Processing shapes grammar

But whose processing are we talking about?

Dirk Pijpops[1,2] & Freek Van de Velde[1]

[1]QLVL Research Unit, University of Leuven

[2]Research Foundation Flanders (FWO)

DGfS workshop 'Coding asymmetries' (Saarbrücken 8-10 March 2017)

Introduction

- Basic tenet of usage-based linguistics: processing -> usage -> grammar

Performance-Grammar Correspondence Hypothesis (Hawkins 2004: 3)

Grammars have conventionalized syntactic structures in proportion to their degree of preference in performance, as evidenced by patterns of selection in corpora and by ease of processing in psycholinguistic experiments.

Introduction

- Building on Hawkins' theories, Rohdenburg proposes the **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)

Construction + Ø	in less complex environments
Construction + $X_{morpheme/word/...}$	in more complex environments

Introduction

- Why?
  - Processing-driven (cognitive complexity)
  - Whose processing?
  - Addressee's processing (Rohdenburg 1996:149)
    - The extra element aids the hearer
    - For the speaker, adding an extra element just adds to the cognitive burden
- Why the hearer? That is counter-intuitive, as:
  - Speaker's altruism is evolutionarily implausible (Kirby 1999)
  - Bottleneck in human communication is in encoding, not decoding (Levinson 2000: 28)

Ok. Now we have our straw men

das Armdrücken
in
Saarbrücken

First a few words on clause structure in Dutch

- Works pretty much like German
- Topological approach with a bipolar structure (Klammerstruktur) (Zifonun 1997: 1498; Zwart 2011: 26)
- Ignoring the left-detached and the right-detached position, the schema for main clauses is:

Prefield	1st pole	Midfield	2nd pole	Postfield (-> extraposition)
Ik	***zoek***	*(naar) een boek over taalkunde*		
Ik	***heb***	*(naar) een boek over taalkunde*	***gezocht***	
Ik	***heb***	*(naar) een boek*	***gezocht***	*over taalkunde*
~~*Ik*~~	~~***heb***~~		~~***gezocht***~~	~~*een boek over taalkunde*~~
Ik	***heb***		***gezocht***	*naar een boek over taalkunde*
XP	V-fin	XP	V-nonfin	any XP that starts with a relator (no bare NP)

The issue

- Verb *zoeken* occurs in two variants: with a DO and with a PO

1. *We zoeken alternatieven.* (WR-P-P-G-0000254655.p.11.s.5)
   'We are looking for alternatives.'

2. *Wij zoeken dan wel naar alternatieven.* (WR-P-P-G-0000488037.p.6.s.3)
   'We, then, look for alternatives.'

- The PO (in 2) is the 'bulkier' variant and may be expected to occur in cognitively more complex contexts, following Rohdenburg (1996)

1. Relevance for corpus linguists: Do psycholinguistic mechanisms of complexity affect language use itself? Do we find their influence in naturally occurring language use, outside of experimental settings? **Does processing shape (probabilistic) grammar?**

