# Investigating Trust Using Natural Language Processing

Austin Bradley

April 2024

## Abstract

Communication matters in economic decision making, even in games with simple, unique equilibria. However, the exact mechanism through which this effect occurs remains uncertain. Previous studies have only been able to handle ad hoc hypotheses due to their relatively small collections of messages. In this chapter, we employ Natural Language Processing (NLP) on data from a large-scale trust game run on Amazon's MTurk system with one-way communication from trustees to trustors. NLP methods, combined with our data set of over 1000 messages, allow us to approach the data with minimal substantive assumptions to investigate the mechanisms underpinning trust. The structure of our data-set clearly links trustors' choices to their received messages, allowing us to identify new features that increase the perceived trustworthiness of a message. Additionally, knowing the ultimate decisions of the agents, we identify features of messages which predict that they are likely not trustworthy.

# 1 Introduction

Communication matters in economic decision making. This remains true even in games with unique, easily discerned equilibria. In these settings, standard theory predicts that, since all parties are fully informed about the payoffs for each outcome, any costless signal sent between players will contain no useful information. In the absence of any sort of commitment mechanism, any rational, payoff-maximizing individual understands that a message they receive does not alter the underlying incentives and thus would not change their behavior. However, experimental evidence consistently shows that this is not the case (see Crawford (1998) for a review).

The increase in cooperation is particularly pronounced when decision makers communicate using free-form messages. In trust games, free-form promises increase trust much more than prewritten messages with the same literal meaning (Charness and Dufwenberg, 2010). Furthermore, a modest communication effect persists even when communication is restricted to only small talk, irrelevant to the experimental task (Fiedler and Haruvy, 2009). Although the impact of free-form communication is well established, the reason for the increase in cooperation remains unknown. The flexibility afforded by free-form communication also increases complexity, making it difficult to study. Free-form messages provide a rich source of information on decision makers' choice processes. However, this complexity also means that there is immediately apparent way to sort through the immense message space.

In this chapter, we investigate whether Natural Language Processing techniques provide an effective means of uncovering new factors which drive the outsized impact of free-form communication. We study one-way messaging data in the context of a principal-agent trust game adapted from Charness and Dufwenberg (2006) where Senders (agents) compose pre-play messages to Receivers (principles). Receivers may earn more by trusting the Sender but run the risk of the Sender keeping all potential gains for themselves, leaving the Receiver with nothing. This anonymous text-based messaging is similar to what people encounter everyday in an ever-increasing share of interactions occurring online. With the large dataset of experimental communication data and the associated choices, we employ a variety of Natural Language Processing techniques, allowing us to identify new features associated with changes in both trust and trustworthiness.

Previous literature establishes the impact of communication in a wide variety of settings including bargaining (Ellingsen and Johannesson, 2004), principal–agent trust games (Charness and Dufwenberg, 2006; Ben-Ner et al., 2007; Chen and Houser, 2017), and Oligopoly games (Fonseca and Normann, 2012; Cooper and Kühn, 2016). As an extreme example, communication remains influential even in dictator games, free of potential strategic considerations, (Andreoni and Rao, 2011). The consistent theme across these settings is that communication promotes cooperation and efficiency.

The fact that cheap talk messages impact choices at all presents a significant obstacle in developing complete theories of behavior; determining what makes a message effective is crucial to furthering this understanding. Previous studies have proposed a wide variety of mechanisms through which cheap talk messages can influence the behavior of others. Potential factors include: an aversion to lying (Lundquist et al., 2009), a reduction in social distance between subjects (Buchan et al., 2006; Charness and Gneezy, 2008), and the content of the messages independent of personal relationships (Mohlin and Johannesson, 2008). In each of these cases, the findings arise from comparing different permissible levels of communication, not an analysis of message contents. The full magnitude of the impact of communication cannot be explained solely as the result of game-relevant signals (Charness and Dufwenberg, 2010).

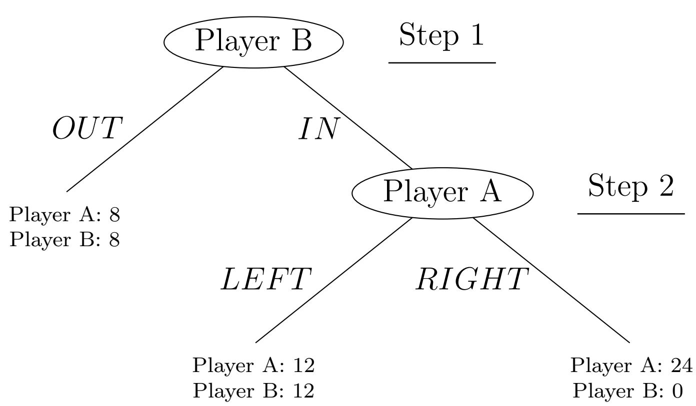
Machine learning offers a new set of tools and techniques we may use to parse unstructured communication data. Previous laboratory experiments employ machine learning to discover novel findings in complex environments to inform theory (Camerer et al., 2019; Fudenberg and Liang, 2019). If properly processed, similar techniques can be used on text data. Natural Language Processing (NLP) has already provided insights into economic decision making, analyzing the impact of investor sentiment on the movements of stock prices (Tetlock, 2007). Much like our setting, this sentiment has no formal informational value yet still contains predictive power for future price movements. In the economics laboratory, NLP is still in its relative infancy. However, the existing literature has shown its potential both to streamline data classification (Penczynski, 2019; Tebbe and Wegener, 2022) and to produce new findings (Hanaki and Ozkes, 2022; Bursztyn et al., 2023; Andres et al., 2023).

The remainder of the chapter proceeds as follows. The remainder of this section presents relevant theory followed by the experimental procedures we use to generate a large data-set of free-form text messages. We then provide a detailed description of this data set along with the methods applied to make this unstructured data useful in Section 2. Section 3 presents our analysis of the impact of features we identify in our message corpus. Finally, we summarize our findings and conclude in Section 4.

## 1.1 Background

In this study, we examine a simple trust game with the following form:

Figure 1: Basic Trust Game




The standard subgame perfect equilibrium from backward induction is trivial. Player A's maximize their monetary payoff by choosing Right if given the opportunity, leaving Player B's with nothing. Knowing this, Player B's maximize their own expected payoffs by distrusting Player A's and playing Out in the first stage. Thus, for the game to allow an alternative outcome, players must have considerations outside of their own monetary gain.

Traditionally, when communication is added to a game, authors omit the messaging step from their formal representation of the game. We pause here to note, however, that when the game allows free-form text communication, the game remains technically finite. Given a maximum message length of $L$ characters with 256 potential ASCII choices for each, the action space for the Sender's pre-play message contains $256^L$ possible choices. This is intractably large and this observation alone does little to further our understanding of communication and the vast majority of this space is filled with random, unintelligible messages. However, it allows us to more carefully formulate the task at hand when we study the contents of subjects' messages. The current state of the art is to manually partition this message space into categories corresponding to available actions in the game (promises to choose Left, requests to choose In, etc.). Our goal is to expand upon this by utilizing more granular partitionings using NLP to capture a greater portion of the message space's complexity.

To represent the existing approach for studying free-form communication, we define a baseline partition of the message space we call literal meaning. The literal meaning of a message is a mapping from the message space to possible actions in the game. We chose this as our benchmark because it can be generalized to other environments besides our trust game to reasonably separate the basic, game-relevant signals from the "purely free-form" elements of a message. Given the actions in our trust game, there are nine possible messages: In, Out, or none for the Receiver's actions and Left, Right, or none for the Sender's actions. We omit Out and Right because they are rarely used, leaving us with four possible messages. Senders may signal their intent to choose Left, suggest the Receiver choose In, both, or neither. Using Natural Language Processing, we examine whether this literal meaning, based solely of available actions, provides a sufficiently full account of the information contained in free-form messages.

Literal meaning forms the benchmark against which we examine the value of NLP. In the analyses that follow, we employ a variety of NLP tools and techniques to generate different reductions of the message space. Across the different formulations of the message space, we also include the literal meanings of a message as control variables. In doing so, when we identify message features which significantly correlate with subjects' choices, we find evidence of predictive information not yet considered.

In the trust game we examine, there are two choices which may correlate with message choice: the Receiver's choice between In and Out and the Sender's choice between Left and Right. We refer to each of these decisions as Trust and Share, respectively.

Within this framework, we define our research hypotheses.

- Hypothesis 1: There are significant predictors of trust after controlling for literal meaning.

- Hypothesis 2: There are significant predictors of trustworthiness after controlling for literal meaning.

In addition, we examine whether the features we identify accurately and effectively signal the Sender's intentions to the Receiver.

- Hypothesis 3: Any significant predictors of trustworthiness are also significant predictors of trust.

# 2 Methods

We recruited our subjects on Amazon Mechanical Turk (MTurk), using task software programmed in the oTree framework for Python (Chen et al., 2016). MTurk allows requesters to post Human Intelligence Tasks (HITs) that can be performed for a small fee by a large pool of workers. Numerous previous studies conducted on MTurk have yielded high-quality results consistent with those from traditional laboratory settings (Snowberg and Yariv, 2021). To further ensure the quality of our subject pool, we restrict eligibility to include only accounts that have at least a 99% approval rate and more than 1,000 submitted HITs, which corresponds to approximately the 75th percentile of MTurk workers (Robinson et al., 2019). Participants earn a $1.00 participation fee as well as a bonus payment based on their choices. The typical bonus payment is $1.10. The median completion time is approximately 8 minutes.

After accepting the HIT, participants provide informed consent and receive the instructions for the study followed by a brief comprehension quiz. Participants are required to correct any incorrect answers before proceeding to the study. Once all questions are answered correctly, participants proceed and play 10 rounds of a trust game first introduced by Berg et al. (1995) with instructions adapted closely from Charness and Dufwenberg (2006).

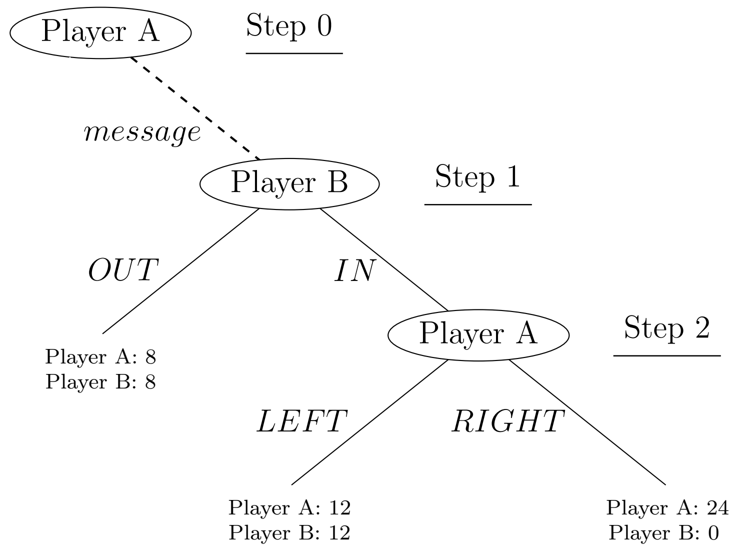
The game consists of two players —– Player A, the Sender and Player B, the Receiver —– and two steps. In the first step, Player B chooses between two actions: In or Out. If they select Out, both players receive 8 tokens of experimental currency and the round ends; otherwise, the game continues to step 2. In step 2, Player A chooses between Left or Right. Choosing Left results in both players receiving 12 tokens while choosing Right results in a payoff of 24 tokens for Player A and 0 for Player B. Figure 2 provides an overview of this game.

