Encoding versus decoding.
Why do language users make sentence structure explicit?

Dirk Pijpops

QLVL, University of Leuven

Research Foundation Flanders (FWO)

Psycholinguistics in Flanders (PiF) conference, Monday 29th of May 2017, Leuven

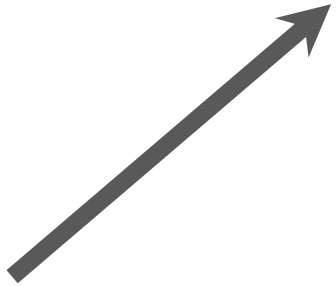
Cognitive Complexity Principle:

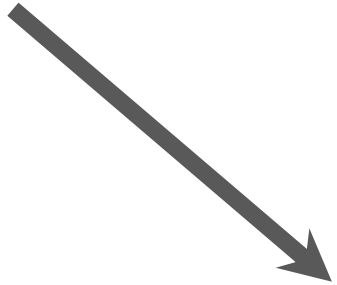
In case of more or less explicit grammatical options the **more explicit one(s)** will tend to be favored in cognitively **more complex environments**. (Rohdenburg 1996: 151)

(1) *Well, I'm not, because I understand **(that)** most of his girlfriends have either been, you know, like the hooker or porn star types.* (COCA, cited in Shank et al. 2016)

(2) *De Indiërs aarzelen **(om)** te investigeren in Uganda.* (Bouma 2017: 65)
The Indians hesitate (for) to invest in Uganda

Probabilistic Grammar





Complexity ~ Explicitness

Online processing:
spoken language

Processing constraints

Complexity ~ Explicitness

Offline processing:
written language & language tasks

Processing constraints **alleviated**

(Gries 2003: 48-66, Jaeger 2010)

Encoding or decoding?

- Implicit decoding-perspective (Hawkins 1990, 1992, 2004; Rohdenburg 1996, Bouma 2017,…)
  - Being explicit is burdersome for encoder
  - Encoder knows syntactic structure
  - Linguists decode
  - Training sessions

Encoding or decoding?

- Encoding-perspective
  - Bottleneck is encoding, not decoding (Levinson 2002: 28)
  - Encoder's altruism is evolutionarily implausible (Kirby 1999)

Encoding or decoding?

- Explicitness also has benefits for the encoder
  - Often allows for more flexible word order (Willems 2017)
  - Buys time (Ferreira & Dell 2000: 299)

(1) *Well, I'm not, because I understand* ***(that)*** *most of his girlfriends have either been, you know, like the hooker or porn star types.* (COCA, cited in Shank et al. 2016)

Encoding or decoding?

- Psycholinguistic experiments: **encoding** (overview in Ferreira & Dell 2000)

- Corpus research: **?** (Colleman 2006, Shank et al. 2016, Bouma 2017,…)
  - Hard to distinguish
  - Community level

Case study

(3) *We* ***zoeken*** *alternatieven*. (SoNaR corpus)

