Distributional Semantic Modelling in Cognitive Sociolinguistics: QLVL probes Semantic Space

Kris Heylen, Dirk Geeraerts & Dirk Speelman

KU Leuven
Quantitative Lexicology and Variational Linguistics

Purpose of the talk

Theoretical: Lexicology in the usage-based and lectally enriched framework of Cognitive Sociolinguistics

Methodological: Summarize QLVL's use of Distributional Semantic Models in last 7 years for analysing lexical semantics in large corpora,

Descriptive: Charting Lexical variation on two levels:

1. Varieties: Measuring differences in word choice on an aggregate level
2. Variants: Word choices for individual concepts / polysemy of lexemes

Overview

1. Cognitive Sociolinguistics
   - Structure of Lexical Variation
   - The Profile-Based Approach
2. Onomasiological Variation
   - Semantic Vector Spaces
   - Different Context Models
   - Lexical Sociolectometry
3. Including Semasiological Variation
   - Measuring Polysemy
   - Analysing Semantic Structure
   - Measuring semantic change
   - Token clouds
4. Conclusion

Overview

1. Cognitive Sociolinguistics

QLVL's research is situated in what has become known as Cognitive Sociolinguistics (Kristiansen & Dirven 2008; Geeraerts, Kristiansen& Peirsman 2010):

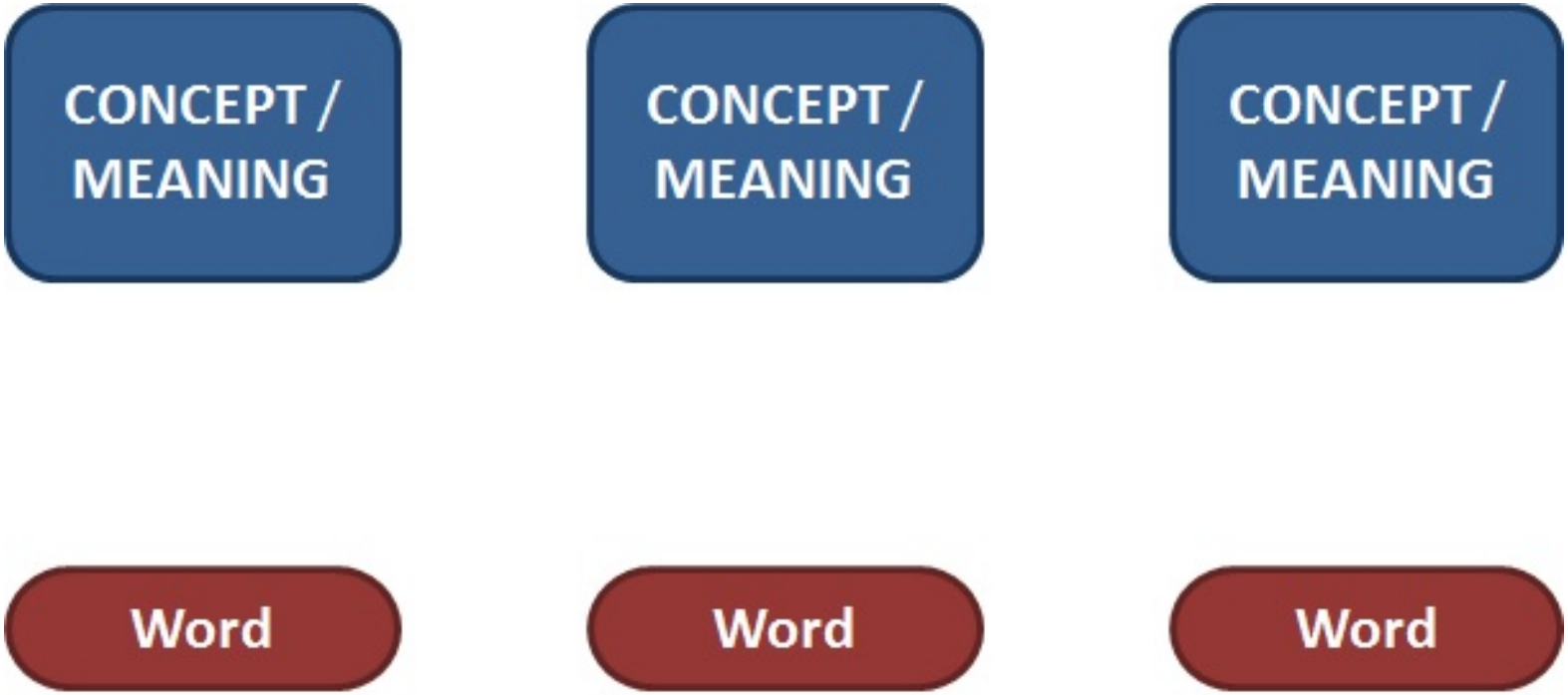
- a meaning-centered theory of language
- a usage-based perspective of language
- emphasis on the a socio-cultural aspects of semantic structure
- commitment to the use of advanced quantitative methods