Prior to step 1, Player A may send a message to influence Player B's decision. We purposefully provide no guidance to subjects on how they should use their message to avoid potential experimenter demand effects. This message is delivered to Player B prior to their decision between In and Out.

Each subject composes only one message as Player A which is sent to all 10 Player Bs they are paired with. Multiple responses allow us to capture a more detailed measure of trust. We assign pairings sequentially, with each participant acting as Player B for the 10 players immediately preceding them and as Player A for the 10 players following them.

Our method of implementing this trust game has two distinct benefits over previous studies. First, conducting our experiment on MTurk allows us to recruit a large pool of subjects rapidly, at low cost. As a result of this, our current dataset is quite large relative to previous experimental communication studies. Second, collecting multiple reactions to each message gives us a more precise measurement of its effectiveness.

Figure 2: MTurk Game Overview




# 3 Data

## 3.1 Message Overview

Our data consist of 1004 unique messages and their corresponding actions collected during the summer of 2020. For each message, we measure both trust and trustworthiness. Trust is measured as the frequency the recipients of a particular message choose to play In while trustworthiness is the binary Left/Right choice of the sender; choosing Left and splitting the pool of tokens is trustworthy. To summarize the general content and variation within our messages, Table 1 shows a selection of the most effective and least effective messages — always and never trusted, respectively. We briefly note that there is an immediately apparent qualitative difference between these groups. Effective messages are longer and display a higher level of effort and understanding on average than ineffective messages. For the sake of space, only a partial selection of effective messages is shown.

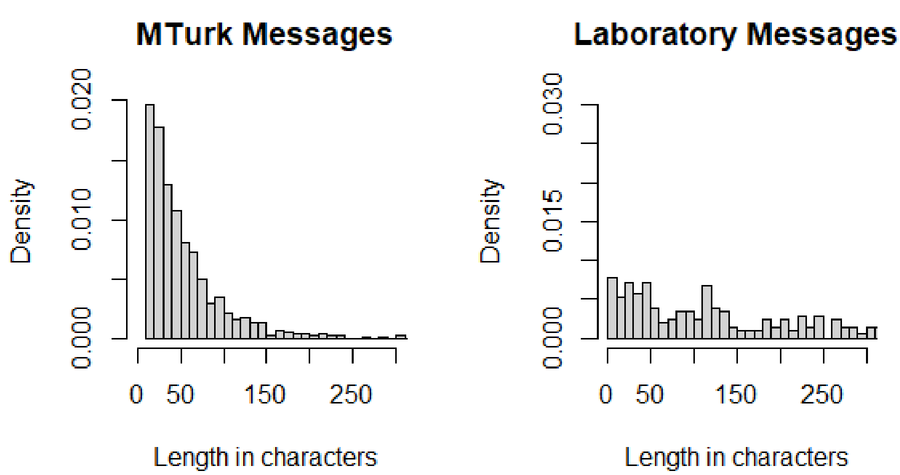
Table 1: Example Messages: Most and Least Effective

| Type | Messages |
|---|---|
| Most Trusted | I'm going LEFT, so please select IN. This will maximize the earnings for both of us equally.<br>choose IN so that we will both get 12 tokens<br>Choose "In" and I'll make sure we both get 12 tokens!<br>Hi there. I would understand why you would choose OUT, but I promise to make the choice ...<br>I promise you that I will always choose "left" so that we both get 12 points each.<br>Helloo!!! Let's make this work for both of us in the best possible way! Have a great day!<br>Hey, I'm not a greedy person and in these situations, I always choose the most beneficial option for ...<br>You can be sure that i will choose left, as it is the best option for both us. Please go ahead and select 'In'<br>I'm good for the 12. I believe that there are real people here, and I'm completely down with splitting the 24 and ...<br>I'm picking left so if you want a few extra tokens, go for in. Totally understand if you don't but that's my choice.<br>I'm always going to choose "left" because I think it's dickish to screw another player over.<br>I will maximize both our outcomes ie LEFT if you stay in. |
| Least Trusted | I'm going to choose Right<br>i get 24 you get 0<br>U GO ON STRITE<br>Hello Let's play the game<br>I want choose Player A<br>hi, how are you doing?<br>Hey. let's work together, let's both choose out.<br>I choose out<br>choose out |

A sample of messages trusted by all/none of their recipients.

In our data, the median message length is 40 characters with a mean of 54.5, so our messages are considerably shorter on average than those seen when similar procedures are conducted in a laboratory setting — median length of 110 and mean of 137.6, (Charness and Dufwenberg, 2006). Figure 3 presents the distribution of length in our sample compared to this laboratory setting

Figure 3: Message Length Distribution




A comparison of the distribution of length in our messages collected through MTurk and messages from a similar, in-person experiment reported in Charness & Dufwenberg (2006).

To ensure the validity of our data, we check that our subjects' behavior conforms with those in previous laboratory experiments. We manually label promises with the help of four research assistants. We label a message a a promise if three out of the four labeled a message as such. For a message to be classified as a promise we require not only that the message contain a statement of a future action, but also that is be coherent within the context of the game. Out of the hundreds of unique messages, a limited number of participants promised to do actions which are either impossible, i.e. "I will choose to be player B," or promised to play Right and take the entire pool for themselves. Although these are clearly promises in a grammatical sense, we do not include them in our classification. Out of 1004 total messages, 389 are promises. We also manually identify requests for In.

## 3.2 Tokenization and Lemmatization

Table 2: Most Common Tokens

| **Tokens** | **Frequency** | **Tokens** | **Frequency** | **Tokens** | **Frequency** |
|---|---|---|---|---|---|
| i | 751 | so | 131 | 'll | 72 |
| choose | 586 | u | 131 | best | 72 |
| in | 498 | player | 120 | please | 69 |
| left | 393 | 's | 116 | more | 67 |
| you | 365 | that | 114 | maximize | 66 |
| to | 359 | can | 106 | point | 65 |
| will | 341 | make | 102 | our | 59 |
| be | 338 | out | 97 | each | 56 |
| the | 291 | hi | 97 | most | 55 |
| both | 264 | hello | 96 | game | 54 |
| we | 253 | 12 | 95 | all | 54 |
| and | 244 | play | 94 | b | 52 |
| a | 235 | do | 88 | always | 51 |
| go | 176 | token | 86 | money | 51 |
| if | 175 | 'm | 82 | pick | 51 |
| get | 174 | fair | 81 | with | 50 |
| of | 153 | this | 81 | have | 49 |
| for | 152 | it | 77 | right | 48 |
| let | 137 | good | 74 | earn | 46 |

To work with our text data, subjects' messages must first be processed into collections of tokens, each representing an individual or short series of words. With natural language, this preprocessing step presents significant challenges as we simultaneously want to make the problem tractable by reducing the number of unique words and retain as much of the richness of the original messages as possible. This preprocessing is conducted using the Natural Language Toolkit package for Python, (Loper and Bird, 2002). To do this, we utilize the WordNet lemmatizer, (Fellbaum, 2010) to reduce the various conjugations and declensions into their base forms. This lemmatizer performs best when the part of speech of a words are included, so we also employ Treebank part of speech tagger (Marcus et al., 1993). The end result of these two algorithms is that each message is split into an ordered set of tokens, each token being a word and part of speech pair.

Usually, messages are further reduced by removing stop-words which provide little additional information about the contents of a document. This is a useful step when trying to categorize documents based on their subject matter, where the typical stop-words would be widely shared between the documents in a corpus. However, we omit this step as standard stop-word dictionaries would remove a great deal of information relevant to our specific environment. The messages in our corpus are largely all about the same topic – the game at hand. As such, our variation largely stems from different rhetorical approaches around the same topic. Words like "I", "we", and "you" would all be filtered out by standard stop-word dictionaries, but are clearly important in our environment. Table 2 presents the number of occurrences of the most frequent tokens in our data.

In addition to the collection of tokens representing each message, we construct a measure of the sentiment of each message. In stock market data, investor sentiment has predictive value of future returns (Tetlock, 2007). Using the OpinionFinder Subjectivity Lexicon (Wilson et al., 2005), we count the number of positive, negative, and total words in each message. This positive and negative word classification requires some minor adjustments to better fit our environment. For example, "right" is identified as a positive word in the sentiment dictionary. As this is an action label in our game, this is removed from the subjectivity dictionary.

# 4 Results

## 4.1 Baseline Approach: Literal Meaning

Our goal in this chapter is to examine whether using Natural Language Processing to analyse the content of participant messages yields any insights that would otherwise not have been obvious. To do so, we create a baseline message model which maps the free-form text messages to their literal meanings, based solely on the game's action space. Before proceeding with the NLP techniques, we first establish the validity of our baseline model by regressing subjects' choices onto dummy variables representing the possible literal meanings.

Table 3 below summarizes the results of both the Trust and Share (trustworthiness) models.

Table 3: Regressions on Baseline Messages

| | Trust | Share |
|---|---|---|
| LEFT | 0.219*** | 0.225*** |
| | (0.015) | (0.039) |
| IN | 0.149*** | 0.041 |
| | (0.018) | (0.046) |
| BOTH | 0.253*** | 0.186*** |
| | (0.016) | (0.041) |
| CONSTANT | 0.462*** | 0.532*** |
| | (0.009) | (0.022) |
| N | 976 | 969 |

Regression coefficient estimates for the literal meaning baseline. $* : p < .05, ** : p < .01, *** : p < .001$

The results agree with what has been seen in previous studies involving communication. In the trust model, where the dependent variable is the frequency with which recipients of a given message choose the In action, statements of intention to choose Left by the Sender, requests that the recipient play In, and combinations of the two have significant positive effects relative to messages which contain none of these features. In the Share model, using the likelihood that the Sender chooses Left after sending a given message as the dependant variable, we find that players who send messages promising to choose Left, sharing the tokens, are significantly more likely to do so. Requesting that the recipient of a message choose In without promising Left has no significant relationship with the Sender's behavior.

## 4.2 Data Validity: Order Effects

Additionally, we check to see if there are significant order effects based on prior messages received. To do this, we divide our subject pool into two groups based on their behavior as Receivers. One group includes all subjects who chose In in the first round, with the other including all who chose Out. We then compare the frequency with which each group chose In in round 2 and in round 10. The assumption is that if there are substantial order effects, the In/Out choice in the second round will be significantly correlated with the first round choice, however, the choice in round 10 will not be.

We find no evidence of substantial order effects (presented in Figure 4). For both groups, the frequency of subjects choosing In is not significantly different between rounds two and ten. However, we do observe evidence of subject-level heterogeneity. Those who chose In (or Out) in the first round tend to do so throughout the duration of the game.

