



Token-based distributional semantics for grammatical alternation research

Stefano De Pascale¹ & Dirk Pijpops²



¹RU Quantitative Lexicology and Variational Linguistics, KU Leuven

²RU Lilith, ULiège

This talk will be about...

1. Meaning differences in grammatical alternations
2. Token-level word embeddings
3. Pilot study: transitive-prepositional alternations in Dutch *grijpen* (*naar*) 'grab (at)'
4. Conclusions

1. Meaning differences in grammatical alternations

some **theoretical** issues with the position of semantics in grammatical alternation research

1.1. alternation: two language structures that exhibit a systematic difference in form across some set of lexical items

e.g. *He loaded/stuffed/sprayed... the truck with paint*

He loaded/stuffed/sprayed... paint unto the truck

→ originally defined from a formal point of view, only afterwards from a functional/semantic point of view

1.2. underlying assumption: linguistic elements that are formally similar enough are better alternation candidates than linguistic elements that do not share many formal similarities

1. Meaning differences in grammatical alternations

some **theoretical** issues with the position of semantics in grammatical alternation research

2.1. conflict between the researcher's bias in determining what counts as an alternation and the natural flow of discourse

2.2. speech planning proceeds through conceptualization of a communicative intent into a message that is articulated into one of many possible different options, and not by a first assessment of which option best captures the message

1. Meaning differences in grammatical alternations

some **methodological** issues with the position of semantics in grammatical alternation research

3.1. by focusing on the semantic interchangeability of the alternating constructions, we procedurally remove the 'non-interchangeable' instances (influence from sociolinguistics)

3.2. strange paradox: you have to blind yourself in order to discover the true underlying basis of an alternation

1. Meaning differences in grammatical alternations

some **methodological** issues with the position of semantics in grammatical alternation research

4.1. extant variationist research treats semantic/lexical predictors (semantic properties of the embedding lexical context) as nuisance factors, with no substantial interest in explanatory power

e.g.

- *animacy* as predictor, only coarse-grained distinctions
- lexical effects as control variables in random-effects structure

1. Meaning differences in grammatical alternations

what can be done and what has been done?

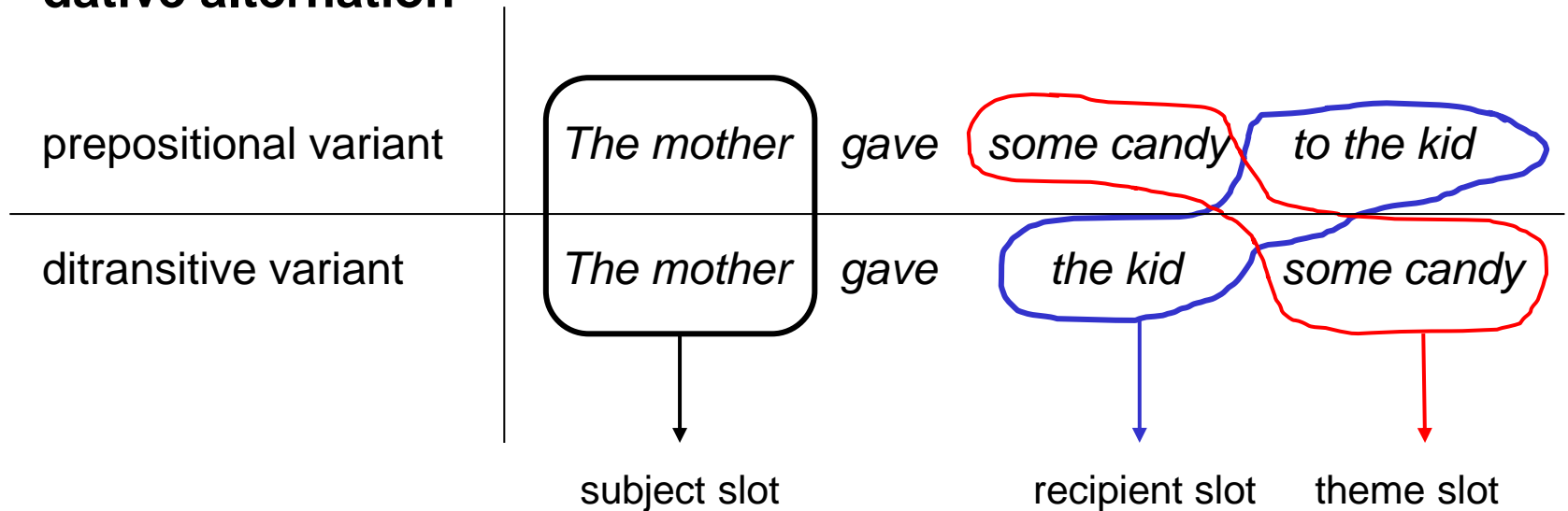
- upsurge in attention for operationalizing semantics in alternation studies (Perek 2018, Gries 2019, Pijpops 2019, Speelman et al. forthcoming)
1. fine-grained properties of concrete lexical context
 2. methodological reflection on the corpus-based modelling of semantics
 3. prefer a comprehensive analysis of the lexical context instead of a restrictive analysis

