# Modeling Coreference in Contexts with Three Referents

Jet Hoek, Andrew Kehler & Hannah Rohde

RAILS, 25 October 2019

LEVERHULME TRUST

# The puzzle

Donald called Rudy. . . .

# Models of coreference

# Models of coreference

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

$$\mathbf{p}(\mathbf{referent}|\mathbf{pronoun}) \sim \mathbf{p}(\mathbf{pronoun}|\mathbf{referent})$$

# Models of coreference

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

$$\mathbf{p}(\mathbf{referent}|\mathbf{pronoun}) \sim \mathbf{p}(\mathbf{pronoun}|\mathbf{referent})$$

**Expectancy Model** (Arnold 2001)

$$\mathbf{p}(\mathbf{referent}|\mathbf{pronoun}) \sim \mathbf{p}(\mathbf{referent})$$

# Models of coreference

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

$$\mathbf{p}(\mathbf{referent}|\mathbf{pronoun}) \sim \mathbf{p}(\mathbf{pronoun}|\mathbf{referent})$$

**Expectancy Model** (Arnold 2001)

$$\mathbf{p}(\mathbf{referent}|\mathbf{pronoun}) \sim \mathbf{p}(\mathbf{referent})$$

**Bayesian Model** (Kehler et al. 2008; Kehler & Rohde 2013; Rohde & Kehler 2014)

$$\mathbf{p}(\mathbf{referent}|\mathbf{pronoun})_{\mathbf{interpretation}} \sim \mathbf{p}(\mathbf{referent})_{\mathbf{prior}} * \mathbf{p}(\mathbf{pronoun}|\mathbf{referent})_{\mathbf{likelihood}}$$

# 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**

# Weak versus strong Bayes

## Bayesian Model

$$\textbf{p}(\textbf{referent}|\textbf{pronoun})_{\textbf{interpretation}} \sim \textbf{p}(\textbf{referent})_{\textbf{prior}} * \textbf{p}(\textbf{pronoun}|\textbf{referent})_{\textbf{likelihood}}$$

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*

# Weak versus strong Bayes

**Bayesian Model**

$$\mathbf{p}(\mathbf{referent}|\mathbf{pronoun})_{\mathbf{interpretation}} \sim \mathbf{p}(\mathbf{referent})_{\mathbf{prior}} * \mathbf{p}(\mathbf{pronoun}|\mathbf{referent})_{\mathbf{likelihood}}$$

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*

→ 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**

$$\mathbf{p(referent|pronoun)_{interpretation}} \sim \mathbf{p(referent)_{prior}} * \mathbf{p(pronoun|referent)_{likelihood}}$$

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*

→ 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)

# 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

# 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

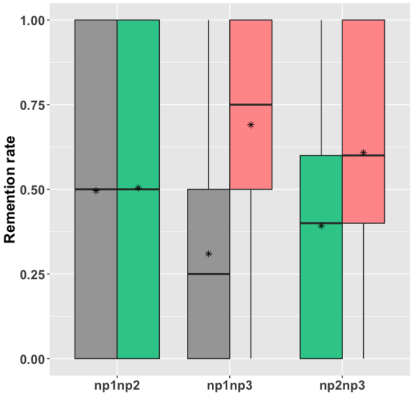
# Story continuation experiment

**Items**

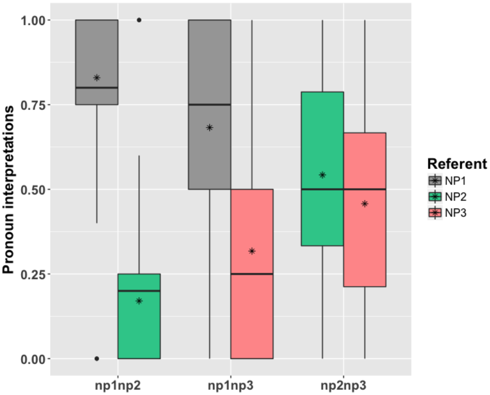
Adam called Diana for Russel. He ______________________ [pronoun prompt]
Adam called Diana for Russel. __________________________ [free prompt]

- 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

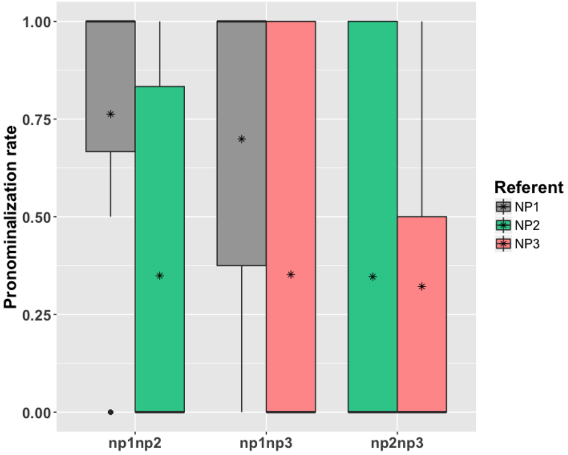



Free prompt




Pronoun prompt

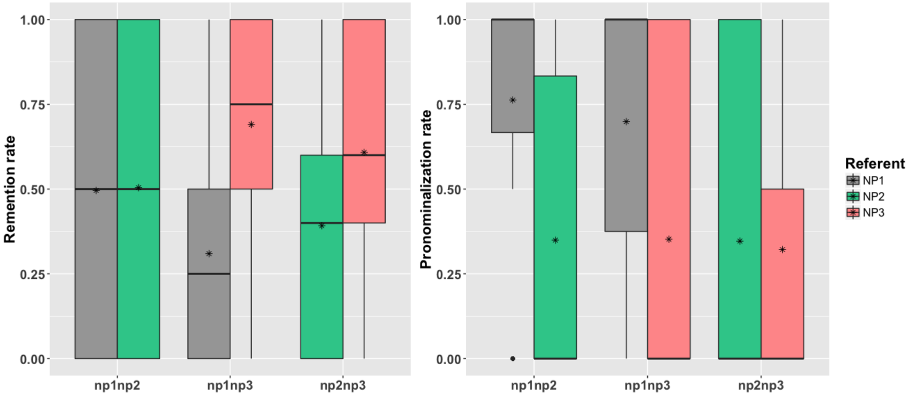
# Results: Subjects are preferentially pronominalized