## 4.3 Regressions on Tokens

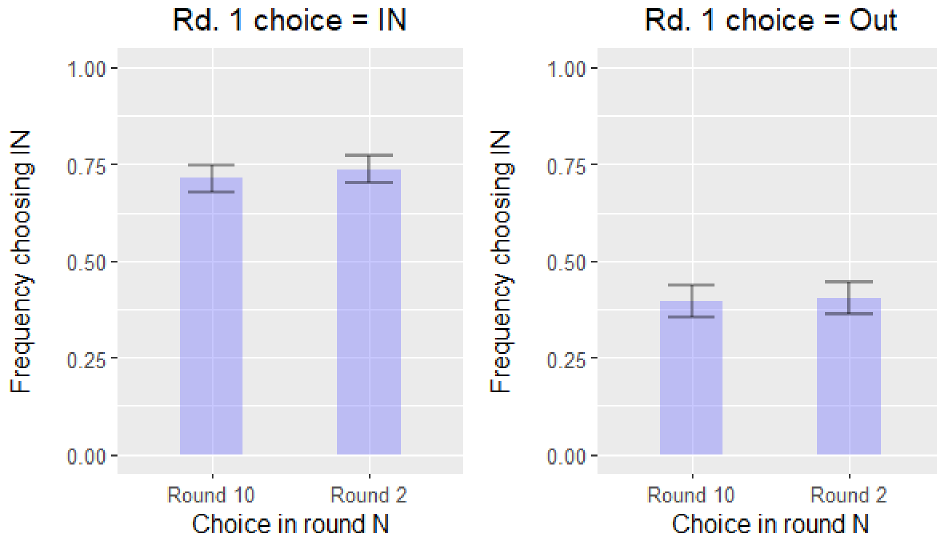
The literal meaning model of communication provides us with a reasonable baseline against which we can compare results using Natural Language Processing. In our first NLP model, we

Table 4: Bag-of-words Model

| **Tokens** | **Trust** | **T-stat** | **Trustworthiness** | **T-stat** |
|---|---|---|---|---|
| **choose** | -0.023 | (-1.512) | -0.099* | (-2.259) |
| **in** | 0.068*** | (3.545) | 0.036 | (0.662) |
| **left** | 0.060** | (3.278) | 0.083 | (1.608) |
| **to** | 0.029* | (2.069) | 0.088* | (2.187) |
| **we** | 0.047* | (2.202) | 0.138* | (2.306) |
| **and** | 0.031* | (1.998) | 0.046 | (1.051) |
| **go** | -0.013 | (-0.732) | -0.114* | (-2.326) |
| **for** | 0.008 | (0.398) | 0.106* | (1.974) |
| **u** | 0.048* | (2.007) | 0.003 | (0.043) |
| **'s** | 0.001 | (0.038) | 0.159* | (2.363) |
| **that** | 0.042* | (2.043) | 0.043 | (0.742) |
| **can** | 0.046* | (2.084) | -0.143* | (-2.275) |
| **make** | 0.066** | (3.015) | 0.010 | (0.165) |
| **out** | -0.101*** | (-4.933) | 0.087 | (1.509) |
| **fair** | 0.052* | (2.316) | 0.113 | (1.793) |
| **this** | 0.063** | (2.730) | 0.099 | (1.510) |
| **maximize** | 0.067* | (2.247) | -0.148 | (-1.692) |
| **always** | 0.053* | (2.068) | 0.136 | (1.869) |
| **right** | -0.130*** | (-4.614) | -0.224** | (-2.767) |

Each coefficient is associated with a dummy variable indicating a message includes the relevant word. Only tokens with significant coefficient estimates are included. Control variables for messages' literal meaning are omitted from the table. $* : p < .05, ** : p < .01, *** : p < .001$

Figure 4: Order Effects




Participant choices in the second and tenth rounds, conditional on their decision in the first round. 95% confidence intervals are plotted in black.

augment our literal message model to include the top 50 most commonly used tokens in our message corpus. By using this model, with control variables included for the literal meaning categories in the baseline model, we examine which, if any, of the most commonly used words are associated with higher (or lower) than expected levels of trust and trustworthiness — the likelihood that recipients of a message choose In and senders choose Left, respectively.

Table 4 presents the results of the Trust and Share frequency regressed on the token dummy variables. As this model assumes that each message is an unordered collection of independent tokens, we refer to this as the bag-of-words model. For brevity, we omit tokens that are not significant in either of the two models from the table. The full results are included in the appendix.

Examining the Trust model first, the tokens with the most significant effects are those which correspond to action labels on the game tree, even after including the literal meaning controls. Each of these tokens effect on the likelihood the recipient of a message chooses In is in the expected direction with "In" and "Left" having positive effects significant at the 1% level and "Out" and "Right" having negative effects significant at the 1% level. In addition, there are several other tokens with significant, positive effects in the Trust model. "Make" and "This" are both significant at the 1% level and positive. The coefficient estimates for the tokens "To", "We", "And", "u", "That", "Can", "Fair", "Maximize", and "Always" are all positive and significant at the 5% level.

With less statistical power from fewer observations, there are fewer significant tokens in the Share model. In the Share model, the only Action label with a significant coefficient estimate is "Right" which has a significant, negative effect at the 1% level. In addition, “choose” “go” and “can” are negatively associated with the likelihood senders choose Left and significant at the 5% level. The coefficient estimates for “To,” “We,” “For,” and “ ’s,” are positive and significant at the 5% level.

This approach shows that there is indeed additional information relevant to predicting players’ choices remaining even after considering the bare meaning of the messages. For many of these tokens, the interpretation of their coefficient estimates is clear. The significance of the Action labels is, perhaps, somewhat unsurprising. However, their significance even after controlling for promises to choose Left and requests to choose In suggests that these sort of statements are most effective when they directly and unambiguously reference the intended or requested action.

For the remaining significant tokens, properly interpreting the results requires us to return to our message corpus to examine the context in which they appear.

Having only one decision associated with each message (instead of ten in the Trust model) we find that there are fewer tokens significantly associated with changes in the likelihood their authors chose Left. Of the action labels present in the game, Right, is the only one with a significant coefficient estimate in the Share model, predicting a much lower likelihood of the sender choosing Left. Although the negative coefficient is coherent with the messages which contain the Right token, it is rather unexpected. Stating an intention to choose Right does not make sense in the usual understanding of the trust game wherein the message senders’ attempt to persuade their partners that they are likely to choose Left. Given the incongruity between stating intent to choose Right and the usual strategy of persuasion, it is tempting to write these messages off as mistakes, written by participants who did not understand the game’s instructions. However, this would not explain why these Senders follow through on their promises to choose Right. This, instead, suggests that there is a subset of participants who are strongly averse to lying but not necessarily averse to unfair outcomes.

The four tokens with positive coefficient estimates in the Share model fall into two categories. The first of these contains the plural first-person pronouns “We” and “ ’s” which is used as for contractions containing “us” almost exclusively. The remaining two positive share coefficient tokens are “To” and “For” which are primarily used in messages that detail a reason why the recipient should trust the sender.

Apart from ”Right,” the remaining significant tokens in the Share choice model do not have readily apparent interpretations in the trust game. Interpreting these tokens, “Go,” “Can,” and “Choose,” requires that we examine them in the context in which they appear. Messages using the “can” token are typically used by Senders to provide a reason why the Receiver should choose In by describing the outcome. A typical message of this type ends with “... so we can get the most tokens.” “Go” messages are used to convey relatively weak promises, using more casual language than other promises. “Choose” messages, although often used in promises, are more often used as commands for the Receiver to choose In.

Comparing the results in each model, we do not find significant discrepancies that indicate effective deception. Tokens indicative of effective deception would be those that have significant, negative coefficients in the Share model but positive, significant coefficients in the Trust model. Only one token, “Can,” fits this definition. Given the large number of regressors in the model, this is indistinguishable from random chance. Therefore, we draw no conclusions from this regression model about potential signs of effective deception.

## 4.4 Cluster Analysis

What the above analysis shows, in short, is that there are features in our corpus of messages that have strong correlations with both trust and trustworthiness choices even after controlling for literal meaning. With this approach — running OLS regressions on dummy variables for common tokens — we establish an initial result that literal meaning does not capture all the useful information contained in this free-form text data. These significant tokens suggest new classes of messages worth further investigation. However, this method of treating each message as linear combinations of independent words leaves out much of the complexity present in natural language.

To address this shortcoming, we employ cluster analysis to identify groups of similar messages which we may then aggregate and examine further. Cluster analysis produces groups of messages such that the within-group similarity is greater than the inter-group similarity. The specifics of this task vary greatly based on the chosen measure of similarity and objective function comparing intra- and inter-group similarity. This permits us a great deal of freedom to choose a similarity measure which captures more of the complexity of natural language than in our previous approach.

Previous authors have employed cluster analysis on laboratory data including subject choices (Ellingsen et al., 2018; Grimm et al., 2021; Woods et al., 2022) and eye tracking data (Polonio and Coricelli, 2019; Devetag et al., 2016).

Our approach involves two steps. First, we employ Affinity Propagation, (Frey and Dueck, 2007) to perform our cluster analysis and group our messages. Affinity Propagation allows for both the number of clusters and the number of elements assigned to each to vary, which we require in the absence of ex ante knowledge of the shape of the message space. Flexible cluster size also allows non sequitur messages to separate themselves into smaller clusters instead of forcing themselves into unsuitable clusters to achieve an arbitrary number of groups. Each cluster also forms around an exemplar message, giving us a natural starting point when we manually interpret and label each cluster. The Affinity Propagation algorithm is described in Appendix **??**.

Following this aggregation, we then employ standard OLS, fitting a linear probability model to test which areas of our message space have significantly higher rates of trust (greater likelihood Receivers choose In) and trustworthiness (greater likelihood Senders choose Left). As with our bag-of-words token regressions, we add a manually coded control variable for literal meanings.

Conducting cluster analysis with affinity propagation requires us to define a similarity matrix $\mathbf{S}$ for the preference of each element for potential exemplars. Given this degree of flexibility, we conduct cluster analysis with two different similarity measures, one using solely the information in our corpus of text (corpus-only approach) and the other informed by outside natural language knowledge (knowledge-based approach). By using multiple similarity measure, we test the robustness of our conclusions against changes in our clustering procedures. If cluster analysis is a useful tool for us to process free-form laboratory text data, we expect the results to be relatively consistent across these choices.

In the Corpus-only approach, we define our similarity matrix $\mathbf{S}_J$ by taking each message, again, as an unordered collection of tokens and calculating the Jaccard index of overlap between sets. For messages $m_i$ and $m_k$, both being sets of tokens:

$$s(i,k)_{\ i\neq k} = j(m_i, m_k) = \frac{|m_i \cap m_k|}{|m_i \cup m_k|}\ ;\ s(k,k) = \text{Median}(\mathbf{J})$$

Applying this measure to each pair in our corpus yields a complete graph in matrix form; each element $j(i,j)$ representing the similarity of message $i$ to message $j$ and vice versa; a higher value implying greater similarity.

## 4.5 Results: Affinity Propagation Using Corpus-only Similarity Measure

Affinity Propagation identifies 112 clusters of similar messages in our corpus. The median cluster size is 8 messages with a mean of 9.1. Figure 5 presents the distribution of the number of messages assigned to each cluster.

