A quantitative measure of Constructional Contamination

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

[1]QLVL, University of Leuven

[2]Research Foundation Flanders

Theory

- 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

Methodology

- How can you measure this effect in your own corpus study?

Case study: partitive genitive

iets leuk(s) [Quantifier Adjective (-s)]$_{NP}$

'something fun'

*Ik heb **iets leuk** bedacht*

*Ik heb **iets leuks** bedacht*.

'I have thought up something fun.'

Ik heb [*iets leuk*]$_{\text{PART GEN}}$ *bedacht.*

‘I have thought up something fun.’

Ik heb [*iets verkeerd*]$_{\text{PART GEN}}$ *gegeten.*

‘I have eaten something wrong.’

Ik heb [*iets*][*verkeerd*]$_{\text{PART GEN}}$ *geïnterpreteerd.*]

Always without *-s*

Adverb

‘I have interpreted something wrong.’

‘I have misinterpreted something.’

Quantifier + Adverb
iets verkeerd
appears without *-s*

iets verkeerd geïnterpreteerd

'misinterpreted something'

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

iets verkeerd gegeten

'eaten something wrong'

Partitive genitive
wat zinnig(s)
remains unaffected

wat zinnigs gehoord

'heared 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*, …)

(Grondelaers et al. 2002)

- **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

Partial String Resemblance

$$\frac{number\ of\ times\ the\ adjective\ occurs\ in\ its\ bare\ form\ outside\ the\ partitive\ genitive}{(number\ of\ times\ the\ adjective\ occurs\ in\ its\ bare\ form\ outside\ the\ partitive\ genitive + number\ of\ times\ it\ occurs\ in\ the\ partitive\ genitive)}$$

(Corpus of Spoken Dutch, Oostdijk et al. 2002)

String Resemblance

$$\frac{number\ of\ times\ the\ quantifier - adjective\ occurs\ outside\ the\ partitive\ genitive}{(number\ of\ times\ the\ quantifier - adjective\ occurs\ outside\ the\ partitive\ genitive + number\ of\ times\ it\ appears\ in\ the\ partitive\ genitive)}$$

Semantic String Resemblance

$$\frac{number\ of\ times\ the\ quantifier - adjective\ occurs\ in\ an\ ambiguous\ context}{(number\ of\ times\ the\ quantifier - adjective\ occurs\ in\ an\ ambiguous\ context + number\ of\ times\ it\ occurs\ in\ an\ unambiguous\ partitive\ genitive)}$$

Evaluation

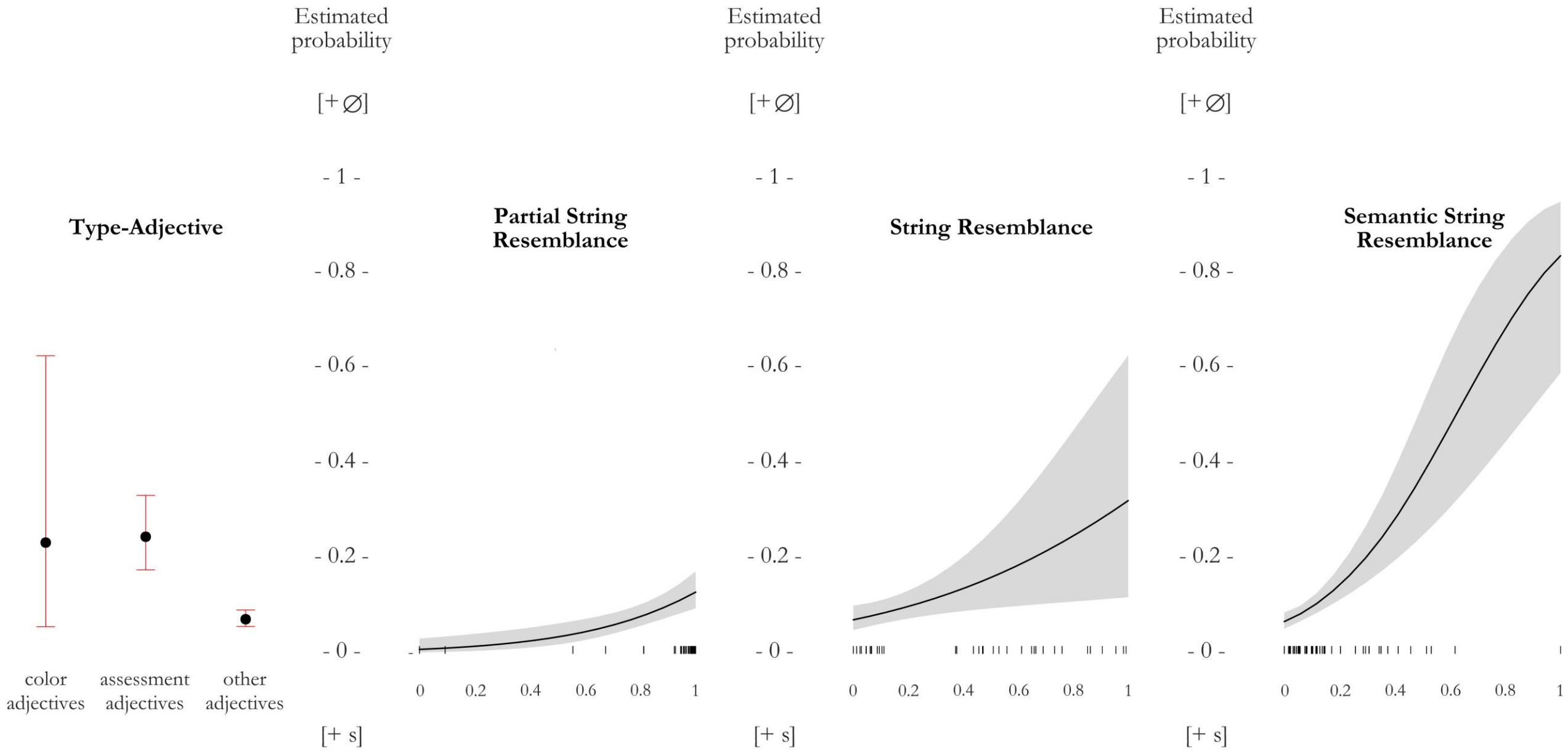
- Blind the dataset for *-s* occurrence
- Throw out any occurrence that has a sniff of ambiguity

