signaling game and the utility of using an ambiguous message like m_D is always lower than that of the other two m.

- IDENTITY-BASED: the system used here is identity-based given the fact that priors over the persona of the speaker have an influence on the interpretation function that will be favored (in our simple example, it fully determines it).

- PLAUSIBLE DENIABILITY: As long as members of the audience acknowledge that others might be from different social groups and use different interpretation functions, the meaning of the dogwhistle is never completely clear.

- SAVVY LISTENER: The savvy listener in our model is an individual who assumes that the speaker is being duplicitous and that they have motives beyond communicating information about the world. Otherwise, they use the same mechanisms as other speakers.

5 Conclusion

We have constructed a model that allows us to describe the processes behind the use of dogwhistles by using mechanisms that are already widely used in pragmatics to describe scalar implicatures and social meaning interpretation. We have assumed that a group of more "naïve" listeners use regular RSA-style computations to infer the probable meaning of a dogwhistles utterance while still taking into account the fact that the words used could have different meanings for other audiences. Duplicitous speakers willing to convey two different messages to audiences with different preferences and biases will use dogwhistles to do so and it is likely that they will be understood in the way they intended given what they were assuming of the crowd.

Meanwhile, savvy listeners assume that speakers are indeed duplicitous and conclude that the most likely interpretation after hearing a dogwhistle term is that the speaker is trying to appeal to two different audiences.

Following a similar path and adding iterations in the reasoning, we could easily model other roles in the dogwhistling game, including for example speakers who use dogwhistles in order to specifically trigger savvy listener reactions and then defend themselves from accusations of duplicity by calling out savvy listeners for *ad hominem/appeal to motive* positions.

We think that the framework used here could be generalized to other cases of identity-based interpretation, including cases outside the realm of political discourse, where the meaning intended by speakers is sometimes vague enough to trigger various interpretations from listeners.

References

Eric Acton. 2020. Pragmatics and the third wave. *Social Meaning and Linguistic Variation: Theorizing the Third Wave*.

Asif Agha. 2003. The social life of cultural value. *Language & communication*, 23(3-4):231–273.

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

Nicholas Asher and Soumya Paul. 2018a. Bias in semantic and discourse interpretation. *arXiv preprint arXiv:1806.11322*.

Nicholas Asher and Soumya Paul. 2018b. Strategic conversations under imperfect information: epistemic message exchange games. *Journal of Logic, Language and Information*, 27(4):343–385.

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

Heather Burnett. 2019. Signalling games, sociolinguistic variation and the construction of style. *Linguist and Philos*, 42:419–450.

Brian Robert Calfano and Paul A Djupe. 2009. God talk: Religious cues and electoral support. *Political Research Quarterly*, 62(2):329–339.

Penelope Eckert. 2008. Variation and the indexical field 1. *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.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Robert E. Goodin and Michael Saward. 2005. Dog whistles and democratic mandates. *The Political Quarterly*, 76(4):471–476.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Robert Henderson and Elin McCready. 2019a. Dogwhistles and the at-issue/non-at-issue distinction. In *Secondary Content*, pages 222–245. Brill.

Robert Henderson and Elin McCready. 2019b. How dogwhistles work. In *JSAI International Symposium on Artificial Intelligence*, pages 231–240. Springer.

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.

Justin Khoo. 2017. Code words in political discourse. *philosophical topics*, 45(2):33–64.

David Kuo. 2006. *Tempting faith: An inside story of political seduction*. Simon and Schuster.

David K. Lewis. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell.

Tali Mendelberg. 2001. The race card: Campaign strategy. *Implicit Messages, and the*.

Jennifer Saul. 2018. Dogwhistles, political manipulation, and philosophy of language. *New Work on Speech Acts*, pages 360–383.

Jason Stanley. 2015. *How propaganda works*. Princeton University Press.

Ismail K 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.