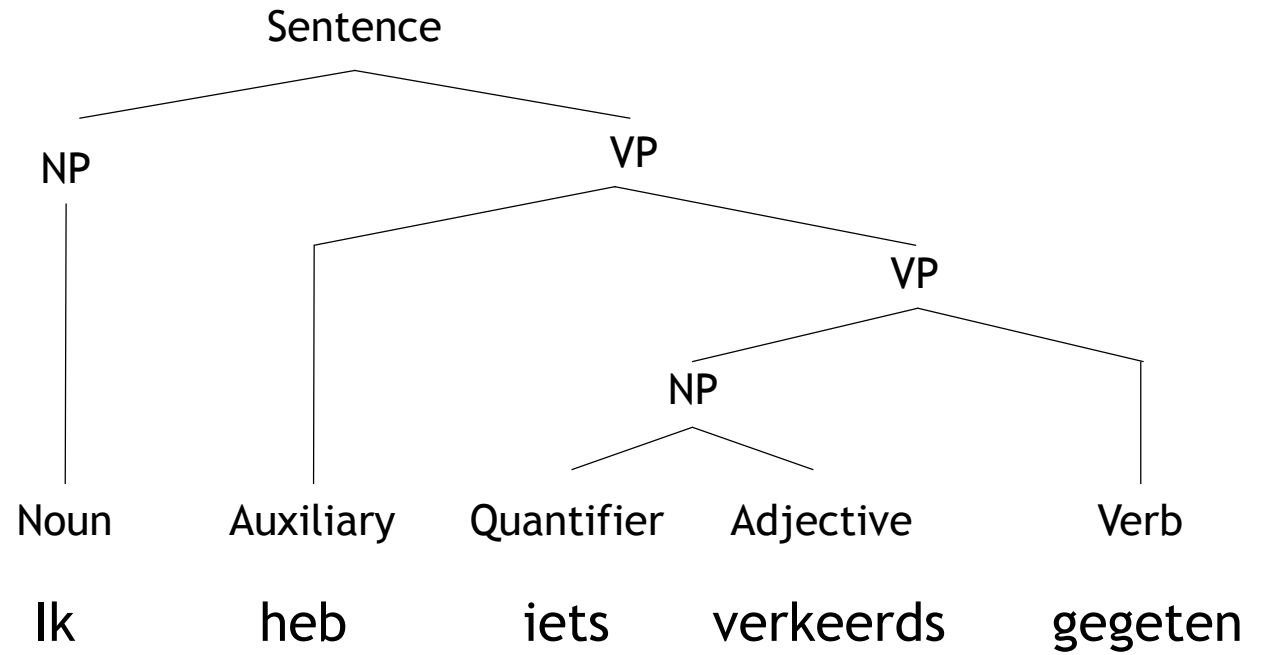


How to measure Constructional Contamination

Dirk Pijpops^{1,2} & Freek Van de Velde¹

¹QLVL, University of Leuven

²Research Foundation Flanders



Ik heb iets verkeers geeten
'I have eaten something wrong'



Ik heb iets verkeers geeten
'I have eaten something wrong'

Ik heb geeten
iets verkeers



Claims

- Language users do not always fully parse. Often, they simply **chunk**

(Ferreira & Patson 2007, Dabrowska 2014, Diessel 2015,...)

- As a by-product, this produces the effect of **constructional contamination**
- **Unrelated constructions** that happen to produce similar strings, quantitatively contaminate each other

Case study: partitive genitive

iets leuk(s)

[Quantifier Adjective (-s)]NounPhrase

‘something fun’

*uit de tijd dat de marechaussee nog wel **iets beter** te doen had*

‘From back when the marechaussee had something better to do’

*Alsof die **iets beters** te doen had*

‘Like she had something better to do’

Ik heb [je berichtje] [verkeerd geïnterpreteerd.]

‘I have misinterpreted your text message.’

Ik heb [iets] [verkeerd geïnterpreteerd.]

‘I have misinterpreted something.’

Adverb

Always without -s

Ik heb [iets leuk]_{PART GEN} bedacht.

