Vlad Niculae, Srijan Kumar, **Jordan Boyd-Graber**, and Cristian Danescu-Niculescu-Mizil. **Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game**. *Association for Computational Linguistics*, 2015, 10 pages.


```
    @inproceedings{Niculae:Kumar:Boyd-Graber:Danescu-Niculescu-Mizil-2015,
Title = {Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game},
Author = {Vlad Niculae and Srijan Kumar and Jordan Boyd-Graber and Cristian Danescu-Niculescu-Mizil},
Booktitle = {Association for Computational Linguistics},
Year = {2015},
Location = {Beijing, China},
Url = {http://umiacs.umd.edu/~jbg//docs/2015_acl_diplomacy.pdf},
}
```


Accessible Abstract: This paper introduces the application of natural language processing techniques to understand the relationships (and their dissolution) in the game of Diplomacy. This popular board game simulates Europe at the eve of World War I and forces players to work with each other to forge alliances and make plans together. However, the game's setup also encourages players to turn against each other. This paper analyzes whether we can predict these betrayals (we can!) and the linguistic and social phenomena (demands, politeness, and planning) that can predict when a betrayal will happen.

Links:

- Code/Data [http://vene.ro/betrayal/]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game

Vlad Niculae,[1] **Srijan Kumar,**[2] **Jordan Boyd-Graber,**[3] **Cristian Danescu-Niculescu-Mizil**[1]
[1]Cornell University, [2]University of Maryland, [3]University of Colorado
vlad@cs.cornell.edu, srijan@cs.umd.edu,
Jordan.Boyd.Graber@colorado.edu, cristian@cs.cornell.edu

Abstract

Interpersonal relations are fickle, with close friendships often dissolving into enmity. In this work, we explore linguistic cues that presage such transitions by studying dyadic interactions in an online strategy game where players form alliances and break those alliances through betrayal. We characterize friendships that are unlikely to last and examine temporal patterns that foretell betrayal.

We reveal that subtle signs of imminent betrayal are encoded in the conversational patterns of the dyad, even if the victim is not aware of the relationship's fate. In particular, we find that lasting friendships exhibit a form of balance that manifests itself through language. In contrast, sudden changes in the balance of certain conversational attributes—such as positive sentiment, politeness, or focus on future planning—signal impending betrayal.

1 Introduction

A major focus in computational social science has been the study of interpersonal relations through data. However, social interactions are complicated, and we rarely have access all of the data that define the relationship between friends or enemies. As an alternative, thought experiments like the prisoner's dilemma (Axelrod and Dion, 1988) are used to explain behavior. Two prisoners—denied communication—must decide whether to cooperate with each other or defect. Such simple and elegant tools initially helped understand many real world scenarios from pricing products (Rosenthal, 1981) to athletes doping (Buechel et al., 2013). Despite its power, the prisoner's dilemma remains woefully unrealistic. Cooperation and betrayal do not happen in a cell cut off from the rest of the world. Instead, real interactions are mediated by communication: promises are made, then broken, and met with recriminations.

To study the complex social phenomenon of betrayal, we turn to data and observe the players of **Diplomacy** (Sharp, 1978), a war-themed strategy game where friendships and betrayals are orchestrated primarily through language. Diplomacy, like the prisoner's dilemma, is a repeated game where players choose to either cooperate or betray other players. Diplomacy is so engaging that it is played around the world, not only casually as a board game but also over the Internet and in formal settings such as world championships.[1] Players converse throughout the game and victory hinges on enlisting others' support through persuasiveness and cunning duplicity. To illustrate the social relations that carry out throughout the game, consider the following exchange between two Diplomacy allies:

> Germany: Can I suggest you move your armies east and then I will support you? Then next year you move *[there]* and dismantle Turkey. I will deal with England and France, you take out Italy.
>
> Austria: Sounds like a perfect plan! Happy to follow through. And—thank you Bruder!

Austria is very polite and positive in its reply, and appreciates Germany's support and generosity. They have been good allies for the better part of the game. However, immediately after this exchange, Austria suddenly invades German territory. The intention to do so was so well concealed that Germany did not see the betrayal coming; otherwise it would have taken advantage first. Indeed, if we follow their conversation after the attack, we find Germany surprised:

> Germany: Not really sure what to say, except that I regret you did what you did.

[1]A recent episode of *This American Life* describes the Diplomacy game in a competitive offline setting: http://www.thisamericanlife.org/radio-archives/episode/531/got-your-back?act=1

Such scenarios suggest an important research challenge: is the forthcoming betrayal signaled by linguistic cues appearing in the (ostensibly friendly) conversation between the betrayer and the eventual victim? A positive answer would suggest not only that the betrayer unknowingly reveals their future treachery, but also that the eventual victim fails to notice these signals. Capturing these signals computationally would therefore mean outperforming the human players.

