Linguistic Harbingers of Betrayal

me, Vlad Niculae, Cornell University
with Srijan Kumar, University of Maryland College Park
Jordan Boyd-Graber, University of Colorado Boulder
and Cristian Danescu-Niculescu-Mizil, Cornell University
Linguistic Harbingers of Betrayal is everywhere.
Are there any 

*linguistic* cues that
foretell betrayal?
What is betrayal?

(And how does it differ from just lying?)
Can this be a betrayal?
Can this be a betrayal?

Deceptive review spam
(Li, Huang, Yang & Zhu, 2011)
(Ott, Choi, Cardie & Hancock, 2011)
(Feng, Banerjee & Choi, 2012)
...
Can this be a betrayal?

Deceptive review spam
(Li, Huang, Yang & Zhu, 2011)
(Ott, Choi, Cardie & Hancock, 2011)
(Feng, Banerjee & Choi, 2012)

Deception in court cases
(Bachenko, Fitzpatrick & Schonwetter, 2008)
(Fornaciari & Poesio, 2013)
Can this be a betrayal?

Deceptive review spam
(Li, Huang, Yang & Zhu, 2011)
(Ott, Choi, Cardie & Hancock, 2011)
(Feng, Banerjee & Choi, 2012)

Deception in court cases
(Bachenko, Fitzpatrick & Schonwetter, 2008)
(Fornaciari & Poesio, 2013)

Elicited deception in essays
(Newman, Pennebaker, Berry & Richards, 2003)
(Mihalcea & Strapparava, 2009)
(Pérez-Rosas & Mihalcea, 2014)
Okay, how about this, can this be a betrayal?

(Feldman and Happ, 2002)
(Hancock, Curry, Goorha & Woodworth, 2011)

...
The exciting game of international intrigue

“The game that ruins friendships”
The exciting game of international intrigue

“The game that ruins friendships”
"The game that ruins friendships"
The game that ruins friendships

Diplomacy

online!
The game that ruins friendships

online!

249 games
~6 months/game
145k messages

diplom.org; usak.asciiking.com
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.
NOW STAND BACK,
I GOTTA PRACTICE MY STABBIN'
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

... stabs !
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

... stabs !
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

Not really sure what to say, except that I regret you did what you did.
Curse your sudden but inevitable betrayal!
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

Not really sure what to say, except that I regret you did what you did.
Identifying Betrayals

(So That We Can Analyze Their Language)
Identifying Betrayals

Solid friendships:
- established (≥ two supports),
- mutual (≥ one from each).

Betrayal:
- friendship followed by
  ≥ two attacks.
Identifying Betrayals

250 such betrayals in our Diplomacy dataset.

Solid friendships:

- established (≥ two supports),
- mutual (≥ one from each).

Betrayal:

- friendship followed by ≥ two attacks.
Matching Friendship

250 such betrayals in our Diplomacy dataset. We find 250 matching friendships.
(9660 msgs., 59 words/msg.)

Solid friendships:
- established (≥ two supports),
- mutual (≥ one from each).

Betrayal.

Matched for:
- length,
- game year.
Matching Friendship

250 such betrayals in our Diplomacy dataset. We find 250 matching friendships. (9660 msgs., 59 words/msg.)

Solid friendships:
- established (≥ two supports),
- mutual (≥ one from each).

Matched for:
- length, game year.
- no attacks (ever).

Betrayal.
Matching Friendship

250 such betrayals in our Diplomacy dataset. We find 250 matching friendships.
(9660 msgs., 59 words/msg.)
Matching Friendship

250 such betrayals in our Diplomacy dataset. We find 250 matching friendships. (9660 msgs., 59 words/msg.)

Linguistic signs of betrayal while they act as friends?
Matching Friendship

250 such betrayals in our Diplomacy dataset. We find 250 matching friendships. (9660 msgs., 59 words/msg.)

Linguistic signs of betrayal while they act as friends?

The betrayers actively hide it. The victims didn’t see it coming.
Insight: Conversational Balance
Insight: Conversational Balance
Insight: Conversational Balance

- Stable marriages are balanced (Gottman, 1993).
- So are effective pair programming teams (Jung, Chong & Leifer, 2012).
Insight: Conversational Balance

• Stable marriages are balanced (Gottman, 1993).

• So are effective pair programming teams (Jung, Chong & Leifer, 2012).

