Modeling Coreference in Contexts with Three Referents

Jet Hoek, Andrew Kehler & Hannah Rohde

RAILS, 25 October 2019
The puzzle

Donald called Rudy. . .
Models of coreference

Mirror Model (Ariel 1990; Gundel et al. 1993)

Expectancy Model (Arnold 2001)

Bayesian Model (Kehler et al. 2008; Kehler & Rohde 2013; Rohde & Kehler 2014)
Models of coreference

**Mirror Model** (Ariel 1990; Gundel et al. 1993)

\[ p(\text{referent} | \text{pronoun}) \sim p(\text{pronoun} | \text{referent}) \]
## Models of coreference

### Mirror Model
(Ariel 1990; Gundel et al. 1993)

\[ p(\text{referent} | \text{pronoun}) \sim p(\text{pronoun} | \text{referent}) \]

### Expectancy Model
(Arnold 2001)

\[ p(\text{referent} | \text{pronoun}) \sim p(\text{referent}) \]
## Models of coreference

<table>
<thead>
<tr>
<th>Model</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mirror Model</strong></td>
<td>(Ariel 1990; Gundel et al. 1993)</td>
</tr>
<tr>
<td>( p(\text{referent}</td>
<td>\text{pronoun}) \sim p(\text{pronoun}</td>
</tr>
<tr>
<td><strong>Expectancy Model</strong></td>
<td>(Arnold 2001)</td>
</tr>
<tr>
<td>( p(\text{referent}</td>
<td>\text{pronoun}) \sim p(\text{referent}) )</td>
</tr>
<tr>
<td><strong>Bayesian Model</strong></td>
<td>(Kehler et al. 2008; Kehler &amp; Rohde 2013; Rohde &amp; Kehler 2014)</td>
</tr>
<tr>
<td>( p(\text{referent}</td>
<td>\text{pronoun})<em>{\text{interpretation}} \sim p(\text{referent})</em>{\text{prior}} \ast p(\text{pronoun}</td>
</tr>
</tbody>
</table>
Interpretation does not equal production

Story continuation

John scolded Bob. He __________________________ [pronoun prompt]
John scolded Bob. ____________________________ [free prompt]
Interpretation does not equal production

Story continuation

John scolded Bob. He __________________ [pronoun prompt]
John scolded Bob. __________________ [free prompt]

The Bayesian model captures this asymmetry
In its **strong form**, the Bayesian model separates the discourse features that influence the prior and the likelihood:

- **meaning** drives the *prior*
- **topicality** drives the *likelihood*
Weak versus strong Bayes

In its **strong form**, the Bayesian model separates the discourse features that influence the prior and the likelihood:

- **meaning** drives the *prior*
- **topicality** drives the *likelihood*

→ Recent work that shows that the likelihood of pronominalization increases for referents with a higher prior (e.g., Rosa & Arnold 2017)
Weak versus strong Bayes

Bayesian Model

\[
p(\text{referent} | \text{pronoun})_{\text{interpretation}} \sim p(\text{referent})_{\text{prior}} \ast p(\text{pronoun} | \text{referent})_{\text{likelihood}}
\]

In its **strong form**, the Bayesian model separates the discourse features that influence the prior and the likelihood:

- **meaning** drives the *prior*
- **topicality** drives the *likelihood*

\[ \rightarrow \text{Recent work that shows that the likelihood of pronominalization increases for referents with a higher prior (e.g., Rosa & Arnold 2017)} \]

In its **weak form**, the Bayesian model states that **pronoun production and interpretation are related by Bayesian principles**.
Current study

- Most of the research on pronoun production / interpretation has focused on sentence frames with two referents.

- Results appear to differ between implicit causality verbs and studies with transfer-of-possession verbs (e.g., Rohde 2008; Fukumura & van Gompel 2010 versus Rosa & Arnold 2017)

In a new context type with three referents, we test:
1. whether predictability influences pronominalization
2. whether Bayes' Rule captures the relationship between pronoun interpretation and production
Current study

- Most of the research on pronoun production / interpretation has focused on sentence frames with two referents.

- Results appear to differ between implicit causality verbs and studies with transfer-of-possession verbs (e.g., Rohde 2008; Fukumura & van Gompel 2010 versus Rosa & Arnold 2017)

In a new context type with three referents, we test:

1. whether predictability influences pronominalization
2. whether Bayes’ Rule captures the relationship between pronoun interpretation and production
Story continuation experiment

Adam called Diana for Russel.
Story continuation experiment

Items

Adam called Diana for Russel. He _________________ [pronoun prompt]
Adam called Diana for Russel. _________________ [free prompt]

- Counterbalanced which referents were gender-matched
  (NP1&NP2, NP1&NP3, NP2&NP3)
Story continuation experiment

Items

Adam called Diana for Russel. He ___________________ [pronoun prompt]
Adam called Diana for Russel. ___________________ [free prompt]

- Counterbalanced which referents were gender-matched
  (NP1&NP2, NP1&NP3, NP2&NP3)
- 83 native speakers of English
- 30 items
Story continuation experiment

Items

Adam called Diana for Russel. He ____________________ [pronoun prompt]
Adam called Diana for Russel. ________________________ [free prompt]

- Counterbalanced which referents were gender-matched
  (NP1&NP2, NP1&NP3, NP2&NP3)

- 83 native speakers of English

- 30 items

- Continuations were coded for:
  - who the continuation is about
  - what form of referring expression is used (free prompt condition only)
Results: More subject continuations in pronoun prompt
Results: Subjects are preferentially pronominalized

Free prompt
Results 1: Does predictability influence pronominalization?
Results 1: Does predictability influence pronominalization?