In this work, we provide a framework for analyzing a dyad's evolving communication patterns and provide evidence of subtle but consistent conversational patterns that foretell the unilateral dissolution of a friendship. In particular, imminent betrayal is signaled by sudden changes in the balance of conversational attributes such as positive sentiment, politeness, and structured discourse. Furthermore, we show that by exploiting these cues in a prediction setting we can anticipate imminent betrayal better than the human players.

After briefly describing the game (Section 2), we focus on how the structure of the game provides convenient, reliable indicators of whether pairs of participants are friends or foes (Section 3). Given these labels, we explore linguistic features that are predictive of whether friendships will end in betrayal (Section 4) and, if so, when the betrayal will happen (Section 5).

While our focus is on a single popular game, we choose methods that generalize to other domains, revealing dynamics present in other social interactions (Section 6). We discuss how automatically predicting stable relationships and betrayal can more broadly help advance the study of trust and relationships using computational linguistics.

2 Communication and Conflict in Diplomacy

A game of Diplomacy begins in 1901 with players casting themselves as the European powers at the eve of the first world war: England, Germany, France, Russia, Austria, Italy, and the Ottoman Empire. The goal of the game (like other war games such as Risk or Axis & Allies) is to capture all of the territories on the game board (Figure 1). The games are divided into years starting from 1901 and each year is divided into two seasons—Spring and Fall. Each season consists of two alternating phases: *diplomacy*—the players communicate to form strategies—and *orders*—the players submit their moves for the season. Seasons are therefore the main unit of game time.

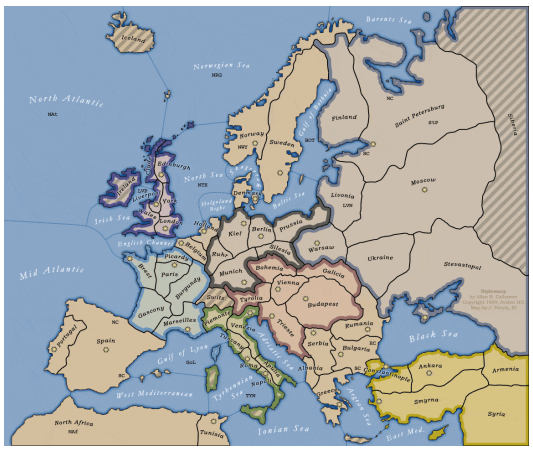


Figure 1: The full Diplomacy board representing Europe circa 1914. The seven nations struggle to control the map.

2.1 Movement, Orders, and Battles

On the board, each player can operate a unit for each city they control. During each turn, these pieces have the option of moving to an adjacent territory. What makes Diplomacy unique is that all players submit their written (or electronic) orders; these orders are executed simultaneously; and *there is no randomness* (e.g., dice). Thus, the outcome of the game depends only on the communication, cooperation, and movements of players.

When two units end their turn in the same territory, it implies a battle. Who wins the battle is decided purely based on numerical superiority (ties go to defenders). Instead of moving, a unit can support another unit; large armies can be created through intricate networks of support. The side with the largest army wins the battle.

The process of *supporting* a unit is thus critical for both a successful offensive move and a successful defense. Often, a lone player lacks the units to provide enough support to his attacks and thus needs the help of others.[2] Because these orders (both movement and support) are machine readable, we have a clear indication of when players are working together (supporting each other) or working against each other (attacking each other); we will use this to define relationships between

[2] While support can come from a player's own units, allies often combine resources. For example, if an English army in Belgium is attacking a Germany Army in Ruhr, a French army in Burgundy could strengthen that attack. This is accomplished by the French player submitting a move explicitly stating "I support England's attack from Belgium to Ruhr".

players (Section 3). However, coordinating these actions between players requires cooperation and *diplomacy*.

2.2 Communication

In the *diplomacy* phase of the game, players talk to each other. These conversations are either public or—more typically—one-on-one. Conversations include greetings, extra-game discussions (e.g., "did you see Game of Thrones?"), low-level tactics ("if you attack Armenia, I'll support you"), and high-level strategy ("we need to control Central Europe"). The content of these messages forms the object of our study.

Because of the centrality of language to Diplomacy, we can learn the rhetorical and social devices players use to build and break trust. Because this language is embedded in every game, it has convenient properties: similar situations are repeated, the goals are clear, and machine-readable orders confirm which players are enemies and which are friends. In the next section, we explore the Diplomacy data.