we search alternatives

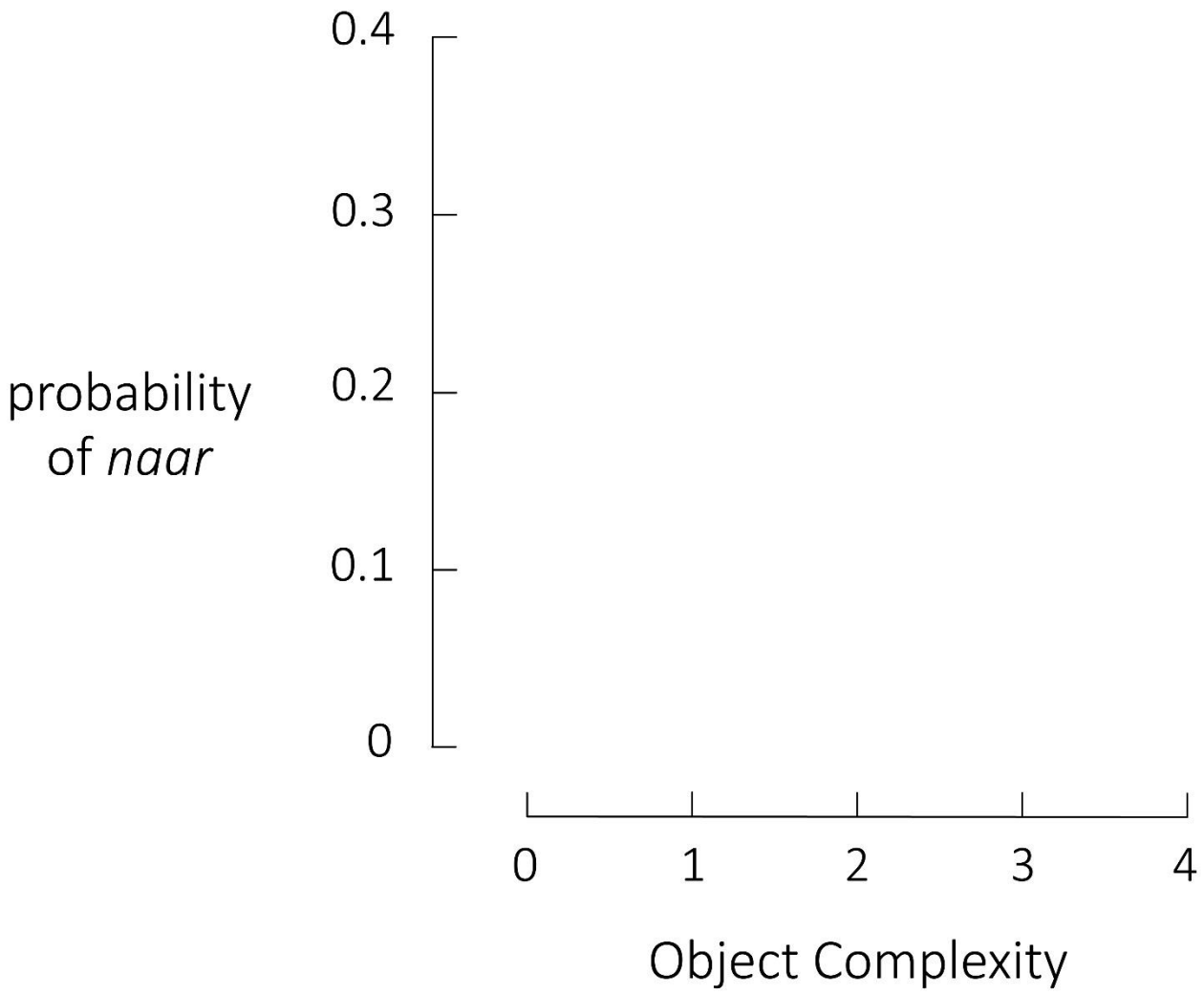
(4) *Wij* ***zoeken*** *dan wel* ***naar*** *alternatieven*. (SoNaR corpus)

we search then well to alternatives

SoNaR corpus (Oostdijk et al. 2010)

- Why **written** language? (Gries 2003: 48-66, Jaeger 2010)
  - Hyperconservative
  - Get enough data
- **Excluded** tweets, text messages, chats, discussion lists: quality of syntactic parses deemed too low
- Extracted all instances of ***zoeken*** **'to search'**, in which the object is overtly expressed: 61998 without *naar* vs. 17440 with *naar*

- Logistic regression model
- Response: presence of *naar*
- Add fixed effect: Object Complexity



Control for lectal factors

- 2 countries: Belgium, the Netherlands
- 20 components: newspapers, subtitles, wikipedia, blogs,…

	Belgium	The Netherlands	unknown
auto cues	4218	535	0
blogs	**0**	**0**	**36**
books	15	6685	0
brochures	162	15	7
e-magazines	1132	384	0
electronic newsletters	**0**	**0**	**1**
guides manuals	2	12	1
legal texts	2	7	85
newspapers	26465	10399	0
periodicals magazines	14898	2411	0
policy documents	29	6	679
printed newsletters	**0**	**7**	**0**
proceedings	23	1	0
reports	53	257	1
subtitles	**5967**	**0**	**2739**
teletext pages	**106**	**0**	**0**
texts for the visually impaired	**0**	**175**	**0**
web sites	143	87	0
wikipedia	**0**	**0**	**1673**