1. Meaning differences in grammatical alternations

- previous studies have turned to distributional semantics models, in particular **type-based vector representations**
 - typically one separate semantic vector for each relevant word type (or argument slot) in the construction, so as to reveal semantic classes
 - disadvantage: the semantics of these words are treated as isolated from the original instance of the construction
- here we propose **token-based vector representations**
 - single semantic vector for a concrete instance (i.e. a token) of the syntactic variant in the alternation
 - by averaging the semantic vectors of the specific context words present in that concrete instance

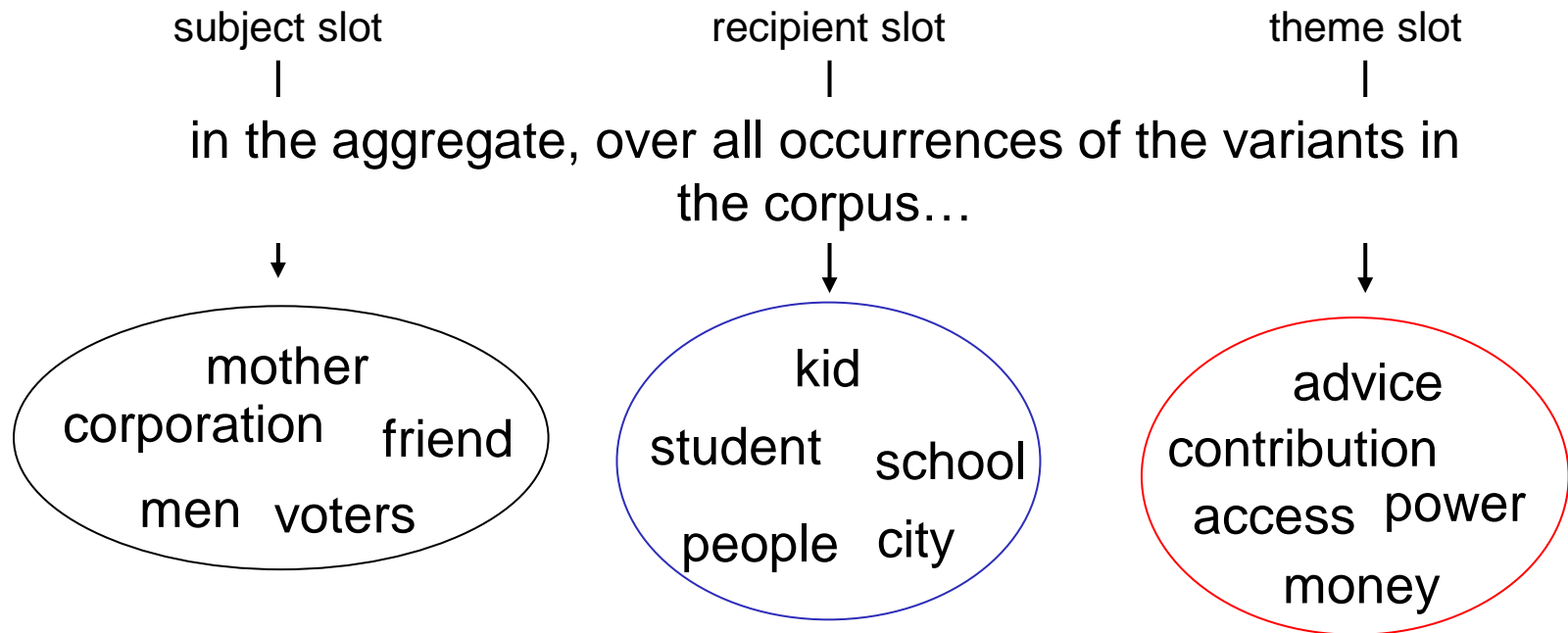
2. Token-level vs. type-level word embeddings

