Constructional contamination

An occasional rarity or a pervasive effect?

Dirk Pijpops, Isabeau De Smet & Freek Van de Velde

Research Foundation Flanders

QLVL, University of Leuven

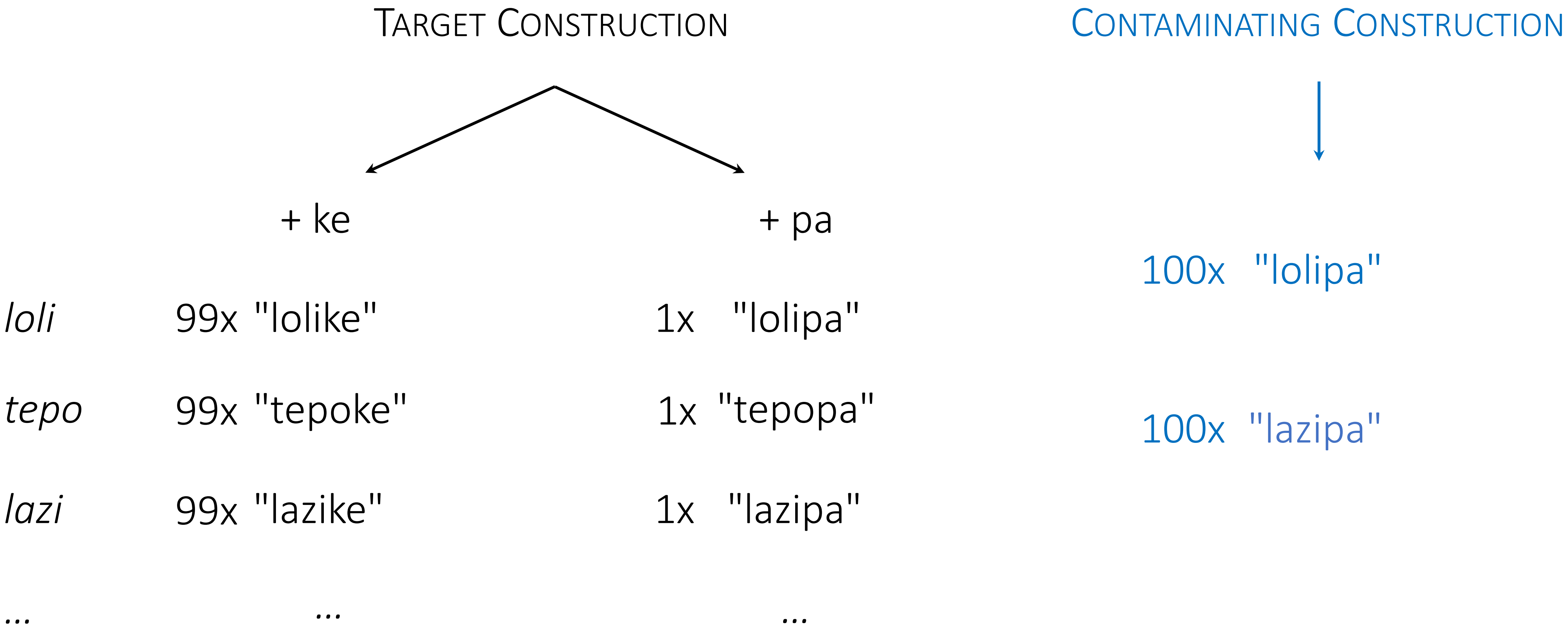
What is constructional contamination?

Is it real?

If so, is it an occasional rarity or a pervasive effect?

Constructional contamination

- Mechanism based on shallow parsing & storage of ready-mades
- Lexical preferences resulting from that mechanism



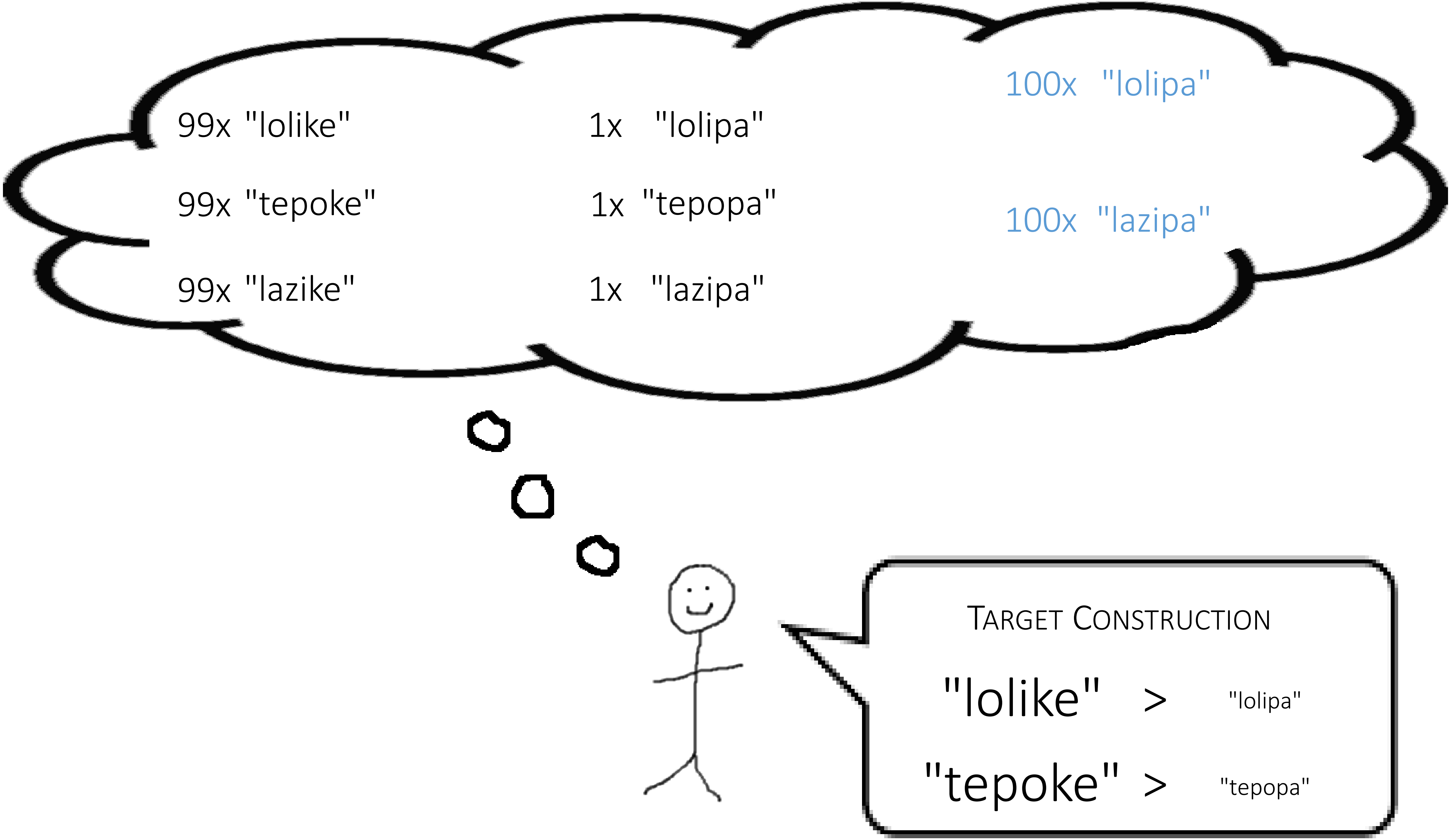
Target Construction

	+ ke	+ pa
loli	99x "lolike"	1x "lolipa"
tepo	99x "tepoke"	1x "tepopa"
lazi	99x "lazike"	1x "lazipa"
…	…	…

Contaminating Construction

100x "lolipa"

100x "lazipa"



99x "lolike" 1x "lolipa" 100x "lolipa"