2.3 Preprocessing

We use games from two popular online platforms for playing Diplomacy.[3] The average season of an online Diplomacy game lasts nine days. We remove non-standard games caused by differences between the two platforms, as well as games that are still in progress. Moreover, in each game, we filter out setup messages, regulatory messages to and from the administrator of the game and messages declaring the state of the game, keeping only messages between the players. This leaves 249 games with 145.000 total messages.

The dataset confirms that communication is an essential part of Diplomacy: half of the games have over 515 messages exchanged between the players, while the top quartile has over 750 messages per game. Also, non-trivial messages (with at least one sentence) tend to be complex: over half of them have at least five sentences, and the top quartile consists of messages with eight or more sentences.

3 Relationships and Their Stability

In this section, we explore how interactions within the game of Diplomacy define the relationships between players. While such dyadic relationships can be undefined (e.g., England and Turkey are in opposite corners of the map), specific interactions between players indicate whether they are friendly or hostile to each other.

Event	Time	What happened
F_1	4	B supports V's army in Vienna
F_2	3	V supports B's attack from Warsaw to Silesia
F_3	3	B again supports V in Vienna
F_4	1	V supports B's move from Venice to Tyrolia
H_5	0	B attacks V in Vienna
H_6	-1	V retaliates, attacking B in Warsaw

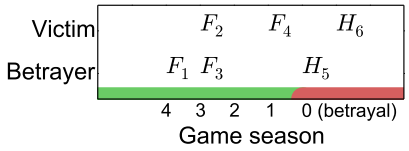


Figure 2: A friendship between Player B (eventual betrayer) and Player V (eventual victim) unravels. For the first four events, the players exchange **F**riendly acts (in green). Eventually B's unilateral hostile act betrays V's trust, leading to hostility (in red). The dissolution takes place at the time of the first hostile act ($t = 0$) and we index game seasons going back from the betrayal, such that lower indices mean betrayal is nearer.

Friendships and hostilities. Alliances are a natural part of the game of Diplomacy. While the best outcome for a player is a *solo victory* against all other players, this is rare and difficult to achieve without any cooperation and assistance. Instead, the game's structure encourages players to form long-term alliances. Allies often settle for (less prestigious) team victories, but these coalitions can also crumble as players seek a (more prestigious) solo victory for themselves. This game dynamic naturally leads to the formation of *friendly and hostile dyads*, which are relatively easy to identify through post-hoc analysis of the game, as explained next.

Acts of friendship. Diplomacy provides a *support* option for players to help each other: this game mechanism (discussed in Section 2) provides unequivocal evidence of friendship. When two players engage in a series of such friendly acts, we will say that the two are in a relation of *friendship*.

Acts of hostility. Unlike support, hostile actions are not explicitly marked in Diplomacy. We consider two players to be hostile if they get involved in any unambiguous belligerent action, such as invading one another's territory, or if one supports an enemy of the other.[4]

[3] Anonymized transcripts and more information available at `http://vene.ro/betrayal/`

[4] In Diplomacy all game actions are simultaneous, and this can lead to ambiguous interpretation of the nature of a

Betrayal. As in real life, friendships can be broken unilaterally: an individual can *betray* his friend by engaging in a hostile act towards her. Figure 2 shows two players who started out as friends (green) but became hostile (red) after a betrayal. Importantly, until the last act of friendship (game season $t = 1$), the *victim* is unaware that she will be betrayed (otherwise she would not have engaged in an act of friendship) and the *betrayer* has no interest in signaling his planned duplicity to his partner.

This setting poses the following research challenge: are there linguistic cues that appear during the friendly conversations and portend the upcoming betrayal? A positive answer would have two implications: the betrayer unknowingly hints at his future treachery, and the victim could have noticed it, but did not. We will explore this question in the following sections.

Relationship stability. Before venturing into the linguistic analysis of betrayals, we briefly explore the dynamics underlying these state transitions. We find that, as in real life, friendships are much more likely to collapse into hostilities than the reverse: in Diplomacy, the probability of a friendship to dissolve into enmity is about five times greater than that of hostile players becoming friends. The history of the relationship also matters. A friendship built on the foundation of many cooperative acts is more likely to endure than friendship with a short history, and long-lasting conflict is less likely to become a friendship. In numbers, the probability that a two season long friendship ends is 35%, while for pairs who have helped each other for ten or more seasons, the probability of betrayal is only 23%. Similarly, the probability that a two season long conflict resolves is 7%, while players at war for over ten seasons have only a 5% chance to make up. These numbers aren't particularly shocking—the idea that the passage of time has an effect on the strength of a relationship is intuitive. For the purposes of this study, we control for such effects in order to capture purely linguistic hints of betrayal.