dative alternation



2. Token-level vs. type-level word embeddings

dative alternation



- type-based vectors for each word in each slot
- cluster analysis → semantic classes in each slot
- no interaction between classes of different slots, no feedback of concrete interplay of specific lexemes in the corpus occurrence

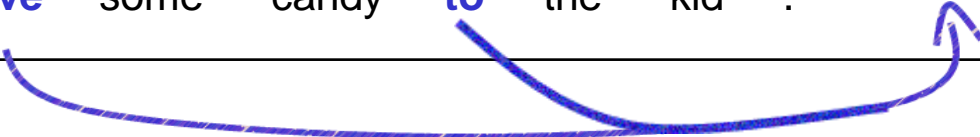
2. Token-level vs. type level word embeddings

foot		1.2		1.4		1.2			
cry		0.4		0.2		3.6			
sugar		0.2	+	2.8	+	0.5			
work		2.1		0.7		0.4			
family		3.2		2.9		2.5			
sweet		0.8		3.3		3.1			
	The	mother	gave	some	candy	to	the	kid	.

step1: type-based representations for each context word

2. Token-level vs. type level word embeddings

foot		1.2		1.4		1.2		1.3	
cry		0.4		0.2		3.6		1.4	
sugar		0.2	+	2.8	+	0.5		1.2	
work		2.1		0.7		0.4		1.1	
family		3.2		2.9		2.5		2.9	
sweet		0.8		3.3		3.1		2.4	
<hr/>									
	The	mother	gave	some	candy	to	the	kid	.
<hr/>									



step2: average type-vectors of the context words,
so to have a single vector representation of a single
realization of the alternation variant

3. transitive-prepositional alternation in Dutch

goal of this pilot study:

1. a comprehensive analysis of the full range of variable and non-variable lexical context in which the alternating variants occur
2. zoom in on variable lexical context: is it possible to arrive at generalizations?

3. transitive-prepositional alternation in Dutch

- alternation that occurs
 - with various verbs and verb classes in Dutch: motoric verbs (*graaien, grabbelen*), tractional verbs (*krabben, likken*) etc.
 - with many different prepositions: *aan, bij, naar, tegen* etc.
- *grijpen* vs. *grijpen naar* ‘grab (at)’
 - e.g. *de inbreker greep (naar) het mes en stak de bewoner in de buik*
‘the burglar grabbed (at) the knife and stabbed the resident in the stomach’