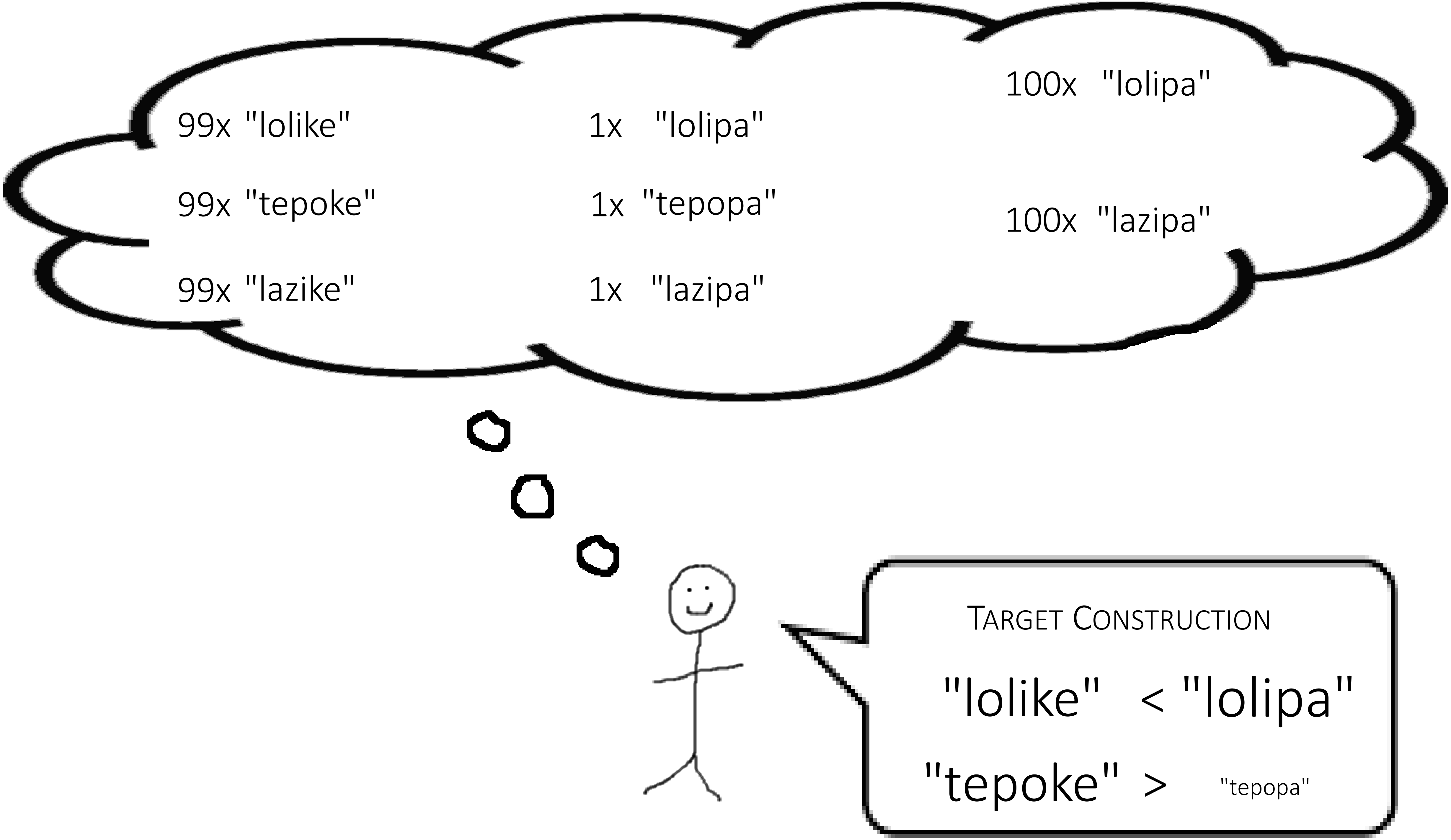
99x "tepoke" 1x "tepopa"

99x "lazike" 1x "lazipa" 100x "lazipa"

TARGET CONSTRUCTION

"lolike" > "lolipa"

"tepoke" > "tepopa"



99x "lolike"

99x "tepoke"

99x "lazike"

1x "lolipa"

1x "tepopa"

1x "lazipa"

100x "lolipa"

100x "lazipa"

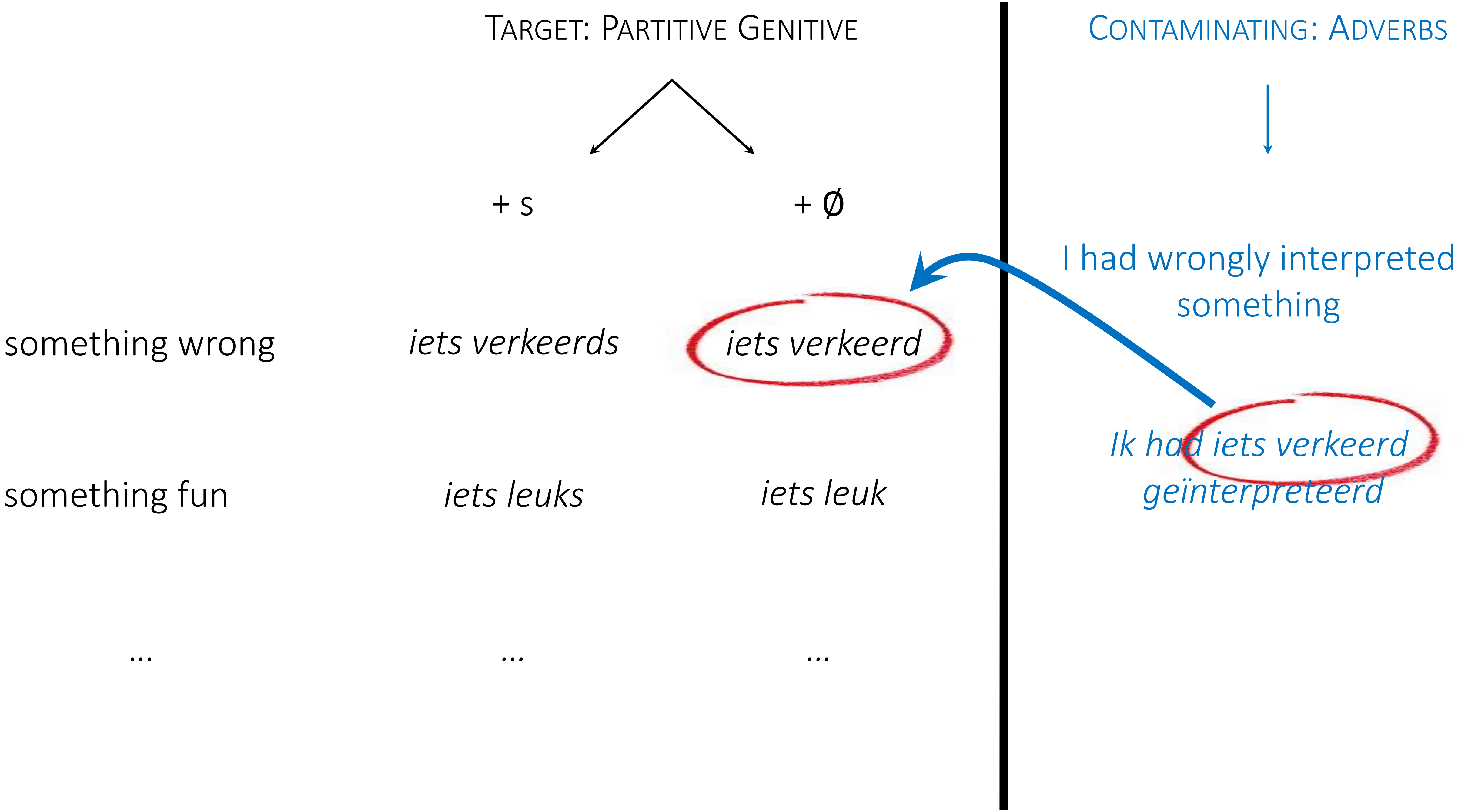
TARGET CONSTRUCTION

"lolike" < "lolipa"

"tepoke" > "tepopa"

Is it real?