Starting from the relationship definitions discussed in this section, in what follows we show how subtle linguistic patterns of in-game player conversations can reveal whether or not a friendship will turn hostile or not.

pair's interactions. Our definition of hostility intentionally discards such ambiguous evidence. For instance, if two players attempt to move into the same unoccupied territory, this is not necessarily aggressive: allies sometimes use this tactic ("bouncing") to ensure that a territory remains unoccupied.

4 Language Foretelling Betrayal

In this section, we examine whether the conversations between two Diplomacy allies contain linguistic cues foretelling if their friendship will last or end in betrayal. We expect these cues to be subtle, since we only consider messages exchanged when the two individuals are being ostensibly friendly; when at least one of them—the eventual victim—is unaware of the relationship's fate.

4.1 What Constitutes a Betrayal

To find betrayals, we must first find friendships. Building on the discussion from Section 3, we consider a friendship to be *stable* if it is ongoing, established, and reciprocal. Thus, we focus on relationships that contain at least two consecutive and reciprocated acts of friendships that span at last at least three seasons in game time. We also check that no more than five seasons pass between two acts of friendships, as friendships can fade.

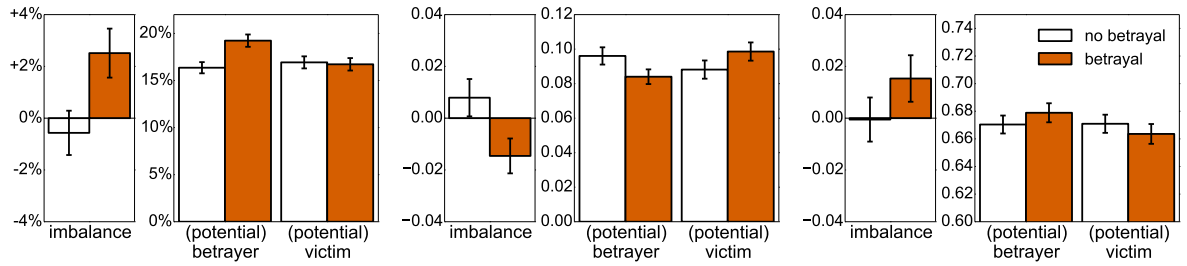
Betrayals are established and reciprocal friendships that end with at least two hostile acts. The person initiating the first of these hostile acts is the *betrayer*, while the other person is the *victim*.[5]

For each betrayal instance, we find the most similar stable friendship that was *never* dissolved by betrayal. Using a greedy heuristic, we select friendships that match the betrayals on two statistics: the length of the friendship and number of seasons since the start of the game. After this matching process, we find no significant difference in either of the two variables (Mann-Whitney $p > 0.3$). Matching betrayals with lasting friendships in this fashion removes historical and relationship-type effects such as those discussed in Section 3, and focuses the comparison on the variable of interest: whether a given stable friendship will end in a betrayal or not.

4.2 Linguistic Harbingers of Betrayal

Now we switch to exploring linguistic features that correlate with future betrayal in the controlled setting described above. We start from the intuition that a stable relationship should be balanced (Jung et al., 2012): friends will help each other

[5] In rare cases, the betrayal can be mutual (i.e., both players start attacking each other in the same season). In such cases, we consider both betrayals.



(a) Positive sentiment (percentage of sentences)

(b) Planning discourse markers (avg. number per sentence)

(c) Politeness (avg. message score)

Figure 3: Friendships that will end in betrayal are imbalanced. The eventual betrayer is more positive, more polite, but plans less than the victim. The white bars correspond to matched lasting friendships, where the roles of potential betrayer and victim are arbitrarily assigned; in these cases, the imbalances disappear. Error bars mark bootstrapped standard errors (Efron, 1979).

while enemies will fight each other. A precarious friendship might feel one-sided, while a conflict may turn to friendship through a magnanimous olive branch. Therefore, we focus our attention on linguistic features that have the potential to signal an imbalance in the communication patterns of the dyad.

To ensure that we are studying conversational patterns that occur *only* when the two individuals in the dyad are ostensibly being friends, we only extract features from the messages exchanged before the last act of friendship, that is, before the season labeled 1 in Figure 2. Considering the nature of this setting, we can only hope for subtle linguistic cues: if there were salient linguistic signals, then the victim would notice and preempt the betrayal. Instead, they are taken by surprise; the following is a typical reaction of a player after having been betrayed by a friend:

> Well that move was sour. I'm guessing France put you up to it, citing my large growth. This was a pity, as I was willing to give you the lion's share of centers in the west. [...] If you voiced your concerns I would have supported you in most of the western centers. Unfortunately now you have jumped out of the pan into the fire.