Free prompt

# Results 1: Does predictability influence pronominalization?

# Results 1: Does predictability influence pronominalization?



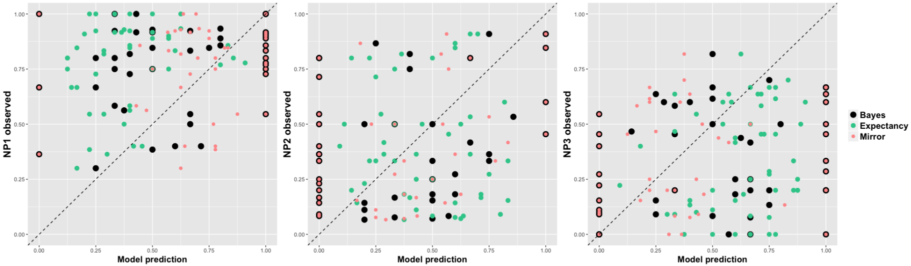

Free prompt

# Results 2: Does Bayes' Rule rule?

Following Rohde & Kehler (2014), we used the free prompt continuations to calculate Bayes-derived estimates of $p(\textit{referent}|\textit{pronoun})$ via the prior $p(\textit{referent})$ and likelihood $p(\textit{pronoun}|\textit{referent})$, 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 $p(referent|pronoun)$ via the prior $p(referent)$ and likelihood $p(pronoun|referent)$, 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:** $\mathbf{R^2 = .377}$
**Participants:** **Bayes:** $\mathbf{R^2 = .084}$, Expectancy: $R^2 = .021$, Mirror: $R^2 = .075$

# Interim discussion

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

# Interim discussion

- We do not find any evidence that pronominalization is affected by predictability

  $\rightarrow$ 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
  $\rightarrow$ In line with strong Bayes
- The Bayesian model outperforms the Expectancy model
- The Bayesian model is outperformed by the Mirror model

$\rightarrow$ Is this due to the construction or does it have something to do with the number of referents?

# Follow-up: 2-human Benefactive prompts

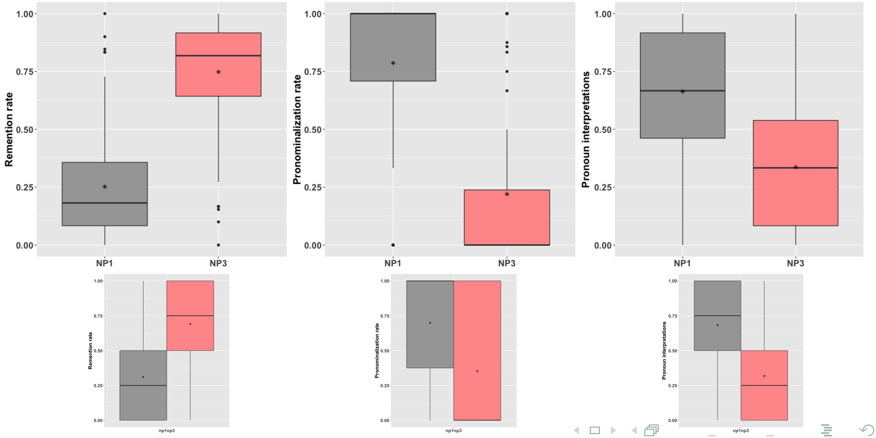
## Items

Adam called the hospital for Russel. He ________________ [pronoun prompt]
Adam called the hospital for Russel. ____________________ [free prompt]

# Follow-up: 2-human Benefactive prompts

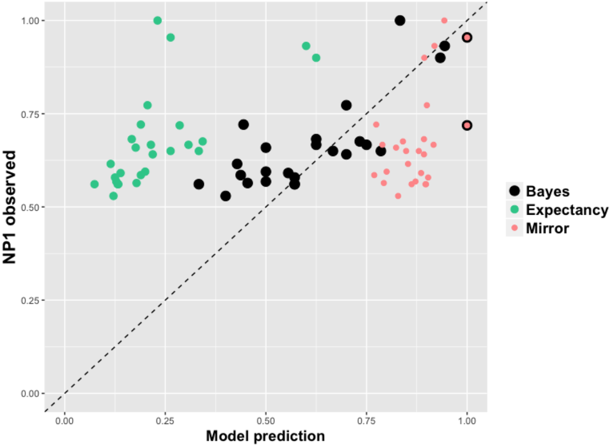
## Items

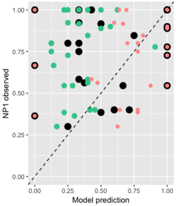
Adam called the hospital for Russel. He ________________ [pronoun prompt]
Adam called the hospital for Russel. ____________________ [free prompt]



# Follow-up: 2-human Benefactive prompts

# Follow-up: 2-human Benefactive prompts




**Items:** **Bayes: $\mathbf{R^2}$ = .719**, Expectancy: $R^2 = .311$, Mirror: $R^2 = .714$
**Participants:** **Bayes: $\mathbf{R^2}$ = .348**, Expectancy: $R^2 = .008$, Mirror: $R^2 = .282$

# Discussion

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

# 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