Control for lectal factors

- Added Text-type as random effect: Belgian subtitles, Netherlandic texts for the visually impaired, Belgian teletexts,…

Control for semantic factors

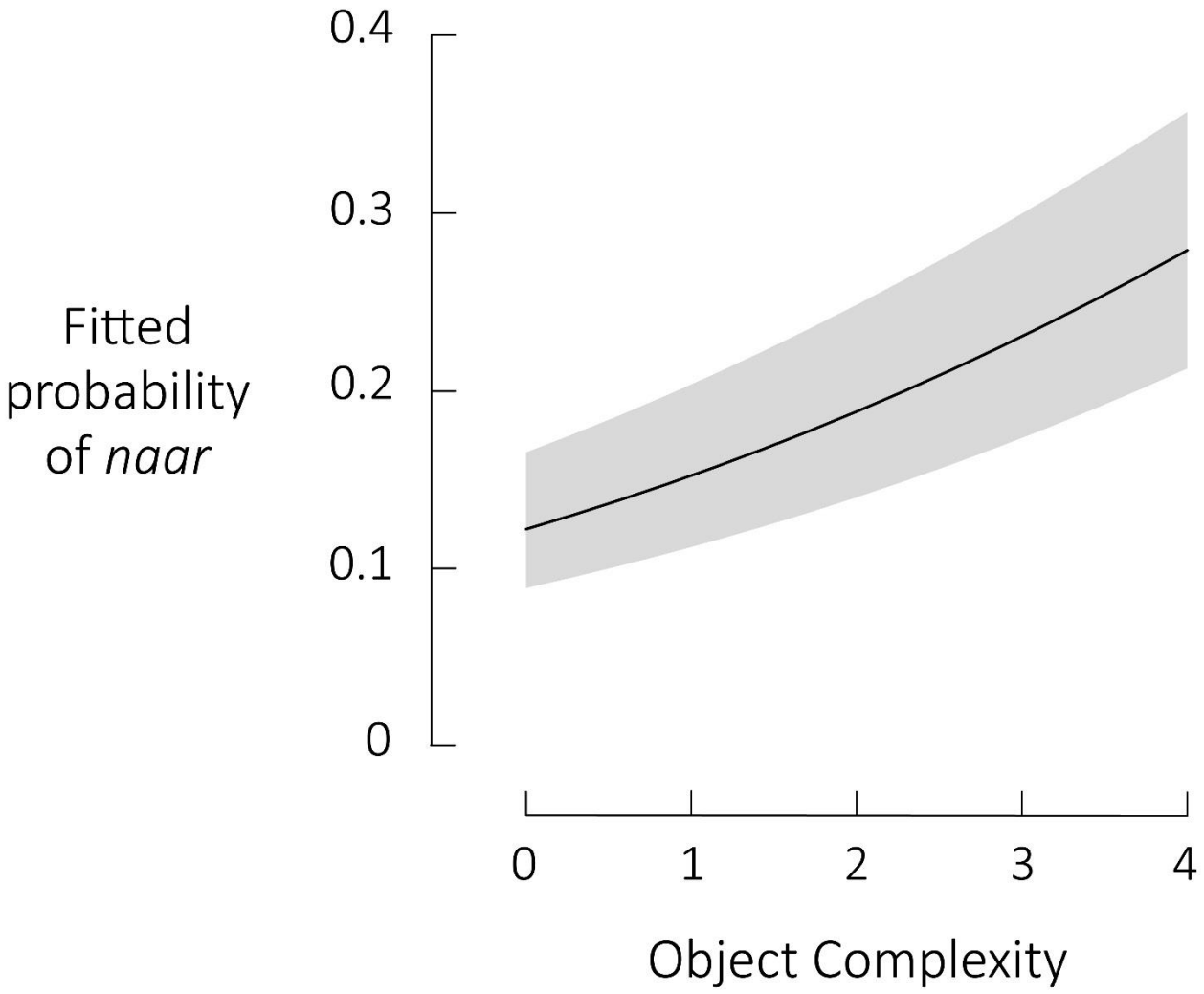
- Figurative 'seek to make/acquire': without *naar*
  - *contact* 'contact'
  - *aansluiting* 'association'
  - *bescherming* 'protection'
  - …
- Literal 'look for': with *naar*
  - *spoor* 'track'
  - *overlevende* 'survivor'
  - *slachtoffer* 'victim'
  - …