3. transitive-prepositional alternation in Dutch

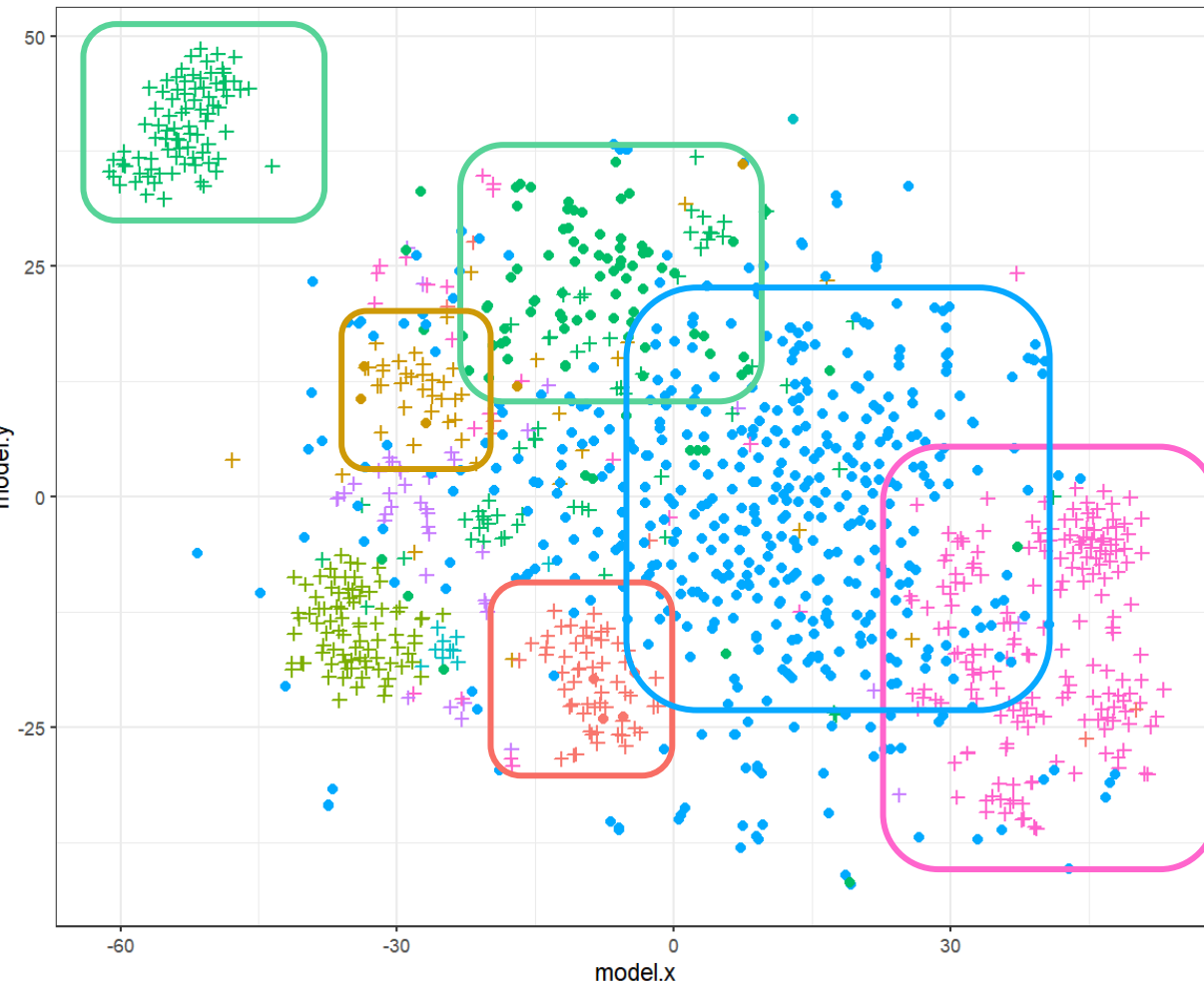
- dataset curated for Pijpops (2019)
 - 11632 sentences with *grijpen (naar)* (and surrounding sentences)
 - manually annotated for inclusion in or exclusion from ‘envelope of variation’
- rich information about reasons for exclusion/inclusion:
 - formal motivation (e.g.: use of other preposition, use as adjective, verb is coordinated etc.)
 - semantic motivation

3. transitive-prepositional alternation in Dutch

wha's next:

1. a comprehensive analysis of the full range of variable and non-variable lexical context in which the alternating variants occur
2. zoom in on variable lexical context: is it possible to arrive at generalizations?