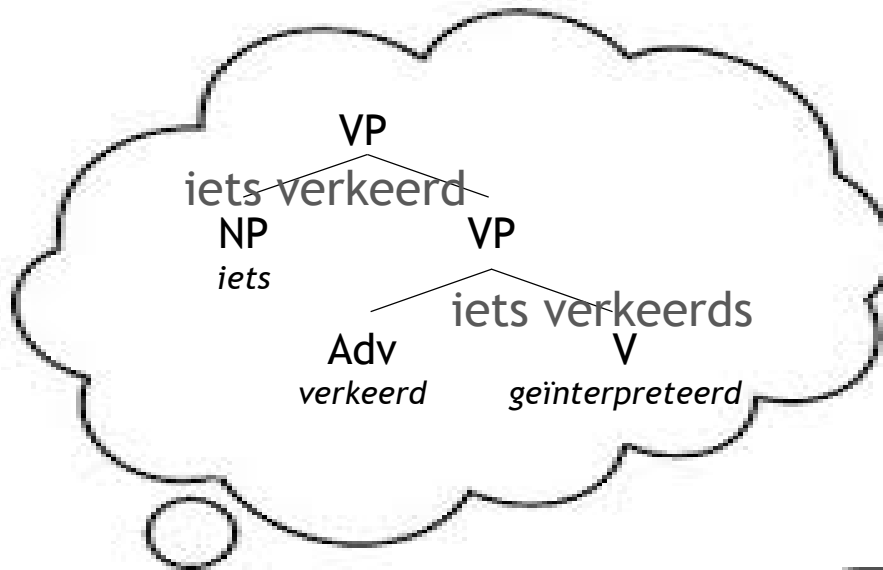
‘I have thought up something fun.’

Ik heb [iets verkeerd]_{PART GEN} gegeten.

‘I have eaten something wrong.’

Partitive genitive

With or without -s



Adverb
iets + verkeerd
appears without -s

Partitive genitive
iets verkeerd(s)
preference for variant without -s

iets verkeerd geïnterpreteerd
'misinterpreted something'

iets verkeerd gegeten
'eaten something wrong'



Partitive genitive
wat zinnig(s)
remains unaffected

wat zinnigs gehoord
'heard something sensible'

Constructional contamination

If language users do not always execute a full parse, then the frequent occurrence of the string *iets verkeerd* in a **different construction**, should cause them to prefer the variant without *-s* of *iets verkeerd(s)* in **the partitive genitive construction**.

How do we measure this?

1. In the contaminating construction, identify the strings that **superficially resemble** strings in the target construction
 - Manually: 1 way
 - Automatically: 3 ways

2. In the target construction, check whether even the **strictly unambiguous** occurrences of these strings are affected
 - Mixed effects regression

(Speelman 2014, Gries 2015)

Manually

- Extract all instances of *Quantifier + Adjective (s)* from ConDiv corpus
- Check all instances: are these partitive genitives?
- Identified color adjectives (*blue, red, green,...*) and assessment adjectives (*wrong, good, better, ...*)

- **Adverbs:**

*Voortaan de spelregels **iets beter** uitleggen.*

‘Next time, explain the rules of the game a bit better.’

- **Predicative constructions:**

*Is net **iets beter**.*

‘it’s just a little better’

- **Color nouns**

***Veel wit**, geïnspireerd op sportthema’s.*

‘a lot of the color white, inspired on sporting themes’

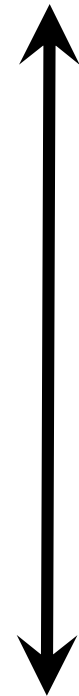
Manually

variable **Type-Adjective**

color adjectives, assessment adjectives, other adjectives

Automatically

- Partial String Resemblance
- String Resemblance
- Semantic String Resemblance



Little resemblance

A lot of resemblance

Evaluation

- Blind the dataset for -s occurrence
- Throw out any occurrence that has a sniff of ambiguity

*Bang dat ze **iets verkeerd** zullen doen.*

‘They are scared to do something wrong’

‘They are scared to do something wrongly’

⇒ 2700 strictly unambiguous partitive genitives

Evaluation

BASE MODEL: *Variety, Register, Quantifier, Frequency* + Random Effect *Phrase*

+ *Type-Adjective*

Estimated
probability

[+ ∅]

