Sticks and Stones

The Pragmatics of Harmful Language Use

1 Introduction

When Donald Trump was asked to explain calling women "fat pigs", "dogs", "slobs" and "disgusting animals" his reply was: "*I've been challenged by so many people, I don't frankly have time for total political correctness. And to be honest with you, this country doesn't have time either*" [13].

In contrast to Mr. Trump, the authors of the present paper had time. We provide a formal framework to analyze the pragmatic phenomenon of harmful use of language. If particular word choices are not a central element in the transmission of information, as tacitly suggested by Mr. Trump, and may even be an impediment to the speaker, why may they still refrain from using certain expressions rather than others? Furthermore, under which circumstances are speakers disposed to alter their way of expressing themselves? We address these questions in the context of game theoretic models of natural language use. Specifically, we analyse situations where a speaker is indifferent to the use of one expression over another, while her audience has a clear preference. That is, we model communication in which the wording of an utterance can have an impact on the listener, despite its irrelevance in terms of information transfer. Cases in point are racial slurs and so-called *trigger words* that cause listeners suffering from posttraumatic stress disorder (PTSD) to re-live past trauma [4, 14]. So although two expressions might be semantically identical, choosing one over the other can have a substantial effect on actual communication. We thus take the pragmatic slogan seriously that there is far more to communication than the literal meaning of the expressions employed (cf. [9]).

We will present three core results: First, we offer a formal explanation for the reason why the offensive use of language is so persistent. Second, we argue that there is potential for better communication, according to efficiency standards presently introduced. Third, we will argue that the degree of empathy language users exhibit is of crucial importance to reap these benefits. To this end, we use a game theoretic model of rational language use between agents that reason about each other's linguistic choices [8]. This approach looks to explain and predict pragmatic inferences of natural language users by modeling a back-and-forth reasoning between two agents engaged in conversation. Our formal contribution is to forgo the common assumption that it is exclusively the player sending a message for whom the word choice is costly. Instead, we will look at situations where the cost a message's recipient faces also varies depending on the expressions employed by the sender.

Our exposition proceeds as follows: In Section 2 we give a brief introduction to the application of game theory to the study of pragmatics as well as the specifics of the model

we use. Section 3 contains the model's application to message related cost and a discussion of its predictions, while Section 4 presents the addition of costly signals to extend the model. Section 5 discusses the results and concludes.

2 Game Theory and Pragmatics

Signaling Games

We model language use as a signaling game. These are sequential games between two players: a sender and a receiver [3]. The game is played in a particular state of the world, formally an element of set T which represents possibles ways the world could be. Only the sender knows which state this happens to be and would like to transfer this information to the receiver. The receiver, who cannot observe the true state, tries to infer the true state by observing a signal which the sender emits.

The game is then played as follows: After the sender observes the true state of the world t she sends some message m_i, $m_i \in M$ where M is a set of messages. The receiver then observes the message m_i and points to some $t \in T$ which she thinks holds. If this is the actual t, communication was successful. The players' payoff depends on whether the communication was successful and which message was used. The different messages bear a specific cost for each player, formally this is a player-specific cost vector.

A sender's *strategy* is a function from T to M; a receiver's strategy is a function M to T. Informally, strategies are plans of actions for the players which tell them what to do whenever they get to make a move. We call a strategy *pure* if a player will choose some action with certainty, that is, always, when confronted with a given situation. In contrast to pure strategies stand *mixed* strategies where players will choose from various actions, each with a given probability. Players are assumed to be payoff maximizers. They will play the strategies which lead to the highest expected utility, denoted by EU.

Extension Towards Pragmatics

Game theory has been used to model a variety of phenomena related to information transmission ([12, 11, 2]). In the field of linguistics, it was applied to obtain a firmer conceptual grip on Gricean pragmatics (cf. [1, 10]). Here the communication between (rational) agents is thought of as a game played between interlocutors. Formulating situations where pragmatic phenomena arise with mathematical precision helps in different ways: On the one hand, it allows one to bridge the gap between theoretical work on pragmatics and experimental data with language users (e.g. [5]). More specifically, it generates testable predictions about language use under a variety of parameterizations, as pertaining to the agent's (limited) rationality, cognitive ability to think multiple steps ahead, their preferences and methods of deciding on what action to take or their beliefs about the world. On the other hand, such models provide insight into how agents actually arrive at the linguistic behavior they exhibit. Indeed, models concerned with the pragmatic behavior of linguistic agents are caught between the pressure of trying to accurately predict and fit experimental data and the hope to realistically capture one or the other aspect of a pragmatic phenomenon - to represent "what actually is going on", so to speak.