Free prompt
Results 2: Does Bayes’ Rule rule?

Following Rohde & Kehler (2014), we used the free prompt continuations to calculate Bayes-derived estimates of $p(\text{referent}|\text{pronoun})$ via the prior $p(\text{referent})$ and likelihood $p(\text{pronoun}|\text{referent})$, as well as estimates for the Expectancy Model (prior) and the Mirror Model (normalized likelihood). We then compared the model estimates with the pronoun interpretations measured in the pronoun prompt condition.
Results 2: Does Bayes’ Rule rule?

Following Rohde & Kehler (2014), we used the free prompt continuations to calculate Bayes-derived estimates of $p(\text{referent}|\text{pronoun})$ via the prior $p(\text{referent})$ and likelihood $p(\text{pronoun}|\text{referent})$, as well as estimates for the Expectancy Model (prior) and the Mirror Model (normalized likelihood). We then compared the model estimates with the pronoun interpretations measured in the pronoun prompt condition.

**Items:**  
Bayes: $R^2 = .122$, Expectancy: $R^2 = .003$, **Mirror:** $R^2 = .377$

**Participants:**  
Bayes: $R^2 = .084$, Expectancy: $R^2 = .021$, Mirror: $R^2 = .075$
Interim discussion

- We do not find any evidence that pronominalization is affected by predictability
  → In line with strong Bayes
Interim discussion

- We do not find any evidence that pronominalization is affected by predictability
  → In line with strong Bayes

- The Bayesian model outperforms the Expectancy model

- The Bayesian model is outperformed by the Mirror model
Interim discussion

- We do not find any evidence that pronominalization is affected by predictability
  → In line with strong Bayes

- The Bayesian model outperforms the Expectancy model

- The Bayesian model is outperformed by the Mirror model

  → Is this due to the construction or does it have something to do with the number of referents?
Follow-up: 2-human Benefactive prompts

<table>
<thead>
<tr>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam called the hospital for Russel. He ____________ [pronoun prompt]</td>
</tr>
<tr>
<td>Adam called the hospital for Russel. ________________ [free prompt]</td>
</tr>
</tbody>
</table>
Follow-up: 2-human Benefactive prompts

Items

Adam called the hospital for Russel. He ____________ [pronoun prompt]
Adam called the hospital for Russel. ____________ [free prompt]
Follow-up: 2-human Benefactive prompts

Participants:
- Bayes: $R^2 = 0.348$
- Expectancy: $R^2 = 0.008$
- Mirror: $R^2 = 0.282$

Items:
- Bayes: $R^2 = 0.719$
- Expectancy: $R^2 = 0.311$
- Mirror: $R^2 = 0.714$
Follow-up: 2-human Benefactive prompts

**Items:**
- Bayes: $R^2 = .719$
- Expectancy: $R^2 = .311$
- Mirror: $R^2 = .714$

**Participants:**
- Bayes: $R^2 = .348$
- Expectancy: $R^2 = .008$
- Mirror: $R^2 = .282$
The models’ poor fit for the observed pronoun interpretation data in our first experiment appears to be due to the number of referents in the experiment with 2-human Benefactive prompts, Bayes is back. But why? Power issue? But no fewer observations per ambiguous pair than earlier work with 2 referents. 3 referents make the task harder? But is it really? In which way? And why would this matter?
Discussion

- The models’ poor fit for the observed pronoun interpretation data in our first experiment appears to be due to the number of referents.
- In the experiment with 2-human Benefactive prompts, Bayes is back.
Discussion

- The models’ poor fit for the observed pronoun interpretation data in our first experiment appears to be due to the number of referents
- In the experiment with 2-human Benefactive prompts, Bayes is back

But why?
Discussion

- The models’ poor fit for the observed pronoun interpretation data in our first experiment appears to be due to the number of referents.
- In the experiment with 2-human Benefactive prompts, Bayes is back.

**But why?**

- Power issue?
Discussion

- The models’ poor fit for the observed pronoun interpretation data in our first experiment appears to be due to the number of referents.
- In the experiment with 2-human Benefactive prompts, Bayes is back.
  
  But why?

- Power issue?
  - But no fewer observations per ambiguous pair than earlier work with 2 referents.
Discussion

- The models’ poor fit for the observed pronoun interpretation data in our first experiment appears to be due to the number of referents.

- In the experiment with 2-human Benefactive prompts, Bayes is back.

But why?

- Power issue?
  - But no fewer observations per ambiguous pair than earlier work with 2 referents.

- 3 referents make the task harder?
Discussion

- The models’ poor fit for the observed pronoun interpretation data in our first experiment appears to be due to the number of referents.
- In the experiment with 2-human Benefactive prompts, Bayes is back.

But why?

- Power issue?
  - But no fewer observations per ambiguous pair than earlier work with 2 referents.
- 3 referents make the task harder?
  - But is it really? In which way? And why would this matter?
Thank you!

jhoek@uni-koeln.de