QLVL has been developing this line of research since the 1990s:

- Structure of Lexical Variation (1994)
- Profile-based approach (1999)

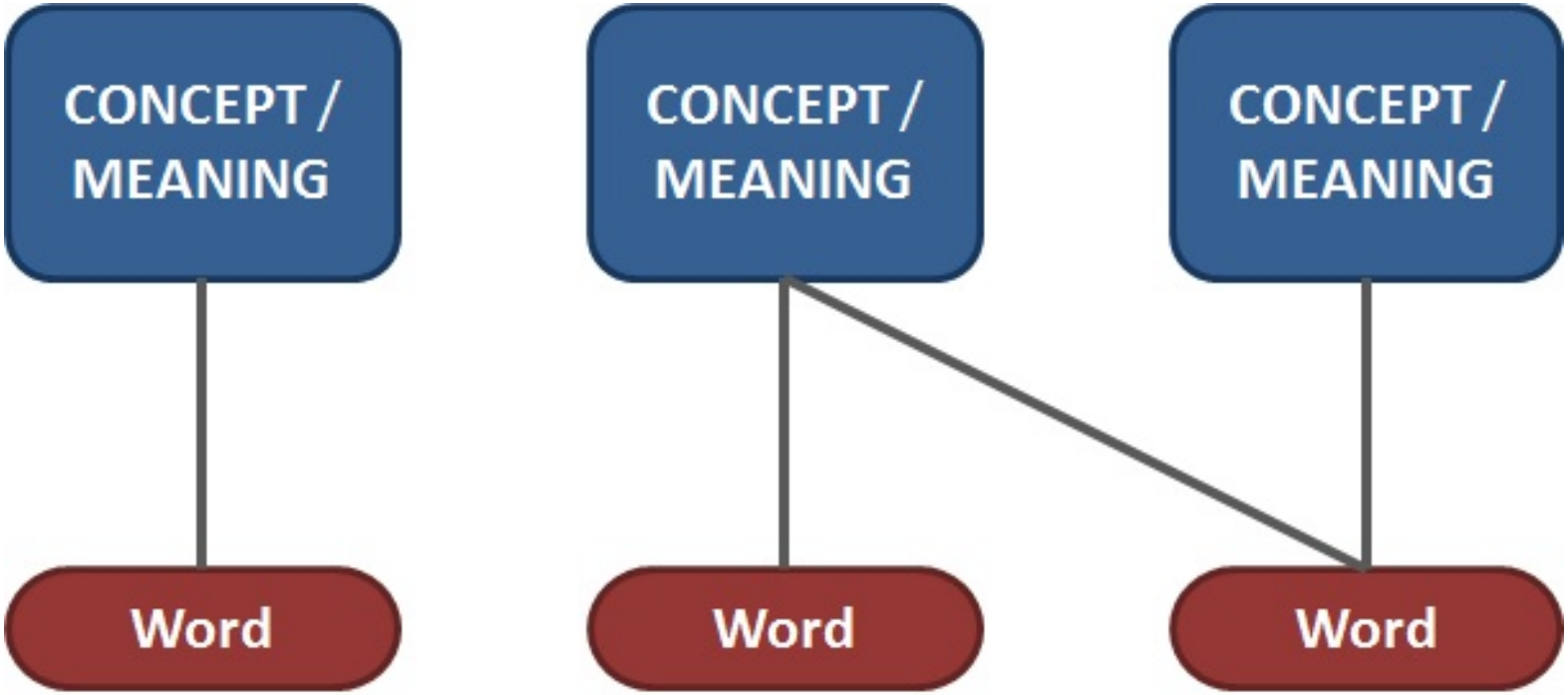
Structure of Lexical Variation (1994)

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):