The model we are concerned with tries to capture the agents' mutual reasoning about each other. Which action the players deem optimal for them and perform depends on a process of reasoning about what the other player believes. The model tries to capture this process of reasoning about what the other player does, believes that her opponent will do, believes what

her opponent will believe that she will do, and so forth.

The IBR Model

We use the *iterated best response* (IBR) model as our formal framework [7], [8]. IBR relies on an explicit representation of the agents' beliefs about each other. The interlocutor's beliefs are spelled out in a round-for-round manner in order to capture what intuitively could be the described as the epistemic dynamic of: "I think that you think that I think that you think...etc.". Language users in the model have some hierarchically ordered level of sophistication which correspond to the number of steps an agent can reason back and forth between herself and the other agents. The reasoning chain formed by such steps is then a sequence of iterated optimal responses. This chain is bounded by the maximal depth a player of a given level of sophistication - or strategic type - can go to.

Entirely unsophisticated types, players who do not take the opponent's reasoning into account at all and are thus completely unstrategic, are assigned the level-0. In the games under discussion in this paper, this will mean that senders of level-0 will be only concerned with saying whatever is true according to the semantics of the language fragment in question, while receivers of that level will always interpret a message they receive literally. An agent of level-k then forms some belief about the other's behavior who, by assumption, is some level l, where $k \geq l$. Intuitively, this can be interpreted as the player looking down the sophistication hierarchy in order to build some rational expectation of what the other of some specified level $l \leq k$ will do. We also make the common assumption that players of level k expect their opponents to reason exactly one step less than themselves and are thus of level $k-1$.

As in standard signaling games, T is the set of states the players can be in, M the set of messages at the sender's disposal. We assume that T and M are finite sets, $|M| = m$, $|T| = n$. A sender strategy is a row-stochastic $(n \times m)$-matrix S, and a receiver strategy is a row-stochastic $(m \times t)$-matrix R. The matrices' entries represent the probability with which the players choose a move in the column - messages for senders, states for receivers - when observing the situation in the row - state for senders, messages for receivers, $P(m|t,S) = S_{tm}$ and likewise $P(t|m,R) = S_{ma}$.

The semantics of the language fragment in question is modeled as a $(n \times m)$-matrix. If a state t_i makes a message m_j true, then the entry x_{ij} in the matrix will be 1, otherwise 0. This matrix reporting the boolean values is called the *Boolean Matrix* and denoted by B. Players of level-0 are supposed to be *naive* with respect to sending and receiving messages. Naive senders always send true messages, while naive receivers always take messages literally. The strategies of level-0 type are represented by the normalized boolean matrix. Normalization here simply means a mapping of some $(m \times n)$-matrix A, onto some other $(m \times n)$-matrix B such that $B_i \propto A_i$ if $\sum_j (A_{ij}) > 0$ and $B_{ij} = \frac{1}{n}$ otherwise. In the case of the receiver, the transposed boolean matrix is normalized. More sophisticated players are modeled to have some belief about the interlocutors behavior. We assume that players view all their opponents' strategies to be equally probable at first. [1]

Finally, we define the notion of *best response* given a set of beliefs. A sender's best response to some receiver strategy R is any strategy that maps each state to that signal which maximizes the expected utility. In case n signals are tied for expected utility, the sender will play each with probability $\frac{1}{n}$. A receiver's best response to a sender's strategy S is defined almost analogously. Except that there may be so-called *surprise messages* m_j for which $Rij = 0$, for all i. These are messages the sender will use with probability 0, that is, never.

[1] This corresponds to unbiased beliefs. We are leaving the details aside for ease of exposition, see [6] for details.

Since strategies are plans of action for every conceivable contingency that might arise in the course of a game, we assume that a receiver will then play each option with probability $\frac{1}{|T|}$.

