

Modelling meaning differences in syntactic alternations with token-based vectors

Stefano De Pascale & Dirk Pijpops



RU Quantitative Lexicology and
Variational Linguistics, KU Leuven



RU Lilith,
ULiège

Structure

1. Challenge in alternation studies in usage-based cognitive linguistics
2. Token-level word embeddings
3. Pilot study: transitive-prepositional alternation in Dutch *grijpen* (*naar*) ‘grab (at)’
4. Conclusions

1. Alternation studies

- English dative alternation: *She gave me a hockey stick* vs. *She gave a hockey stick to me*
- English genitive alternation: *The president's hockey stick* vs. *the hockey stick of the president*
- English *at*-alternation: *she bit her lower lip* vs. *she bit at her lower lip*
- Estonian adessive case vs. *peal*: *Raamat on laual.* vs. *Raamat on laua peal* 'the book is on the table'.
- Dutch transitive-reflexive alternation: *Elizabeth ergert John* vs. *John ergert zich aan Elizabeth* 'Elizabeth annoys John
- ...

1. Alternation studies

- Workhorse technique: logistic regression
- Requirement: interchangeable instances vs. categorical instances

→ Undesirable:

The ‘categorical’ instances are hugely interesting, e.g. *verlangen (naar)* ‘desire’

Hij miste de eigenschappen die deze functie verlangde.

‘He lacked the qualities that this function demanded.’

1. Alternation studies

- Workhorse technique: logistic regression
- Requirement: interchangeable instances vs. categorical instances

→ Undesirable:

Only makes theoretical sense if there are strict distinctions between grammatical (and semantic) categories and if grammar is uniform throughout the population

- Categorical rules: (generative) syntacticians
- Variable rules: sociolinguists

1. Alternation studies

- Workhorse technique: logistic regression
- Requirement: interchangeable instances vs. categorical instances

→ Undesirable:

Only makes theoretical sense if there are strict distinctions between grammatical (and semantic) categories and if grammar is uniform throughout the population

↔ Usage-based cognitive linguistics: prototype structure of grammatical categories

- Words that have properties of several categories, e.g. participles
- Language users don't always care about 'deep' structural differences if the form and meaning are close enough, e.g. constructional contamination
- Diachronic fluctuation and synchronic variation, e.g. noun incorporation, grammaticalization
- ...

1. Alternation studies

e.g. *zoeken (naar)* ‘search (for)’

We zochten contact met Marijn Gelten, voorzitter van de MD-vereniging.

‘We tried to contact Marijn Gelten, president of the MD-association.’