grijp.pos-all.lemmapath.foc-cont-none.ass-foc-ppmi0.soc-FOC.ass-soc-ppmi.union



- random selection of 600 PO and 600 DO tokens and no formal reasons for exclusion
- *shape coding*
prepositional variant: •
ditransitive variant: +
- *color coding*
manually-defined semantic categories (prior to distributional modelling):
 - body parts
 - *macht* ('power')
 - prizes & valuables
 - *kans* ('chance')
 - abstract/concrete objects ('opt for')

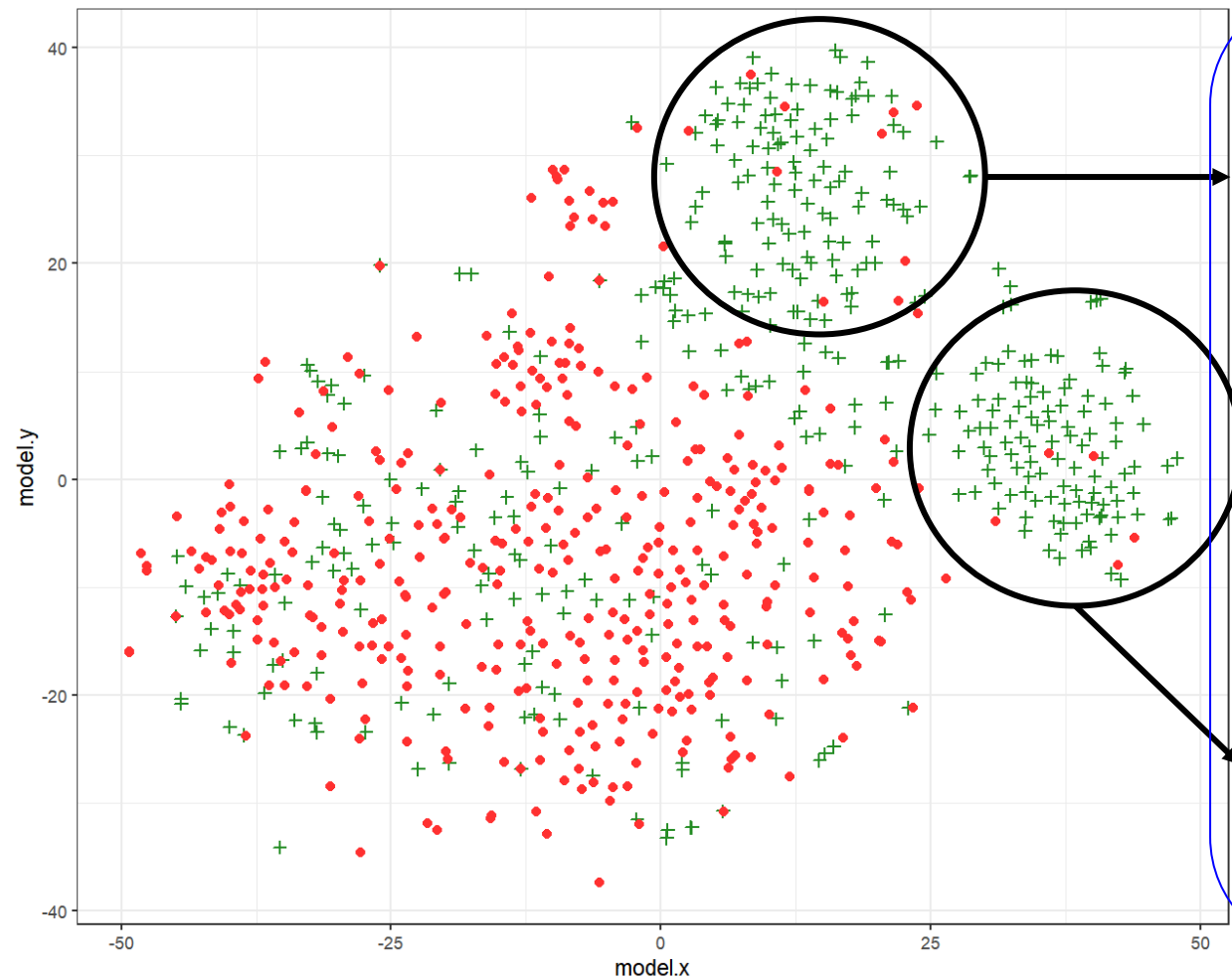
3. transitive-prepositional alternation in Dutch