Type-Adjective

- 0.8 -

- 0.6 -

- 0.4 -

- 0.2 -

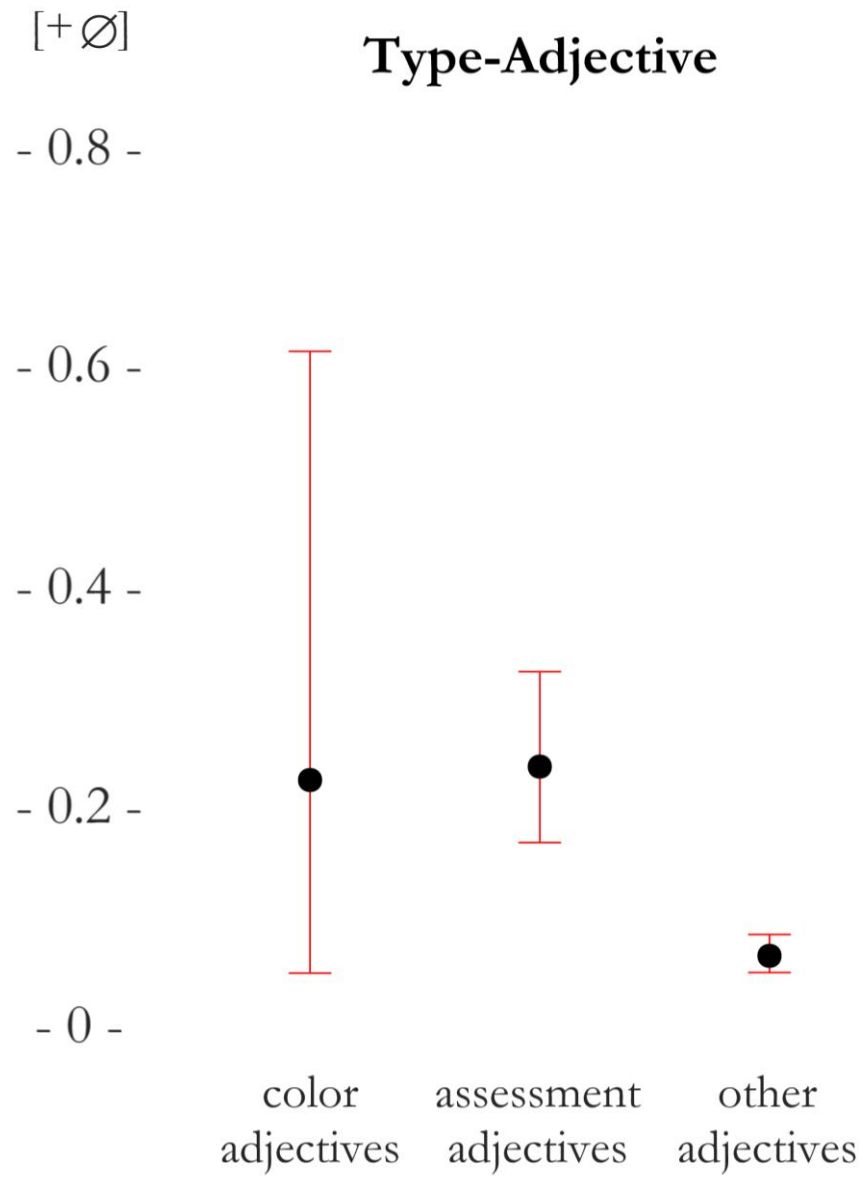
- 0 -

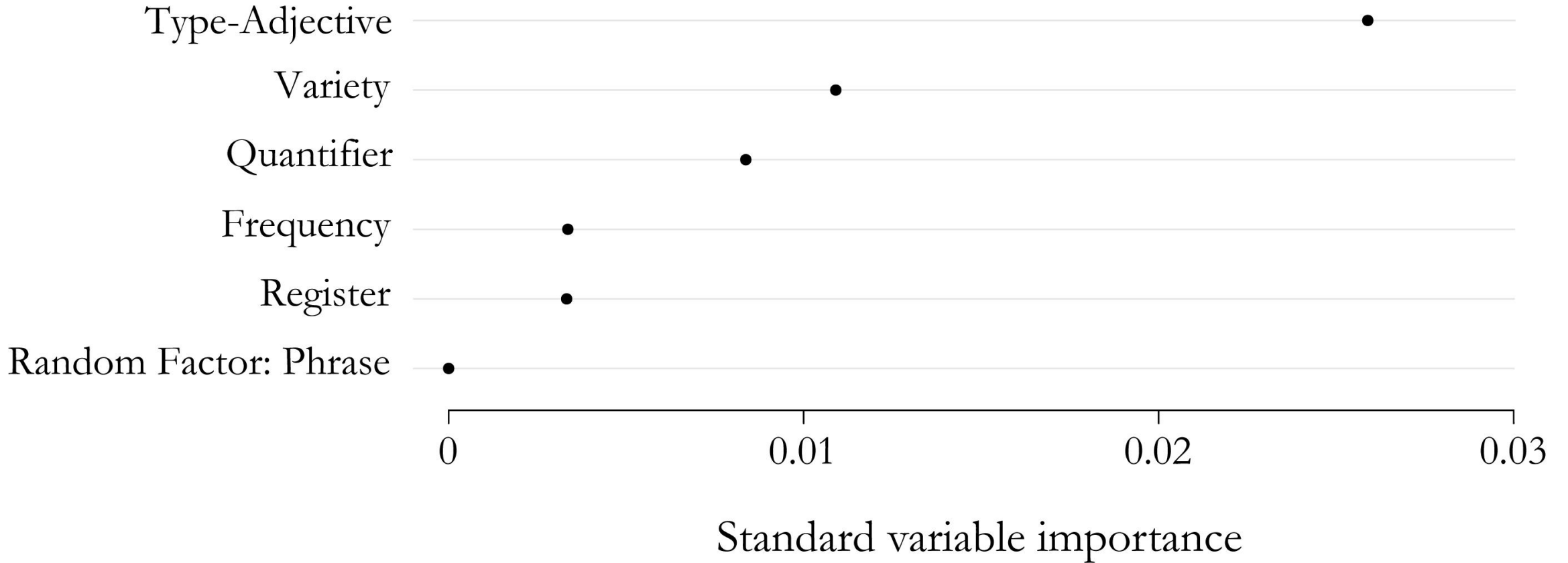
color
adjectives

assessment
adjectives

other
adjectives

[+ s]

