Explanation and Rationality
in Models of Language Use

Kees van Deemter

Information & Computing Sciences
Utrecht University

NLP/NLG: why are we doing it?

Kees van Deemter

Information & Computing Sciences
Utrecht University

Plan of the talk

1. Rationality as an explanatory principle in NLP
2. Models of referring
   - A Bayesian model (RSA)
   - The PRO model
   - Simplified presentation: models used as illustrations
3. How explanatory are these models?
4. How about your work?

NLP: science or engineering

NLP can be a branch of engineering, aiming to build systems for

- Machine Translation
- Robotics
- Decision support,
- etc.

NLP: science or engineering

- NLP can also be a science, aiming to construct computational models of:
  - Human language evolution
  - Human language acquisition (1st or 2nd)
  - Human language comprehension
  - **Human language production** : our focus today

- My group: "We aim to understand human language"
  - But what does that mean?

NLP: science or engineering

A difference of emphasis:

- NLP-as-engineering emphasizes applications and usefulness.
- NLP-as-science emphasizes …

NLP: science or engineering

A difference of emphasis:

- NLP-as-engineering emphasizes applications and usefulness.
- NLP-as-science emphasizes explanation?

NLP: science or engineering

Stanford Encyclopedia of Philosophy.
Entry: "Scientific Explanation" (Woodward & Ross 2021):

"The Deductive-Nomological Model"

- Showing that your observation follows from known principles/laws/insights
- Reducing the unknown to the known

NLP: science or engineering

One known human trait is **rationality**

Early use in linguistics: Gricean Maxims (1975)

- Not detailed enough to make predictions

Rationality as an explanatory principle. Example: chess

Rationality as an explanatory principle

Example: Wang Jing plays chess move x.

x is the best move.

Why did Wang Jing play x?

Explanation: Wang Jing was behaving rationally.

Plausible if (1) Wang Jing has great analytical skills, and (2) these skills suffice to make him find move x

Accurate if x is predicted with high probability.

➔ explanation must be **detailed**

Irrationality as an explanatory principle

Example: Wang Jing plays move y (accepting a gambit)

But y is **bad** move.

Why did Wang Jing play y?

Explanation: Wang Jing is only a beginner, and beginners tend to focus on short-term gain.

Lesson: Rationality has limits!

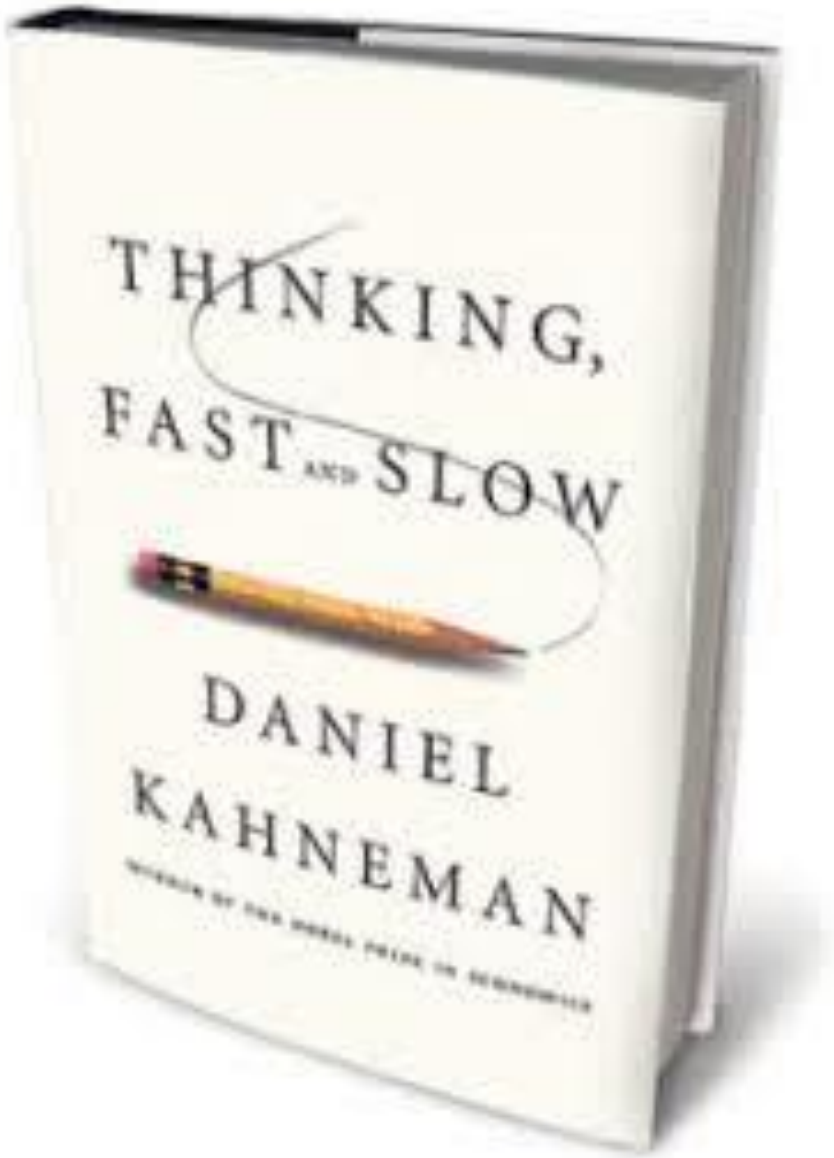
Bounded rationality
in behavioural economics