> *Bang dat ze **iets verkeerd** zullen doen.*
>
> ‘They are scared to do something wrong’
>
> ‘They are scared to do something wrongly’

Evaluation

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

- MODEL 1: + *Type-Adjective*
- MODEL 2: + *Partial String Resemblance*
- MODEL 3: + *String Resemblance*
- MODEL 4: + *Semantic String Resemblance*


Estimated probability
[+∅]
- 1 -
- 0.8 -
- 0.6 -
- 0.4 -
- 0.2 -
- 0 -
Type-Adjective
color adjectives
assessment adjectives
other adjectives
Partial String Resemblance
String Resemblance
Semantic String Resemblance
0
0.2
0.4
0.6
0.8
1
[+ s]

	Akaike Information Criterion (AIC)	
▪ Base model	1818	
Variety, Register, Quantifier, Frequency		
Random Effect *Phrase*		
▪ Base model + *Type Adjective*	- 22	p < 0.0001
▪ Base model + *Partial String Resemblance*	- 15	p = 0.0002
▪ Base model + *String Resemblance*	- 4	p = 0.0159
▪ Basic model + *Semantic String Resemblance*	- 22	p < 0.0001

- Between the contaminating and the target construction, there should be a bridge formed by **ambiguous occurrences**.

- Once that bridge is in place, constructional contamination may **affect even strictly unambiguous instances**.

- The explanation for this effect follows naturally from an **exemplar-based view of language processing**.

Theory

- 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

Methodology

- How can you measure this effect in your own corpus study?
  - Manually
  - (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: Univerisity of Leuven MA thesis.

Extra slides

Preference for the [+ ø] variant Total number of occurrences: 2388			
Adjective	[+ Ø] occ.	[+ s] occ.	Collostr. strength
verkeerd ‘wrong’	150	76	53.48
groen ‘green’	41	0	28.35
goed ‘good’	75	167	4.13
wit ‘white’	7	1	3.96
geel ‘yellow’	4	0	2.72
beter ‘better’	62	152	2.65
blauw ‘blue’	4	1	2.10
zwart ‘black’	4	1	2.10
apart ‘separate’	8	11	1.53
fout ‘incorrect’	2	0	1.36
oranje ‘orange’	2	0	1.36
deftig ‘decent’	9	17	1.13
raar ‘weird’	11	27	0.82
rood ‘red’	2	2	0.71
gemakkelijk ‘easy’	1	0	0.68
warm ‘warm’	3	5	0.65
speciaal ‘special’	35	115	0.60
interessant ‘interesting’	29	98	0.49