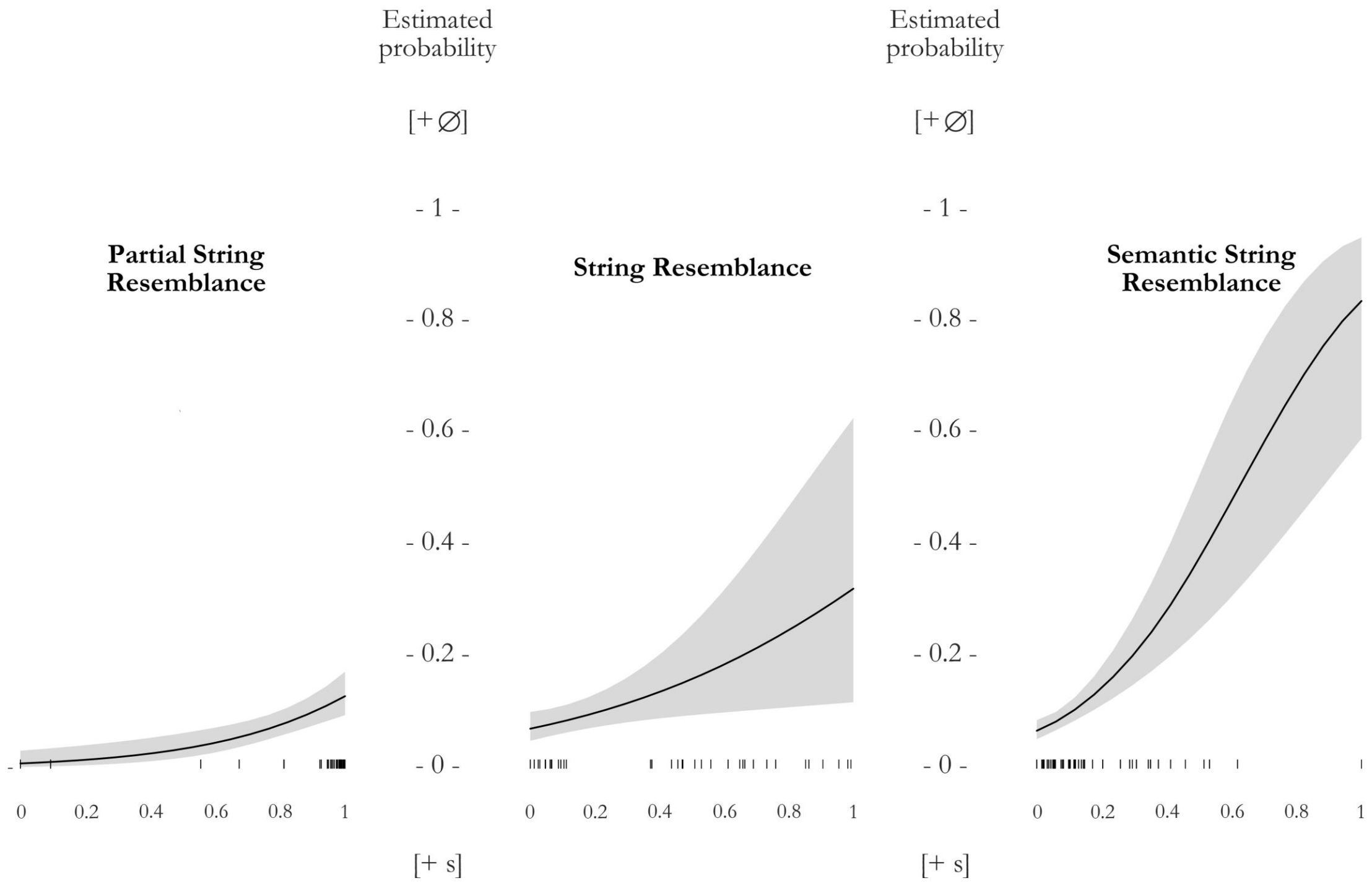




Evaluation

BASE MODEL: *Variety, Register, Quantifier, Frequency* + Random Effect *Phrase*

- MODEL 2: + *Partial String Resemblance*
- MODEL 3: + *String Resemblance*
- MODEL 4: + *Semantic String Resemblance*



- If two unrelated constructions have a number of instances that **superficially resemble** each other, we expect constructional contamination between the constructions.
- Constructional contamination may **affect even strictly unambiguous instances**.
- The explanation for this effect follows naturally from an **exemplar-based view of language processing**.

How do we measure this?

- Manually dig through your data
- (Semi-)automatically: Semantic String Resemblance

Thanks!

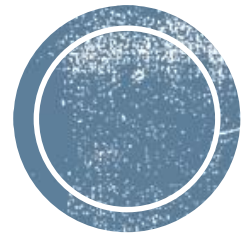
Pijpops, Dirk & Freek Van de Velde. Forthcoming. 'Constructional contamination: How does it work and how do we measure it?' *Folia Linguistica. Special Issue*. 50(2)

dirk.pijpops@kuleuven.be

freek.vandevelde@kuleuven.be

References

- Dąbrowska, Ewa. 2014. Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics* 25(4). 617-653.
- Diessel, Holger. 2015. Usage-based construction grammar. In Ewa Dąbrowska & Dagmar Divjak (eds.), *Handboek of Cognitive Linguistics*, 296-321. Berlin: De Gruyter Mouton.
- Ferreira, Fernanda and Nikole Patson. 2007. The "good enough" approach to language comprehension. *Language and Linguistics Compass* 1. 71-83.
- Gries, Stefan Thomas. 2015. The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10(1). 95-125.