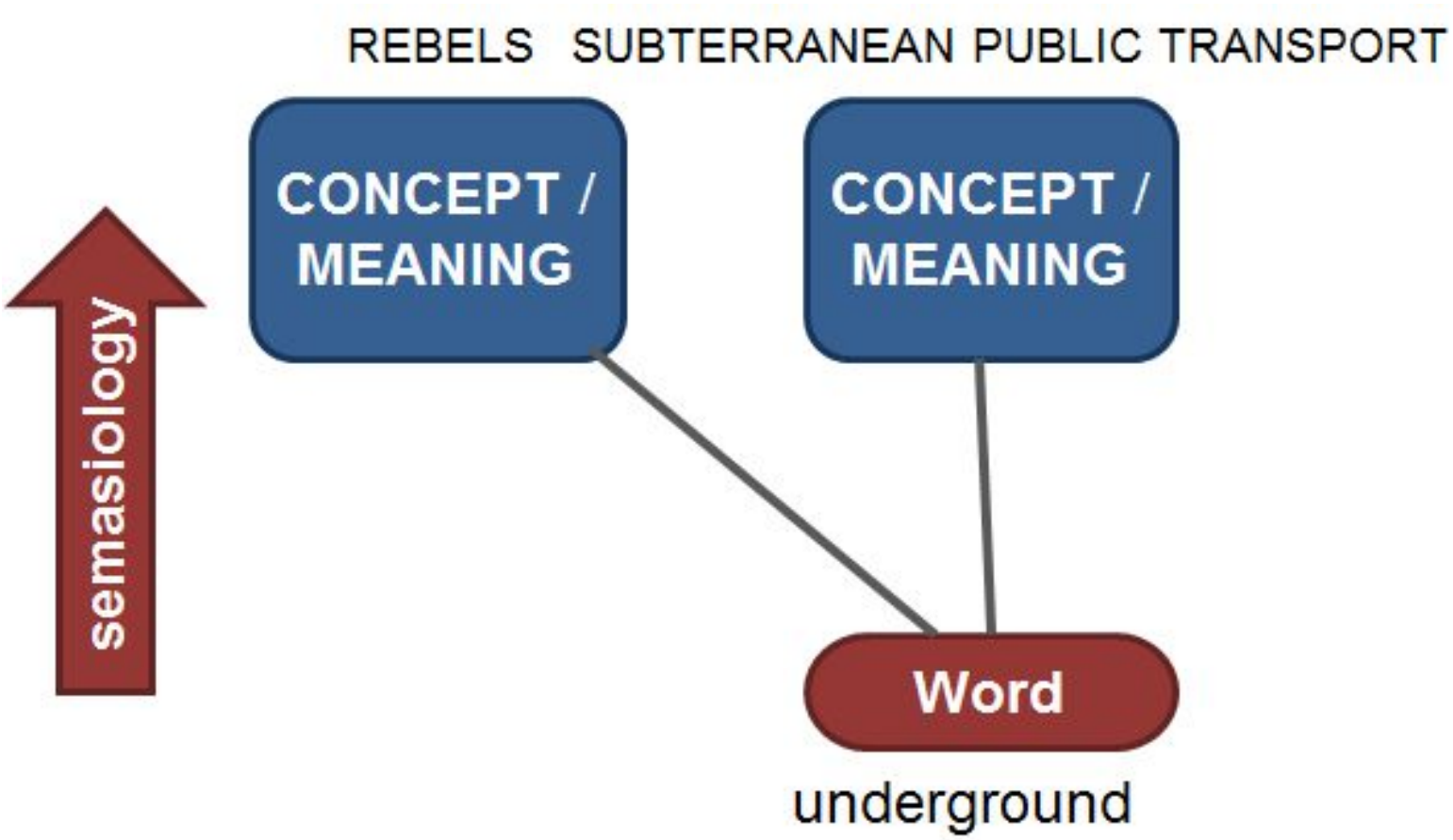
Structure of Lexical Variation (1994)

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):