Herbert Simon > 1957

- Need to calculate 10 ply → players use shortcuts
- Models need to take such resource limitations into account

Kahneman & Tversky > 1975

- Prospect Theory, Nobel Prize Economics 2002
- Experiments with investment decisions
- Humans do not maximise *expected payoff*



Obstacles against (naïve) rationality:

- Shortcuts
- Optimism bias
- Anchoring
- …

Example: Referring Expressions Generation

Refer = identify an entity for a hearer/reader

Referring Expressions Generation (REG)

- Input = domain + target referent
- Output = a definite NP ("the …")
- Generating "one-shot" REs (No anaphora!)

First a Bayesian model, called RSA:
*Frank & Goodman (2012) Predicting Pragmatic Reasoning in Language Games. Science **336**: 998.*

The RSA algorithm

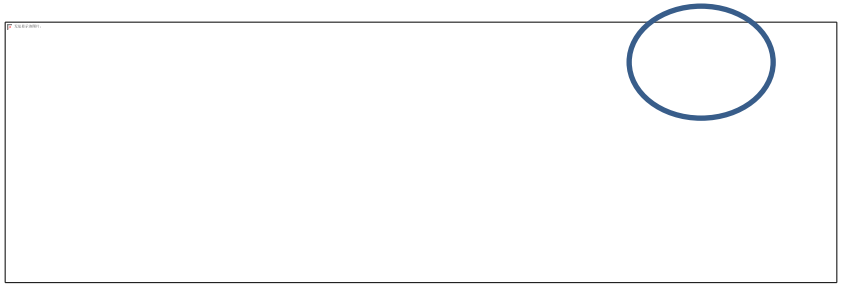
Rational Speech Act theory: In the spirit of Grice, but far more detailed

- Speakers optimize the probability of correct interpretation
- Hearers interpret by assuming that speakers optimize their utterances in that way

RSA (hearer model)

Hearer's interpretation depends on:

- The prior probability that the referent will be referred to.
- The probability that word *w* will be used for that referent in a particular context.

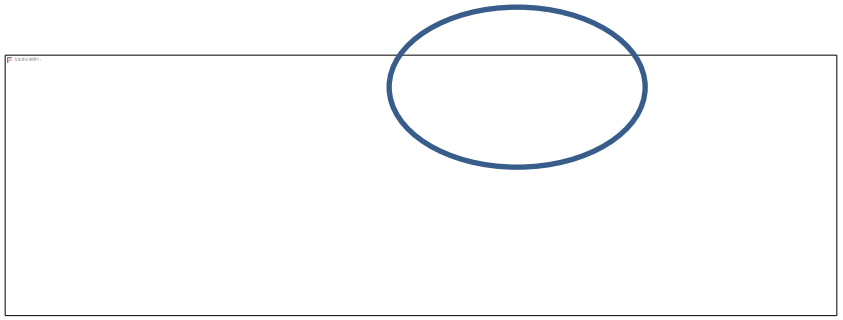


***P(r)* is a model of the salience degree of *r*.**

RSA (hearer model)

Hearer's interpretation depends on:

- The prior probability that the referent will be referred to.
- The probability that a word *w* will be used for that referent in a particular context.