- Grondelaers, Stefan, Katrien Deygers, Hilde Van Aken, Vicky Van den Heede and Dirk Speelman. 2000. Het CONDIV-corpus geschreven Nederlands [The CONDIV-corpus of written Dutch]. *Nederlandse Taalkunde* 5(4). 356-363.
- Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij and Maarten van den Toorn. 1997. *Algemene Nederlandse Spraakkunst [General Dutch Grammar]*. Groningen: Nijhoff.
- Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat and Harald Baayen. 2002. Experiences from the Spoken Dutch corpus project.
- Pijpops, Dirk and Freek Van de Velde. Constructional contamination: How does it work and how do we measure it? *Folia Linguistica*.
- Pijpops, Dirk and Freek Van de Velde. 2015. Ethnolect speakers and Dutch partitive adjectival inflection. A corpus analysis. *Taal en Tongval* 67(2). 343-371.
- Pijpops, Dirk and Freek Van de Velde. 2014. A multivariate analysis of the partitive genitive in Dutch. Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory*. Published online, ahead of print.
- Speelman, Dirk. 2014. Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In Dylan Glynn & Justyna A. Robinson (eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, 487-533. (Human Cognitive Processing [HCP]). Amsterdam: John Benjamins.
- Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(307).
- Van de Velde, Freek. 2001. *Iets taalkundig(s): een functioneel georiënteerde analyse van deflexie en de genitiefontwikkeling in het Nederlands [Something linguistic: a functionally oriented analysis of deflexion and the development of the genitive in Dutch]*. Leuven: University of Leuven MA thesis.