Case study 1: partitive genitive



TARGET: PARTITIVE GENITIVE

	+ s	+ Ø
something wrong	*iets verkeerds*	*iets verkeerd*
something fun	*iets leuks*	*iets leuk*
…	…	…

CONTAMINATING: ADVERBS

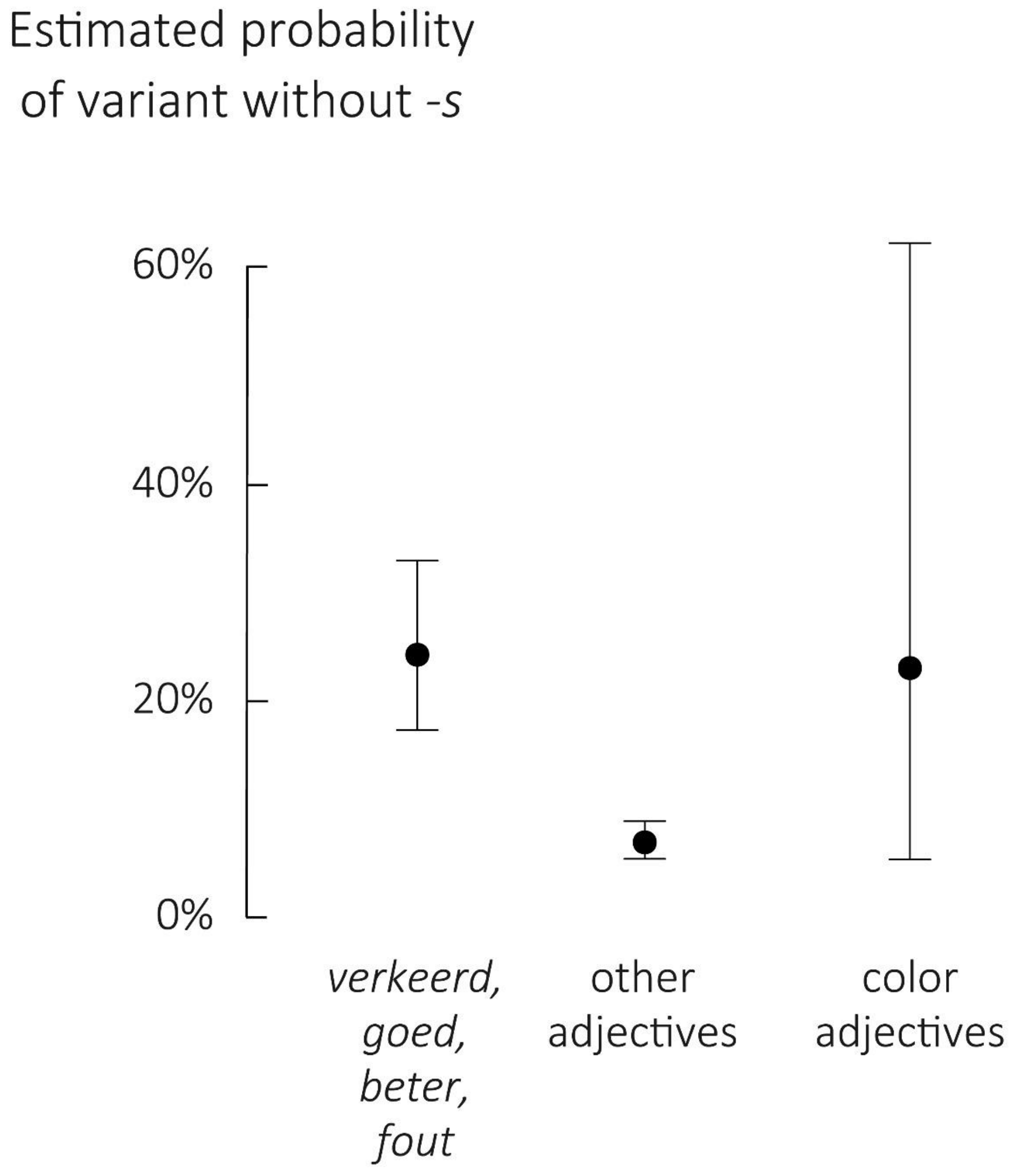
Ik had iets verkeerd geïnterpreteerd

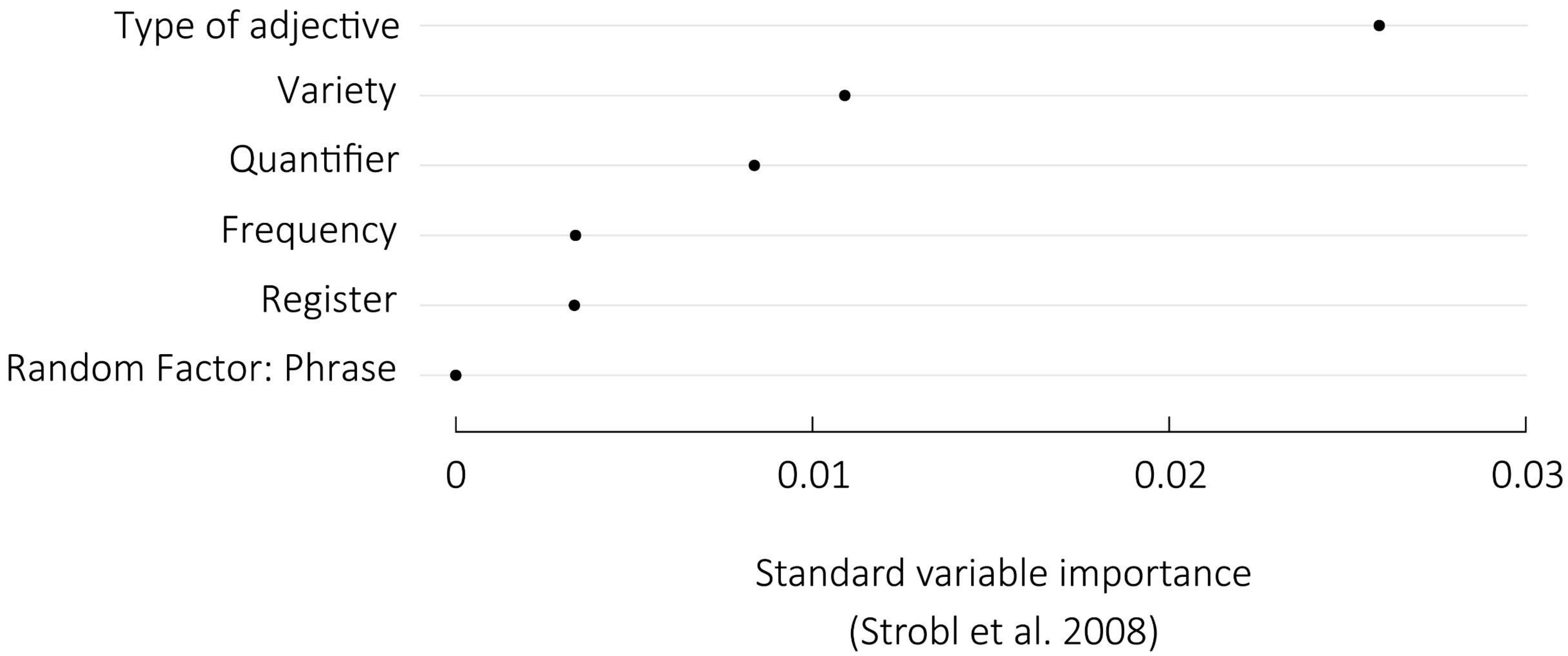
I had wrongly interpreted something

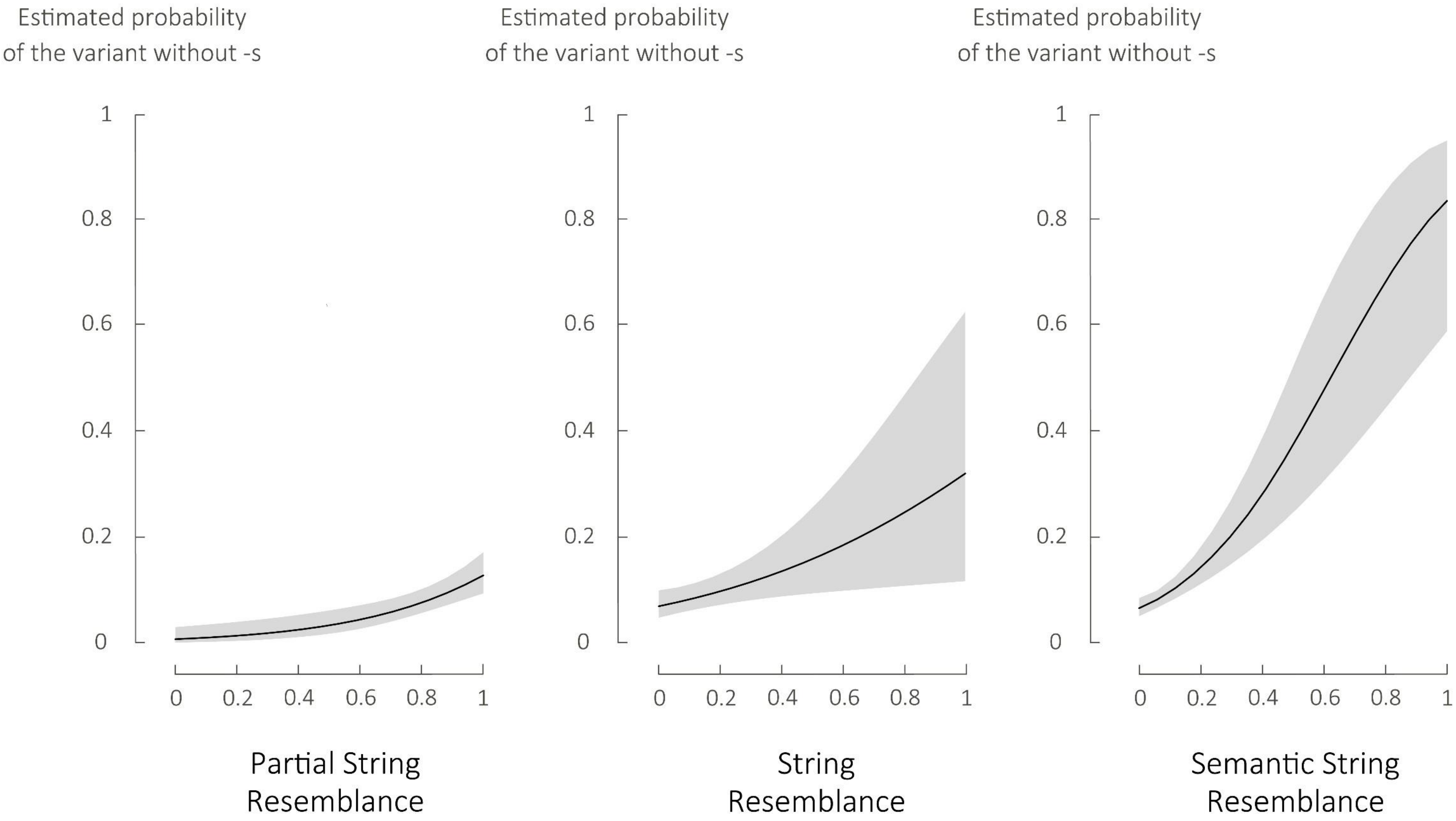
Case study 1: partitive genitive

- Prediction: among the partitive genitives, **the variant without -*s* will be much more dominant with adjectives that often appear as adverbs resembling partitive genitives** without -*s*, viz. *verkeerd* 'wrong', *goed* 'good', *beter* 'better' and *fout* 'incorrect'

- Only look at strictly unambiguous partitive genitives