***P(w|r)* is a model of the speaker's referential choices when referring to *r*.**

RSA (speaker model)

Speaker model:

The fewer objects *w* is true of, the likelier *w* is used.

➔ RSA asserts that **Discriminatory Power** is crucial.

Rationality of the speaker = utility for the reader!

Our own experiments say:

1. Discriminatory power plays a limited role in speaking.

2. Preferences between properties interfere:
   - Based on perceptual salience/codability
     Pechmann 1989, Dale & Reiter 1995, Belke & Meyer 2002, Murray 2006, Fang et al. 2008, Schwartzkopf et al. 2010.
   - Colour > size.
   - Up/down > left/right *(Arts 2004)*.

3. Speakers routinely "overspecify".

To model these findings:

PRO: **P**robabilistic **R**eferential **O**verspecification.

Van Gompel, van Deemter, Gatt, Snoeren, Krahmer (2019): Conceptualisation in Reference Production: Probabilistic Modelling and Experimental Testing. Psychological Review.

PRO models the distribution of REs.

- a very limited role for **discriminatory power**
- an additional mechanism for **over-specification**

The PRO algorithm (sketch)

- If a fully discriminating property exists, then select one.
- Select properties with probability *x*

  Higher x for more preferred properties
- Further properties may be added, based on parameter y for eagerness to over-specify

  Higher y for greater eagerness

Factors influencing x and y

Earlier research:

- y is influenced by whether the task is **fault critical** *(Arts et al. 2012)*
- y is influenced by the **complexity of the scene** *(Koolen et al. 2013, Paraboni et al 2007, 2014)*

Experimental Materials & RSA's predictions

Fig. 1a. Size-only fully discriminating condition

Fig. 1b. Colour-only fully discriminating condition

Fig. 1c. Colour-or-size fully discriminating condition

Condition S

RSA predicts:
"the small candle" 0.67
"the grey candle" 0.33

Condition C

RSA predicts:
"the grey candle" 0.67
"the small candle" 0.33

Condition C+S

RSA predicts:
"the grey candle" 0.5
"the small cande" 0.5

Proportion of choices of RE types in each of the three conditions (C / S / C&S)

		RSA	Human	PRO
C+S	Choose C&S	0	0.17	0.17
	Choose S	0.5	0.04	0.02
	Choose C	0.5	0.79	0.81
S	Choose C&S	0	0.83	0.77
	Choose S	0.67	0.17	0.23
	Choose C	0.33	0.01	0.00
C	Choose C&S	0	0.08	0.08
	Choose S	0.33	0.00	0.00
	Choose C	0.67	0.92	0.92

Which model is best?