Extra slides

Akaike Information Criterion (AIC)

▪ Base model	1818	
<i>Variety, Register, Quantifier, Frequency</i>		
Random Effect <i>Phrase</i>		
▪ Base model + <i>Type Adjective</i>	- 22	p < 0.0001
▪ Base model + <i>Partial String Resemblance</i>	- 15	p = 0.0002
▪ Base model + <i>String Resemblance</i>	- 4	p = 0.0159
▪ Basic model + <i>Semantic String Resemblance</i>	- 22	p < 0.0001



2700 strictly unambiguous partitive genitives

*Als ik **iets verkeerd** gegeten heb, heb ik buikpijn*

Partitive genitive: 'If I have eaten something wrong, I have a stomach ache.'

Adverb: 'If I have eaten something in the wrong way, I have stomach ache.'





Preference for the [+ ø] variant

Total number of occurrences: 2388

Adjective	[+ ø] occ.	[+ s] occ.	Collostr. strength
<i>verkeerd</i> ‘wrong’	150	76	53.48
<i>groen</i> ‘green’	41	0	28.35
<i>goed</i> ‘good’	75	167	4.13
<i>wit</i> ‘white’	7	1	3.96
<i>geel</i> ‘yellow’	4	0	2.72
<i>beter</i> ‘better’	62	152	2.65
<i>blauw</i> ‘blue’	4	1	2.10
<i>zwart</i> ‘black’	4	1	2.10
<i>apart</i> ‘separate’	8	11	1.53
<i>fout</i> ‘incorrect’	2	0	1.36
<i>oranje</i> ‘orange’	2	0	1.36
<i>deftig</i> ‘decent’	9	17	1.13
<i>raar</i> ‘weird’	11	27	0.82
<i>rood</i> ‘red’	2	2	0.71
<i>gemakkelijk</i> ‘easy’	1	0	0.68
<i>warm</i> ‘warm’	3	5	0.65
<i>speciaal</i> ‘special’	35	115	0.60
<i>interessant</i> ‘interesting’	29	98	0.49

Preference for the [+ s] variant

Total number of occurrences: 630

Adjective	[+ ø] occ.	[+ s] occ.	Collostr. strength
<i>dergelijk</i> ‘similar’	3	183	15.18
<i>leuk</i> ‘fun’	23	331	14.53
<i>nieuw</i> ‘new’	38	377	11.15
<i>bijzonder</i> ‘extraordinary’	2	101	8.05
<i>mooi</i> ‘beautiful’	11	116	3.86
<i>zinnig</i> ‘sensible’	28	163	1.81
<i>lekker</i> ‘tasty’	10	73	1.59
<i>gek</i> ‘crazy’	0	14	1.43
<i>nuttig</i> ‘useful’	22	124	1.35
<i>vreemd</i> ‘weird’	4	33	1.05
<i>positief</i> ‘positive’	8	47	0.80
<i>concreet</i> ‘concrete’	8	40	0.52
<i>spannend</i> ‘exciting’	7	33	0.42
<i>klein</i> ‘small’	1	8	0.39
<i>erg</i> ‘awful’	6	25	0.28
<i>aardig</i> ‘nice’	2	10	0.28
<i>verschrikkelijk</i> ‘horrible’	1	6	0.26
<i>belangrijk</i> ‘important’	7	27	0.23
<i>gestreept</i> ‘striped’	0	1	0.10