We conduct cluster analysis with the goal of identifying key features in our corpus, without a priori knowledge of the number of expected clusters or their contents. Instead, we must investigate each cluster and assess whether or not the clusters share a common meaning to a sufficient degree. We must also manually assign a label to each cluster, describing the common theme shared among the messages assigned to it, using the exemplar element for each cluster as our starting point. In Table 5 below, we present two example clusters and their exemplars, demonstrating the similarity between the messages grouped by affinity propagation.

Figure 5: Distribution of Cluster Sizes: Jaccard Distance

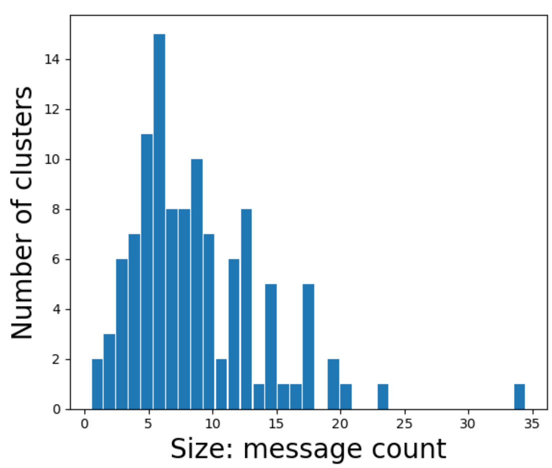


Table 5: Example Clusters

| Cluster | Message (**exemplar**) |
|---|---|
| 55 | **I believe in fair play** |
| 55 | I believe in equality |
| 55 | I am playing in |
| 55 | I believe in playing fair and having an equal outcome. |
| 55 | I believe in fair, even-steven financial deals. I like a win-win situation. ... |
| 55 | I believe in being fair with people. |
| 58 | **Hey, I want to make sure we both get the highest amount of bonus!...** |
| 58 | Hopefully we both want to get the highest points possible |
| 58 | I will make sure we get the same amount of tokens |
| 58 | Hello. I hope we both make a good amount of bonus money today! Good luck! |
| 58 | Hi there! I always split these 50-50 because I would want someone to do the ... |
| 58 | Hello. Lets make this to where we both can get the largest bonuses out of th... |
| 58 | To make sure you guys get a higher bonus, please choose IN in each step so w... |

In both example clusters, the elements are sufficiently similar to lead to a clear substantive interpretation. Cluster 55 messages state that the sender believes in fairness, which is captured well by the exemplar message, "I believe in fair play." Messages in cluster 58 state the author's hopes for the game outcome, speaking in terms of the amount of tokens. Again, the exemplar, "Hey, I want to make sure we both get the highest amount of bonus..." represents the typical message in the cluster well.

### 4.5.1 Regressing Trust Choices on Clusters

Our primary research question is whether free-form message text contains additional useful information after accounting for a message's literal meaning. To do so, we fit an OLS regression model regressing players' choices (Receivers' frequency choosing In and Senders' likelihood of choosing Left following a given message) on dummy variables for each of the four possible literal meanings. Following this, we regress the residuals from the literal meaning regressions on dummy variables for cluster assignments. Using this two step approach, the coefficients and t-statistics reported are for tests for heteroskedacity in the literal meaning model. Messages in clusters with significant, positive coefficient estimates are trusted (trustworthy) more often than predicted by literal meaning alone.

We first examine the most effective messages in gaining the recipient's trust. For each cluster, we include both the exemplar message and a brief description of the contents of the common themes.

Again, we are cautious of over-inference, and avoid delving too deeply into any individual cluster. Instead, we focus on the common features across these groups of highly effective messages. The most effective messages largely fall into two broad categories. First, two of the top ten clusters (numbers 78 and 45) follow a common three part structure — explicitly

Table 6: Most Trusted Corpus-only Clusters

| Cluster | **Description**<br>Exemplar Message | Size | Trust<br>(t-stat) | Share<br>(t-stat) |
|---|---|---|---|---|
| 84 | **Promises 'what's best for everyone'** | 13 | 0.19 | 0.01 |
| | Hi, I will make the choice that is beneficial for all of us :) | | (4.2) | (0.06) |
| 78 | **Requests In w 'please', promises Left, specifies '12 token' outcome** | 13 | 0.16 | -0.07 |
| | Please choose "IN" so that we can both get 12 points. I will choose "LEFT." | | (3.48) | (-0.51) |
| 96 | **Requests In so both can 'maximize'** | 9 | 0.19 | -0.18 |
| | If you choose In we can both maximize our tokens. | | (3.19) | (-1.09) |
| 58 | **Suggests that both 'get the highest bonus'** | 7 | 0.19 | -0.03 |
| | Hey, I want to make sure we both get the highest amount of bonus! ... | | (2.96) | (-0.19) |
| 40 | **"Let's try and be fair"** | 6 | 0.20 | 0.06 |
| | lets try and be fair | | (2.95) | (0.29) |
| 103 | **Promises with justification. Also states explicit outcome** | 15 | 0.11 | 0.16 |
| | I want to be as fair as possible for both of us. I am not greedy ... | | (2.6) | (1.31) |
| 81 | **Promises 'most for both of us'** | 12 | 0.12 | 0.32* |
| | Trying to make the most for both of us | | (2.49) | (2.38) |
| 45 | **Request In, promise Left, & specifies outcome for 'both of us'** | 18 | 0.10 | 0.11 |
| | Hello, I will chose left if you decide to go in. I want this to be a win/win ... | | (2.45) | (0.95) |
| 75 | **A greeting plus a promise of the 'most' or 'best' outcome** | 5 | 0.18 | 0.06 |
| | Hey! Let's try and make as much as we can for everyone! | | (2.43) | (0.29) |
| 19 | **Promises left with an explicitly stated 'reason to trust'** | 5 | 0.17 | 0.09 |
| | I know you have to trust me, but I promise to pick left for both our benefits. | | (2.33) | (0.45) |

The top ten most significant, positive coefficients when regressing Trust on cluster dummies, controlling for promise labels and length, along with the corresponding coefficient estimates in our Share model. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. $*: p < .05, **: p < .01, ***: p < .001$

requesting the receiver plays In, promising to choose Left, and stating the positive result should both these choices be made. These messages conform with our definition of a promise, but go a step further in describing the outcome and requesting specific action from the Receiver. An additional two clusters (19, 58, and 96) do not follow this three-part formula but are also detailed promises.

Second, four of the ten clusters are 'soft promises' (clusters 103, 40, 81, 84). These messages fail to meet our definition of a promise, requiring the author to makes a clear statement of a future intended action. Instead, they make softer claims: "Let's try to be fair," for example. Although these are not strictly promises in the grammatical sense, the implication is clear in the context of the game. Recipients of these messages interpret them as statements of intent to choose Left even without an explicit action indicated.

Table 7: Least Trusted Corpus-only Clusters

| Cluster | Exemplar Message | Size | Trust (t-stat) | Share (t-stat) |
|---|---|---|---|---|
| 97 | choose out | 6 | -0.29 (-4.29) | 0.05 (0.24) |
| 32 | choose right! | 10 | -0.21 (-4.05) | -0.34* (-2.27) |
| 41 | I chose out. | 12 | -0.18 (-3.50) | -0.08 (-0.59) |
| 34 | hello..let's play | 10 | -0.15 (-2.62) | 0.13 (0.80) |
| 1 | Please Choose In | 15 | -0.10 (-2.20) | -0.02 (-0.13) |
| 86 | i will be fair | 10 | -0.11 (-2.03) | -0.01 (-0.07) |
| 29 | I think you should choose In | 8 | -0.12 (-2.02) | -0.34* (-2.03) |
| 21 | I will choose Left | 24 | -0.07 (-2.01) | 0.00 (0.04) |

Significant, negative coefficients when regressing Trust on cluster dummies. Controls for literal meaning and length are omitted from the table. Clusters are described with each cluster's exemplar message chosen by the Affinity Propagation algorithm. For each cluster, the exemplar message adequately represents the messages included. $* : p < .05, ** : p < .01, *** : p < .001$

None of the increased rates of trust in these clusters have a corresponding significant increase in the author's trustworthiness. However, we cannot reliably interpret this as evidence of effective deception. Our design focuses on the In/Out decision and thus has lower statistical power to estimate Left/Right coefficients; each message has ten readers but only one author.

Examining the least effective clusters (those with the most significant negative coefficients in the Trust regression) we can again identify two broad themes. As with the most effective clusters, we also present each cluster's coefficient in the Share regression, modeling the Senders' choices given the message they sent.

Table 8: Significant Corpus-only Clusters: Share

| Cluster | **Description**<br>Exemplar Message | Size | Share<br>(t-stat) | Trust<br>(t-stat) |
|---|---|---|---|---|
| 79 | **"I like this game"** | 8 | 0.46 | 0.06 |
| | I like this game. | | (2.77) | (1.07) |
| 35 | **Suggestion to 'work together'** | 11 | 0.37 | 0.10* |
| | Hiya. I hope we can work together. | | (2.61) | (1.98) |
| 81 | **Promises 'most money for both of us'** | 12 | 0.32 | 0.12* |
| | Trying to make the most for both of us | | (2.38) | (2.49) |
| 55 | **"I believe in fair play"** | 6 | 0.46 | 0.16* |
| | I believe in fair play | | (2.38) | (2.3) |
| 9 | **"Good player"** | 8 | -0.29 | -0.03 |
| | GOOD PLAYER | | (-1.73) | (-0.47) |
| 29 | **"I think you should choose In"** | 8 | -0.34 | -0.12* |
| | I think you should choose In | | (-2.03) | (-2.02) |
| 61 | **"Going left"** | 6 | -0.43 | -0.06 |
| | Going left | | (-2.26) | (-0.94) |
| 32 | **"choose right!"** | 10 | -0.34 | -0.21*** |
| | choose right! | | (-2.27) | (-4.05) |
| 52 | **Appeal to 'maximization'** | 9 | -0.45 | 0.02 |
| | let's maximize OUR bonuses | | (-2.84) | (0.43) |

The most significant, positive coefficients when regressing Share on cluster dummies. Controls for literal meaning and length are omitted from the table. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. $* : p < .05, ** : p < .01, *** : p < .001$

As suggested by the token regressions, Receivers distrust messages that seem to go against the Senders' best interests (clusters 32, 97, and 41). There is no immediately apparent explanation for sending a message stating your intent to play Right or asking the reader to choose Out, although these messages do not violate the rules of the game.

The second theme among the most distrusted messages is newly revealed by our Cluster Analysis; brief messages containing greetings but no other content (cluster 34) have highly significant, negative coefficients after controlling for literal meaning. The remaining message clusters (1, 86, 29, and 21) are all the briefest possible messages one can send that have a meaning corresponding to game actions. Compared to messages with the same literal meaning, their recipients are significantly less likely to choose In.

### 4.5.2 Regressing Share choices on clusters

Moving to the Senders' Left/Right choices, we fit an OLS model, using dummy variables for each message's cluster assignment and human-coded literal meaning as independent variables to estimate the likelihood that its author chooses to share the tokens (choose Left in step 2). As a consequence of our experiment's design, this model has significantly less statistical power than our Trust model; there are ten trust choices for each message but only one associated Share choice. We focus primarily on relating the Share coefficient estimates to their Trust model counterparts, asking whether Receivers' choices are rational responses to accurate predictions about Senders' behavior. Table 8 presents clusters with the most significant positive and most significant negative Share model coefficient estimates.