Sentiment. Changes in the sentiment expressed in conversation can reflect emotional responses, social affect, as well as the status of the relationship as a whole (Gottman and Levenson, 2000; Wang and Cardie, 2014). We quantify the proportion of exchanged sentences that transmit positive, neutral and negative sentiment using the Stanford Sentiment Analyzer (Socher et al., 2013).[6] Example sentences with these features, as well as all other features we consider, can be found in Table 1.

We find that an imbalance in the amount of positive sentiment expressed by the two individuals is a subtle sign that the relation will end in betrayal (Figure 3a, left; one-sample t-test on the imbalance, $p = 0.008$). When looking closer at who is the source of this imbalance (Figure 3a, right), we find that that it is the eventual betrayer that uses significantly *more positive sentiment* than the control counterpart in the matched friendship (two-sample t-test, $p = 0.001$). This is somewhat surprising, and we speculate that this is the betrayer overcompensating for his forthcoming actions.

Argumentation and Discourse. Structured discourse and well-made arguments are essential in persuasion (Cialdini, 2000; Anand et al., 2011). To capture discourse complexity, we measure the average number of explicit discourse connectors per sentence (Prasad et al., 2008).[7] These markers belong to four coarse classes: *comparison*, *contingency*, *expansive*, and *temporal*. To capture *planning*, we group temporal markers that refer to the future (e.g.,"next", "thereafter") in a separate category. To quantify the level of argumentation, we calculate average number of claim and premise markers per sentence, as identified by Stab and Gurevych (2014). We also measure the number of request sentences in each message, as identified by the heuristics in the Stanford Politeness classifier (Danescu-Niculescu-Mizil et al., 2013).

The structure of the discourse offers clues to whether the friendship will last. For example, Figure 3b shows that in friendships doomed to end in betrayal, the victim uses planning discourse markers significantly more often than the betrayer (one-sample t-test on the imbalance, $p = 0.03$), who is

[6] We collapse the few examples classified as *extreme positive* and *extreme negative* examples into *positive* and *negative*, respectively.

[7] We remove the connectors that appear in over 20% of the messages (*and*, *for*, *but*, *if*, *as*, *or*, and *so*).

Feature	Example sentence from the data
Positive sentiment	I will still be thrilled if it turns out you win this war.
Negative sentiment	It's not a great outcome, but still an OK one.
Neutral sentiment	Do you concur with my assumption?
Claim	But **I believe** that E/F have discarded him and so **I think** he might bite.
Premise	I put Italy out **because** I wanted to work with you.
Comparison	We can trade centers **as much as** we like **after** that.
Contingency	He did not, **thus** we are **indeed** in fine shape to continue as planned.
Expansion	Would you **rather** see WAR-UKR, or GAL-UKR?
Temporal	I think he can **still** be effective to help me take TUN **while** you take ROM.
Planning	HOL should fall **next** year, and **then** MUN and KIE shortly **thereafter**.
Number of requests	
Politeness	I wonder if you shouldn't try to support Italy into MAR ... What do you think?
Subjectivity	I'm **just curious** what you **think**.
Talkativeness	

Table 1: Summary of the linguistic cues we consider.

likely to be aware that the cooperation has no future. (More argumentation and discourse features will be discussed in the following sections.)

Politeness. Pragmatic information can also be informative of the relation between two individuals; for example Danescu-Niculescu-Mizil et al. (2013) show that differences in levels of politeness can echo differences in status and power. We measure the politeness of each message using the Stanford Politeness classifier and find that friendships that end in betrayal show a slight imbalance between the level of politeness used by the two individuals (one-sample t-test on the imbalance, $p = 0.09$) and that in those cases the future victim is the one that is less polite.

Subjectivity. We explored phrases expressing opinion, accusation, suspicion, and speculation taken from an automatically collected lexicon (Riloff and Wiebe, 2003), but did not find significant differences between betrayals and control friendships.

Talkativeness. Another conversational aspect is the amount of communication flowing between the players, in each direction. To quantify this, we simply use the number of messages sent, the average number of sentences per message, and the average number of words per sentence. Abnormal communication patterns can indicate a relationship breakdown. For example, friendships that dissolve are characterized by an imbalance in the number of messages exchanged between the two players (one-sample t-test, $p < 0.001$).

These results show that there are indeed subtle linguistic imbalance signals that are indicative of an forthcoming betrayal, even in a setting in which the victim is not aware of the impending betrayal.