Control for semantic factors

- Added lemma of the syntactic head of the object as random effect
- Collapse levels if < 100 hits (Wolk et al. 2013) → 99 levels
- Removed *heil* 'salvation', *niets* 'nothing', *toenadering* 'overture', *toevlucht* 'refuge', *verkoeling* 'cooling': no variation

- 58444 without *naar* vs. 17439 with *naar*

- Coëfficient: 0.26, $p < 0.0001$



Encoding vs. decoding

- Encoding Hypothesis 1: *naar* allows for a more flexible word order

	Ik	***heb***	*gisteren een boek*	***gezocht***	
*	*Ik*	***heb***	*gisteren*	***gezocht***	*een boek*
	Ik	***heb***	*gisteren naar een boek*	***gezocht***	
	Ik	***heb***	*gisteren*	***gezocht***	*naar een boek*

Encoding vs. decoding

- Encoding Hypothesis 1: *naar* enables the encoder to extrapose complex objects to the postfield
- Encoding Hypothesis 2: *naar* buys time for the encoder to formulate a complex object
- Decoding Hypothesis: *naar* acts as a signpost in decoding, and marks "what follows now, is the object"

⇨ Remove all observations where the object is extraposed to the postfield

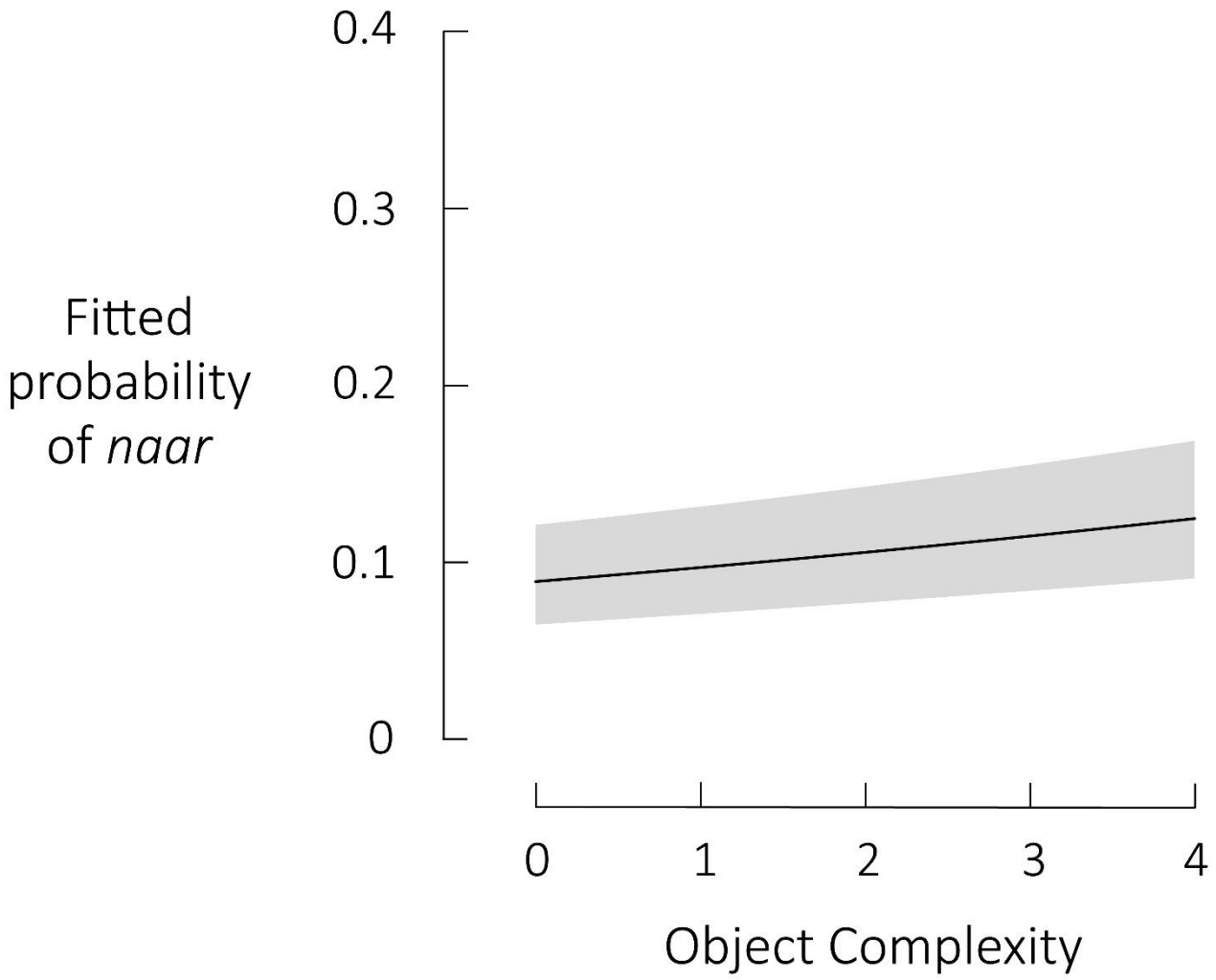
Encoding vs. decoding

- Encoding Hypothesis 1: word order, effect should disappear or reverse
- Encoding Hypothesis 2: buy time, effect should remain
- Decoding Hypothesis: signpost, effect should remain

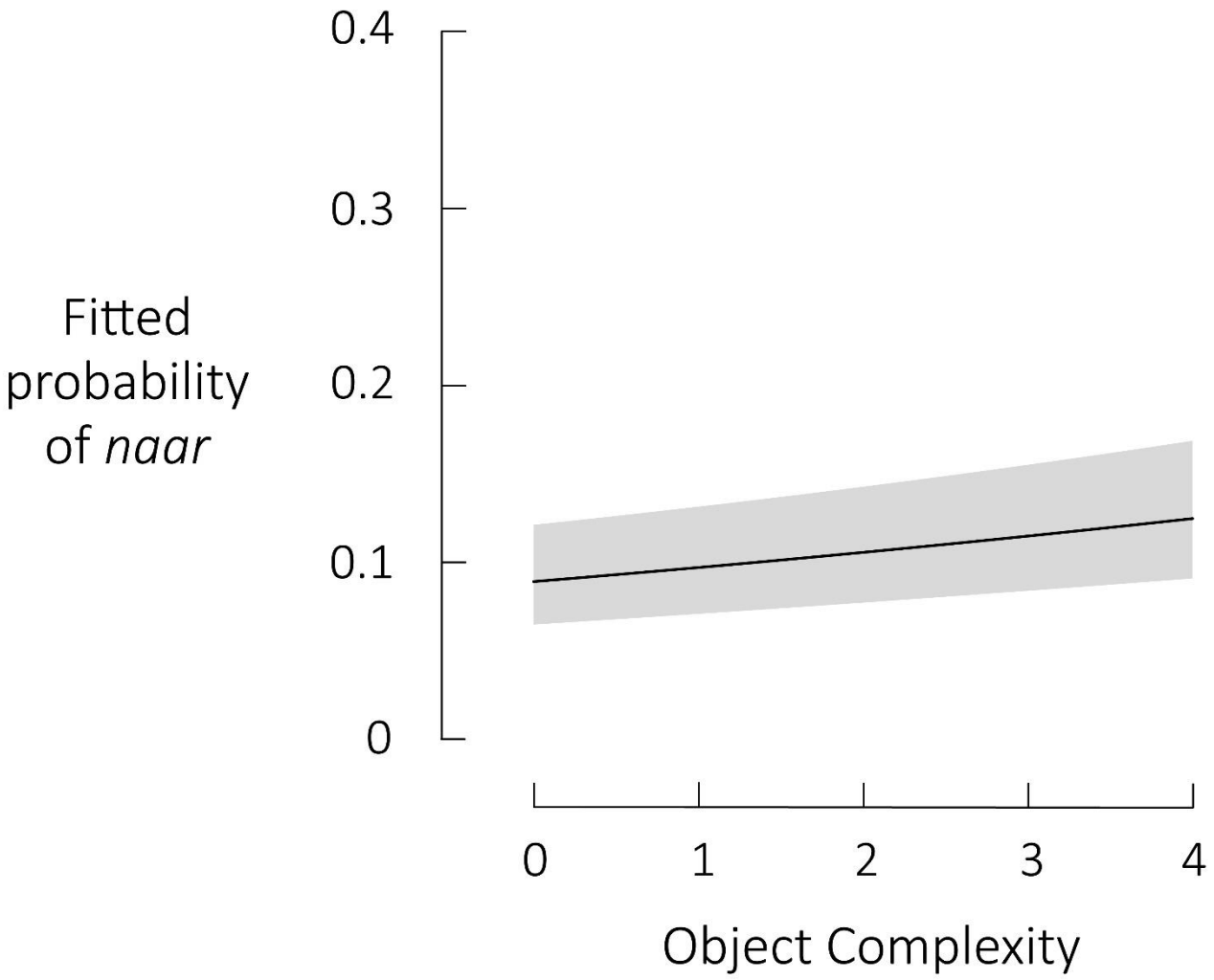
⇨ Remove all observations where the object is extraposed to the postfield