3 Harmful Signals

Say two people share some language and engage in conversation, one as a speaker, the other as a listener. The speaker would like to communicate what kind of work a mutual acquaintance of the two does. The listener does not yet know what her occupation happens to be. The acquaintance could either be an ecologist or a geologist. To transfer this knowledge the speaker can utter either "She's an ecologist", "She's a professional treehugger, stupidly trying to protect plants and stuff", "She's a geologist" or "She's a professional stonehugger, stupidly looking at rocks and stuff". Assume that uttering any of the first two is true just in case the person is an ecologist and the last two if she's a geologist. As both of the two well know, the listener cares dearly for the environment, while the speaker does not. The listener, although she understands perfectly well what the speaker is trying to convey to her, would be deeply offended by her uttering the word "treehugger". The speaker herself couldn't care less what words are used when talking about ecology, as long as the message comes across and the hearer adopts the right belief about the mutual acquaintance's profession. However, the speaker has a deep-rooted fear of rocks. Hearing the word "stonehugger" sends chills down her spine.

Using the IBR framework we can model this as follows. There are two players, P_1 and P_2. For now, let P_1 be the sender and P_2 the receiver. Let $T = \{t_{geo}, t_{eco}\}$, where t_{geo} is a world where the acquaintance is a geologist and in t_{eco} an ecologist. We let both worlds be equally likely. Both receiver and sender get a payoff of a if communication is successful.

The set of messages is $M = \{m_{geo}, m_{stone}, m_{eco}, m_{tree}\}$. Its semantics are given by B.

$$B = \begin{array}{c} \\ t_{geo} \\ t_{eco} \end{array} \begin{array}{c} \begin{array}{cccc} m_{geo} & m_{stone} & m_{eco} & m_{tree} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

m_{stone} being played inflicts a cost of α to P_1, while m_{tree} bears cost β for P_2, where $a > \alpha, \beta > 0$.[2] Using the same ordering as above, the cost vectors are $c^{P_1} = (0, \alpha, 0, 0)$ and $c^{P_2} = (0, 0, 0, \beta)$. As a first step, we determine what the naive players' strategies are:

Level-0 Players

$$S_0 = Norm(B) = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix} \quad R_0 = Norm(B^T) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

We can see that the naive sender plays a mixed strategy when communicating the state, sending either applicable message half of the time. The receiver's actions are clearly determined by a pure strategy. The best response of k-level players is based on the interlocutor's $k-1$ level when message cost is taken into account.

[2] Communicating is always preferable to non-communication and no expressions can only be harmful.

We obtain the best response of the level-1 sender by analysing her reasoning about a level-0 receiver. She looks at what a naive receiver would do and then chooses the best answer to that strategy, while accounting for the cost of her own moves. We proceed analogously for the receiver. This then yields strategies for both level-1 players: **Level-1 Players**

$$S_1 = BR_S(R_0^T - c^{P_1}) = BR_S(\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} - \begin{pmatrix} 0 & \alpha & 0 & 0 \end{pmatrix})$$

$$= BR_S(\begin{pmatrix} 1 & 1-\alpha & 0 & 0 \\ 0 & 0-\alpha & 1 & 1 \end{pmatrix}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}$$

$$R_1 = BR_R({S_0}^T - c^{{P_2}^T}) = BR_R(\begin{pmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 0.5 \\ 0 & 0.5 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ \beta \end{pmatrix}) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

We see that $R_1 = R_0$, i.e. the receiver did not change strategy from 0-level to 1-level, thus we conclude that $S_2 = S_1$. But $S_1 \neq S_0$, so we calculate:

$$R_2 = BR_R({S_1}^T - c^{{P_2}^T}) = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$S_3 = S_2$ and $R_3 = R_2$. Since neither strategy changed from level 2 to level 3 we have reached a fixed point of the reasoning process. That is, $< S_2, R_2 >$ is a stable strategy pair.

Some remarks about these strategies are in order. First, the sender will never send the message she finds offensive, given a non-offensive alternative. Second, although the sender will never play m_{stone} the receiver still needs has some plan of action for that contingency. Thirdly, while a sender of level zero still harms herself by blindly sending a costly message, senders above level zero will stop doing so. Thus, m_{stone} will not be seen on higher levels. Fourth, and most relevant to our discussion, is that even though the m_{tree} is costly for the receiver, the sender will still send it half of the time. This point is the main focus of the next section.