4.3 Predictive Power

To test whether these linguistic cues have any predictive power and to explore how they interact, we turn to a binary classification setting in which we try to detect whether a player V will be betrayed by a player B. (We will call player V the potential victim and player B the potential betrayer.) Expert humans—the actual victims—performed poorly on this task and were not able to tell that they will be betrayed: by virtue of how the dataset is constructed, the performance of the human players is at chance level.

We use the same balanced dataset of matched betrayals and lasting friendships as before and consider as classification instances all the seasons coming from each of the two classes (663 betrayal seasons and 712 from lasting friendships). As features, we use the cues described above and summarized in Table 1, differentiated by source: V or B. We use logistic regression after univariate feature selection. The best setting for the model parameters[8] is selected via 5-fold cross validation, ensuring that instances from the same game are never found in both train and validation folds. The resulting model achieves a cross-validation accuracy of 57% and a Matthews correlation coefficient of 0.14, significantly above chance (52% accuracy and 0 Matthews correlation coefficient), with 95% bootstrapped confidence. This indicates

[8] We optimize the number of features selected, the scoring function used (ANOVA or χ^2), whether to automatically reweigh the classes, the regularizer (ℓ_1 or ℓ_2), and the value of the regularization parameter C between 10^{-12} and 10^{12}.

From	Positive feature	From	Negative feature
B	Positive sentiment	B	Expansion
B	Sentences	B	Comparison
		B	Contingency
		B	No. Words
		B	Planning
		B	Negative sentiment

Table 2: Selected features for recognizing upcoming betrayal, in decreasing order of the absolute value of their coefficients. The *From* column indicates whether the message containing the feature was sent by the potential **B**etrayer or the potential **V**ictim. (In this case, only betrayer features were selected.) Positive features indicate that a friendship is more likely to end in betrayal.

that, unlike the actual players, the classifier is able to exploit subtle linguistic signals that surface in the conversation.[9]

The selected features and their coefficients are reported in Table 2. On top of the observations we previously made, the feature ranking reveals that writing more sentences per message is more common when one will betray. Discourse features also prove relevant: more complex discourse indicates a lower likelihood of the player betraying (e.g., Figure 3b).

Overall, the selected linguistic features capture a consistent signal that characterizes people's language when they are about to betray: they tend to plan less than their victims, use less structure in their communication, and are overly positive.

5 Sudden yet Inevitable Betrayal

The results from Section 4 suggest that language cues can be subtle signs of future relationship disruption. Even though people are aware that most relationships eventually end, one would still prefer to reap their benefits as long as possible. In Diplomacy, despite the common knowledge that everyone prefers to win alone, players still take chances on long-lasting alliances. This leads to an alternate research question: assuming that a relationship will be disrupted, how soon can one expect to be betrayed? This is still just as challenging for the expert human players, as they were not able to anticipate and thereby avoid betrayal.

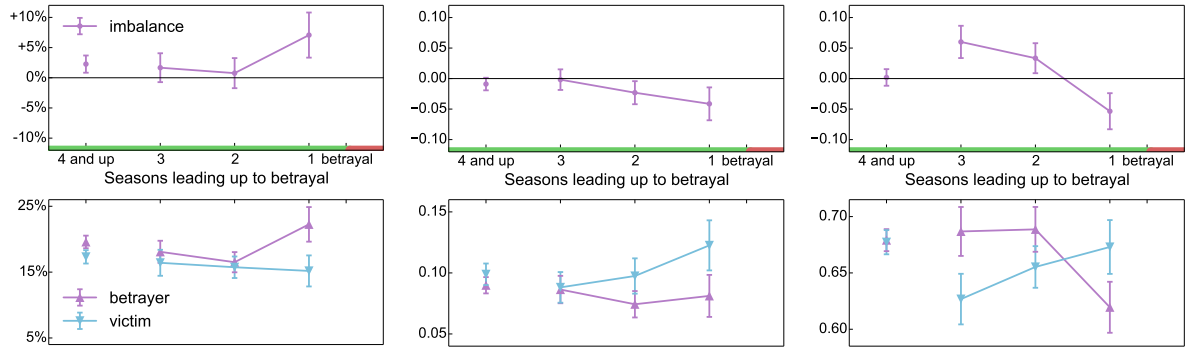
Next we investigate if the variation of the linguistic cues over time can predict imminent change in the relationship. We consider only the subset of betrayals used in Section 4, and label each individual game season with its distance from the end of the friendship (as in Figure 2). We prevent short alliances of circumstance from distorting the features close to betrayal by keeping only friendships lasting at least four seasons.