- Mixed-effects regression model
- Control for all factors known to influence alternation and random lexical preferences




Type of adjective
Variety
Quantifier
Frequency
Register
Random Factor: Phrase
0
0.01
0.02
0.03
Standard variable importance
(Strobl et al. 2008)



Pijpops, Dirk & Freek Van de Velde. 2016. **Constructional contamination: How does it work and how do we measure it?** *Folia Linguistica* 50(2). 543–581.

So is it an occasional rarity or a pervasive effect?

Case study 2: verbal clusters

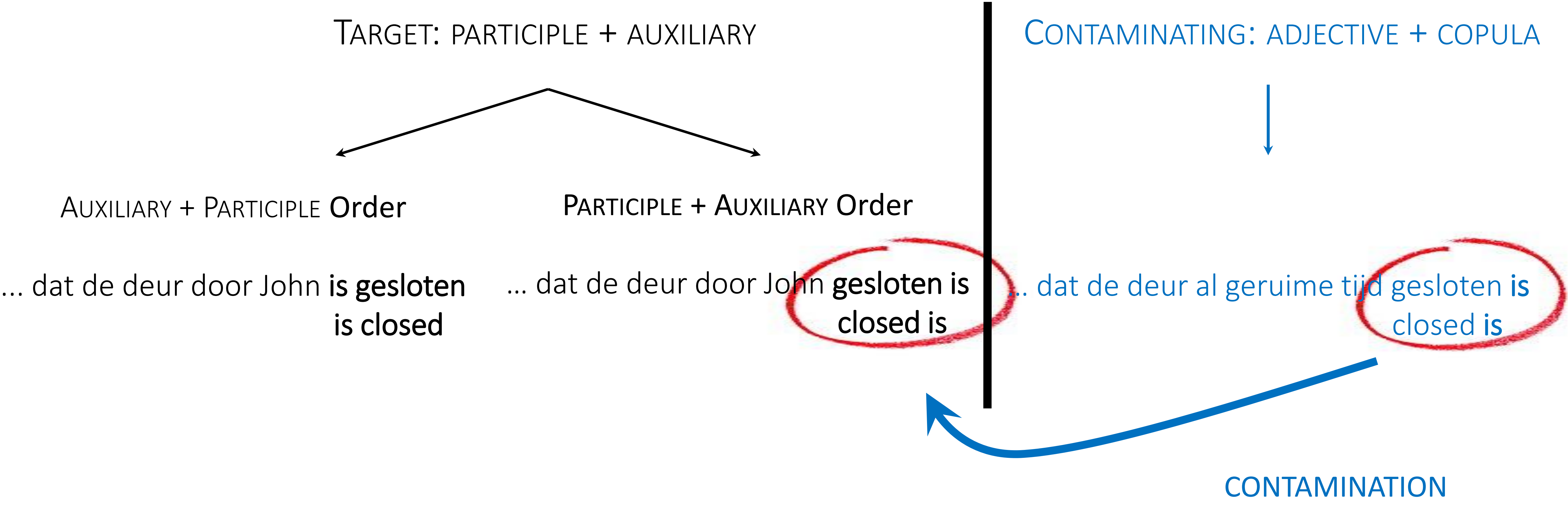
Case study 2: verbal clusters

De deur ***moet*** *door John* ***gesloten zijn.***

The door **must** by John **closed be**

… dat de deur door John ***gesloten is.***

… that the door by John **closed is**.