3.1 Room for Improvement

The players' reasoning leads to a situation where communication is successful. Using the stable strategy pair $< S_2, R_2 >$ the *sender's payoff* is always a. For the receiver, however, the story is different. When S_2 is played and t_{eco} holds, m_{tree} will be sent with probability 0.5. Since the probability of t_{eco} is 0.5, m_{tree} is sent in $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ of all cases on average, inflicting a cost of β. The *receiver's payoff* is a mere $a - \frac{\beta}{4}$ as shown in the table below.

To provide some intuition, consider the positions the agents are facing in this game. Whether the receiver likes the message or not, she can always reason what her interlocutor wants to communicate to her. If the messages m_{geo} or m_{eco} are sent, she will choose the

Table 1: Expected Payoff for Level of Sophistication

	Sender	Receiver
Level-0	$a-\frac{\alpha}{4}$	$a-\frac{\beta}{4}$
Level-k, $k>0$	a	$a-\frac{\beta}{4}$

corresponding action, communication is successful, so both players get their payoff and everyone is happy. If the message m_{tree} is sent, however, she pays the associated cost. Still, she knows what the sender is trying to communicate and her best response will be to cooperate. Refusing to understand is not an option, for it would only add insult to injury; she would both pay the cost of hearing m_{tree} and lose the payoff of understanding correctly. Because the sender can anticipate the receiver to behave this way, she does not need to care about sending either true message.

Usually, language users will encounter the expressions of their language many times over the course of their communicative lives - both as listeners as well as speakers. To capture this we'll posit a third player - nature - who decides uniformly at the beginning of the game who talks and who listens. Thus their expected utility in repeated interactions is $\frac{1}{2}\times a+\frac{1}{2}\times(a-\frac{i}{4})=a-\frac{i}{8}$, for $i\in\{\alpha,\beta\}$.

The crucial observation here is that the agents could have done better. To see this, imagine the agents only used m_{eco} and m_{geo} to communicate. That is, if they do not inflict cost on each other and communication is still successful, then both receive a full payoff of a. Never playing costly messages formally means that the respective sender plays strategy S_* where neither m_{tree} nor m_{stone} are ever played. This leads us to the strategy pair $<S_*,R_*>$:

$$S_*=\begin{pmatrix}1&0&0&0\\0&0&1&0\end{pmatrix}\quad R_*=\begin{pmatrix}1&0\\0.5&0.5\\0&1\\0.5&0.5\end{pmatrix}$$

If this sender strategy is still able to convey the correct state, then everyone is better off or at least as well off as before. Indeed, such an improvement would make the players' payoff meet the standards for *Pareto optimality*. Pareto optimality captures the fact that no player could possibly be better off without the other player losing out. In the present context, it means that the potential for better communication is only exhausted, once the players reach Pareto optimality. Technically speaking, some strategy is Pareto optimal if it is not *Pareto dominated* by any other strategy. We say that the strategy s' *Pareto dominates* s, in turn, if s' gives any player a higher utility than s, while no other player's utility decreases. It is straightforward to see that the strategy pair $<S_*,R_*>$ is Pareto optimal since the highest possible utility for either player is a and $<S_*,R_*>$ has a utility of a for both players, therefore there is no other strategy which *Pareto dominates* it. Furthermore, we see that $<S_*,R_*>$ *Pareto dominates* the IBR strategy which has expected utility of $a-\frac{i}{8}$, for $i\in\{\alpha,\beta\}$. In the next section we show that a small adjustment to our model suffices to reach the strategy pairs $<S_*,R_*>$.

4 Reaching the Optimum

So far it has been supposed that players in the role of the sender do not care about the harm they inflict when choosing one over the other message. We have seen that the players then arrive at non-optimal strategies because they are indifferent to the respective other's cost. However, if this assumption concerning the indifferent stance of the sender is given up, then the employment of different strategies is predicted. We can gain clarity under which circumstances agents will play strategies that are Pareto optimal.