• Can we apply this to linguistic conversational features?
**(Im)balance: Sentiment**

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

(Proportion of sentences showing positive sentiment.)
(Error bars show standard error.)
(Im)balance: Sentiment

Imbalance: \( f(\text{betrayer}) - f(\text{victim}) \)

(Proportion of sentences showing positive sentiment.)
(Error bars show standard error.)
(Im)balance: Sentiment

Imbalance: \( f(\text{betrayer}) - f(\text{victim}) \)

(Proportion of sentences showing positive sentiment.)
(Error bars show standard error.)
(Im)balance: Politeness

Imbalance: f(betrayal) - f(victim)

(Average 0-1 politeness score of requests: http://politeness.mpi-sws.org)
(Error bars show standard error.)
(Im)balance: Future Planning

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

(Average number of planning connectors per message, e.g. “next”, “after”)

(Error bars show standard error.)
(Im)balance: Future Planning

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

Demand-Withdraw pattern pre-divorce. (Gottman & Levenson, 2000)

(Average number of planning connectors per message, e.g. “next”, “after”)
(Error bars show standard error.)
Conversational (Im)balance

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

Friendships that break exhibit imbalance through language cues.

(Error bars show standard error.)
Are backstabbing friendships doomed from the start?
Are backstabbing friendships doomed from the start?

Or do the dynamics change over time?
(Im)balance Over Time

Imbalance: \( f(\text{betrayer}) - f(\text{victim}) \). Looking only at betrayals.

(Error bars show standard error.)
(Im)balance Over Time

Imbalance: \( f(\text{betrayer}) - f(\text{victim}) \)

Positive sentiment

Politeness

Future planning

(time until betrayal)

(Error bars show standard error.)
(Im)balance Over Time

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

- Positive sentiment
- Politeness
- Future planning

(Error bars show standard error.)
(Im)balance Over Time

Imbalance: \( f(\text{betrayer}) - f(\text{victim}) \)

Positive sentiment

Future planning

(Error bars show standard error.)
(Im)balance Over Time

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

Positive sentiment

Imbalance:

Politeness

Future planning

(Error bars show standard error.)
(Im)balance Over Time

Imbalance: \( f(\text{betrayer}) - f(\text{victim}) \)

positive sentiment

politeness

future planning

(time until betrayal)

(amount)

(Error bars show standard error.)
(Im)balance Over Time

Imbalance: \( f(\text{betrayer}) - f(\text{victim}) \)

Positive sentiment

Politeness

Future planning

Imbalance:
\[ f(\text{betrayer}) - f(\text{victim}) \]

(Error bars show standard error.)
(Im)balance Over Time

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

- Positive sentiment
- Politeness
- Future planning

(Error bars show standard error.)
(Im)balance Over Time

**Imbalance**: \( f(\text{betrayer}) - f(\text{victim}) \)

**positive sentiment**

---

**politeness**

---

**future planning**

---

(Error bars show standard error.)

**Demand-Withdraw pattern pre-divorce.**

(Gottman & Levenson, 2000)
**Demand-Withdraw pattern pre-divorce.**

\[
\text{Imbalance: } f(\text{betrayer}) - f(\text{victim})
\]

(Error bars show standard error.)
(Im)balance Over Time

Imbalance: $f(\text{betrayed}) - f(\text{victim})$

Demand-Withdraw pattern pre-divorce.
(Gottman & Levenson, 2000)

Positive sentiment

Politeness

Future planning

(Error bars show standard error.)
As betrayal draws nearer, balance is broken.

Attributes change at different rates.

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

Error bars show standard error.
As betrayal draws nearer, balance is broken. Attributes change at different rates. Are these cues predictive?

Imbalance: $f(\text{betrayer}) - f(\text{victim})$

*positive sentiment*

*politeness*

*future planning*

(Error bars show standard error.)
Predicting Betrayal

“Would it be ok with you if I took Denmark? I think I'm going to need it if I am going to hold France back.”

“Hi Germany, How about I give you back Denmark next year. This is because I probably won't get a centre this year and would rather not disband a unit.”

“I am supporting you into Sweden this turn if you want. If you want to be able to keep Sweden I suggest moving into Finland. Cheers, Harriet Jones, PM.”

“Thanks, I accept the support. I'll decide what I want to do with the army.”
Predicting Betrayal

“Would it be ok with you if I took Denmark? I think I'm going to need it if I am going to hold France back.”

“Hi Germany, How about I give you back Denmark next year. This is because I probably won't get a centre this year and would rather not disband a unit.”