Target: participle + auxiliary

Auxiliary + Participle Order

... dat de deur door John **is gesloten**
is closed

Participle + Auxiliary Order

... dat de deur door John **gesloten is**
closed is

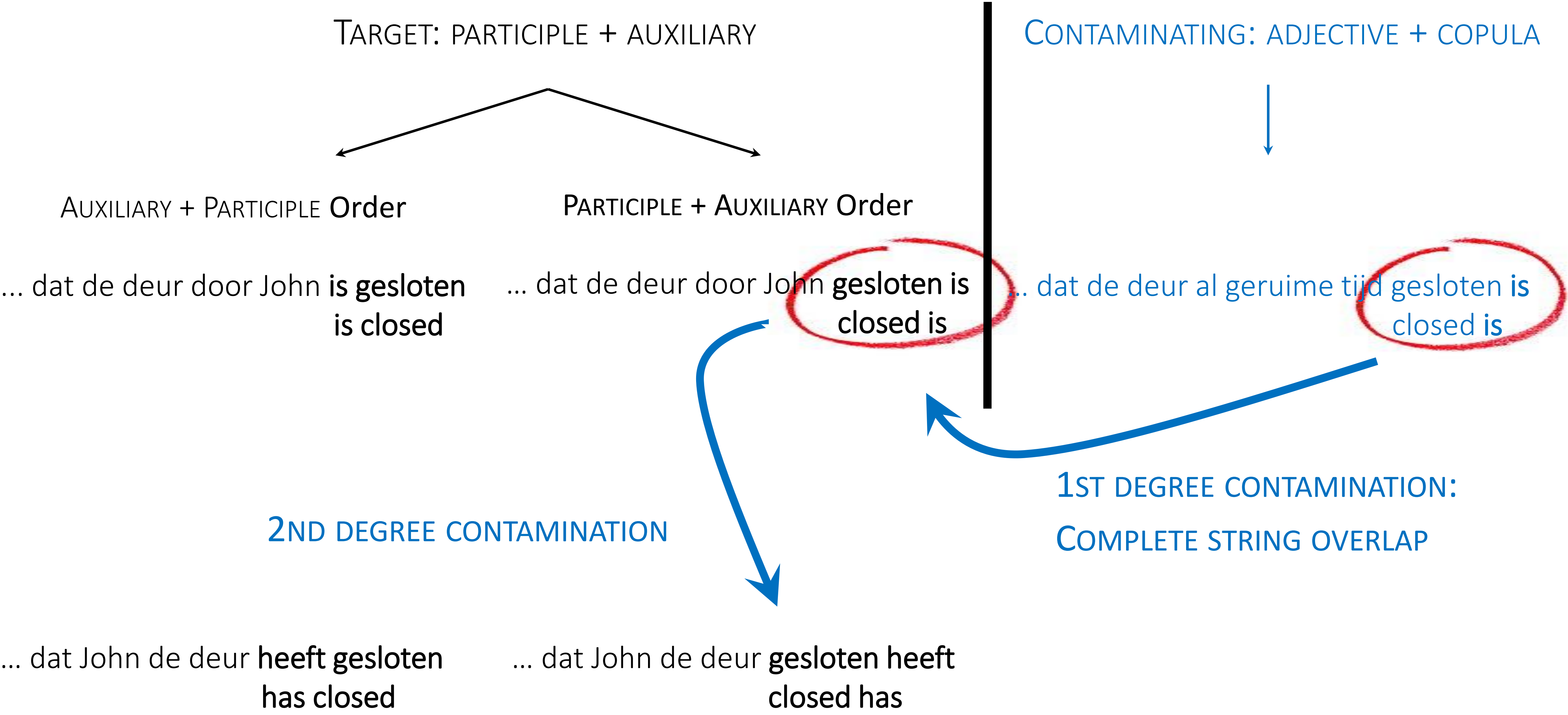
Contaminating: adjective + copula

... dat de deur al geruime tijd gesloten **is**
closed **is**

CONTAMINATION

- Prediction 1: The more often a participle is used as an adjective, the more often it will appear in the Participle + Auxiliary order in unambiguous verbal contexts

- Prediction 2: This effect will be stronger among the auxiliaries that can be used as copula, viz. *zijn* 'be' and *worden* 'become', and weaker among other auxiliaries, such as *hebben* 'have'



TARGET: PARTICIPLE + AUXILIARY

AUXILIARY + PARTICIPLE Order

... dat de deur door John **is gesloten**
is closed

PARTICIPLE + AUXILIARY Order

... dat de deur door John **gesloten is**
closed is

CONTAMINATING: ADJECTIVE + COPULA

... dat de deur al geruime tijd gesloten **is**
closed **is**

1ST DEGREE CONTAMINATION:
COMPLETE STRING OVERLAP

2ND DEGREE CONTAMINATION

... dat John de deur **heeft gesloten**
has closed

... dat John de deur **gesloten heeft**
closed has

Case study 2: verbal clusters

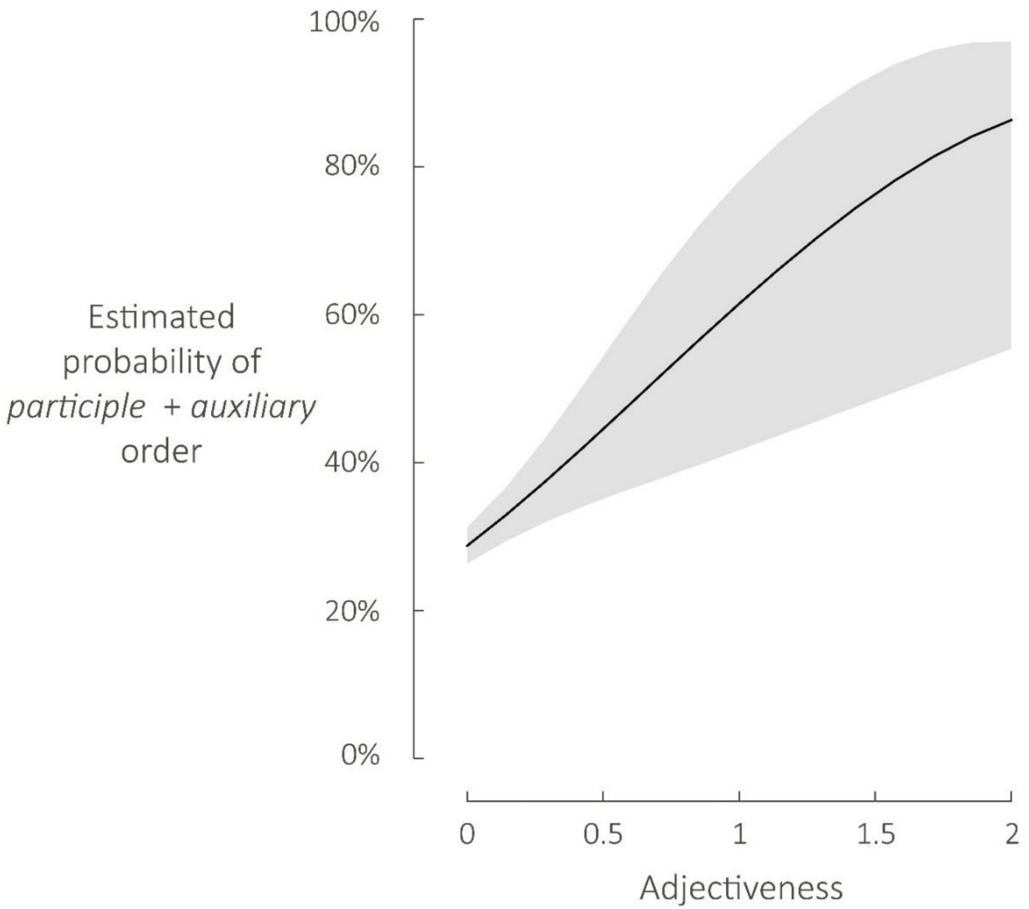
- PREDICTION 1: The more often a participle is used as an adjective, the more often it will appear in the PARTICIPLE + AUXILIARY order in unambiguous verbal contexts
- PREDICTION 2: This effect will be stronger among the auxiliaries that can be used as copula, viz. *zijn* 'be' and *worden* 'become', and weaker among other auxiliaries, such as *hebben* 'have'

Case study 2: verbal clusters

- Dataset from Gert De Sutter
- De Sutter distinguished between ambiguous & unambiguous verbal clusters
- Only looked at unambiguous verbal clusters
- Added variable $Adjectiveness = \ arsin(\sqrt{\frac{adjectival\ occurrences}{total\ occurrences}})$