- shape of token space reveals different semantic representations for objects in the DO-variant and PO-variant
 - range of objects for the DO-variant is smaller (*macht* ‘power’, *kans* ‘chance’, *keel* ‘throat’), but each object type is relatively frequent
 - multiple identifiable pockets
 - “tendency of quasi noun incorporation” (Pijpops 2019: 253)
 - range of objects for PO-variant is larger, and it is harder to find internal semantic structure
 - one larger blob of tokens (blue)
 - infrequent and/or less similar nouns

3. transitive-prepositional alternation in Dutch

what's next:

1. a comprehensive analysis of the full range of variable and non-variable lexical context in which the alternating variants occur
2. zoom in on variable lexical context: is it possible to arrive at generalizations?



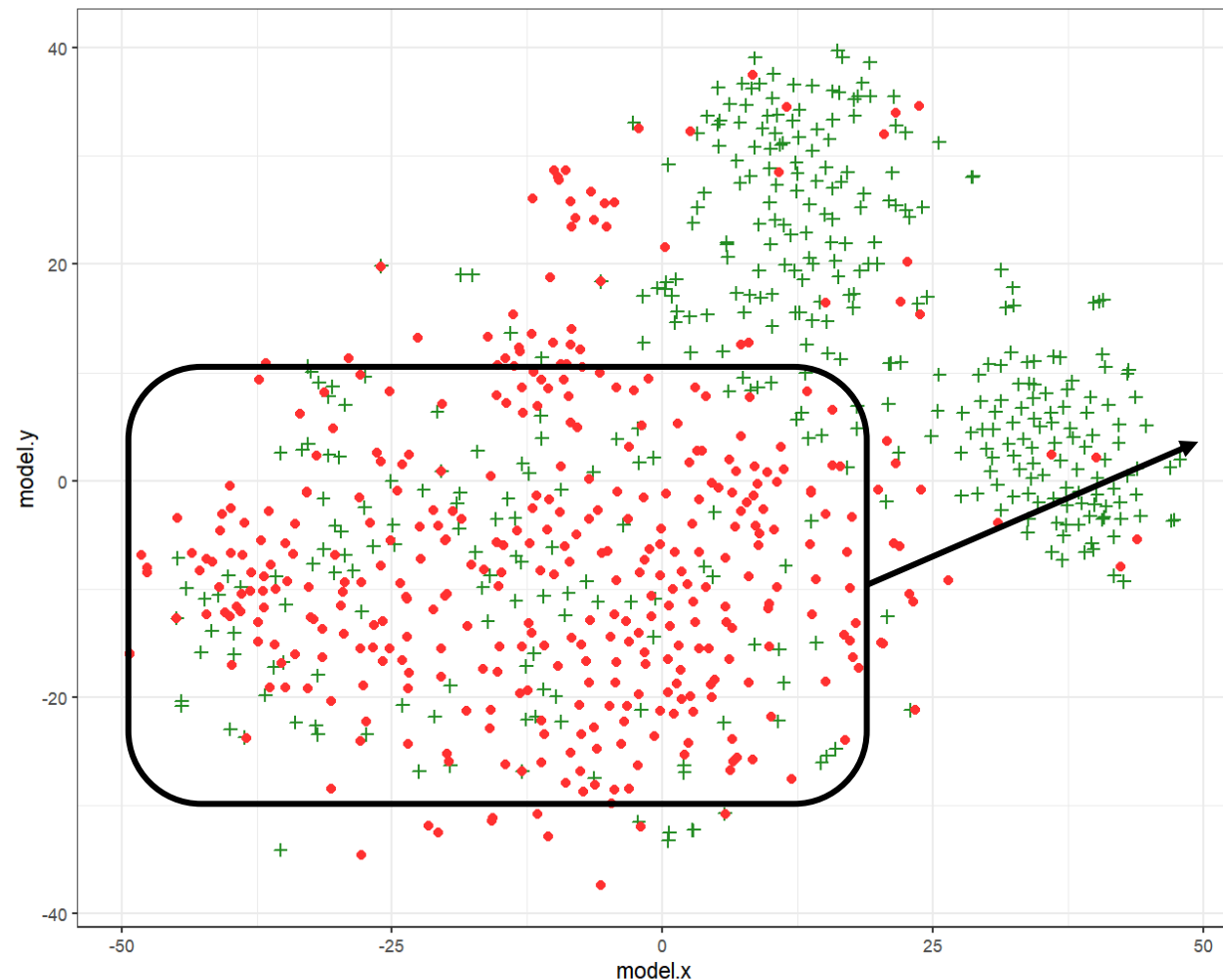
1. semi-schematic level of alternation

'secure [AN ACHIEVEMENT]'
(naar) de titel/prijs/zege grijpen

2. most concrete instantiation of alternation

'seize power'
(naar) de macht grijpen

really variable contextual slots of (many) DO and (few) PO variants?



3. large, but semantically unstructured space of consistent overlap between DO and PO variants

→ presence variable microcontexts

- [...] **greep** de man **naar** het (sic) brandblusser en sloeg de chauffeur [...] 'the man **grabbed at** the fire exstinguisher and and hit the driver'
- [...] **greep** het stuurslot en sloeg de man ermee [...] '**grabbed** the steering wheel and hit the man with it'

for further information:

stefano.depascale@kuleuven.be

<http://wwwling.arts.kuleuven.be/qlvl>

dirk.pijpops@uliege.be

[https://www.uliege.be/cms/c_9054334/fr/repertoire?uid=
u235546](https://www.uliege.be/cms/c_9054334/fr/repertoire?uid=u235546)

3. methodological afterthought

- what is the most useful transformation of raw frequencies, if the intention is to reveal cognitively and semantically plausible information?