We zochten [contact met Marijn Gelten, voorzitter van de MD-vereniging]_{DO}

→ Interchangeable instance

We [zochten contact] [met Marijn Gelten, voorzitter van de MD-vereniging]_{PO}

→ Categorical instance

1. Alternation studies

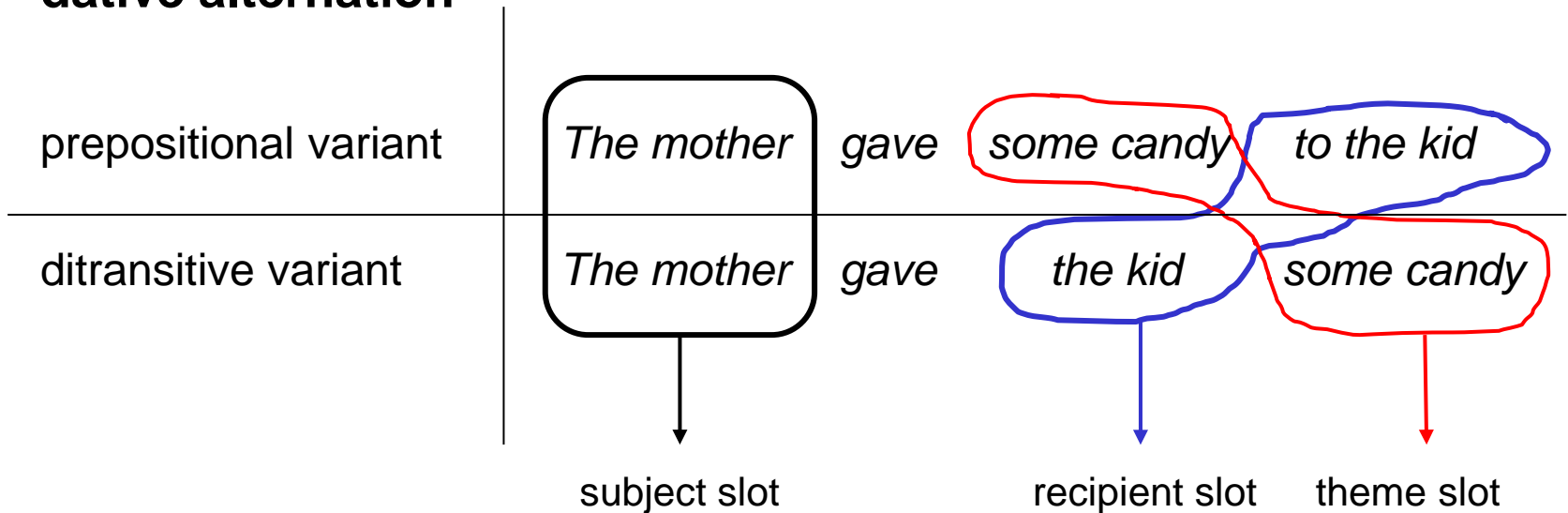
- Workhorse technique: logistic regression
 - Requirement: interchangeable instances vs. categorical instances
- This requirement is forcing us to throw out the highly interesting ‘categorical’ instances
- This requirement is forcing us to make choices that are theoretically badly motivated
- ⇒ Methodological problem requires a methodological solution