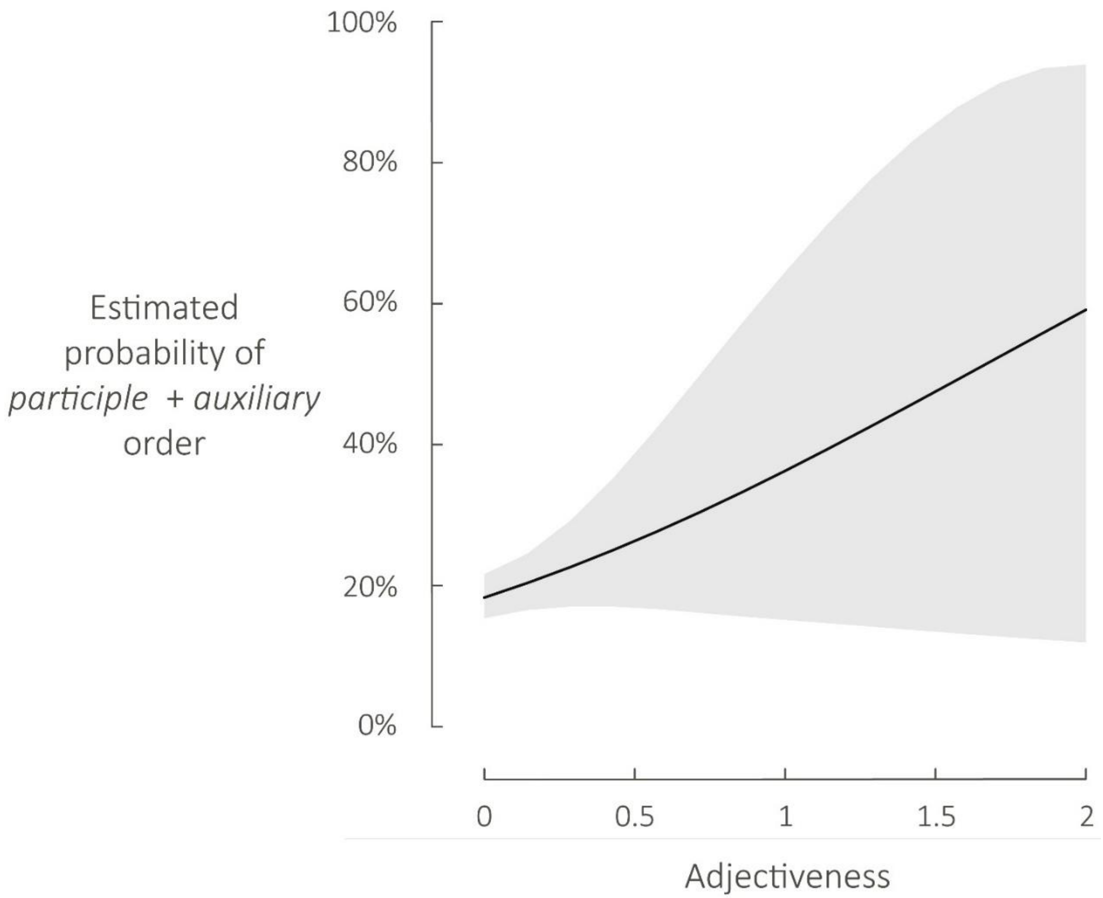
- Prediction 1: *Adjectiveness* will correlate positively with preference for the PARTICIPLE + AUXILIARY order
- Prediction 2: This effect will be stronger for auxiliaries *zijn* 'be' and *worden* 'become' than for *hebben* 'have'

zijn ‘be’ & *worden* ‘become’



>

hebben ‘have’



So is it an occasional rarity or a pervasive effect?

Case study 3: weak vs. strong preterites

Case study 3: weak vs. strong preterites

- Germanic languages: two morphological strategies to form preterite
  - strong inflection
    - vowel change (‘ablaut’)
    - ***zwem-zwom*** (‘swim’ – ‘swam’)
  - weak inflection
    - dental suffix
    - ***speel-speelde*** (‘play’ – ‘played’)

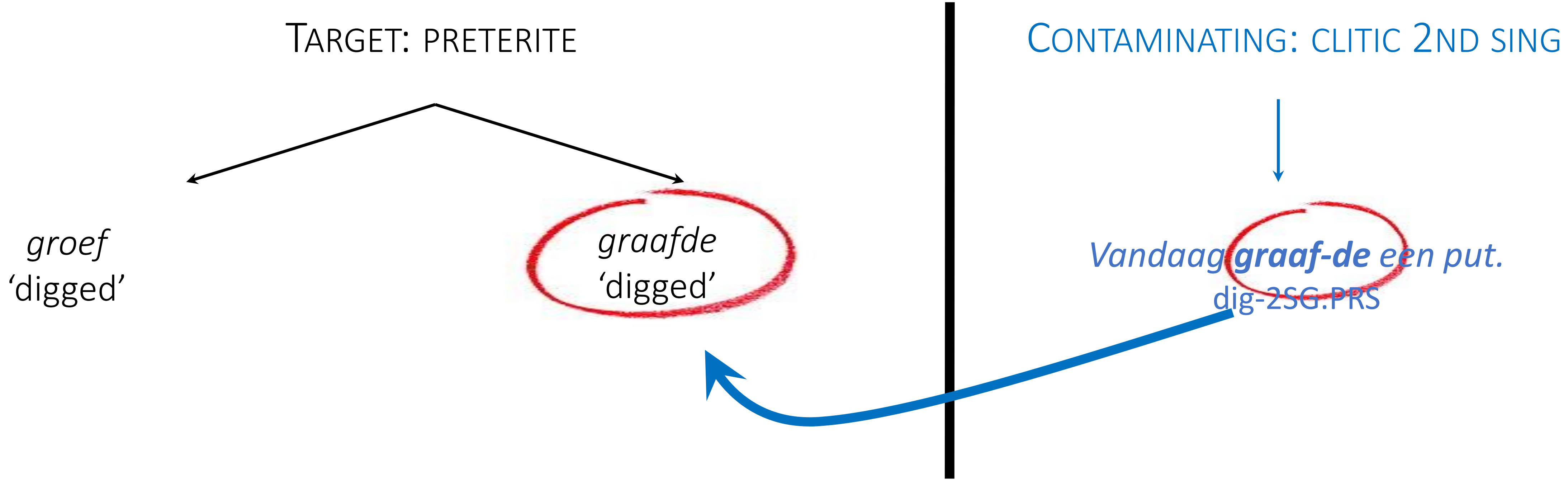
Case study 3: weak vs. strong preterites

- Contaminating construction: clitic realization of the 2nd person singular subject pronoun (cfr. Vosters 2012)

  Vandaag ***graaf-de*** *een put.* (Vosters 2012: 242)

  Today dig-2SG.PRS a hole

  ‘You will dig a hole today.’



TARGET: PRETERITE

groef
‘digged’

graafde
‘digged’