2. Relevance for psycholinguists: **Whose processing are we talking about?** The producer's or the addressee's?

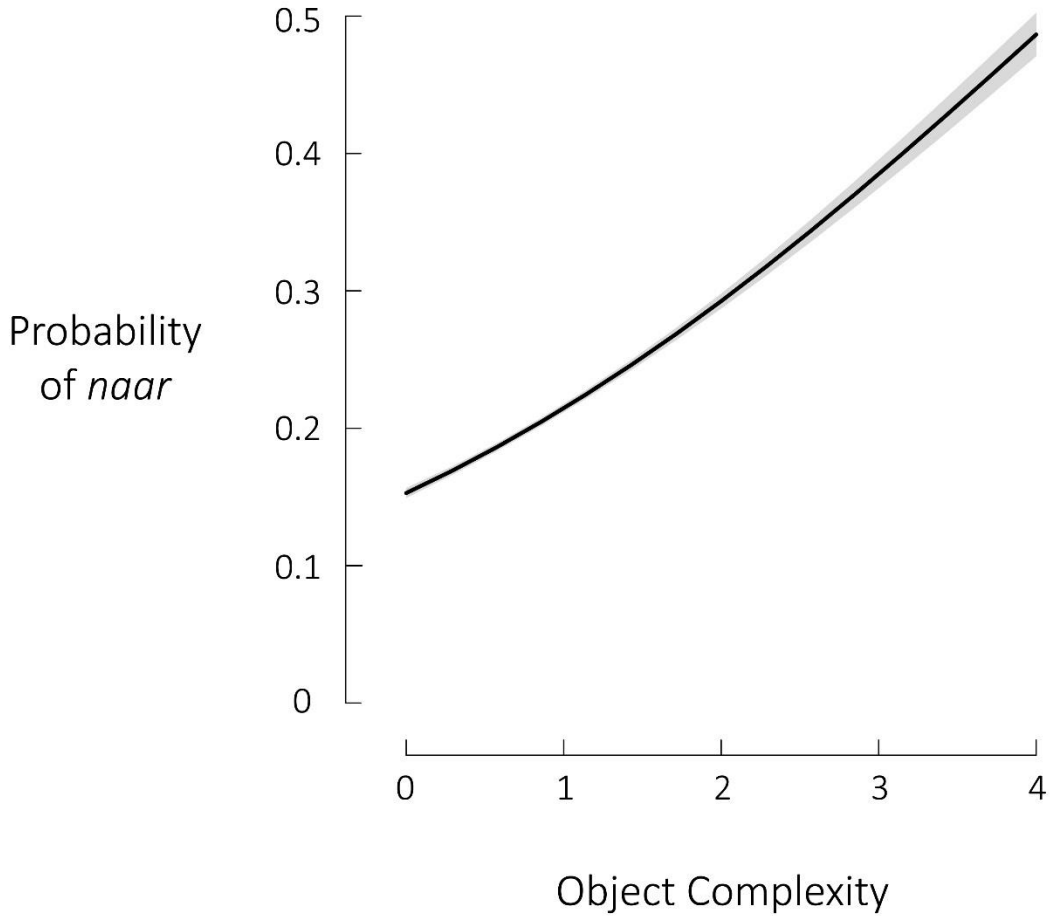
Does processing shape (probabilistic) grammar?

Sonar corpus of written Dutch (Oostdijk et al 2013, cf. Gries 2003: 48-66, Jaeger 2010,…)

- Why written language? To be **hyperconservative** (Ford & Bresnan 2013)
- **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*

Does processing shape (probabilistic) grammar?

- 61998 without *naar* ↔ 17440 with *naar*
- As the object becomes more complex, the probability of *naar* increases (positive estimate for Object Length: 0.41)
- Highly significant: < 0.0001



Whose processing are we talking about?

- Producer-driven Hypothesis 1: *naar* allows the producer to extrapose long objects to the postfield
- Producer-driven Hypothesis 2: *naar* functions as a grammatical *uh*, buying time for the producer to formulate a complex object
- Addressee-driven Hypothesis: *naar* functions as a grammatical signpost for the addressee. It marks 'what follows now, is the object of the verb'

Remove the observations where the object is extraposed to the postfield

Remove: *Het stadsbestuur* ***heeft*** *daarom* ***gezocht*** *naar een efficiëntere en goedkopere oplossing*

Keep: *Nijmegen* ***zoekt*** *naar een oplossing*

Prediction: as the object becomes more complex, the probability of *naar* will **no longer** increase

Whose processing are we talking about?