The clusters with the significant coefficient estimates in the Share model repeat themes seen in the Trust model. The majority of clusters with significant, positive coefficient estimates are soft promises (clusters 35, 55, and 81). The Sender implies they will be trustworthy, but does not explicitly state an intended action. Additionally, messages in which the Sender states that they "like this game" are significantly correlated with choosing Left. Every Sender of messages in cluster 79 chose Left in step 2.

Also similar to the Trust model, the messages in these least trustworthy clusters are almost all noticeably lower quality than the typical message in our corpus. This includes both nonsensical messages like those in cluster 32 and the briefest possible promises or requests in cluster 61 and 29. The cluster with the strongest negative coefficient estimate is the exception to this pattern of poor quality — cluster 52, which contains explicit appeals to payoff maximization. Unlike the low-quality message clusters, there is no corresponding significant negative coefficient in our Trust model. Recipients of these appeals to maximization continue to choose In, entrusting the tokens to their partner at the same rate as other messages with the same literal meaning.

Considering both the Trust and Share models formed from our corpus-only clusters, we find five broad classes of messages which warrant further study. Our primary goal is to study the determinants of trust and we find two themes associated with increased trust and two associated with decreased likelihood of trust. Both detailed promises — explicitly promising an action, giving their partner a call to action, and detailing the promised outcome — and soft promises (e.g. "I think fairness is important,") are associated with increased trust. These messages tend to have corresponding positive coefficient estimates in out Share model, however, we cannot make strong claims about this agreement due to the Share model's relative lack of statistical power. In contrast, Receivers trust poor-quality messages and brief greetings significantly less often. Our coefficient estimates in the Share model generally support their skepticism, with estimates agreeing in direction if not also in significance.

The final notable cluster, appeals to maximization, provides our only evidence of possible effective deception. Despite our Share model estimates indicating that these messages are the least likely to be trustworthy, we find no significant effect in our Trust model, despite it having much greater statistical power.

## 4.6 Affinity Propagation Using Knowledge-based Similarity

In the preceding section, we show the potential of cluster analysis as a tool for making sense of unstructured text data. From over 1000 messages, Affinity Propagation successfully aggregates individual messages into clusters which have substantial within-cluster similarly. These clusters allow us to conduct careful, limited inference and identify previously unknown message features which impact their recipients' choices.

However, these results are not without caveats. The clusters themselves are sensitive to our choice of similarity measure and other tuning parameters in the affinity propagation model. The Jaccard index as a similarity measure is both easily interpretable and based solely on the text contained in our message corpus. This simplicity makes it an attractive baseline similarity measure but, in choosing this measure, we ignore the large body of existing Natural Language Processing work. To illustrate this weakness, consider the brief greeting clusters, 34 and 4 in the above model (exemplar messages "Hello, let's play" and "Hi, how are you?" respectively.) Although the within-cluster similarity is satisfactory, the division between the two is based solely on the use of "Hi" versus "Hello" to convey nearly identical information. We make no claim about this specific partition being the definitive description of the message space, but this over-clustering raises concerns about whether our results are robust to our specification of similarity measure.

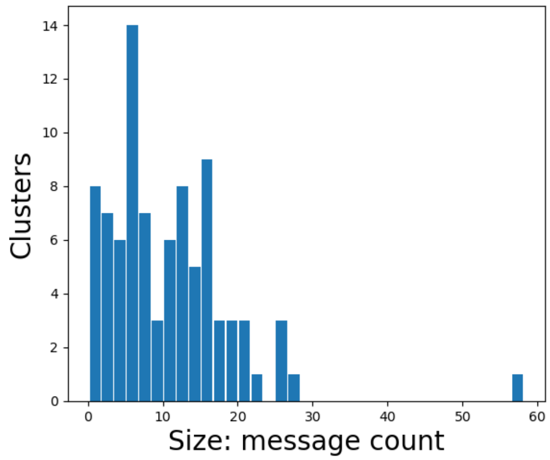
We test the robustness of our corpus-only Affinity Propagation results by conducting the same clustering and regression methods but using an alternative similarity. This Knowledge-based approach is trained on existing information outside of our corpus of messages. In this second approach, we measure the similarity between two messages using an approach first introduced by (Mihalcea et al., 2006) to compare short text documents. For two sets of words $T_1$ and $T_2$, the MCS similarity is:

$$S_{mcs} = \frac{\sum_{w \in T_1} \text{maxsim}(w, T_2) \cdot \text{idf}(w)}{2 \sum_{w \in T_1 \text{idf}(w)}} + \frac{\sum_{w \in T_2} \text{maxsim}(w, T_1) \cdot \text{idf}(w)}{2 \sum_{w \in T_2 \text{idf}(w)}}$$

maxsim(w,T) is the maximum semantic similarity between a word $w$ and any word in the sentence $T$, the precise details of which vary depending on the choice of semantic similarity measure. We measure word-to-word semantic similarity with the Google word2vec neural network, (Mikolov et al., 2013). After training, word2vec takes individual words as inputs and outputs a vector representation reduction of latent meaning. The network we implement here is pre-trained on a corpus of articles available through Google News, using the Gensim python package (Rehurek and Sojka, 2011). The resultant neural network encodes individual words into a 300-dimension vector; the similarity between any two words is the cosine distance between their word2vec vector forms.

idf(w) is the inverse document frequency for word $w$, $idf(w) = ln(\frac{N}{n(w)})$ where N is the total number of messages in our corpus and $n(w)$ is the number of messages containing $w$ (Sparck Jones, 1972). A higher idf value for a word implies greater specificity and thus receives greater emphasis when calculating similarity.

Figure 6: Distribution of Cluster Sizes using MCS-w2v Distance




After calculating the similarity matrix using the MCS and w2v, we again apply affinity propagation, resulting in 89 unique clusters. The mean size is 11.3 messages with a median size of 10. Figure 6 shows the distribution of cluster sizes.

Table 9 presents two example clusters formed by Affinity Propagation using the MCS-w2v measure of message similarity. As in the corpus-only approach, there is a high degree of within-cluster similarity around a common theme. For the two selected examples, cluster 3 and 59, we label them as greetings and emphatic promises to choose Left respectively. As before, the exemplar elements serve as reliable summaries of each cluster's contents. We also observe that the greater flexibility afforded by the Knowledge-based similarity measure achieves the desired result of softening the boundaries between clusters while maintaining within-group similarity. Cluster 3 now includes short greetings regardless of whether they use "Hi" or "Hello."

With the new, reduced-size clusters, we again fit Ordinary Least-Squares models, regressing Trust and Share choices on dummy variables for each message's literal meaning. Following this, we regress the residuals of the Trust and Share models on dummy variable for each cluster. Our goal is to assess the robustness of our previous approach and examine whether the same communication strategies are correlated with changes in behavior across the two partitions of the message space. Table 11 and Table 10 present the most significant positive and most significant negative coefficient estimates in the Trust model.

Compared to the Corpus-only model, we find fewer clusters with significant, positive coefficients the Knowledge-based Trust model. However, the same message features stand out as the most effective. The clusters significant at the 5% level follow the same two patterns we observed in the previous model. We see detailed promises represented in cluster 39 and soft promises in cluster 31 and 15. Additionally, we see clusters featuring suggestions to cooperate without explicitly promising an action or requesting one from its recipient (clusters 64, and 66). Finally, cluster 32, which includes appeals to 'maximization' has a significant, positive association with trust. Notably, this cluster also has a significant, negative correlation with the likelihood their authors choose Left. Cluster 32 is the only cluster in our model with significant coefficients in both models in opposing directions.

Table 9: Example MCS-word2vec Clusters

| Cluster | Message (**exemplar**) |
|---|---|
| 3 | **Hi, how are you?** |
| 3 | a bit confussed but attemp either way. hi |
| 3 | HELLO! HOW ARE YOU? |
| 3 | Are you ready? |
| 3 | Hello, what are you going to choose? |
| 3 | hi how are you |
| 3 | HI HOW ARE YOU |
| 3 | hi hoe are you |
| 3 | hi, how are you? |
| 3 | hello how are you doing |
| 3 | Hello, how are you doing? |
| 59 | **I will always choose LEFT.** |
| 59 | I will always choose Left. I want to split the bonus in a fair way. |
| 59 | Hello. Let's try to maximize our earnings. I will always choose LEFT as I be... |
| 59 | I will always choose LEFT. That means we are both assured of getting a fair ... |
| 59 | I always split evenly with my fellow turkers. |
| 59 | I will always play Left if you play In. |
| 59 | I will choose Left. I always try to split earnings with other workers when I... |
| 59 | I'll always choose Left. |
| 59 | I promise to always play left so we all benefit. |

Table 10: Least Trusted MCS-word2vec Clusters

| Cluster | Exemplar Message | Size | Trust (t-stat) | Share (t-stat) |
|---|---|---|---|---|
| 80 | I choose out | 16 | -0.20 | 0.02 |
| | | | (-4.49) | (0.14) |
| 27 | choose right! | 15 | -0.19 | -0.40** |
| | | | (-4.17) | (-3.13) |
| 61 | it really good and nice | 6 | -0.21 | -0.20 |
| | | | (-2.9) | (-1.06) |
| 42 | out player a-8 player -b 8 | 7 | -0.16 | 0.32' |
| | | | (-2.41) | (1.79) |
| 74 | Please choose IN. | 27 | -0.08 | 0.01 |
| | | | (-2.37) | (0.13) |
| 85 | I AM PLAYER A | 8 | -0.15 | -0.16 |
| | | | (-2.37) | (-0.98) |
| 3 | Hi, how are you? | 21 | -0.08 | 0.05 |
| | | | (-2.06) | (0.47) |

Significant negative coefficient estimates when regressing In/Out choices on cluster dummies using a knowledge-based measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with each cluster's exemplar message chosen by the Affinity Propagation algorithm.
$* : p < .05, ** : p < .01, *** : p < .001$

Table 11: Most Trusted MCS-word2vec Clusters

| Cluster | **Description**<br>Exemplar Message | Size | Trust (t-stat) | Share (t-stat) |
|---|---|---|---|---|
| 31 | **State intent to do 'best for both of us'** | 26 | 0.11 | 0.09 |
| | I'll do the best for both of us! | | (3.12) | (0.93) |
| 66 | **"Let's work together"** | 15 | 0.11 | 0.13 |
| | Lets work together | | (2.53) | (1.06) |
| 64 | **Suggest 'both earning as much as possible'** | 5 | 0.18 | 0.06 |
| | Hey! Let's try and make as much as we can for everyone! | | (2.33) | (0.29) |
| 32 | **Appeal to 'maximization'** | 6 | 0.10 | -0.35** |
| | lets maximize | | (2.08) | (-2.79) |
| 39 | **Propose 'maximizing both' player's earnings** | 5 | 0.09 | -0.06 |
| | If you choose IN, I will choose LEFT to maximize both of our earnings. | | (1.96) | (-0.53) |
| 15 | **"I believe in equality"** | 5 | 0.17 | 0.35 |
| | I believe in equality | | (1.94) | (1.49) |

Significant positive coefficient estimates when regressing In/Out choices on cluster dummies using a knowledge-based measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with each cluster's exemplar message chosen by the Affinity Propagation algorithm.
$* : p < .05, ** : p < .01, *** : p < .001$

Table 12: MCS-word2vec Clusters: Share

| Cluster | **Description**<br>Exemplar Message | Size | Share<br>(t-stat) | Trust<br>(t-stat) |
|---|---|---|---|---|
| 73 | **Promises Left conditional on In & Specifies the outcome**<br>If you choose in, I will choose left. | 59 | 0.12*<br>(1.97) | 0.00<br>(0.02) |
| 32 | **Appeal to Maximization**<br>lets maximize | 14 | -0.35<br>(-2.79) | 0.10*<br>(2.08) |
| 27 | **Promise to choose Right**<br>choose right! | 15 | -0.40<br>(-3.13) | -0.19***<br>(-4.17) |

The most significant coefficient estimates when regressing Left/Right choices on cluster dummies using a Knowledge-based measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. $* : p < .05, ** : p < .01, *** : p < .001$

The least trusted messages in the knowledge-based model are effectively identical to those found in the corpus-based model. The bulk of these clusters contain messages that suggest the author does not understand the game. They state intent to either choose actions that are impossible within the bound of the game, e.g. "I choose player A," or state their intention to choose Right. Additionally, we again see that recipients of short, simple greetings are significantly more likely to distrust their sender, even relative to other non-promise messages.