Preference for the [+ s] variant Total number of occurrences: 630			
Adjective	[+ Ø] occ.	[+ s] occ.	Collostr. strength
dergelijk ‘similar’	3	183	15.18
leuk ‘fun’	23	331	14.53
nieuw ‘new’	38	377	11.15
bijzonder ‘extraordinary’	2	101	8.05
mooi ‘beautiful’	11	116	3.86
zinnig ‘sensible’	28	163	1.81
lekker ‘tasty’	10	73	1.59
gek ‘crazy’	0	14	1.43
nuttig ‘useful’	22	124	1.35
vreemd ‘weird’	4	33	1.05
positief ‘positive’	8	47	0.80
concreet ‘concrete’	8	40	0.52
spannend ‘exciting’	7	33	0.42
klein ‘small’	1	8	0.39
erg ‘awful’	6	25	0.28
aardig ‘nice’	2	10	0.28
verschrikkelijk ‘horrible’	1	6	0.26
belangrijk ‘important’	7	27	0.23
gestreept ‘striped’	0	1	0.10

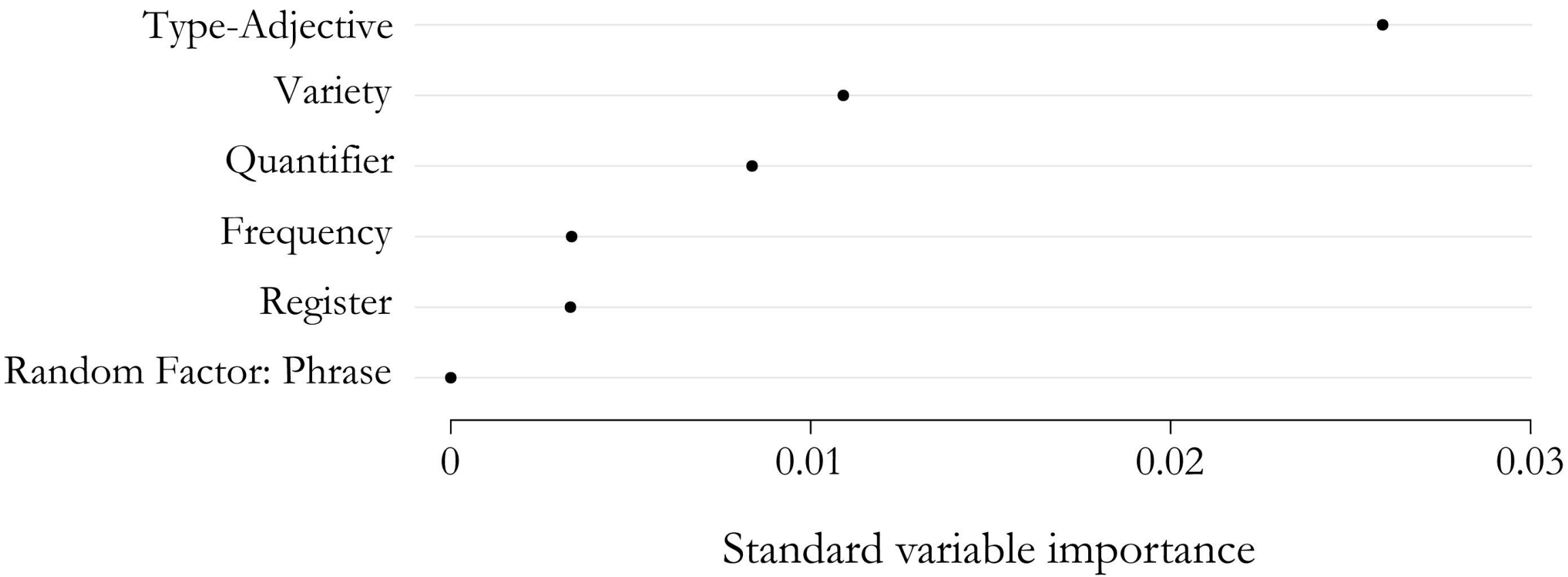
Extraction of data

Quantifiers

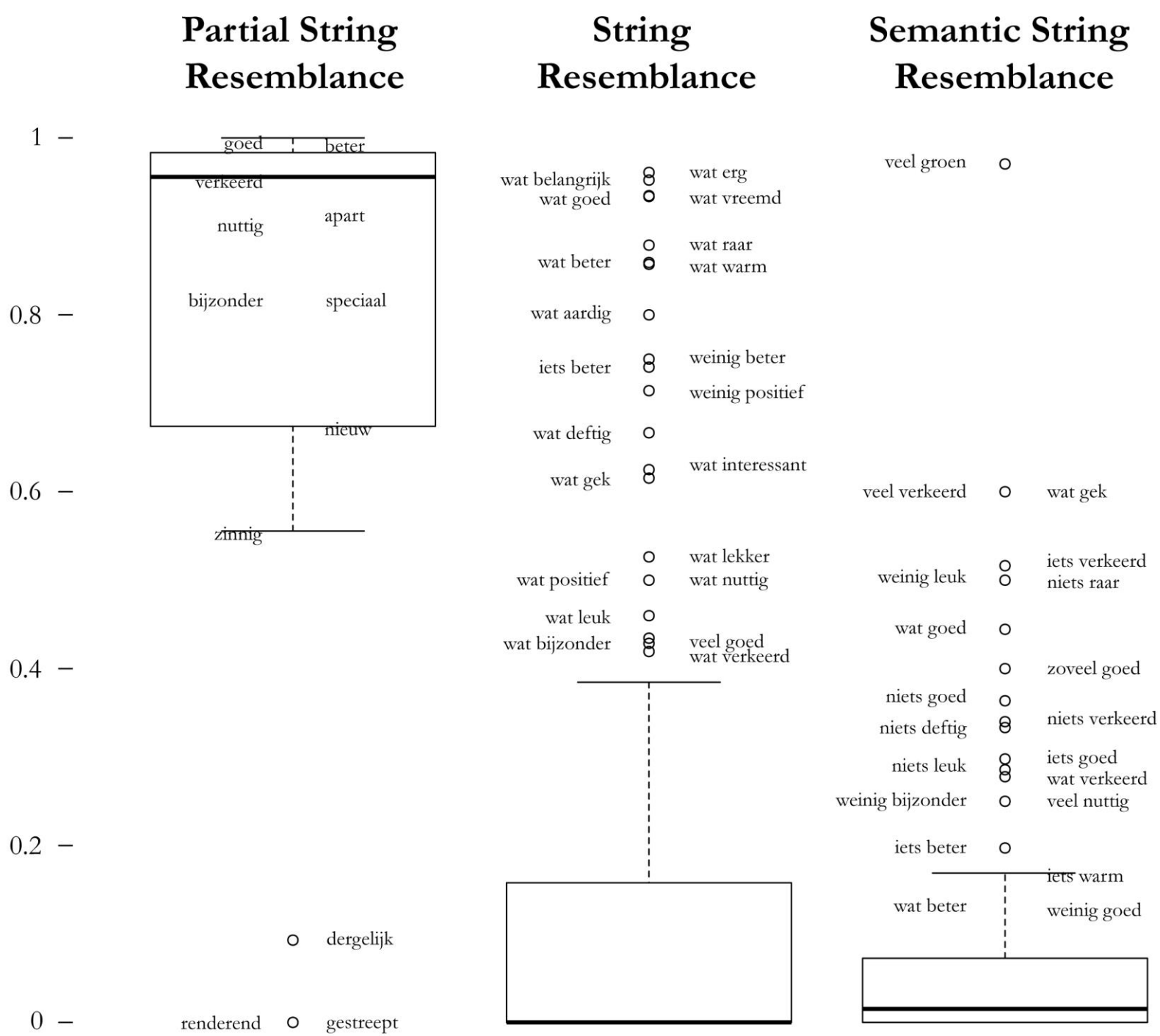
- Listed as indefinite pronoun or numeral in Haeseryn et al. (1997, p.356, 432)
- Occur 14x in a partitive genitive in the Corpus of Spoken Dutch (CGN)
- Not *iemand* or *niemand*

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)



Strobl et al. (2008)


Partial String Resemblance
String Resemblance
Semantic String Resemblance
1
0.8
0.6
0.4
0.2
0
goed
beter
verkeerd
nuttig
apart
bijzonder
speciaal
nieuw
zinnig
dergelijk
renderend
gestreept
wat belangrijk
wat goed
wat erg
wat vreemd
wat beter
wat raar
wat warm
wat aardig
iets beter
weinig beter
weinig positief
wat deftig
wat gek
wat interessant
wat lekker
wat positief
wat nuttig
wat leuk
wat bijzonder
veel goed
wat verkeerd
veel groen
veel verkeerd
wat gek
weinig leuk
iets verkeerd
niets raar
wat goed
zoveel goed
niets goed
niets deftig
niets verkeerd
niets leuk
iets goed
wat verkeerd
weinig bijzonder
veel nuttig
iets beter
iets warm
wat beter
weinig goed

2700 strictly unambiguous partitive genitives

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

‘If I have eaten something wrong, I have a stomach ache.’