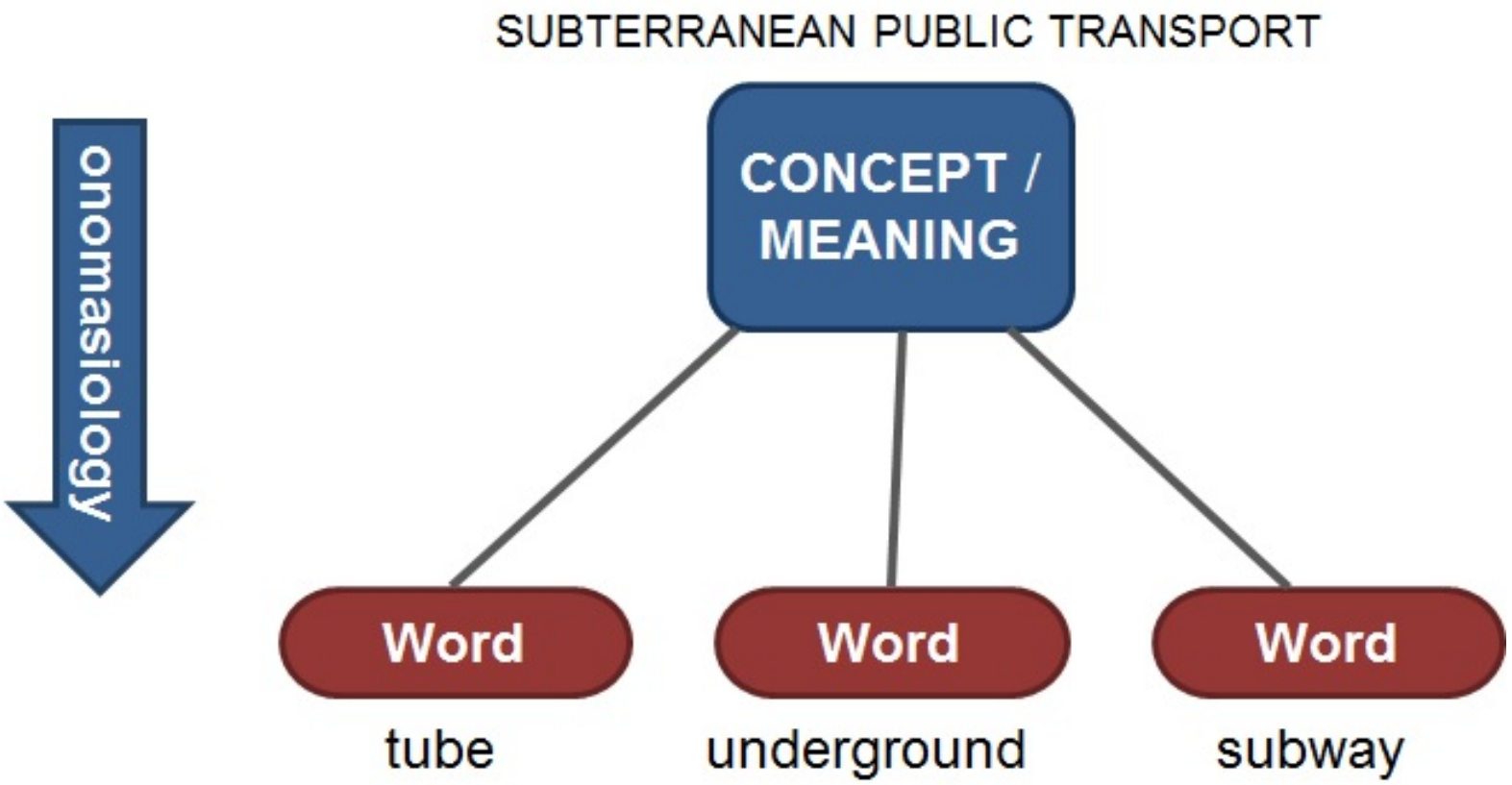
Structure of Lexical Variation (1994)

SEMASIOLOGY:



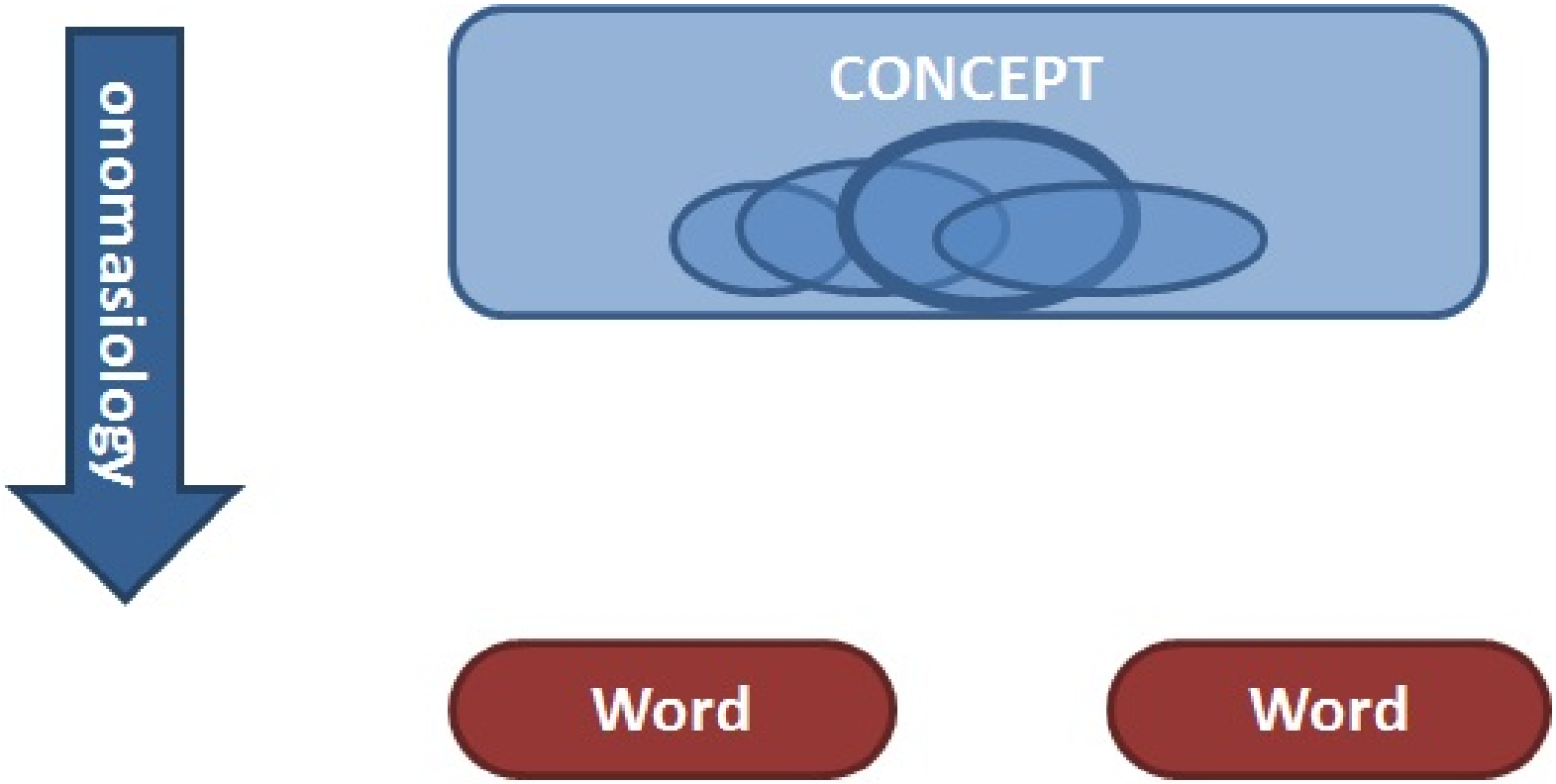
Structure of Lexical Variation (1994)

ONOMASIOLOGY:



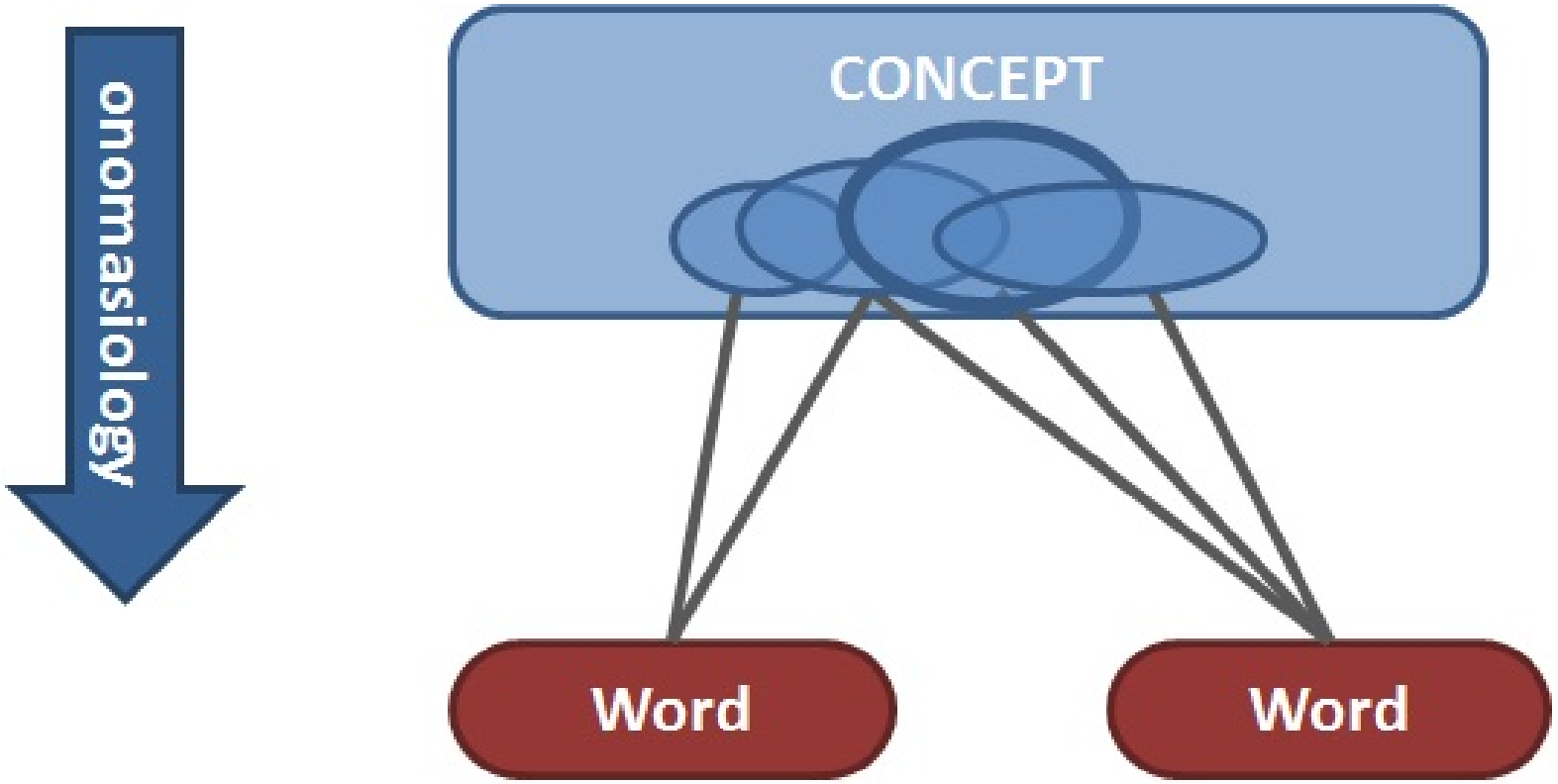
Structure of Lexical Variation (1994)

PROTOTYPE STRUCTURE:



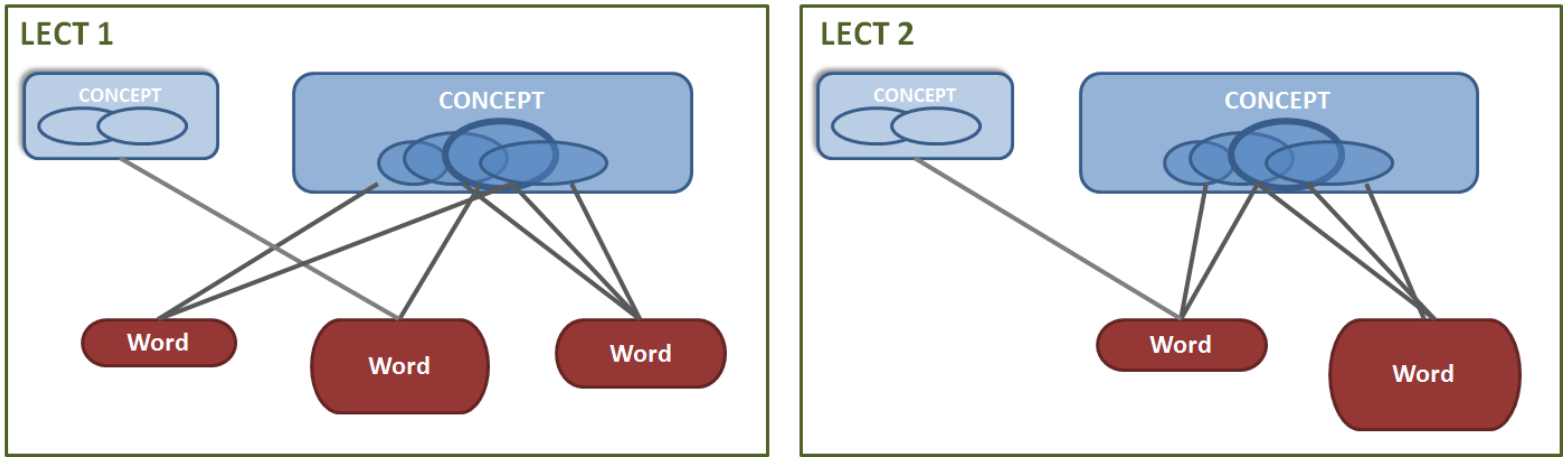
Structure of Lexical Variation (1994)