In the Share model (Table 12) we see once again that the broad themes from our model using a knowledge-based measure of similarity match with the previous, corpus-only approach. The clusters associated with higher likelihoods of their senders choosing Left are detailed promises (cluster 73) specifying an outcome should the recipient choose In.

The clusters with significant, negative coefficients also support our conclusions from the Corpus-only approach. Senders who state their intention to play Right tend, strongly, to do so and recipients respond appropriately to these messages by choosing Out. Our only significant evidence of effective deception, explicit appeals to payoff maximization, is also robust to the alternative measure of similarity. The authors of these messages choose Right 34 percentage points more often, however, there is no significant change in their recipients' In/Out choices.

Overall, the conclusions from our knowledge-based approach agree with the previous, Corpus-only method. Although the resulting partitions are, indeed, sensitive to the choice of similarity measure, the same message features stand out in both the Trust and Share models. Receivers of emphatic, detailed promises are significantly more likely to choose In, relative to the typical promise. Similarly, soft promises, which imply an intention to choose Left without an explicit statement of intent, yield greater trust than other non-promise messages.

With less statistical power, the Share model provides less insight into Senders' behavior, with two coefficient estimates significant at the 5% level. This includes both promises to choose Right and appeals to payoff maximization, both identified in the Corpus-only model.

Despite the strong, negative coefficient estimate in the Share model, Receivers are more likely to trust an appeal to payoff maximization than other messages with the same literal meaning. There are no clusters with significant, positive coefficient estimates.

## 4.7 Reducing Cluster Size

Given the size of our corpus, containing a little over 1000 unique messages, having nearly one cluster for every ten messages raises concerns of over-clustering. We again reiterate that our purpose requires only that the within cluster similarity is sufficiently high. Our purpose is to identify new, relevant message features, not to determine a definitive partitioning of the message space which captures all of the potential complexity and nuance of written communication. Still, with these cluster sizes being implications of largely arbitrary rule-of-thumb parameter choices, we want to examine whether our results are robust to different tuning parameter decisions.

To address this concern, we conduct the same cluster analysis followed by OLS regression procedure as above except that we adjust our tuning parameter, a message's affinity towards itself, to reduce the number of resulting clusters. Specifically, we select this parameter with the objective of maximizing the BIC of the resulting linear model independently for both the Coprus-only and word2vec similarity measures.

This optimization reduces the number of clusters in the Corpus-only and word2vec model to 27 and 16 down from 112 and 89 respectively. As before, using cluster assignments as dummy variables, we fit a linear probability model, estimating the frequency recipients choose Left, trusting the sender, and the frequency with which senders choose Left, sharing the tokens if given the opportunity. Figures 13 and 14 present the coefficient estimates for the Corpus-only and Knowledge-based models, respectively.

## 4.8 Reduced Corpus-only Clusters

Out of the 27 clusters grouped using only information found in our corpus, 6 clusters (6, 10, 11, 18, 20, & 25) have positive, significant coefficient estimates in the trust model. We again see the two common themes from the previous cluster regressions. Clusters 10, 11, 18, and 25 include detailed promises, requesting In and specifying an outcome in addition to the stated intention to choose Left. The remaining two — clusters 6 and 20 — contain suggestions to work together, implying both a request for In and an intent to play Left without explicitly stating this.

In contrast, five clusters are associated with lower trust frequencies among their recipients (clusters 0, 1, 3, 9, and 12). These messages fall into two distinct themes. The first group consists of errors, messages that are either entirely nonsensical or that are not in line with a basic understanding of the game structure (clusters 3 and 9). Additionally, clusters 0 and 12 contain the briefest possible requests that the recipient choose In. The remaining cluster (cluster 1) contains brief greetings with no other content. Brief greetings and nonsense messages are also found in the previous cluster specifications.

In the reduced corpus-only Share model, we largely find agreement between the two models, with the signs of significant coefficient estimates in the Share model agreeing, at least in direction with those in the Trust models. There are no clusters which have coefficient

Table 13: Reduced Size Corpus-only Clusters

| Cluster | Size | **Description**<br>Exemplar Message | Trust<br>(t-stat) | Share<br>(t-stat) |
|---|---|---|---|---|
| 0 | 44 | **Brief request for In** | -0.06* | 0.03 |
| | | Please Choose In | (-2.07) | (0.45) |
| 1 | 21 | **"Hi, how are you?"** | -0.08* | -0.06 |
| | | Hi, how are you? | (-2.07) | (-0.61) |
| 3 | 18 | **"I choose right"** | -0.18*** | -0.30** |
| | | I choose right. | (-4.37) | (-2.70) |
| 5 | 22 | **Good luck / good morning** | -0.04 | -0.25* |
| | | good luck! | (-1.15) | (-2.52) |
| 6 | 27 | **Suggestions to 'work together'** | 0.12*** | 0.20* |
| | | I hope we can work together to win. | (3.57) | (2.28) |
| 9 | 25 | **"I choose out"** | -0.20*** | 0.04 |
| | | I chose out. | (-5.46) | (0.41) |
| 10 | 44 | **Request In, promise Left, & specify the 'maximum/best' outcome** | 0.06* | 0.03 |
| | | If you choose IN, I will choose LEFT to maximize both of our earnings. | (2.40) | (0.45) |
| 11 | 103 | **Request In, promise Left, & specify '12 token' outcome** | 0.05** | 0.05 |
| | | If you choose In I will choose Left and we will both get 12 tokens | (2.84) | (1.01) |
| 12 | 37 | **Brief promise of Left and request for In** | -0.07* | -0.11 |
| | | Choose IN. I'll choose LEFT. | (-2.33) | (-1.47) |
| 17 | 83 | **Promise Left conditional on In** | -0.01 | 0.14** |
| | | I will be choosing left if you choose in | (-0.74) | (2.75) |
| 18 | 64 | **Request In, promise Left, specify '12 token' outcome & uses 'please'** | 0.05* | -0.11' |
| | | Please choose "IN" so that we can both get 12 points. I will choose "LEFT." | (2.22) | (-1.89) |
| 19 | 18 | **"I like this game"** | 0.02 | 0.34** |
| | | I like this game. | (0.41) | (2.92) |
| 20 | 42 | **States intent to make the 'most for both of us'** | 0.09*** | 0.25*** |
| | | Trying to make the most for both of us | (3.39) | (3.47) |
| 22 | 24 | **Suggests 'best payoff' without a request or promise** | 0.01 | -0.24* |
| | | ALL THE BEST | (0.16) | (-2.36) |
| 24 | 34 | **"I am Player A"** | -0.03 | -0.18* |
| | | I AM PLAYER A | (-1.04) | (-2.25) |
| 25 | 48 | **Request In, promise Left, and specifies 'most money'** | 0.06* | 0.06 |
| | | You should choose IN so I can choose LEFT and we both make the most money. | (2.30) | (0.93) |

Coefficient estimates when regressing Left/Right and In/Out choices on cluster dummies using the corpus-only measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. Only clusters with significant coefficient estimates in either model are shown for brevity. $*: p < .05, **: p < .01, ***: p < .001$

estimates significantly different from zero in the Share model with significant, opposing coefficient estimates in the Trust model.

In the Share model, four clusters have significant positive coefficient estimates. Clusters 6 and 20 both consist of suggestions to "work together" or statements of intent to do so. Cluster 17 includes promises to choose Left, conditional on the Receiver's choice to choose In. This conforms to the previous results showing that detailed promises outperform simple ones. Finally, messages in which the Sender states that they "like this game" (cluster 19) are significantly more likely to come from trustworthy senders than other messages with no mention of game actions.

Negative, significant clusters in the reduced, corpus-only Share model also agree with the results of the previous approaches. Nonsensical messages, promising to choose Right or to choose Player A are significantly, negatively associated with their senders choosing Left (clusters 3 and 24). Senders of messages in cluster 22 state their intent to "maximize" and are significantly more likely to choose Right.

## 4.9 Reduced word2vec Clusters

Clustering using the word2vec similarity measure yields fewer clusters compared to the Corpus-only approach. As intended, this approach mitigates some of the potential over-clustering seen in the first approach and we have fewer groups of closely related messages divided into multiple categories. After this reduction, the same general themes remain common among the messages in clusters with significant coefficient estimates, both positive and negative, in the Trust model. Of the 16 clusters, 4 have significant positive coefficient estimates, controlling for literal meaning (clusters 5, 9, and 11). These clusters capture the same patterns observed in the models using the previous similarity measures. Clusters 5 includes 'soft promises' in which the Sender implies that they will choose Left without explicitly stating intended actions. Similarly, messages in cluster 11 suggest that the players "work together." This is also a soft promise, but also implies a request that the Receiver choose In as well. Finally, cluster 9 includes detailed promises, with requests for In and a specified outcome of "12 tokens."

The clusters with significant, negative coefficient estimates in the reduced word2vec Trust model are again familiar. All significant, negative clusters (clusters 1, 3, and 14) are either incoherent (cluster 1 messages promise to "choose Player A") or make little strategic sense in the context of the game. Clusters 3 and 14 promise to choose Right or request the Receiver chooses Out, respectively.

In the Share model, there is one cluster with a significant coefficient estimate, positive or negative. Messages in cluster 3 are promises to choose Right. Consistent with our previous results, Senders of these messages keep their word and choose Right at higher frequencies.