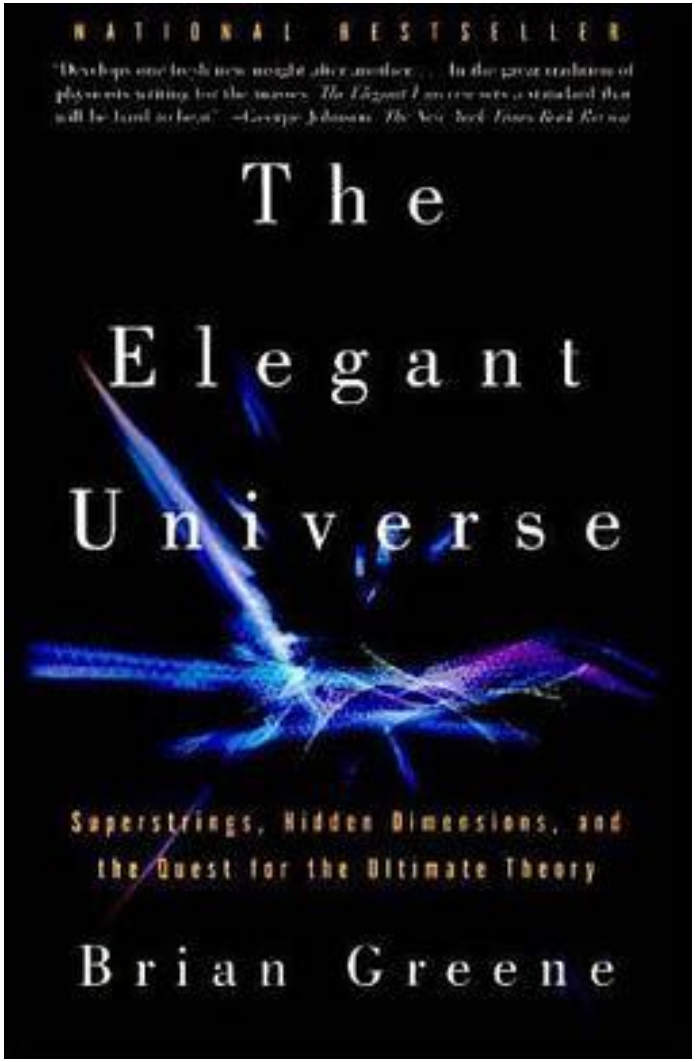
Criteria discussed so far:

- Plausibility
- Accuracy / Detail

Other criteria often mentioned:

- (Generality)
- Elegance

"Elegance" a huge factor in physics



- Elegance is difficult to define.
- Hinges on simplicity, brevity, symmetry?
- debate: has physics relied too much on elegance?

Empirical vs. explanatory adequacy

PRO:

- **Plausibility**: high. Rests on years of previous research (Codability, Gestalt, etc. see *Levelt 1989*, *Speaking*)
- **Accuracy**: High
- **Elegance**: Low

RSA 2012:

- **Plausibility**: Rationality seems plausible. However, there is not much basis in previous theory
- **Accuracy**: Low (does not match experimental data)
- **Elegance**: High. Simple. Covers both speaking and hearing

Lessons from behavioural economics (e.g. Kahneman & Tversky)

Investment decisions ➔ Three lessons:

1. Shortcuts
2. Optimism bias
3. Anchoring

Applying these lessons to future REG models:

Lessons from behavioural economics

Heeding the lessons of behavioural economics:

1. Shortcuts ✓

Finding the shortest RE is NP-complete, so every realistic model has to make shortcuts. *Dale and Reiter 1995, Computational Interpretation of the Gricean Maxims in the generation of referring expressions.*

Lessons from behavioural economics

Heeding the lessons of behavioural economics:

1. Shortcuts v
2. Optimism v

- Underspecification is surprisingly common.

  (PhD Thesis Guanyi Chen 2022)

- Speakers over-estimate listeners' knowledge.

  ("If I know it, then you know it." *Fussell & Krauss 1992*, Coordination of Knowledge in Communication.)

Lessons from behavioural economics

Heeding the lessons of behavioural economics:

1. Shortcuts ∨
2. Optimism bias ∨
3. Anchoring ∨

- Preference for perceptually salient properties
  (e.g., colour, ≧ *Pechmann 1989*)
- Speakers re-use properties uttered by interlocutors
  (referential alignment, *Goudbeek & Krahmer 2012*)

Explanation in NLP today

Explanation not a popular concern in NLP today

- The use of similarity metrics for evaluation (BLEU, DICE, BertScore, MoverScore) suggest that most researchers pursue NLP as science
  - If NLP as engineering, then evaluate usefulness ➔ use task-based evaluation (or questionnaires)
- Yet models are usually evaluated on performance, not on explanatory value

Why is accurate prediction not enough?

Without explanation, no answers to "*What if* questions", such as:

- What if the domain was **larger**?
- What if we changed the **properties** (colours, sizes, etc.)?
- Which lessons carry over to **other types of expressions**?

My question to you

Do you care about “understanding” human language?

a. What makes a *good explanation*?