We consider the same cues described in Table 1, and train a classifier to discriminate between the season preceding the last friendly interaction and all the older seasons. This learning task is imbalanced, with only 14% of the seasons being immediately before the betrayal. Thus, we optimize F_1 score and also measure the Matthews correlation coefficient, which takes a value of 0 for uninformative predictions (random or majority). The best model achieves an F_1 score of 0.31 and a Matthews correlation coefficient of 0.17, significantly better than chance with 95% bootstrapped confidence. This shows that we can capture signs of imminent betrayal, something that even the skilled human players have failed to do. Furthermore, 39% of the predicted false positives are within two seasons of the last friendly act. This suggests that sometimes the warning signs can appear slightly earlier.

The selected features, displayed in Table 3, reflect some of the effects identified in Section 4, such as the importance of positive sentiment and planning discourse markers. Betrayers have a tendency to use more positive sentiment during the last moment of purported friendliness (Figure 4a). Also, expressing more opinions through claims is a sign that one will not betray right away. Three of the discourse features (comparison, contingency and expansion) are selected as *imbalance* features (they have near-opposite coefficients for the betrayer and for the victim), indicating that as betrayal approaches, victims are less eloquent than betrayers. Interestingly, some predictive signals come only from the victim: a partner using increasingly more planning words is at higher risk of being betrayed (Figure 4b). This could be explained by the pressure that making plans for the future can put on a relationship. A similar reasoning applies for making many requests.

We also find that a decrease in a partner's politeness presages their imminent betrayal. The change in politeness over time (Figure 4c) reveals a reversal in the politeness imbalance of the pair. This explains why politeness is not a good enough feature in detecting long-term betrayal. The behav-

[9]Since our focus is on understanding linguistic aspects of betrayal, rather than on achieving the best possible performance on this particular Diplomacy task, we do not use game-specific information, such as the players' position on the map, or any information not accessible to both players.



(a) Positive sentiment (percentage of sentences)

(b) Planning discourse markers (avg. number per sentence)

(c) Politeness (avg. message score)

Figure 4: Changes in balance can mark imminent betrayal. As the breakdown approaches, the betrayer becomes more positive but less polite, and the victim tends to make more requests and become more polite. Error bars mark bootstrapped standard errors (Efron, 1979).

From	Positive feature	From	Negative feature
V	Comparison	B	Claims
V	Positive sentiment	B	Politeness
V	Contingency	B	Contingency
V	Planning	B	Subjectivity
V	Requests	B	Expansion
V	Expansion	B	No. Sentences
		B	Comparison

Table 3: Selected features for recognizing imminent betrayal, in decreasing order of the absolute value of their coefficients. The *From* column indicates whether the message containing the feature comes from the potential **B**etrayer or the potential **V**ictim. Positive features indicate that an exchange is more likely to be followed by immediate betrayal.

ior could have two intuitive explanations. On one hand, if the betrayer has planned the act in advance, politeness can be a strategy for deception. On the other hand, if the betrayer receives impolite requests, the value of the relationship can decrease, hastening a betrayal. We observe a similar dynamic for the average number of sentences per message sent by the betrayer; the feature is selected in both prediction tasks, but with opposite signs: more complex messages suggest that betrayal *will* happen, but *not right away*.

Studying language change as betrayal draws nearer uncovers effects that cannot be seen when looking at an entire friendship on average. For example, while excessively positive and polite partners are potential betrayers, people who have themselves suddenly become more polite are likely to become victims soon.

6 Relevance Beyond the Game

While discovering betrayal in one online game is a fun and novel task, our work connects with broader research in computational social science. In this section we describe how our work tackles issues that previous research on alliances, negotiation, and relationships have faced.

Cooperation and relationship building are an essential part of many activities: completing a group project, opening a business, or forging a new relationship. Each of these has been the subject of extensive research to understand what makes for effective relationships. Jung et al. (2012) show that a balanced working relationship is more likely to lead to better performance on tasks like pair programming. Imai and Gelfand (2010) show that understanding cultural norms improves negotiations. While these data are elicited in the lab, our "found" data are inexpensive because Diplomacy games are fun and inherently anonymized.

Romance is a popular and more real-world phenomenon that helps us understand how relationships form and dissolve. The research that tells us how language shapes early dating (Ranganath et al., 2009) and whether an existing relationship will continue (Slatcher and Pennebaker, 2006; Gottman and Levenson, 2000; Ireland et al., 2011) is formed from an incomplete sample of a course of a relationship. In contrast, a game of Diplomacy is shorter than almost any marriage and we have a complete account of all interactions throughout the entire relationship. Furthermore, this work focuses on the unilateral and asymmetric act of betrayal, rather than on the question of whether a relation will last.