- Producer-driven Hypothesis 1: *naar* allows the producer to extrapose long objects to the postfield
- Producer-driven Hypothesis 2: *naar* functions as a grammatical *uh*, buying time for the producer to formulate a complex object
- Addressee-driven Hypothesis: *naar* functions as a grammatical signpost for the addressee. It marks 'what follows now, is the object of the verb'

Remove the observations where the object is extraposed to the postfield

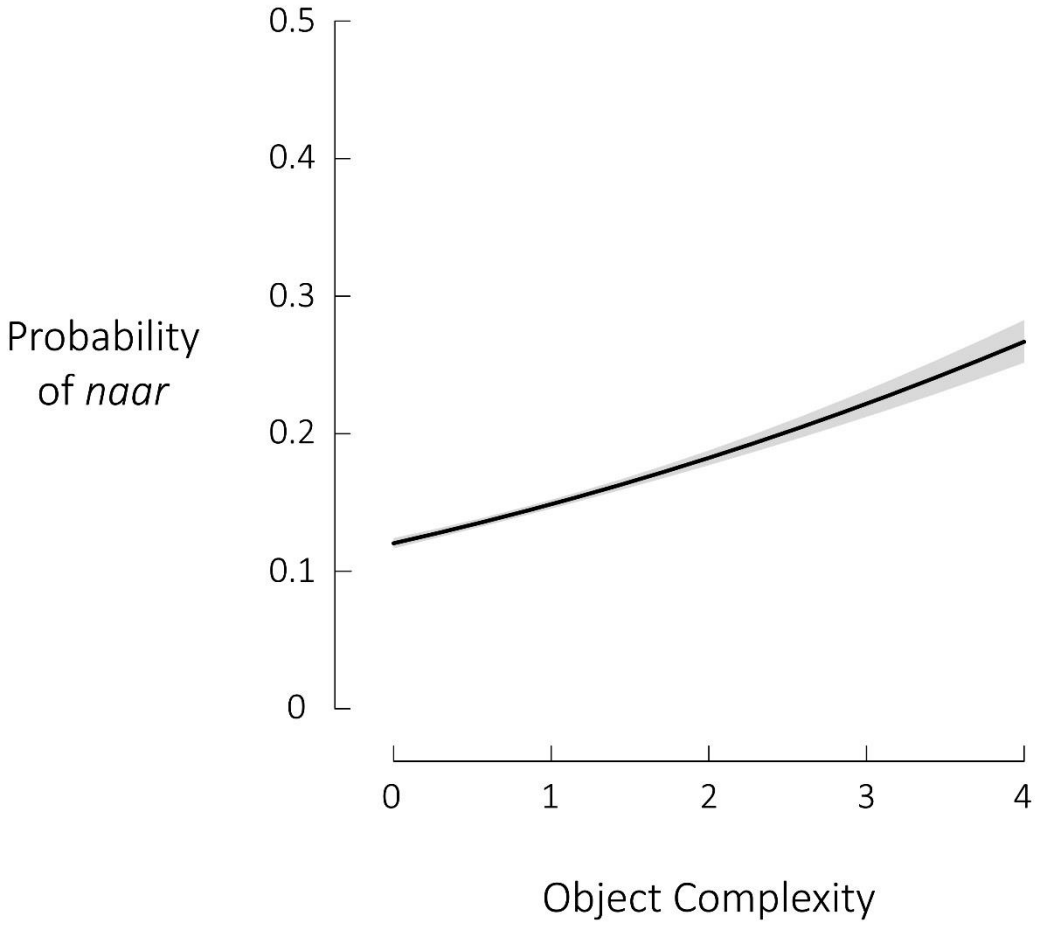
Remove: *Het stadsbestuur* ***heeft*** *daarom* ***gezocht*** *naar een efficiëntere en goedkopere oplossing*

Keep: *Nijmegen* ***zoekt*** *naar een oplossing*

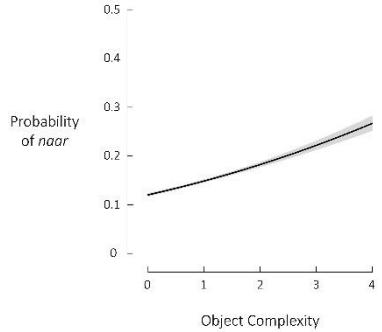
Prediction: as the object becomes more complex, the probability of *naar* will **still** increase

Whose processing are we talking about?