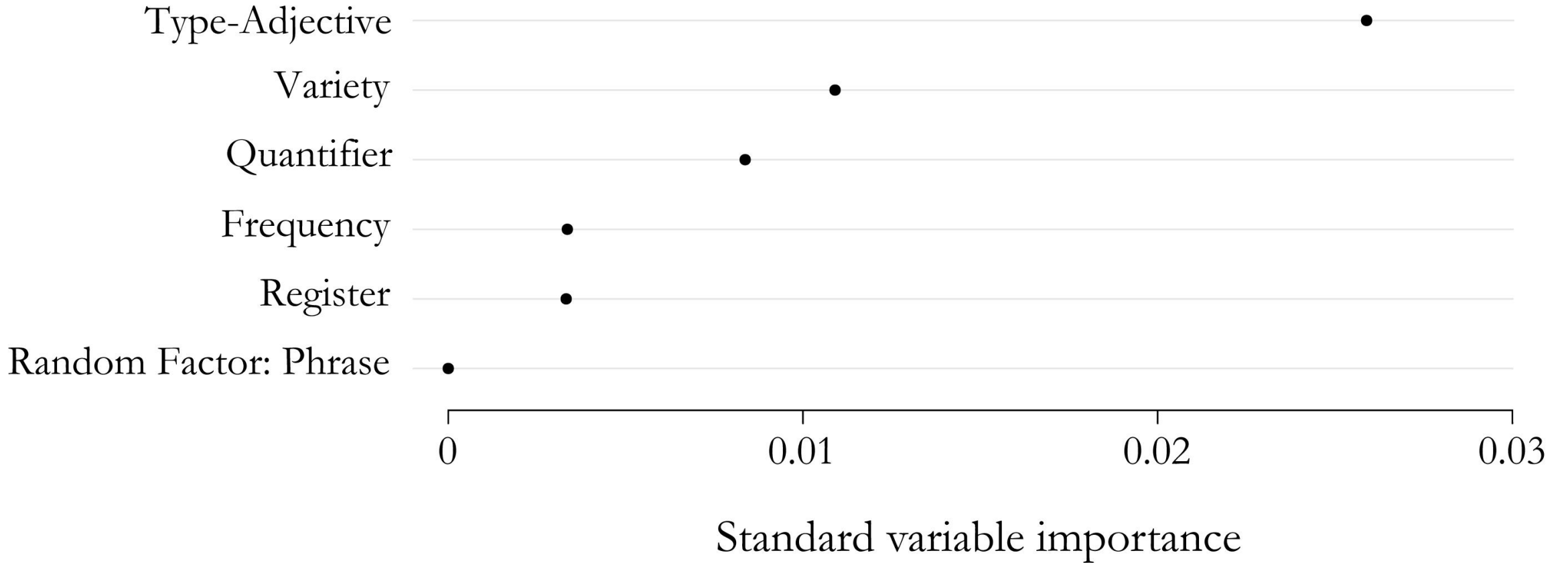
Extraction of data

Quantifiers

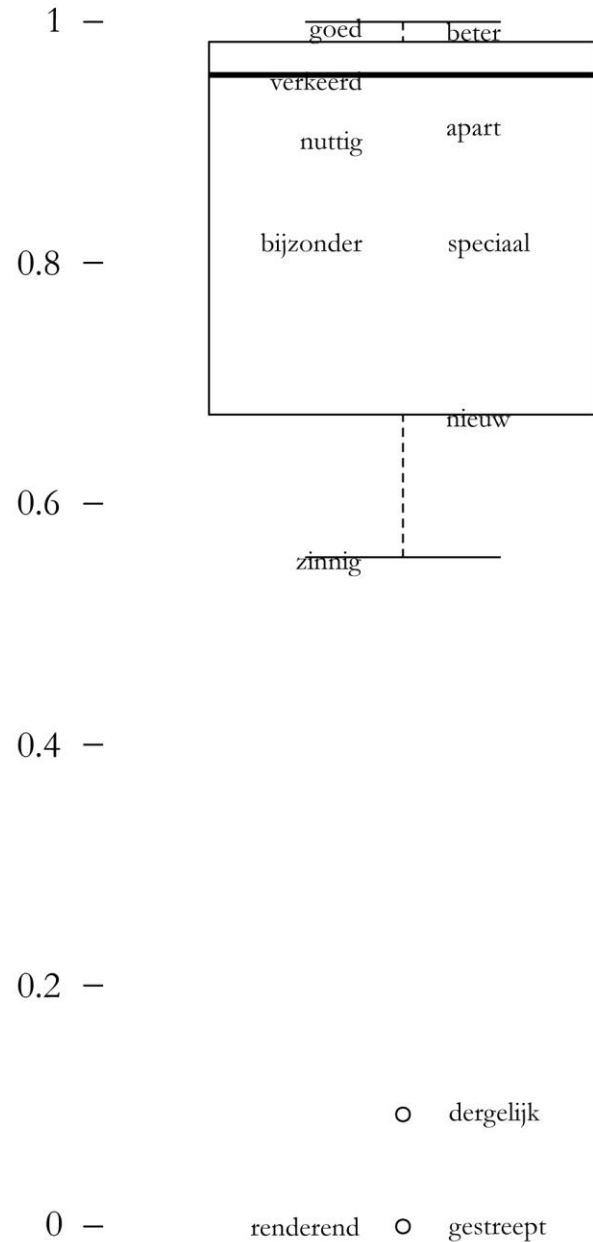
- Listed as indefinite pronoun or numeral in Haeseryn et al. (1997, p.356, 432)
- Occur 14x in a partitive genitive in the Corpus of Spoken Dutch (CGN)
- Not *iemand* or *niemand*

Adjectives

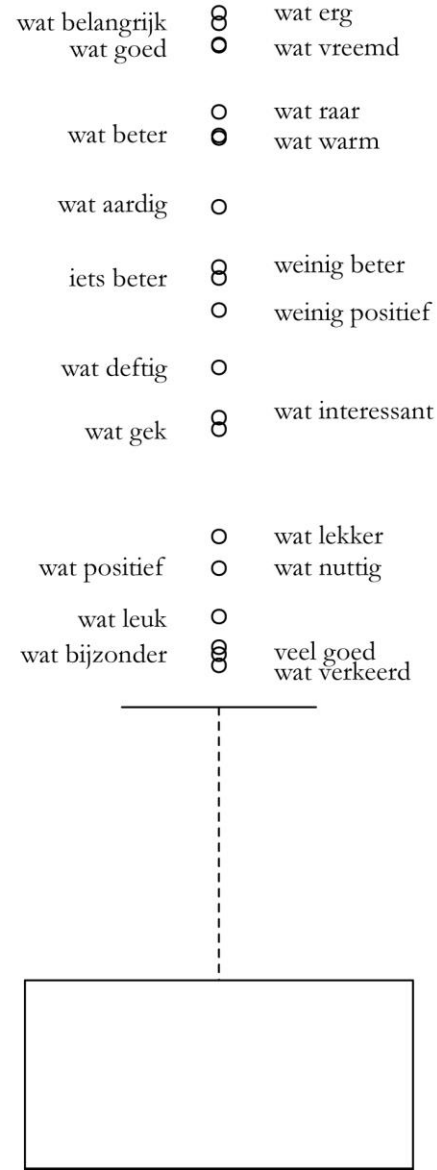
- Occur 7x with any of the selected quantifiers in a partitive genitive in the CGN
- Not homographic with the plural form of a noun, e.g. veel ouders, veel extra's
- + color adjectives, *beter* (Van de Velde 2001)



Partial String Resemblance



String Resemblance



Semantic String Resemblance



Partitive Genitive

ik heb iets leuk bedacht

‘I have thought up something fun’

ik heb veel interessants gehoord

‘I have heard a lot of interesting things’

ik heb iets verkeerd gegeten

‘I have eaten something wrong’

ik heb wat ergs meegemaakt

‘I have experienced something bad’

Adverbs

ik heb herhaaldelijk gebeld

‘I have called repeatedly’

ik heb dat gebaar verkeerd begrepen

‘I have misunderstood that gesture’

ik heb iets verkeerd geïnterpreteerd

‘I have misinterpreted something’