- 58444 without *naar* vs. 10959 with *naar*
- Coëfficient: 0.09, $p < 0.0001$

⇨ Effect greatly weakens



- Encoding Hypothesis 1: word order, effect should disappear CONFIRMED
- Encoding Hypothesis 2: buy time, effect should remain
- Decoding Hypothesis: signpost, effect should remain



Complex object preceeds the verb

(5) *<u>Naar politiek als roeping, of zelfs maar als ethos</u>,* ***zoekt*** *de lezer tevergeefs.*

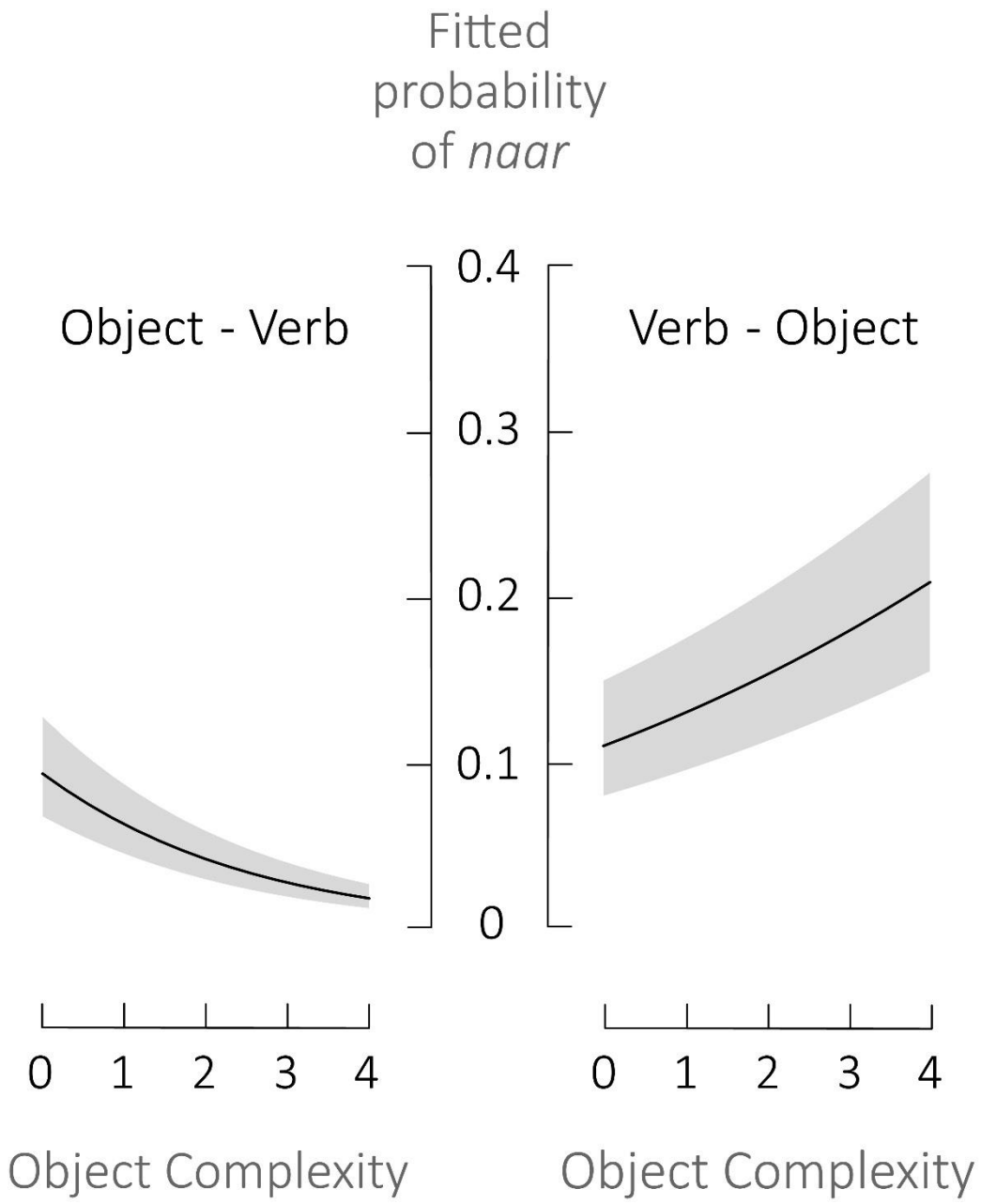
⇨ Encoder: *naar* 

⇨ Decoder: *naar* 

- Add interaction:
  Object Complexity & Object-Verb Order

- Encoding Hypothesis confirmed

- Decoding Hypothesis unconfirmed



Encoding versus decoding. Why do language users make sentence structure explicit?

To facilitate encoding

This dovetails with findings in psycholinguistic experiments, e.g. Ferreira & Dell's *that*-omission study (2000), and references cited therein.

Thanks!

Dirk Pijpops

Pijpops, Dirk and Dirk Speelman. 2017. **Alternating argument constructions of Dutch psychological verbs. A theory-driven corpus investigation.** *Folia Linguistica* 51(1): 207–251.

References

Bates, Douglas, Martin Maechler, Ben Bolker and Steven Walker. 2013. *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.4*.

Ferreira, Victor and Gary Dell. 2000. Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production. *Cognitive Psychology* 40(4). 296–340.

Ford, Marilyn and Joan Bresnan. 2013. Using convergent evidence from psycholinguistics and usage. In Manfred Krug & Julia Schlüter (eds.), *Research Methods in Language Variation and Change*, 295–312. Cambridge: Cambridge University Press.

Fox, John, Sanford Weisberg, Michael Friendly, Jangman Hong, Robert Andersen, David Firth and Steve Taylor. 2016. Effect Displays for Linear, Generalized Linear, and Other Models. R package version 3.2.

Gries, Stefan Thomas. 2003. *Multifactorial analysis in corpus linguistics : a study of particle placement*. New York: Continuum.

Hawkins, John. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Kirby, Simon. 1999. *Function, selection, and innateness : the emergence of language universals*. Oxford: Oxford University Press.

Levinson, Stephen. 2000. *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge: Cambridge : MIT press,.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste and Ineke Schuurman. 2013. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. *Theory and Applications of Natural Language Processing*. 219–247.

Rohdenburg, Günter. 1996. Cognitive Complexity and Increased Grammatical Explicitness in English. *Cognitive Linguistics* 7(2). 149–182.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.