1. Meaning differences in grammatical alternations

- previous studies have turned to distributional semantic modelling, in particular **type-based vector representations**
 - typically, one separate semantic vector for each relevant word type in the argument slots in the construction, to reveal semantic classes (a.o. Perek & Hilpert, Pijpops)
 - 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 (~ BERT embeddings; Fonteyn & Karsdorp 2020; Madabushi, Romain, Divjak & Milin 2020)
 - by averaging the semantic vectors of the specific context words present in that concrete instance

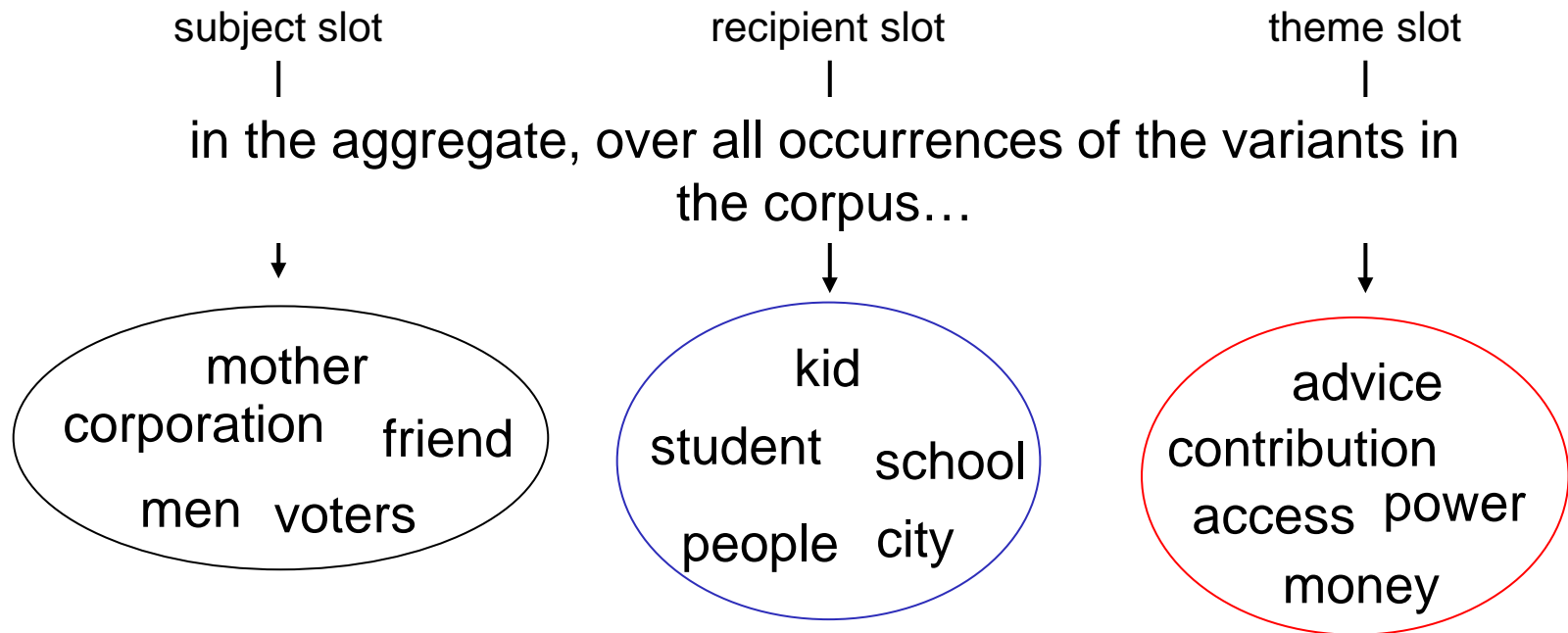
2. Token-level vs. type-level word embeddings