## 4.10 Sentiment

Using the bag-of-words and affinity propagation approaches allow us to gain valuable insights to the message topics and their association with trust behaviors. Although this has already allowed us to discover new potentially impactful features of persuasive messages, there remains a great deal of nuance left out by this reduction. To capture a portion of

Table 14: Reduced MCS-word2vec Clusters

| Cluster | Size | **Description**<br>Exemplar Message | Trust<br>(t-stat) | Share<br>(t-stat) |
|---|---|---|---|---|
| 0 | 43 | **Request for In including ‘please’** | -0.05’ | 0.05 |
| | | Please Choose In | (-1.89) | (0.66 ) |
| 1 | 46 | **“I choose Player A”** | -0.09*** | 0.09 |
| | | I choose player A | (-3.48) | (1.31) |
| 2 | 140 | **Requests In, promises Left, & specifies the ‘best’ outcome** | 0.02 | 0.04 |
| | | I will choose left if you choose IN | (1.56) | (1.01) |
| 3 | 26 | **Brief Promises for Right** | -0.14*** | -0.36*** |
| | | choose right | (-3.96) | (-3.81) |
| 4 | 115 | **Brief promises for Left only** | -0.02 | -0.03 |
| | | i will choose left | (-0.97) | (-0.57) |
| 5 | 113 | **States intent for what’s ‘best for everyone’** | 0.05** | 0.01 |
| | | Hello, I want to do what’s best for the both of us so I’m ... | (3.24) | (0.24) |
| 6 | 34 | **Greeting including ‘good luck’** | -0.02 | -0.03 |
| | | good luck! | (-0.56) | (-0.38) |
| 7 | 69 | **Greeting including ’let’s play’** | -0.03 | -0.02 |
| | | Hello Let’s play the game | (-1.4) | (-0.42) |
| 8 | 70 | **“Let’s make the most money”** | 0.03 | -0.08 |
| | | lets go in | (1.59) | (-1.47) |
| 9 | 69 | **Requests In, promises Left, & specifies ‘12 token’ outcome** | 0.05* | 0.06 |
| | | If you choose In I will choose Left and we will both get 1... | (2.34) | (1.1) |
| 10 | 80 | **Request for In only** | 0.02 | -0.01 |
| | | Choose The In | (0.89) | (-0.13) |
| 11 | 29 | **Proposal to ‘work together’** | 0.11*** | 0.08 |
| | | Lets work together | (3.51) | (0.95) |
| 12 | 64 | **Promises Left only** | -0.0 | -0.11’ |
| | | im going left | (-0.19) | (-1.77) |
| 13 | 38 | **Brief greetings** | -0.05’ | 0.02 |
| | | HELLO! HOW ARE YOU? | (-1.7) | (0.29) |
| 14 | 37 | **Promises Out** | -0.15*** | 0.05 |
| | | I choose out | -(5.07) | (0.67) |
| 15 | 37 | **States intent to be ‘fair’** | 0.05’ | 0.13’ |
| | | I will be fair | (1.77) | (1.71) |

Coefficient estimates when regressing Left/Right and In/Out choices on cluster dummies using the word2vec measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with a manual description in bold as well as each cluster’s exemplar message chosen by the Affinity Propagation algorithm. ‘ : $p < .1, * : p < .05, ** : p < .01, *** : p < .001$

Table 15: Regressions on Sentiment

| | Trust | Share | Trust w Control | Share w Control |
|---|---|---|---|---|
| Positive Words | 0.004 | 0.014 | 0.006 | 0.012 |
| | (0.006) | (0.015) | (0.006) | (0.015) |
| Negative Words | -0.036$^{**}$ | -0.047 | -0.006 | -0.033 |
| | (0.013) | (0.033) | (0.012) | (0.033) |
| Total Words | 0.010$^{***}$ | 0.006$^{*}$ | 0.005$^{***}$ | 0.003 |
| | (0.001) | (0.002) | (0.001) | (0.003) |
| Left_Only | | | 0.195$^{***}$ | 0.210$^{***}$ |
| | | | (0.015) | (0.040) |
| In_Only | | | 0.135$^{***}$ | 0.029 |
| | | | (0.017) | (0.047) |
| Both | | | 0.198$^{***}$ | 0.144$^{**}$ |
| | | | (0.017) | (0.045) |
| CONSTANT | 0.473$^{***}$ | 0.545$^{***}$ | 0.413$^{***}$ | 0.498$^{***}$ |
| | (0.010) | (0.025) | (0.010) | (0.027) |
| N | 976 | 969 | 976 | 969 |

Regression coefficient estimates for the number of tokens with positive or negative sentiment as defined by the OpinionFinder Subjectivity Lexicon. $* : p < .05, ** : p < .01, *** : p < .001$

this additional nuance, we conduct further analysis into our messages by examining their sentiment.

To do this, we re-code our set of tokens into positive, negative, and neutral words based on the classifications provided by the OpinionFinder Subjectivity Lexicon (Wilson et al., 2005). As with previous steps, we must alter this procedure slightly to better fit our environment, removing 'right' from the list of words with positive sentiment as it has a neutral meaning as a potential action in our game; all usages of this word in our data refer to the game action. We construct three variables for each message: a count of positive words, negative words, and a control for total word count.

We regress trust and trustworthiness on positive, negative, and total word count both with and without controlling for a message's literal meaning. Table 15 presents the results of these regressions.

Without controlling for literal meaning, we find that messages with negative sentiment have a significant negative relationship with trust, with each additional positive word being associated with a 3.6 percentage point decrease in trust frequency. There is a additional, positive impact of increased message length on trust. However, after controlling for literal meaning, the significance of this sentiment effect disappears. Neither positive nor negative words have a significant impact on trust or trustworthiness. In the data from our trust game, message sentiment is not a useful predictor of behavior.

# 5 Conclusion

Free-form communication increases cooperation between economic decision makers. However, it is unclear why this effect is so pronounced relative to similar information conveyed through limited communication mechanisms. Investigating this phenomenon is difficult as the same free-form structure which allows for increased cooperation also makes any data collected difficult to parse. Free-form communication increases the action space available to Senders to an intractably large magnitude. Analysing the contents of free-form messages requires that researchers divide the space of all possible actions into a more manageable partition. Typically, the message space is reduced based on human coders assigning each message a meaning corresponding to actions available in the game. In this chapter, we employ Natural Language Processing to generate alternative partitions of the messages space. We then examine whether these alternative message structures reveal any new features of effective communication which would not otherwise have been captured by a message model based solely on the relevant game's actions.

Our first approach, we consider each message as a series of independent tokens, reducing the message space into a manageable collection of binary variables that we can then approach with standard OLS regression methods. Using this approach, we establish a baseline result that, after controlling for the literal meaning of a message, our message data contain information useful in predicting trust and trustworthiness. Explicit reference to promised/requested actions ("Left" or "In") results in greater trust compared to less-explicit promises or requests. Further, emphatic promises containing the word "Always" yield greater trust than other promises and declarations of "Fairness" yield greater trust than other non-promises. The bag-of-words approach yields useful insights into free-form communication but comes with obvious drawbacks. The assumption of linear independence is unrealistic and interpreting results requires one to examine tokens in their original context to give meaning to the regression results. To address the drawbacks of the bag-of-words approach, we conduct cluster analysis, better preserving the complex relations between tokens.

In our second approach, we conduct cluster analysis using Affinity Propagation to divide the message space into groups with similar content. For Affinity Propagation to group similar messages, we must select a distance measure which quantifies the similarity between pairs of messages. We perform Affinity Propagation multiple times with different tuning parameters and distance measures. For distance measures, we use the Jaccard index of set overlap and a measure based on the Google word2vec similarity measure between pairs of words. The Jaccard index is calculated solely based on information within our collection of messages whereas word2vec incorporates existing NLP knowledge from external corpora.

The specifics vary depending on the distance measure and tuning parameters used, but across our four Affinity Propagation groupings, several common themes emerge in the most/least effective clusters after controlling for literal meaning. Promises and requests which specify an outcome (e.g., "12 tokens" or "the most bonus for both of us") are trusted more often than other comparable messages. We also find that clusters containing 'soft promises' — messages such as "I like being fair" which imply cooperation without an explicitly stated intention — elicit trust more often than other non-promise messages. Together, these results suggest that promises are not binary, but instead very in strength. Detailed, emphatic promises perform better than simple promises while implicit promises increase trust, but to a lesser extent than explicit ones.

In contrast, messages consisting only of greetings result in lower rates of trust as do low quality messages (message which contain grammatical errors, are brief, or display apparent confusion about the game). This is notable because brief, low quality messages are not significantly less trustworthy than others with the same literal meaning. Receivers are either inferring incorrect information about the senders of poor quality messages or considering factors beyond their own monetary payoffs when choosing between In and Out.

With much less statistical power (one sender versus ten receivers of each message) our insights into Senders' choices are limited. However, when we do find clusters significantly correlated with higher/lower rates of trustworthiness, they tend to also be significant predictors of trust in the same direction. One notable exception to this rule are appeals directly to optimization, e.g. "Let's maximize our payoff." Appeals to optimization are significantly less trustworthy, but Receivers do not choose In at a significantly lower rate. This is the only evidence we find of potential, effective deception in our analysis.

Finally, we use sentiment analysis to examine the effect of a message's tone on trust. We find no significant correlation between a message's tone and trust or trustworthiness.

The Natural Language Processing techniques we use require delicate interpretation and we must limit the conclusions we draw from our analyses. This work does not claim to establish a definitive model of free-form communication. Our cluster analysis successfully groups sufficiently similar messages together to allow us to interpret each cluster, but the specifics rely heavily on largely arbitrary choices of similarity measures. Our approach is intended to demonstrate a process for systematically parsing unstructured message data to identify features worth further investigation.

In our trust game data, Natural Language Processing using cluster analysis suggests that future research into free-form communication should consider that promises may come in varying strengths. Detailed, emphatic promises outperform simple statements of intention while implied, 'soft' promises increase trust to a lesser extent. The techniques we employ are not well suited to definitively establish that the strength of a promise matters in building trust, but our results may motivate additional research. Receivers' aversion to low-quality messages also warrants further investigation.

Overall, what the analyses we report establish is that free-form communication data contain useful information beyond their basic, literal meanings. Using Natural Language Processing, we can effectively parse unstructured message data to identify new features associated with higher rates of cooperation. Although the approaches we use do not allow us to draw definitive conclusions about the impact of specific message features, the results suggest directions for future research. Larger studies employing the same tools used here can use Natural Language Processing to generate hypotheses about relevant communication features which can then be tested with greater precision in follow-up experiments.

# References


Andreoni, J. and Rao, J. M. (2011). The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of public economics*, 95(7-8):513–520.

Andres, M., Bruttel, L., and Friedrichsen, J. (2023). How communication makes the difference between a cartel and tacit collusion: A machine learning approach. *European Economic Review*, 152:104331.

Ben-Ner, A., Putterman, L., and Ren, T. (2007). Lavish returns on cheap talk: Non-binding communication in a trust experiment. *Available at SSRN 1013582*.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Buchan, N. R., Johnson, E. J., and Croson, R. T. (2006). Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences. *Journal of Economic Behavior & Organization*, 60(3):373–398.

Bursztyn, L., Egorov, G., Haaland, I., Rao, A., and Roth, C. (2023). Justifying dissent. *The Quarterly Journal of Economics*, 138(3):1403–1451.

Camerer, C. F., Nave, G., and Smith, A. (2019). Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning. *Management Science*, 65(4):1867–1890.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Dufwenberg, M. (2010). Bare promises: An experiment. *Economics letters*, 107(2):281–283.

Charness, G. and Gneezy, U. (2008). What's in a name? anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization*, 68(1):29–35.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chen, J. and Houser, D. (2017). Promises and lies: can observers detect deception in written messages. *Experimental Economics*, 20:396–419.