More to the point, we reach the optimal solution if both players must consider receiver's preferences. We do this by extending the sender's cost vector:

$$c^{P_1*} = c^{P_1} + d \cdot c^{P_2}$$

Where c^1 is the original sender's cost, c^2 the receiver's cost and $d \in \mathbb{R}$ a parameter representing the players' disposition to "feel" the other's cost. Intuitively, if $d > 0$ the players are empathetic, if $d < 0$ they are malicious and $d = 0$ is the indifferent case. By extending the model in this way we can make the sender's interest in the receiver's cost explicit. Setting $d = 0$ yields the original model because then $c^{1*} = c^1$, for all other values of d we get an extended model. Let us now play the previous game with this extension. S_0 and R_0 remain unchanged. So let us start with S_1:

$$S_1 = BR_S(R_0^T - c^{1*}) = BR_S(\begin{pmatrix} 1 & 1-\alpha & 0 & 0-d\cdot\beta \\ 0 & 0-\alpha & 1 & 1-d\cdot\beta \end{pmatrix})$$

Assume, for example, the sender fully considers the receiver's cost - $d = 1$. Notice, however, that we would get the same prediction for any $d > 0$.

$$S_1 = BR_S(\begin{pmatrix} 1 & 1-\alpha & 0 & 0-\beta \\ 0 & 0-\alpha & 1 & 1-\beta \end{pmatrix}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} = S_*$$

$R_1 = R_0$ thus $S_2 = S_1$, we end up with $S_2 = S^*$ and $R_2 = R^*$ as stable strategies. As already shown, this strategy pair is Pareto optimal.

5 Discussion and Conclusion

Equivalent expressions in terms of information transfer can differ greatly when it comes to pragmatics. What terms a language user perceives as harmful alters the way she speaks and how beneficial conversation is too her. We represented the harm inflicted by the use of certain expressions as costly messages in a signaling game and analysed under which circumstances players choose costly signals when there are other free signals available. Our IBR model showed that rational language users fail to behave optimally because they do not consider each other's interest. However, they could improve their communication, according to standards of Pareto efficiency, if they were to show even a minimal amount of consideration for each other.

To make this case, we extended standard IBR model with an empathy parameter - d - that captures how the players view the interlocutor's cost. For any positive value for d the extended IBR model yields the prediction that players receive an optimal payoff. Viewing

our results in an informal way suggests that simply caring a little bit about the effect that one's choice of words has, aside from pure information transfer, may produce great benefits for all. Time invested in political correctness might not be wasted after all.

Due to constraints of space, we left relevant aspects out of our model. For example, we did not model that omitting expressions might imply cost for the sender, as she actively needs to alter her speech and figure out what the receiver's preferences are. Another aspect not taken into account is that the degree of empathy could vary between players. Furthermore, it is subject to empirical research to what extent rational speakers really do behave as our model predicts.

References

[1] A. Benz, G. Jäger, and R. Van Rooij. An introduction to game theory for linguists. In *Game theory and pragmatics*, pages 1–82. Springer, 2006.

[2] I.-K. Cho and D. M. Kreps. Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221, 1987.

[3] L. David. Convention: a philosophical study, 1969.

[4] N. Fagan and K. Freme. Confronting posttraumatic stress disorder. *Nursing2016*, 34(2):52–53, 2004.

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

[6] M. Franke. Game theoretic pragmatics. *Philosophy Compass*, 8(3):269–284, 2013.

[7] M. Franke et al. *Signal to act: Game theory in pragmatics*. Institute for Logic, Language and Computation, 2009.

[8] M. Franke and G. Jäger. Pragmatic back-and-forth reasoning. In *Pragmatics, Semantics and the Case of Scalar Implicatures*, pages 170–200. Springer, 2014.

[9] H. P. Grice. Logic and conversation. *1975*, pages 41–58, 1975.

[10] G. Jäger and R. Van Rooij. Language structure: psychological and social constraints. *Synthese*, 159(1):99–130, 2007.

[11] J. M. Smith. *Evolution and the Theory of Games*. Cambridge university press, 1982.

[12] M. Spence. Job market signaling. *The quarterly journal of Economics*, 87(3):355–374, 1973.

[13] M. Weigel. Political correctness: how the right invented a phantom enemy.

[14] R. Yehuda. Post-traumatic stress disorder. *New England journal of medicine*, 346(2):108–114, 2002.