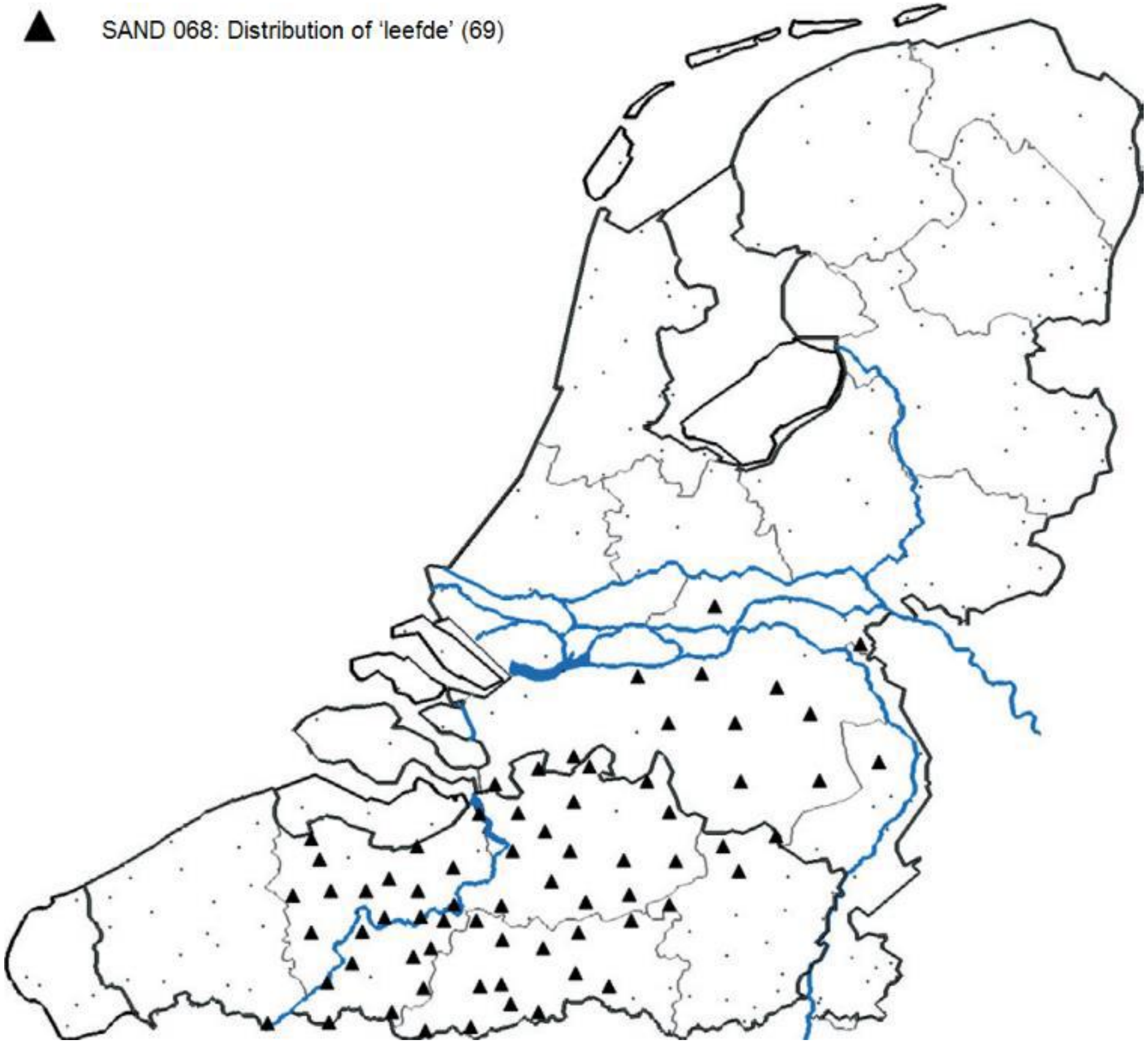
CONTAMINATING: CLITIC 2ND SING

*Vandaag **graaf-de** een put.*
dig-2SG.PRS

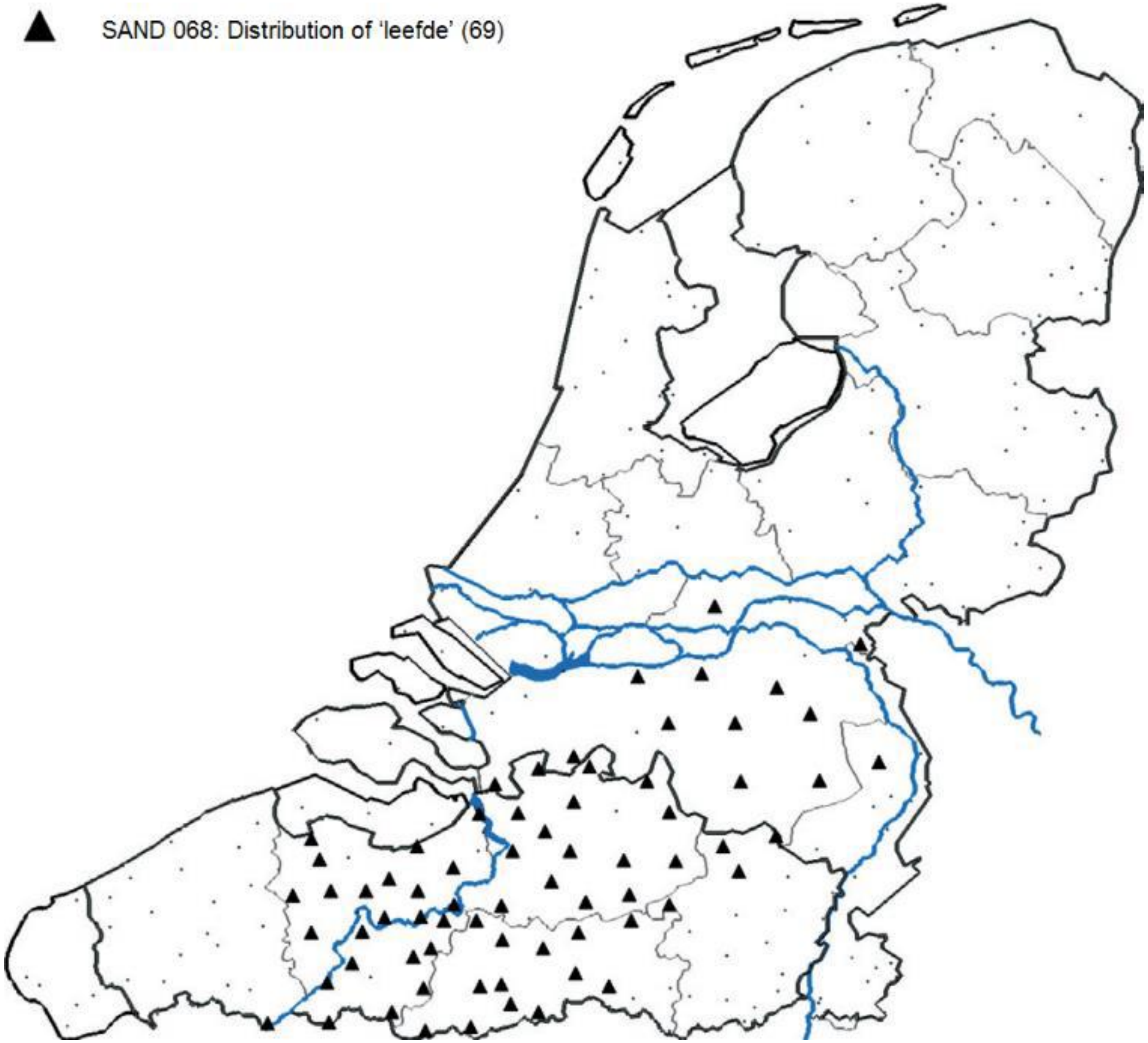
Case study 3: weak vs. strong preterites

- Two predictions:
  - (i) Weak preterites will be more prevalent in the regions known for their enclitic realization of the subject pronoun, compared to the other Dutch-speaking regions of the Low Countries.
  - (ii) Verbs that are more often realized with an enclitic subject tend to weaken more than verbs that are less often realized with an enclitic subject.

Prediction I: more weak forms in Antwerp, Flemish-Brabant and East-Flanders compared to the other Dutch speaking regions



Prediction I: more weak forms in Antwerp, Flemish-Brabant and East-Flanders compared to the other Dutch speaking regions ($p=0.031$)



Prediction II: more weak forms for verbs that
are more likely to appear with clitic

graaf-de
dig-2SG.PRS
‘Do you dig?’

vs.

?slinkt-te
lessen-2SG.PRS
‘Do you lessen?’

Prediction II: more weak forms for verbs that are more likely to appear with enclitic ($p>0.05$)

graaf-de
dig-2SG.PRS
‘Do you dig?’

vs.

?slinkt-te
lessen-2SG.PRS
‘Do you lessen?’

So is it an occasional rarity or a pervasive effect?