Cooper, D. J. and Kühn, K.-U. (2016). Communication and cooperation: A methodological study. *Southern economic journal*, 82(4):1167–1185.

Crawford, V. (1998). A survey of experiments on communication via cheap talk. *Journal of Economic theory*, 78(2):286–298.

Devetag, G., Di Guida, S., and Polonio, L. (2016). An eye-tracking study of feature-based choice in one-shot games. *Experimental Economics*, 19:177–201.

Ellingsen, T. and Johannesson, M. (2004). Promises, threats and fairness. *The Economic Journal*, 114(495):397–420.

Ellingsen, T., Östling, R., and Wengström, E. (2018). How does communication affect beliefs in one-shot games with complete information? *Games and Economic Behavior*, 107:153–181.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Fiedler, M. and Haruvy, E. (2009). The lab versus the virtual lab and virtual field—an experimental investigation of trust games with communication. *Journal of Economic Behavior & Organization*, 72(2):716–724.

Fonseca, M. A. and Normann, H.-T. (2012). Explicit vs. tacit collusion—the impact of communication in oligopoly experiments. *European Economic Review*, 56(8):1759–1772.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

Fudenberg, D. and Liang, A. (2019). Predicting and understanding initial play. *American Economic Review*, 109(12):4112–4141.

Grimm, S., Kocher, M. G., Krawczyk, M., and Le Lec, F. (2021). Sharing or gambling? on risk attitudes in social contexts. *Experimental Economics*, 24:1075–1104.

Hanaki, N. and Ozkes, A. I. (2022). Strategic environment effect and communication. *Experimental Economics*, pages 1–34.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Lundquist, T., Ellingsen, T., Gribbe, E., and Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization*, 70(1-2):81–92.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank.

Mihalcea, R., Corley, C., Strapparava, C., et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mohlin, E. and Johannesson, M. (2008). Communication: Content or relationship? *Journal of Economic Behavior & Organization*, 65(3-4):409–419.

Penczynski, S. P. (2019). Using machine learning for communication classification. *Experimental Economics*, 22(4):1002–1029.

Polonio, L. and Coricelli, G. (2019). Testing the level of consistency between choices and beliefs in games using eye-tracking. *Games and Economic Behavior*, 113:566–586.

Rehurek, R. and Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2):2.

Robinson, J., Rosenzweig, C., Moss, A. J., and Litman, L. (2019). Tapped out or barely tapped? recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. *PloS one*, 14(12):e0226394.

Snowberg, E. and Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, 111(2):687–719.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Tebbe, E. and Wegener, B. (2022). Is natural language processing the cheap charlie of analyzing cheap talk? a horse race between classifiers on experimental communication data. *Journal of Behavioral and Experimental Economics*, 96:101808.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Woods, D., Abdallah, M., Bagchi, S., Sundaram, S., and Cason, T. (2022). Network defense and behavioral biases: an experimental study. *Experimental Economics*, 25(1):254–286.

# Appendix

## 5.1 Affinity Propagation agorithm description

Affinity Propagation begins with a similarity matrix $\mathbf{S}_{n\times n}$ and iteratively updates two matrices: $\mathbf{R}_{n\times n}$, the responsibility matrix and $\mathbf{A}_{n\times n}$, the availability matrix until stable clusters emerge. Both $\mathbf{A}$ and $\mathbf{R}$ initialize with all zero elements. The algorithm proceeds as follows:

1) $\mathbf{R}$ updates following:

$$r(i,k) \leftarrow s(i,k) - \max_{\forall k \neq k'}\{a(i,k') + s(i,k')\}$$

2) $\mathbf{A}$ updates following:

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \notin i,k} \max(0, r(i',k))\}$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k))$$

Final clusters assignments are given by the criteria matrix $\mathbf{C} = \mathbf{A} + \mathbf{S}$; element $i$ chooses $k$ as its exemplar if the highest value in row $i$ in $\mathbf{C}$ is in column $k$. This process repeats, updating $\mathbf{A}$ and $\mathbf{R}$ until cluster assignments remain unchanged between iterations.

Intuitively, the Affinity Propagation algorithm considers, first, each element's choice for its own exemplar out of the other elements in the set, $\mathbf{R}$. After this, these preferences are updated to account for the overall appropriateness of a given message to serve as an exemplar in the final cluster assignments. if a given message is chosen as a most preferred exemplar by relatively few other messages, it will incur a penalty reflecting its overall unsuitability to act as an exemplar.

Table 1 Expanded: Example Messages: Most and Least Effective

| Type | Messages |
|---|---|
| Most Effective | I'm going LEFT, so please select IN. This will maximize the earnings for both of us equally.<br>Hello, I want to do what's best for the both of us so I'm going to be choosing left so that we can both get 12 tokens<br>choose IN so that we will both get 12 tokens<br>I'll choose LEFT if you choose IN.<br>Choose "In" and I'll make sure we both get 12 tokens!<br>I give you my word I will get each of us 12 if you go IN. That way we will both make more.<br>Hi there. I would understand why you would choose OUT, but I promise to make the choice ...<br>I promise you that I will always choose "left" so that we both get 12 points each.<br>Helloo!!! Let's make this work for both of us in the best possible way! Have a great day!<br>Hey, I'm not a greedy person and in these situations, I always choose the most beneficial option for ...<br>You can be sure that i will choose left, as it is the best option for both us. Please go ahead and select 'In'<br>If pick in, I guarantee you will get 12 instead of 8 if you pick out<br>I'm good for the 12. I believe that there are real people here, and I'm completely down with splitting the 24 and ...<br>I'm picking left so if you want a few extra tokens, go for in. Totally understand if you don't but that's my choice.<br>Let's work together to make things fair and so that we can both earn as many tokens as possible.<br>I'm always going to choose "left" because I think it's dickish to screw another player over.<br>So we can maximize our winnings, I will choose left on each turn if you choose In on each turn. That way we both...<br>I will maximize both our outcomes ie LEFT if you stay in.<br>Cooperation is better than being selfish. Love reigns. |
| Least Effective | I'm going to choose Right<br>i get 24 you get 0<br>U GO ON STRITE<br>Hello Let's play the game<br>I want choose Player A<br>hi, how are you doing?<br>Hey. let's work together, let's both choose out.<br>I choose out<br>choose out |

Table 16: Token regressions controlling for the four message types

| Tokens | Trust | T-stat | Trustworthiness | T-stat |
|---|---|---|---|---|
| **i** | 0.003 | (0.212) | 0.063 | (1.411) |
| **choose** | -0.023 | (-1.512) | -0.099* | (-2.259) |
| **in** | 0.068*** | (3.545) | 0.036 | (0.662) |
| **left** | 0.060** | (3.278) | 0.083 | (1.608) |
| **you** | 0.004 | (0.217) | 0.076 | (1.618) |
| **to** | 0.029* | (2.069) | 0.088* | (2.187) |
| **will** | -0.005 | (-0.282) | 0.081 | (1.773) |
| **be** | -0.012 | (-0.847) | -0.077 | (-1.942) |
| **the** | 0.028 | (1.761) | -0.069 | (-1.517) |
| **both** | -0.013 | (-0.635) | 0.042 | (0.764) |
| **we** | 0.047* | (2.202) | 0.138* | (2.306) |
| **and** | 0.031* | (1.998) | 0.046 | (1.051) |
| **a** | -0.009 | (-0.529) | -0.064 | (-1.326) |
| **go** | -0.013 | (-0.732) | -0.114* | (-2.326) |
| **if** | 0.004 | (0.188) | -0.001 | (-0.022) |
| **get** | 0.033 | (1.689) | -0.023 | (-0.430) |
| **of** | 0.016 | (0.754) | -0.020 | (-0.333) |
| **for** | 0.008 | (0.398) | 0.106* | (1.974) |
| **let** | 0.011 | (0.487) | -0.116 | (-1.815) |
| **so** | -0.026 | (-1.196) | -0.108 | (-1.726) |
| **u** | 0.048* | (2.007) | 0.003 | (0.043) |
| **player** | -0.023 | (-0.863) | 0.060 | (0.794) |
| **’s** | 0.001 | (0.038) | 0.159* | (2.363) |
| **that** | 0.042* | (2.043) | 0.043 | (0.742) |
| **can** | 0.046* | (2.084) | -0.143* | (-2.275) |
| **make** | 0.066** | (3.015) | 0.010 | (0.165) |
| **out** | -0.101*** | (-4.933) | 0.087 | (1.509) |
| **hi** | -0.001 | (-0.055) | 0.047 | (0.866) |
| **hello** | 0.029 | (1.464) | 0.069 | (1.243) |
| **12** | 0.041 | (1.664) | -0.007 | (-0.102) |
| **play** | -0.022 | (-1.007) | 0.028 | (0.462) |
| **do** | -0.021 | (-0.947) | -0.062 | (-1.009) |
| **token** | 0.018 | (0.720) | -0.074 | (-1.063) |
| **’m** | 0.010 | (0.411) | 0.009 | (0.137) |
| **fair** | 0.052* | (2.316) | 0.113 | (1.793) |
| **this** | 0.063** | (2.730) | 0.099 | (1.510) |
| **it** | -0.046 | (-1.856) | -0.120 | (-1.708) |
| **good** | 0.004 | (0.189) | -0.039 | (-0.615) |
| **’ll** | -0.028 | (-1.127) | 0.046 | (0.655) |
| **best** | 0.028 | (1.170) | 0.005 | (0.076) |
| **please** | 0.013 | (0.549) | 0.070 | (1.065) |
| **more** | 0.030 | (1.181) | -0.139 | (-1.924) |
| **maximize** | 0.067* | (2.247) | -0.148 | (-1.692) |
| **point** | -0.031 | (-1.285) | 0.025 | (0.357) |
| **our** | -0.007 | (-0.198) | -0.004 | (-0.042) |
| **each** | -0.015 | (-0.545) | 0.008 | (0.100) |
| **most** | -0.023 | (-0.808) | 0.121 | (1.532) |
| **game** | -0.050 | (-1.763) | -0.082 | (-1.000) |
| **all** | -0.033 | (-1.191) | -0.024 | (-0.307) |
| **b** | 0.016 | (0.452) | -0.056 | (-0.575) |
| **always** | 0.053* | (2.068) | 0.136 | (1.869) |
| **money** | 0.008 | (0.263) | -0.047 | (-0.570) |
| **pick** | 0.021 | (0.719) | 0.008 | (0.096) |
| **with** | 0.006 | (0.229) | 0.000 | (0.001) |
| **have** | 0.053 | (1.788) | 0.088 | (1.040) |
| **right** | -0.130*** | (-4.614) | -0.224** | (-2.767) |
| **LENGTH** | -0.000 | (-1.510) | -0.000 | (-0.803) |
| **LEFTONLY** | 0.584*** | (22.134) | 0.620*** | (8.308) |
| **INONLY** | 0.506*** | (18.908) | 0.543*** | (7.178) |
| **BOTH** | 0.544*** | (16.800) | 0.534*** | (5.828) |
| **NEITHER** | 0.430*** | (29.454) | 0.504*** | (12.189) |

Each coefficient is associated with a dummy variable indicating a message includes the relevant word. Control variables for messages' literal meaning are omitted from the table. $* : p < .05, ** : p < .01, *** : p < .001$