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

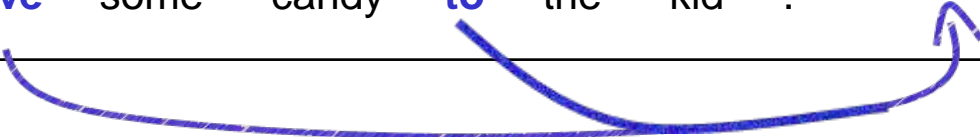
2. Token-level vs. type level word embeddings

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

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?

3. transitive-prepositional alternation in Dutch

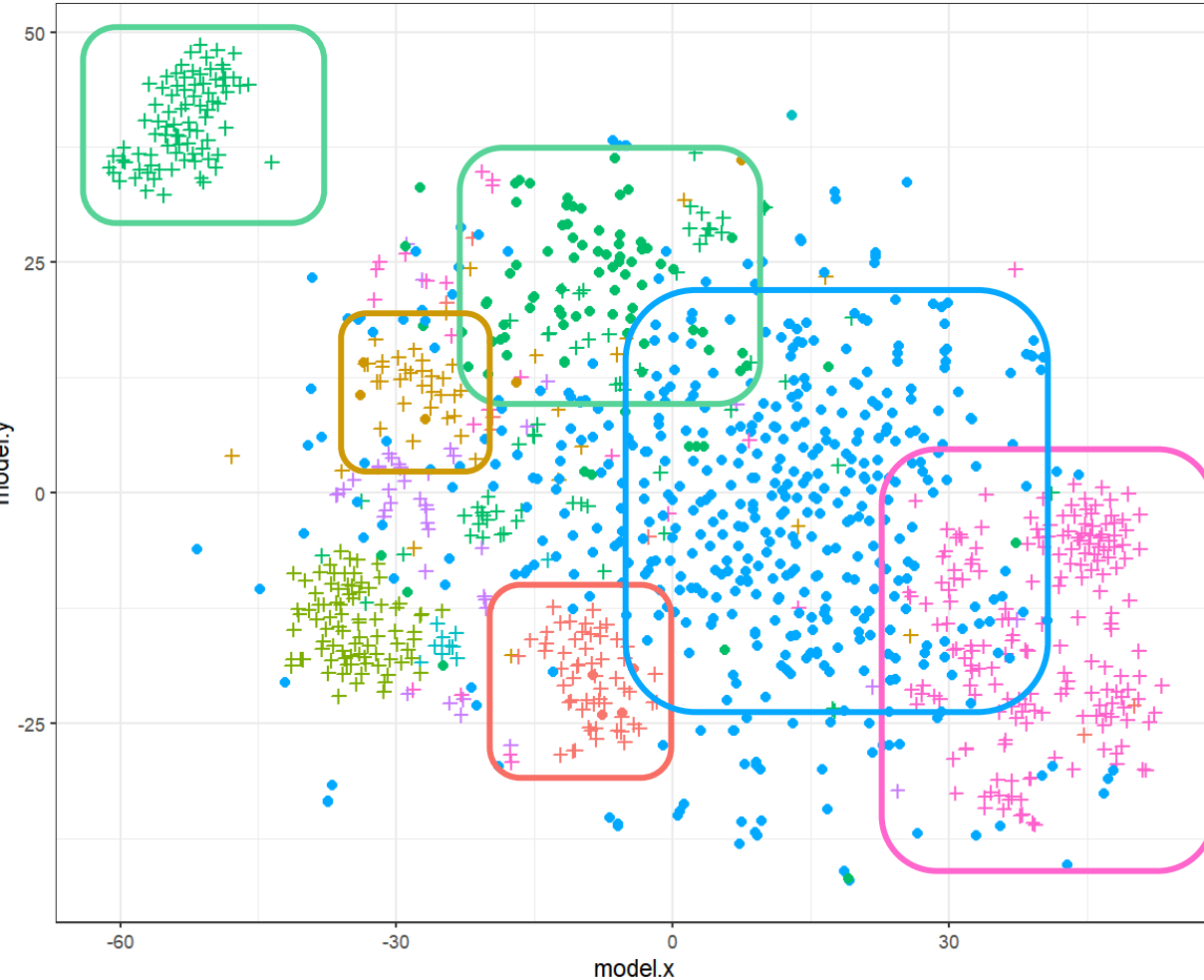
- alternation that occurs
 - with various verbs and verb classes in Dutch: motoric verbs (*graaien, grabbelen*), tractional verbs (*krabben, likken*) etc.
 - with many different prepositions: *aan, bij, naar, tegen* etc.
- *grijpen vs. grijpen naar* ‘grab (at)’
e.g. *de inbreker greep (naar) het mes en stak de bewoner in de buik*
‘the burglar grabbed (at) the knife and stabbed the resident in the stomach’
- dataset curated for Pijpops (2019)
 - 11632 sentences with *grijpen (naar)* (and surrounding sentences)
 - manually annotated for inclusion in or exclusion from ‘envelope of variation’

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: •
transitive variant: +
- *color coding*
manually-defined semantic categories (prior to distributional modelling):
 - body parts
 - *macht* ('power')
 - prizes & valuables
 - *kans* ('chance')
 - abstract/concrete objects ('opt for')