“Germany, Well that move was sour. This was a pity. Unfortunately now you have jumped out of the pan into the fire.”
Predicting Betrayal

Toss in a few more features:

- **Sentiment**
  (Stanford Sentiment Analysis)

- **Argumentation & discourse**
  (Penn Discourse Treebank)
  (Stab & Gurevich, 2014)

- **Politeness**
  (http://politeness.mpi-sws.org)

- **Subjectivity**
  (Riloff & Wiebe, 2003)

- **Talkativeness**
Predicting Betrayal

Prediction tasks:

- Will this friendship break?
- Is betrayal imminent?
Predicting Betrayal

Prediction tasks:

- Will this friendship break?
  (1375 seasons, 48% betrayals)
  Accuracy: (players: 52%)
  MCC:* (players: 0)

- Is betrayal imminent?
  (663 seasons from betrayals, 14% immediately before betrayal)
  F₁: (players: 0)
  MCC:* (players: 0)

*Matthews Correlation Coefficient: 0 = uninformative, 1 = perfect correlation.
Predicting Betrayal

Prediction tasks:

• Will this friendship break?
  (1375 seasons, 48% betrayals)
  Accuracy: (players: 52%) 57%
  MCC:* (players: 0) 0.14

• Is betrayal imminent?
  (663 seasons from betrayals,
  14% immediately before betrayal)
  F₁: (players: 0)
  MCC:* (players: 0)

*Matthews Correlation Coefficient: 0 = uninformative, 1 = perfect correlation.
Predicting Betrayal

Prediction tasks:

- **Will this friendship break?**
  (1375 seasons, 48% betrayals)
  Accuracy: (players: 52%) **57%**
  MCC:* (players: 0) **0.14**

- **Is betrayal imminent?**
  (663 seasons from betrayals, 14% immediately before betrayal)
  \( F_1 \): (players: 0) **0.31**
  MCC:* (players: 0) **0.17**

*Matthews Correlation Coefficient: 0 = uninformative, 1 = perfect correlation.
Predicting Betrayal

Prediction tasks:

- Will this friendship break?
  (1375 seasons, 48% betrayals)
  Accuracy: (players: 52%) 57%
  MCC:* (players: 0) 0.14

- Is betrayal imminent?
  (663 seasons from betrayals, 14% immediately before betrayal)
  F$_1$: (players: 0) 0.31
  MCC:* (players: 0) 0.17

- Outperforming the players!

*Matthews Correlation Coefficient: 0 = uninformative, 1 = perfect correlation.
The intention to betray can leak through words.

Good friendships are balanced.

Imbalance changes as betrayal draws near.
extra slides
<table>
<thead>
<tr>
<th>Feature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive sentiment</td>
<td>I will still be trilled if you win this war.</td>
</tr>
<tr>
<td>Negative sentiment</td>
<td>It’s not a great outcome, but still an OK one.</td>
</tr>
<tr>
<td>Neutral sentiment</td>
<td>Do you concur with my assumption?</td>
</tr>
<tr>
<td>Claim</td>
<td>I believe that E/F have discarded him.</td>
</tr>
<tr>
<td>Premise</td>
<td>I put italy out because I wanted to work with you.</td>
</tr>
<tr>
<td>Comparison</td>
<td>We can trade centers as much as we like.</td>
</tr>
<tr>
<td>Contingency</td>
<td>He did not, thus we are indeed in fine shape.</td>
</tr>
<tr>
<td>Expansion</td>
<td>Would you rather see A or B?</td>
</tr>
<tr>
<td>Temporal</td>
<td>I think he can still be effective while you take ROM.</td>
</tr>
<tr>
<td>Planning</td>
<td>HOL should fall next year, and then MUN after.</td>
</tr>
<tr>
<td>Subjectivity</td>
<td>I’m just curious what you think.</td>
</tr>
<tr>
<td>Politeness</td>
<td>I wonder if you shouldn’t try to support Italy into MAR... What do you think?</td>
</tr>
</tbody>
</table>
## Selected Features

<table>
<thead>
<tr>
<th>Sender</th>
<th>Positive feature</th>
<th>Sender</th>
<th>Negative feature</th>
<th>Sender</th>
<th>Positive feature</th>
<th>Sender</th>
<th>Negative feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Positive sentiment</td>
<td>B</td>
<td>Expansion</td>
<td>V</td>
<td>Comparison</td>
<td>B</td>
<td>Claims</td>
</tr>
<tr>
<td>B</td>
<td>No. Sents</td>
<td>B</td>
<td>Comparison</td>
<td>V</td>
<td>Positive sentiment</td>
<td>B</td>
<td>Politeness</td>
</tr>
<tr>
<td>B</td>
<td>Contingency</td>
<td>V</td>
<td>Contingency</td>
<td>B</td>
<td>Contingency</td>
<td>B</td>
<td>Contingency</td>
</tr>
<tr>
<td>B</td>
<td>No. Words</td>
<td>V</td>
<td>Planning</td>
<td>B</td>
<td>Subjectivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Planning</td>
<td>V</td>
<td>Requests</td>
<td>B</td>
<td>Expansion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Negative sentiment</td>
<td>V</td>
<td>Expansion</td>
<td>B</td>
<td>No. Sentences</td>
<td>B</td>
<td>Comparison</td>
</tr>
</tbody>
</table>