- 61998 without *naar* ↔ 10949 with *naar*
- As the object becomes more complex, the probability of *naar* **still** increases
  (be it less so, positive estimate for Object Length: 0.25)
- Highly significant: < 0.0001



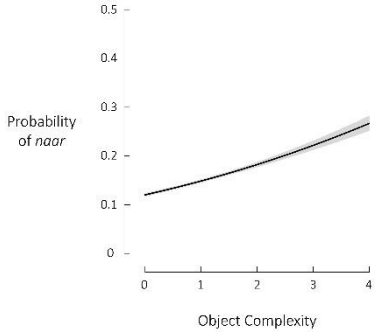
Whose processing are we talking about?



- ~~Producer-driven Hypothesis 1: *naar* allows the producer to extrapose long objects to the postfield~~
- Producer-driven Hypothesis 2: *naar* functions as a grammatical *uh*, buying time for the producer to formulate a complex object
- Addressee-driven Hypothesis: *naar* functions as a grammatical signpost for the addressee. It marks 'what follows now, is the object of the verb‘.

Prediction confirmed: as the object becomes more complex, the probability of *naar* will **still** increase

Whose processing are we talking about?



- Producer-driven Hypothesis 2.1: *naar* buys time to formulate a complex object. However, if it limits the producer's future choice of verb, he/she'd rather not express it.
- Producer-driven Hypothesis 2.2: *naar* buys time to formulate a complex object. Even if it limits the producer's future choice of verb, that's a price he/she is willing to pay.
- Addressee-driven Hypothesis: *naar* functions as a grammatical signpost for the addressee. It marks 'what follows now, is the object of the verb'.

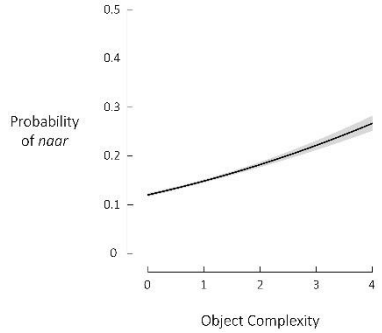
Remove the observations where the verb preceeds the object

Remove: *Nijmegen* ***zoekt*** *naar een oplossing*

Keep: *Naar politiek als roeping, of zelfs maar als ethos,* ***zoekt*** *de lezer tevergeefs*

Prediction: as the object becomes more complex, the probability of *naar* will **no longer increase**, or even **decrease**

Whose processing are we talking about?



- Producer-driven Hypothesis 2.1: *naar* buys time to formulate a complex object. However, if it limits the producer's future choice of verb, he/she'd rather not express it.
- Producer-driven Hypothesis 2.2: *naar* buys time to formulate a complex object. Even if it limits the producer's future choice of verb, that's a price he/she is willing to pay.
- Addressee-driven Hypothesis: *naar* functions as a grammatical signpost for the addressee. It marks 'what follows now, is the object of the verb'.

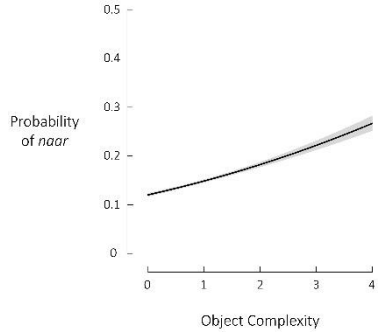
Remove the observations where the verb preceeds the object

Remove: *Nijmegen* ***zoekt*** *naar een oplossing*

Keep: *Naar politiek als roeping, of zelfs maar als ethos,* ***zoekt*** *de lezer tevergeefs*

Prediction: as the object becomes more complex, the probability of *naar* will **still** increase

Whose processing are we talking about?