1. Semantic Vector Space literature:

- Bullinaria & Levy 2007; Levy, Goldberg & Dagan 2015; Heylen et al. 2015; De Pascale 2019
- recommendation to use *effect size measures* (Positive Pointwise Mutual Information)
- only looks at the magnitude of a certain association, but is insensitive to how much evidence there is for that magnitude
- consequence: low-frequent associations tend to be scored higher than high-frequent associations

3. methodological afterthought

- what is the most useful transformation of raw frequencies, if the intention is to reveal cognitively and semantically plausible information?

2. Collostructional analysis literature

- Stefanowitsch & Gries 2003; Stefanowitsch & Flash 2016; Gries 2012, 2015; Pijpops 2019, but see Gries 2019
- recommendation to use *significance-based measures* (G^2 -value from LLR-test; p-value from Fisher-Yates Exact Test)
- combine information about the effect size and the amount of evidence for that effect size
- important to distinguish between large associations based on few data from large associations based on lots of data



previous space: vectors of context words with **Positive Pointwise Mutual Information (PPMI)**
 → bias towards *low-frequent* semantic associations

this space: vectors of context words with **G2-statistic** (from loglikelihood ratio test)
 → bias towards *high-frequent* semantic associations

differences w.r.t. PPMI:

- unstructured group of PO-objects is fragmented
 - pockets of tokens are heterogeneous w.r.t. semantics of the object
- bad semantic representation!

3. transitive-prepositional alternation in Dutch

- semantic vector exploration → hypothesis generation
 - what is the relation between literal and figurative uses of body parts and how does that relate to use of DO and PO variants?