Case study 4: long vs. bare infinitives

Case study 4: long vs. bare infinitives

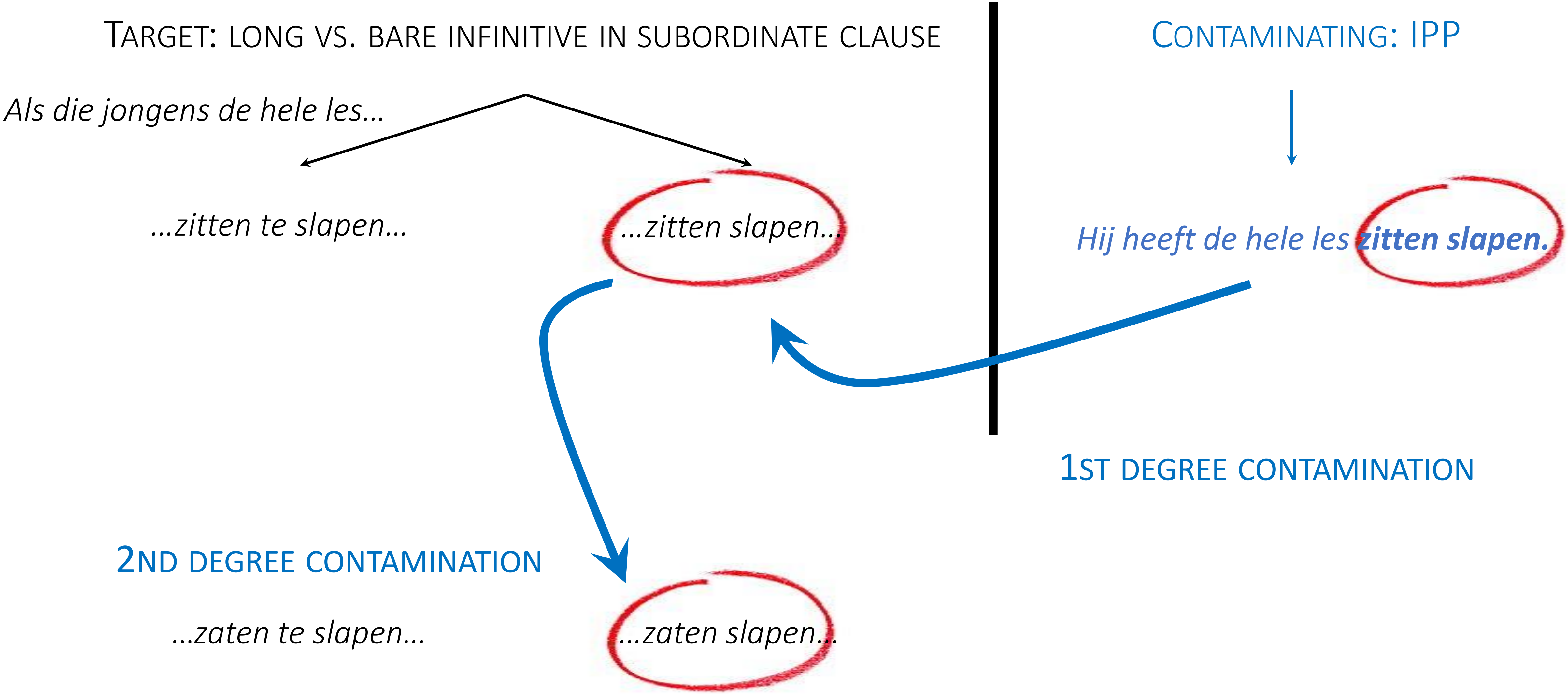
- Auxiliaries can be classified according to the type of complement they take:
  - participle
  - infinitival complement
    - bare infinitive: *Dat moet* ***Ø/*te*** *werken. (*‘That must **Ø** work.’)
    - long infinitive (or: to-infinitive): *Dat lijkt* ****Ø/te*** *werken.* (‘That seems **to** work.’)

Case study 4: long vs. bare infinitives

- Posture verbs (*zitten* ‘sit’, *staan* ‘stand’, *liggen* ‘lie’)
  - finite auxiliary takes long infinitive: *Hij zit* ***te/*Ø*** *slapen*. (‘He is sleeping’.)
  - infinite auxiliary
    - Infinitivus Pro Participio (IPP or ‘Ersatzinfinitiv’)
    - when used in the perfect, auxiliaries may occur in the infinitive instead of the past participle
    - *Hij heeft de hele les* ***zitten Ø slapen****.* (‘He has been sleeping throughout the entire class.’)

Case study 4: long vs. bare infinitives

- Posture verbs (*zitten* 'sit', *staan* 'stand', *liggen* 'lie')
  - finite auxiliary takes long infinitive: *Hij **zit te/*Ø slapen***. ('He is sleeping'.)
    - Exception: if the auxiliary is present simple plural in a subordinate clause, bare infinitive is possible too (Haeseryn et al. 1997: 970; Klooster 2001: 61)
    - *Als die jongens de hele les **zitten Ø slapen**, zullen ze niet veel opsteken.* ('If those boys are sleeping throught the entire class, then they won't learn much') (Haeseryn et al. 1997: 970)
  - infinite auxiliary
    - Infinitivus Pro Participio (IPP or 'Ersatzinfinitiv')
    - when used in the perfect, auxiliaries may occur in the infinitive instead of the past participle
    - *Hij heeft de hele les **zitten Ø slapen**.* ('He has been sleeping throughout the entire class.')



TARGET: LONG VS. BARE INFINITIVE IN SUBORDINATE CLAUSE

Als die jongens de hele les…

…zitten te slapen…

…zitten slapen…

2ND DEGREE CONTAMINATION

…zaten te slapen…

…zaten slapen…

CONTAMINATING: IPP

Hij heeft de hele les ***zitten slapen.***

1ST DEGREE CONTAMINATION

Prediction: Group I is strongly affected by constructional contamination,
group II less so and group III even less so, or not at all.

Group (i): superficial formal identity (1st degree contamination)
e.g. *Als die jongens de hele les* ***zitten Ø slapen****, zullen ze niet veel opsteken.*
('If those boys are sleeping throughout the entire class, then they won't learn much')

Group (ii): superficial formal resemblance (2nd degree contamination)
e.g. *Als die jongens de hele les* ***zaten Ø slapen****, hebben ze niet veel opgestoken.*
('If those boys were sleeping throughout the entire class, they haven't learned much.')

Group (iii): no resemblance
e.g. *De jongen* ***zit*** *al heel de les* ***(te) slapen.***
('The boy has been sleeping the entire class')

Prediction: Group I is strongly affected by constructional contamination, group II less so and group III even less so, or not at all.

Out of 2766 bare infinitives…

Group (i): superficial formal identity (1st degree contamination)
7 instances (<-> 2622 long infinitives)

Group (ii): superficial formal resemblance (2nd degree contamination)
3 instances (<-> 11978 long infinitives)

Group (iii): no resemblance
1 instance (<-> 13576 long infinitives)

Conclusions

- Constructional contamination is a **pervasive effect**
- It follows naturally from a **usage-based** view on language processing, in particular **shallow parsing and ready-mades**
- If we can so easily find four case studies in a single language, **you** should be able to **find many more in other languages**

Special thanks to

- **Gert De Sutter,** for generously sharing dataset of verbal clusters
- **Tom Ruette**, for giving us access to his Twitter-corpus

References

Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Magda Devos & Margreet van der Ham. 2006. *Syntactic atlas of the Dutch dialects. Vol. 1: Pronouns, Agreement and Dependencies.* Amsterdam: Amsterdam university press.

Bates, Douglas, Martin Maechler, Ben Bolker & Steven Walker. 2013. *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.4*. http://cran.r-project.org/package=lme4.

Bloem, Jelke, Arjen Versloot & Fred Weerman. 2017. Verbal cluster order and processing complexity. *Language Sciences* 60. 94–119. doi:10.1016/j.langsci.2016.10.009.

Carroll, Ryan, Ragnar Svare & Joseph Salmons. 2012. Quantifying the evolutionary dynamics of German verbs. *Journal of Historical Linguistics* 2(2). 153–172.

Dąbrowska, Ewa. 2014. Recycling utterances: A speaker’s guide to sentence processing. *Cognitive Linguistics* 25(4). 617–653.

De Sutter, Gert. 2005. Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen. Dissertation University of Leuven.

Ferreira, Fernanda & Nikole Patson. 2007. The “good enough” approach to language comprehension. *Language and Linguistics Compass* 1. 71–83.

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

Grondelaers, Stefan, Katrien Deygers, Hilde Van Aken, Vicky Van den Heede & 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 & Maarten van den Toorn. 1997. *Algemene Nederlandse Spraakkunst [General Dutch Grammar]*. Groningen: Nijhoff.

Harrell, Frank. 2013. rms: Regression Modeling Strategies. R package version 4.0-0. http://cran.r-project.org/package=rms.

Lemmens, Maarten. 2005. Aspectual Posture Verb Constructions in Dutch. *Journal of Germanic Linguistics* 17(3). 183–217. doi:10.1017/S1470542705000073.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Ineke Schuurman. 2013. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, 219–247. Heidelberg: Springer.

Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat & Harald Baayen. 2002. Experiences from the Spoken Dutch corpus project. *Proceedings of the third international conference on language resources and evaluation (LREC)*, 340–347. http://www.lrec-conf.org/proceedings/lrec2002/.

Pijpops, Dirk & 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.

Pijpops, Dirk & Freek Van de Velde. 2016. Constructional contamination: How does it work and how do we measure it? *Folia Linguistica* 50(2). 543–581.

R Core Team. 2014. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna. http://www.r-project.org/.

Vosters, Rik. 2012. Geolinguistic data and the past tense debate. Linguistic and extralinguistic aspects of Dutch verb regularization. In Gunther De Vogelaer & Guido Seiler (eds.), *The dialect laboratory. Dialects as a testing ground for theories of language change*, 227–248. Amsterdam/Philadelphia: John Benjamins.

Wickham, Hadley & Romain Francois. 2015. dplyr: A Grammar of Data Manipulation. http://cran.r-project.org/package=dplyr.