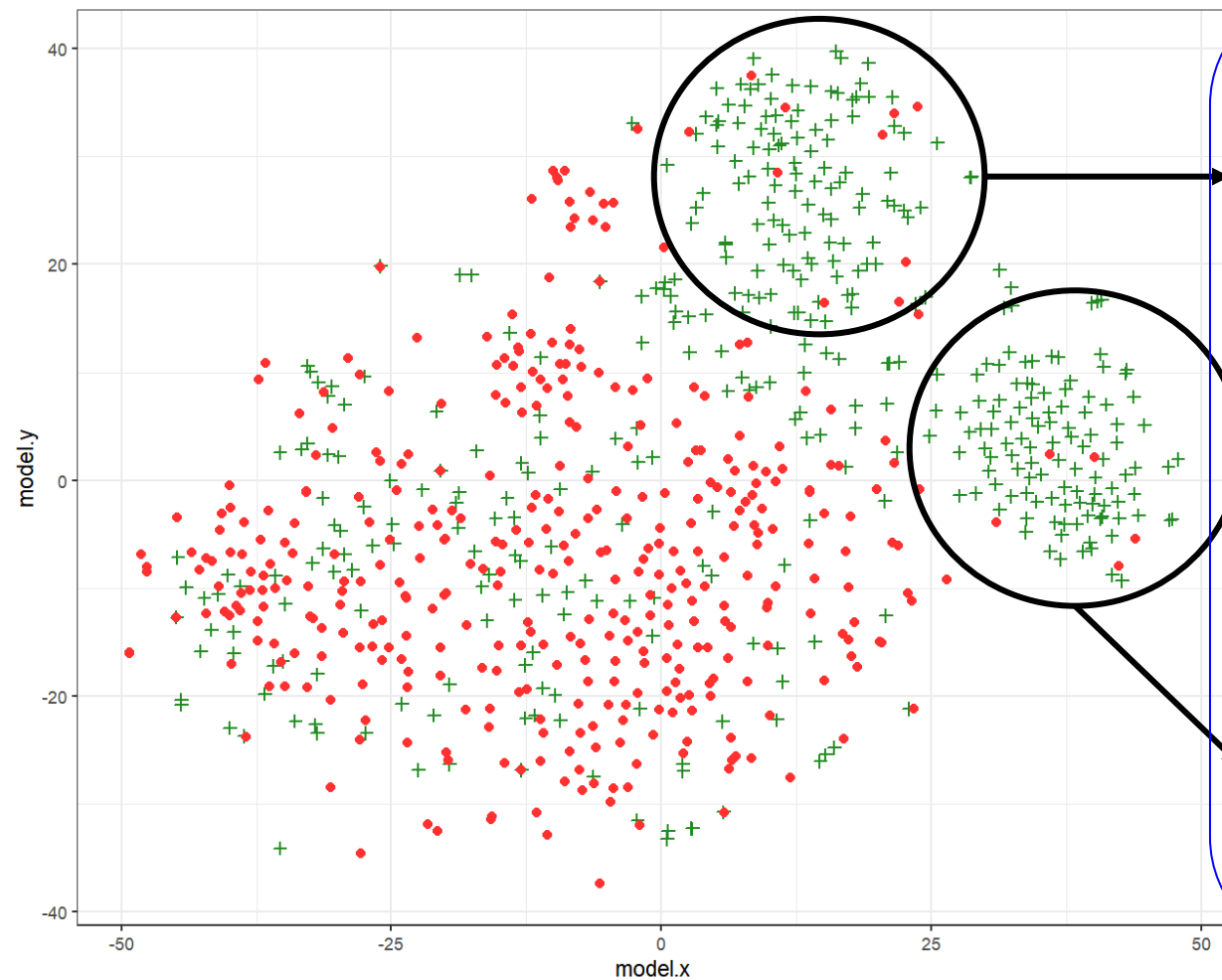
3. transitive-prepositional alternation in Dutch

