Methodological triangulation in corpus-based distributional semantics

Mariana Montes
Supervisor: Dirk Geeraerts
Co-supervisor: Dirk Speelman

RU Quantitative Lexicology and Variational Linguistics

Nephological semantics

Vector space models for variational linguistics

Also starring:

- Dirk Geeraerts (PI)
- Dirk Speelman
- Kris Heylen
- Tao Chen (Informatician)
- Stefano De Pascale (PhD)
- Stefania Marzo
- Christian Andersen (PhD)
- Benedikt Szmrecsanyi

Vector space semantics

Vector: (numeric) profile of a word

Similar contexts ➡ similar profile

	fast	eats	sleeps
dog	1	1	1
cat	0	1	1
cars	1	0	0

- ✓ My **dog** runs *fast* if he *eats* first, and then he just *sleeps*.
- ✓ That **cat** only *eats* and *sleeps*.
- ✓ I don't like *fast* **cars**.

Generalized over hundreds or thousands of occurrences.

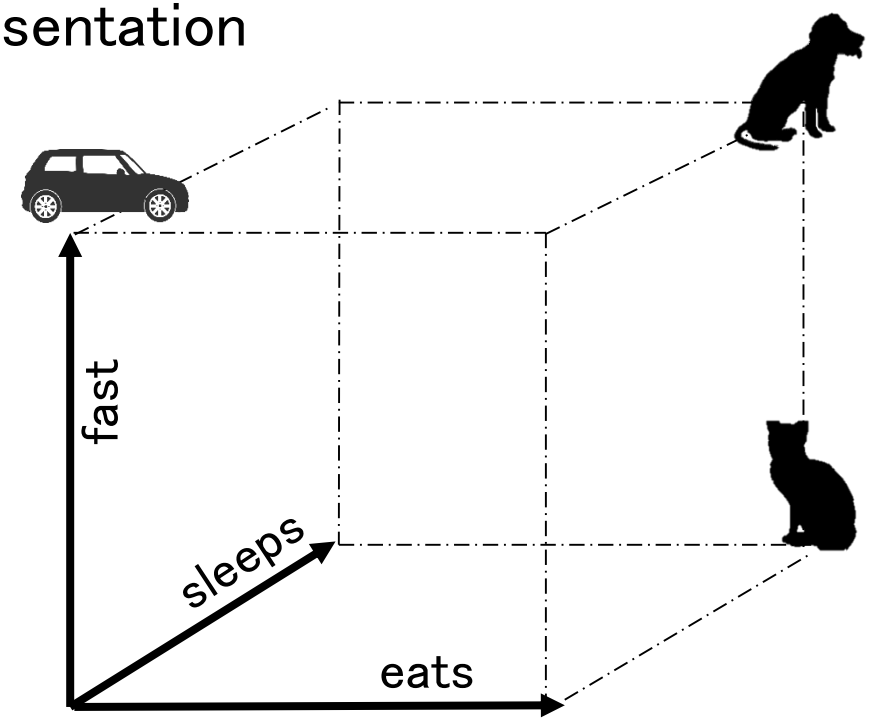
Vector space semantics

Vector: (numeric) profile of a word

Similar contexts ➡ similar profile ➡ similar vectors

➡ close in spatial representation

	fast	eats	sleeps
dog	1	1	1
cat	0	1	1
cars	1	0	0



Vector space semantics

Type vectors

- Each item is a type
- Features: context words across occurrences

Token vectors

- Each item is a token
- Features: context words OF words in context

*I can't bring the **kids** to school today.*

*There are many **children** in school today.*

*My aunt's goat gave birth to three **kids** today.*

- Hides variation within a type.
- Highlights variation within a type.