- Producer-driven Hypothesis 2.1: *naar* buys time to formulate a complex object. However, if it limits the producer's future choice of verb, he/she'd rather not express it.
- Producer-driven Hypothesis 2.2: *naar* buys time to formulate a complex object. Even if it limits the producer's future choice of verb, that's a price he/she is willing to pay.
- Addressee-driven Hypothesis: *naar* functions as a grammatical signpost for the addressee. It marks 'what follows now, is the object of the verb'.

Remove the observations where the verb preceeds the object

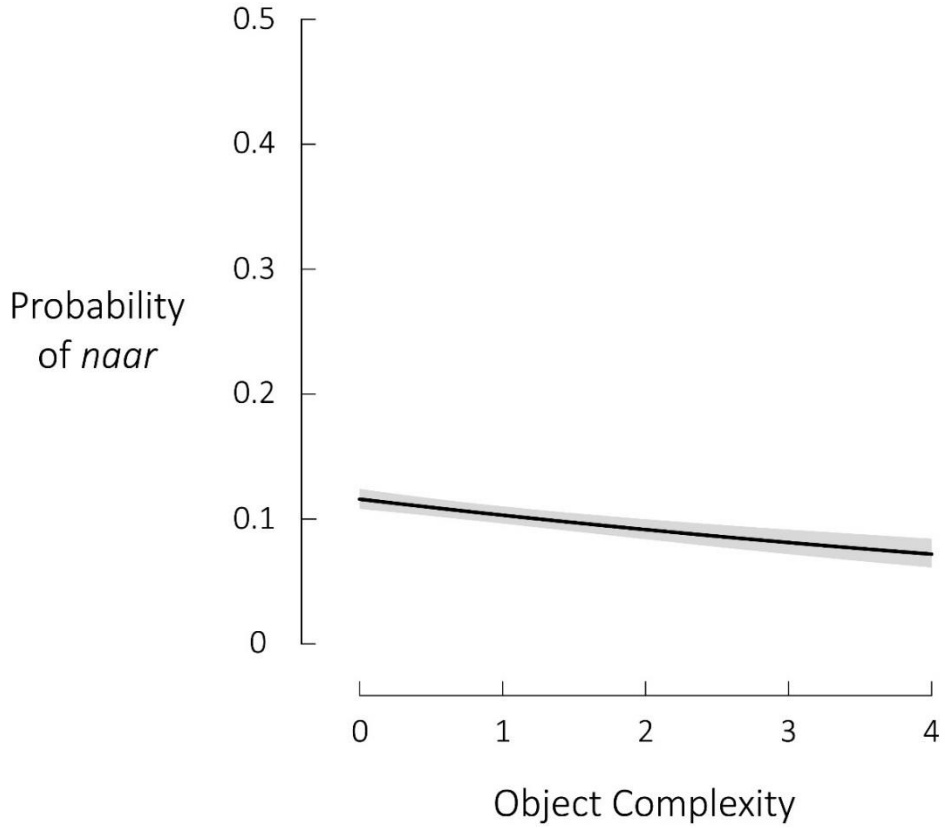
Remove: *Nijmegen* ***zoekt*** *naar een oplossing*

Keep: *Naar politiek als roeping, of zelfs maar als ethos,* ***zoekt*** *de lezer tevergeefs*

Prediction: as the object becomes more complex, the probability of *naar* will **still** increase, perhaps even more so

Whose processing are we talking about?

- 35089 without *naar* ↔ 4288 with *naar*
- As the object becomes more complex, the probability of *naar* **decreases**
  (negative estimate for Object Length: -0.13)
- Highly significant: < 0.0001



Prediction confirmed: as the object becomes more complex, the probability of *naar* **decreases**

Processing shapes grammar

But whose processing are we talking about?

The producer's

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

Thanks!

Dirk Pijpops & Freek Van de Velde

References

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

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

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

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

Jaeger, Florian Tim. 2010. Redundancy and Reduction: Speakers Manage Syntactic Information Density. Cognitive Psychology 61(1). 23–62.

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.

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

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