- 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

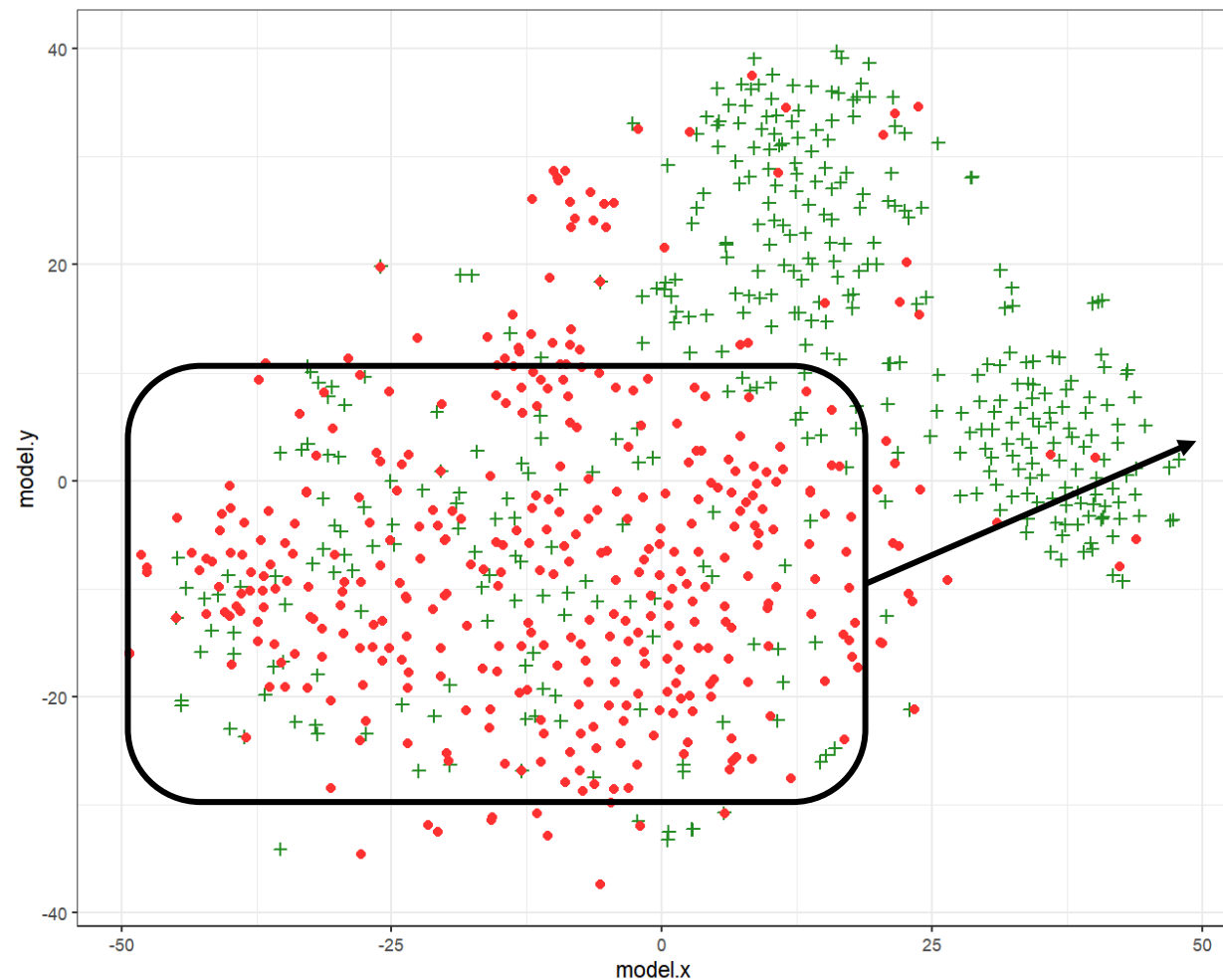
'seize power'
(naar) de macht grijpen

shape/color coding

prepositional variant: ●

transitive variant: +

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 lock and hit the man with it'

4. Conclusions

- Complementary to logistic regression:
 - token-based vectors do not force a distinction between categorical and interchangeable instances, allowing a more comprehensive analysis
 - derive semantic predictors from cloud structure
- Advantages over type-based representations:
 - concrete interplay between lexical slots in a specific corpus attestation of the alternation
 - shape of the semantic space defined by the variants of the construction



Tools & packages

Python 3.6

- nephosem: <https://qlvl.github.io/nephosem/>
- semasioFlow: <https://montesmariana.github.io/semasioFlow/>

R

- semcloud: <https://montesmariana.github.io/semcloud/>

Thank you!

for further information:

stefano.depascale@kuleuven.be, dirk.pijpops@uliege.be

References

- Fonteyn, L. 2020. What about Grammar? Using BERT Embeddings to Explore Functional-Semantic Shifts of Semi-Lexical and Grammatical Constructions. *Computational Humanities Research CEUR-WS.* : 257–268.
- Perek, Florent. 2018. Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory* 14(1). 65–97.
- Pijpops, Dirk. 2019. *How, why and where does argument structure vary? A usage-based investigation into the Dutch transitive-prepositional alternation.* Leuven: KU Leuven dissertation.
- Tayyar Madabushi, H., L. Romain, D. Divjak & P. Milin. 2020. CxGBERT: BERT meets Construction Grammar. *Proceedings of the 28th International Conference on Computational Linguistics*, 4020–4032. Barcelona, Spain (Online): International Committee on Computational Linguistics.