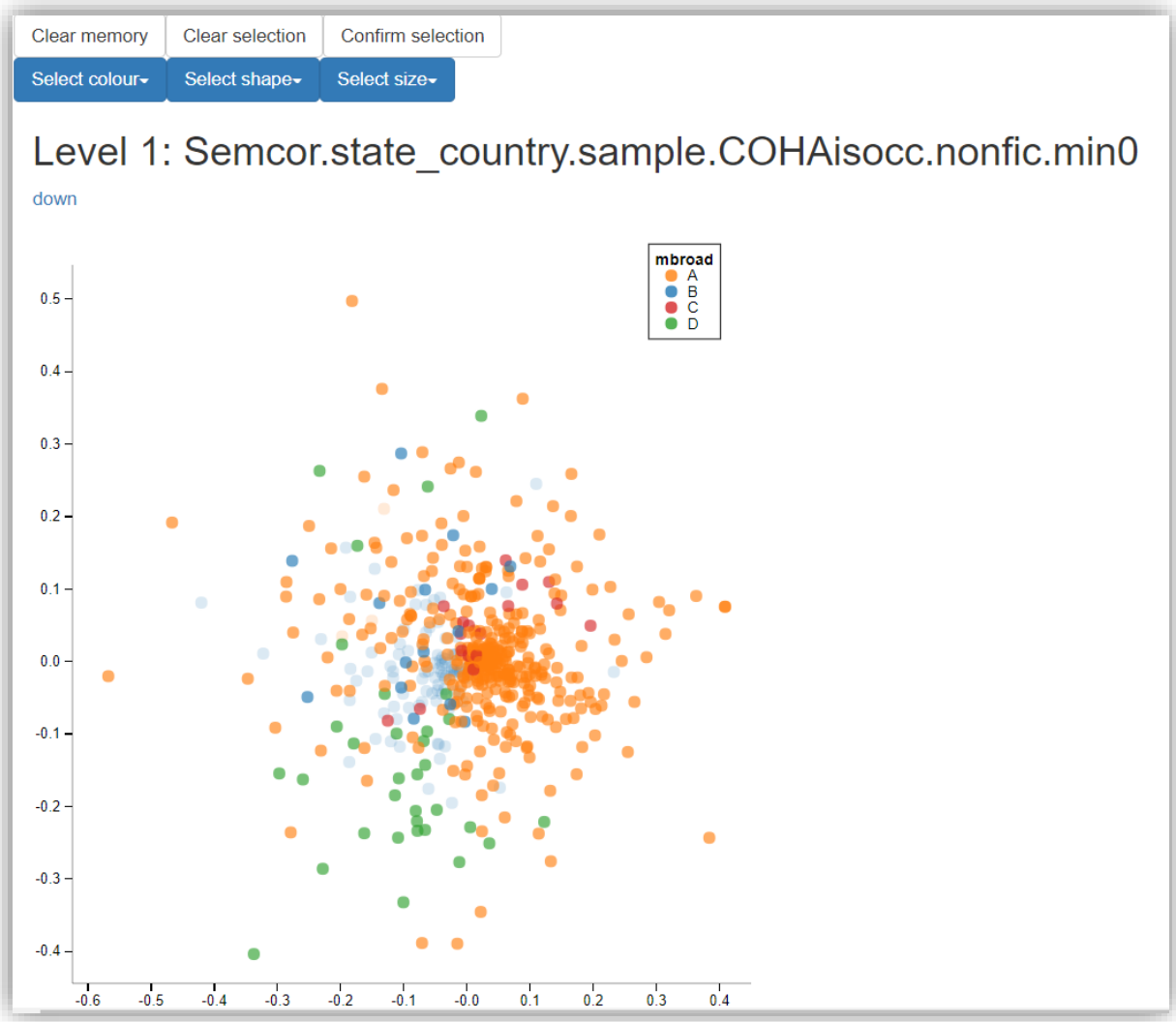
Goals of my project

- Compare different methodologies

		Identification of features	
		manual	statistical
Classification of observations	manual	*Traditional lexicography*	*Collocations*
	statistical	*Behavioral profiles*	*Vector space models*

- What is actually reflected by VSMs?
- How can we apply the strengths of other methods to VSMs?

Clouds



- Types:

  state,

  country,

  church
- Semcor 3.0

Thank you

mariana.montes@kuleuven.be