Playing Diplomacy online is less tangible than a romantic relationship, but understanding trust and deception in online interactions (Riegelsberger et

al., 2003; Newman et al., 2003; Hancock et al., 2007; Ott et al., 2011; Feng et al., 2012) is particularly important because the Internet marketplace is a growing driver of economic growth (Boyd, 2003). Diplomacy offers a setting in which deception occurs spontaneously in the context of complex relationships.

7 Conclusions

Despite people's best effort to hide it, the intention to betray can leak through the language one uses. Detecting it is not a task that we expect to be solvable with high accuracy, as that would entail a reliable "recipe" for avoiding betrayal in relationships; in this unrealistic scenario, betrayals would be unlikely to exist. While the effects we find are subtle, they bring new insights into the relation between linguistic balance and stability in relationships.

Although we use one game to develop our methodology, the framework developed here can be extended to be applied to a wide range of social interaction. Social dynamics in collaborative settings can bear striking similarities to those present in war games. For example, in Wikipedia "edit wars"—where attacks correspond to edit reverts—are common on issues relating to politics, religion, history and nationality, among others (Kittur et al., 2007). As in Diplomacy, Wikipedia editors form alliances, argue and negotiate about possible compromises. A challenge for future work is to find reliable linguistic cues that generalize well between such settings.

Acknowledgements

This work is dedicated to all those who betrayed us. We thank Mario Huys, Chris Babcock, and Christopher Martin for providing the Diplomacy dataset. We are grateful to Flavio Chierichetti, Malte Jung, Sendhil Mullainathan and the annonymous reviewers for their helpful comments. This work was conducted in part while Cristian Danescu-Niculescu-Mizil and Vlad Niculae were at the Max Planck Institute for Software Systems. Jordan Boyd-Graber is supported by NSF Grants CCF-1409287, IIS-1320538, and NCSE-1422492. Cristian Danescu-Niculescu-Mizil is supported by a Google Faculty Research Award. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Douglas W. Oard, and Philip Resnik. 2011. Believe me: We can do this! In *Proceedings of the AAAI 2011 Workshop on Computational Models of Natural Argument*.

Robert Axelrod and Douglas Dion. 1988. The further evolution of cooperation. *Science*, 242(4884):1385–1390.

Josh Boyd. 2003. The rhetorical construction of trust online. *Communication Theory*, 13(4):392–410.

Berno Buechel, Eike Emrich, and Stefanie Pohlkamp. 2013. Nobody's innocent: The role of customers in the doping dilemma. MPRA paper, University Library of Munich, Germany.

Robert B. Cialdini. 2000. *Influence: Science and Practice (4th Edition)*. Allyn & Bacon.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the Association for Computational Linguistics*.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, (1979):1–26.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the Association for Computational Linguistics*.

John M. Gottman and Robert W. Levenson. 2000. The timing of divorce: Predicting when a couple will divorce over a 14-year period. *Journal of Marriage and Family*, 62(3):737–745.

Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.

Lynn Imai and Michele J. Gelfand. 2010. The culturally intelligent negotiator: The impact of cultural intelligence (CQ) on negotiation sequences and outcomes. *Organizational Behavior and Human Decision Processes*, 112(2):83–98.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Malte Jung, Jan Chong, and Larry Leifer. 2012. Group hedonic balance and pair programming performance: Affective interaction dynamics as indicators of performance. In *International Conference on Human Factors in Computing Systems*.

Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: Conflict and coordination in Wikipedia. In *International Conference on Human Factors in Computing Systems*.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the Association for Computational Linguistics*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The Penn Discourse TreeBank 2.0. In *International Language Resources and Evaluation*.

Rajesh Ranganath, Dan Jurafsky, and Dan McFarland. 2009. It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *Proceedings of Emperical Methods in Natural Language Processing*.

Jens Riegelsberger, M. Angela Sasse, and John D. McCarthy. 2003. The researcher's dilemma: Evaluating trust in computer-mediated communication. *International Journal of Human-Computer Studies*, 58(6):759–781.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of Emperical Methods in Natural Language Processing*.

Robert W. Rosenthal. 1981. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory*, 25(1):92–100.

Richard Sharp. 1978. *The Game of Diplomacy*. Arthur Barker Publishing.

Richard B. Slatcher and James W. Pennebaker. 2006. How do I love thee? Let me count the words: The social effects of expressive writing. *Psychological Science*, 17(8):660–664.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of Emperical Methods in Natural Language Processing*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of Emperical Methods in Natural Language Processing*.

Lu Wang and Claire Cardie. 2014. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of the Association for Computational Linguistics*.