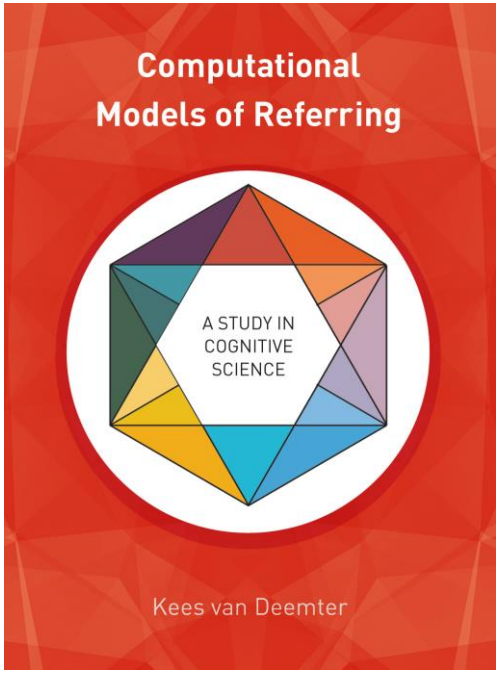
b. *How successful* is current NLP at finding good explanations?

谢谢你们给我这个机会!

Reserve slides

Computational Models of Referring: a Study in Cognitive Science. MIT Press 2016.

Download from https://aura.abdn.ac.uk/handle/2164/8956



Rationality in Language Science

Oldest (?) example: Grice's Maxims

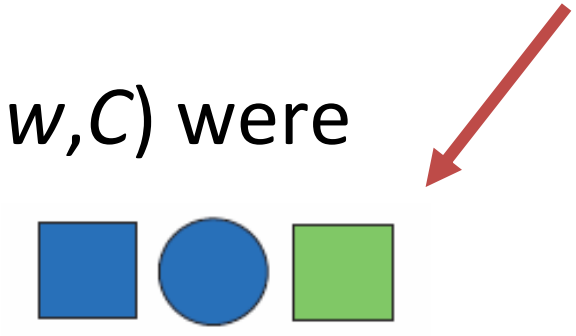
Gricean Maxims (1975): Cooperative Principle essentially says "speak rationally". Shaped into 4 Maxims

- Quantity, Quality, Relation, Manner

- Speaker appears not to speak rationally ➔ Hearer salvages rationality by making additional assumptions ("conversational implicatures")
- The Gricean approach was **plausible**, but not very **accurate**
  - Predictions were not detailed at all

Testing RSA's predictions (2012)

Predictions for $P(w|r,C)$ and $P(r|w,C)$ were tested in a simple experiment:



- 3 objects, 1 target
- 2 shapes, 2 sizes
- "Which word would you use?" (green? square?)

- This experiment: Participants chose 1 word.
- Other experiments use complete Noun Phrases

PRO may be less "elegant" than RSA

If one property in **P** removes all distractors then
    P-choose one such property
    add this property to **D**
While True do
    If **P** = ∅ then Return **D**
    Else if **D** is not distinguishing yet then
        P-choose a property from **P**
        Update **D**, **P** and **M**
    Else *R-choose* between STOP and the prop's in **P**
        If a P from **P** is chosen then update **D, P, M**
        Else Return **D**

Probabilistic choice

Probabilistic Overspecification

2019 experiment

Scenes composed of 3 objects

36 experimental items

- Condition S: Only Size (S) suffices to identify the referent
- Condition C: Only Colour (C) suffices
- Condition C+S: Colour (C) suffices. Size (S) suffices

30 participants described each target object

Evaluation used maximum likelihood:

- For each model *m*, compute ***P*(*data*|*m*)** .
- Busemeyer's *Generalisation Criterion Methodology*

PRO "outperformed"

- the "classic" pre-2012 models (e.g. Incremental Algorithm)
- probabilistic versions of these models
- Rational Speech Act (RSA) theory, etc.

Details: *Van Gompel et al. (2019)*

A recent refinement of RSA

Degen et al, (2020) When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, 127(4).

- Each property P has a "precision" parameter x_P :

  $x_{red} = 0.9$ and $x_{small} = 0.8$ says:
  "red" is more precise than "small"
  → "red" is 9/8 more likely to be misinterpreted.
  → "red" is 9/8 more likely to be used.

- No comparison with performance of other algorithms.
  - lack of what we called detail/accuracy