PROTOTYPE STRUCTURE:

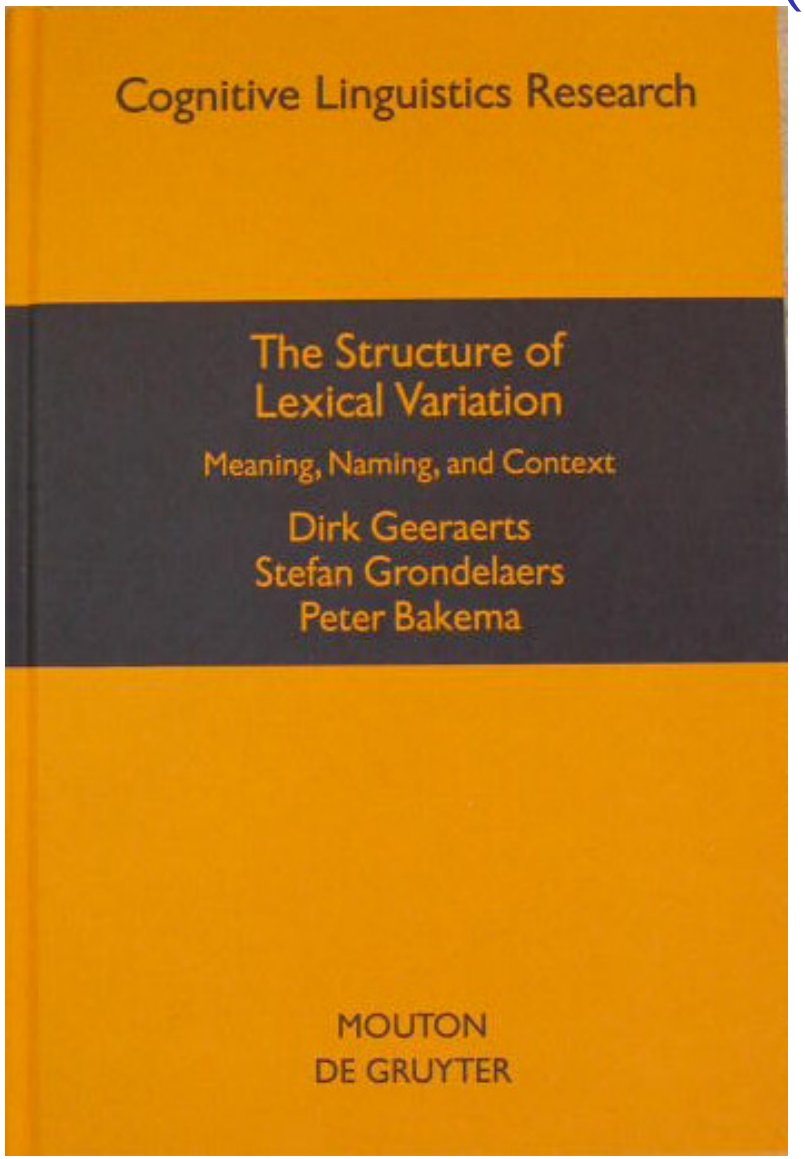


Structure of Lexical Variation (1994)

LECTAL VARIATION:

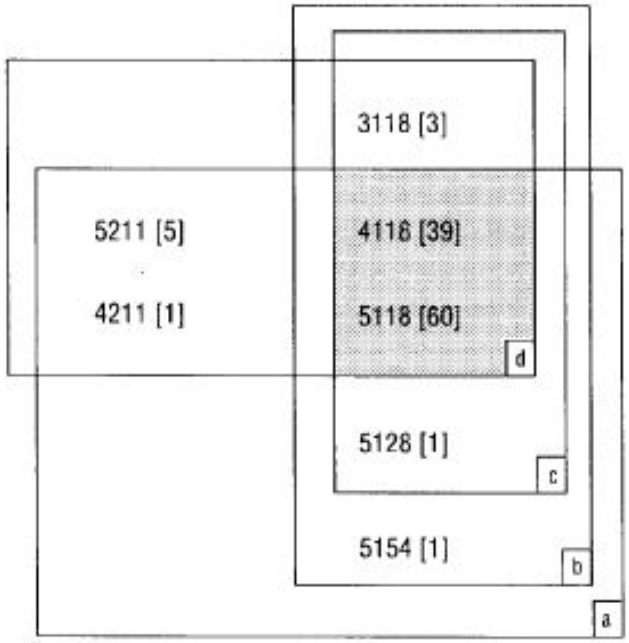


Structure of Lexical Variation (1994)



Structure of Lexical Variation (1994)

Structure of Lexical Variation (1994)



a Reaching down to the ankles or the calves

b Tight-fitting

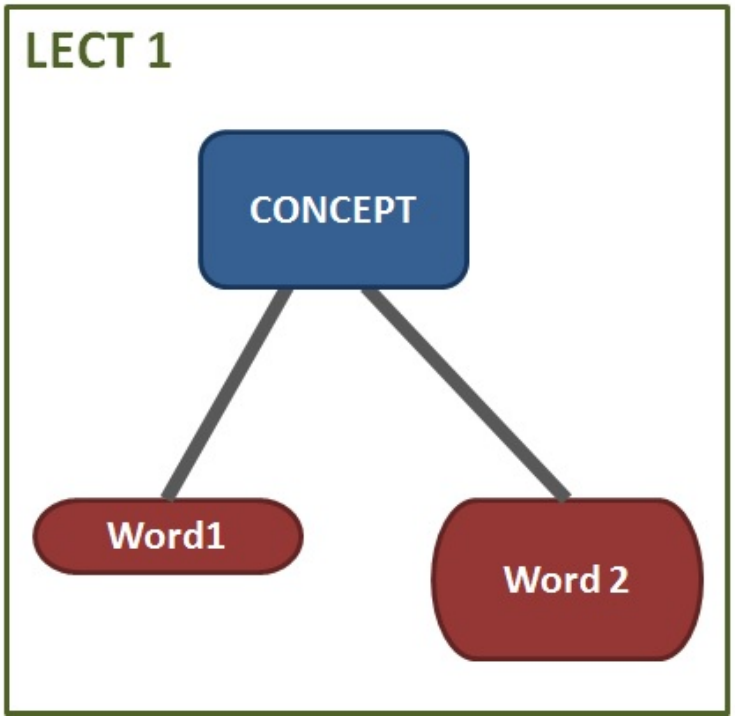
c Made of elastic material

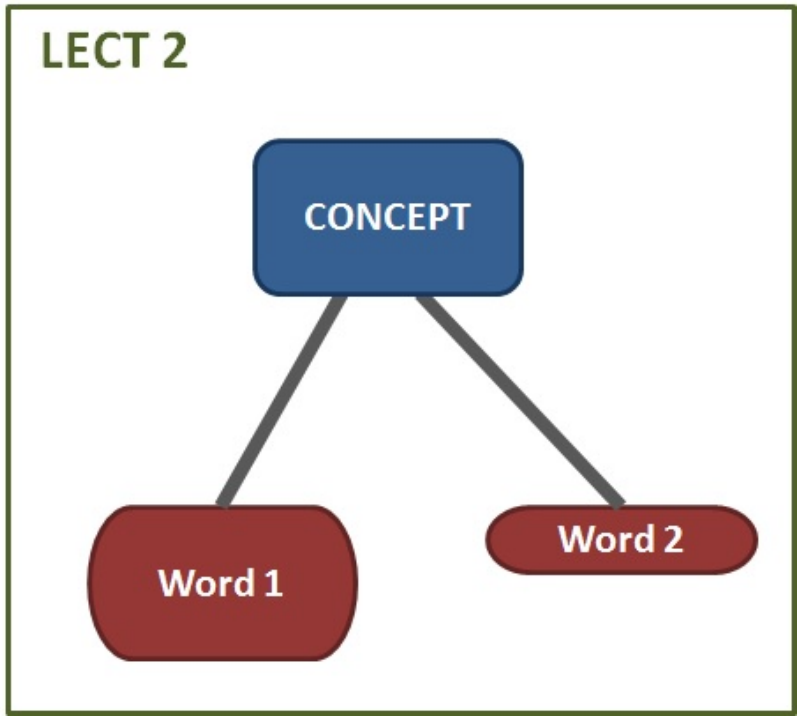
d Without fastening on the end of the legs

Figure 3.4(1)
The semasiological structure of *legging*

The Profile-Based Approach (1999)

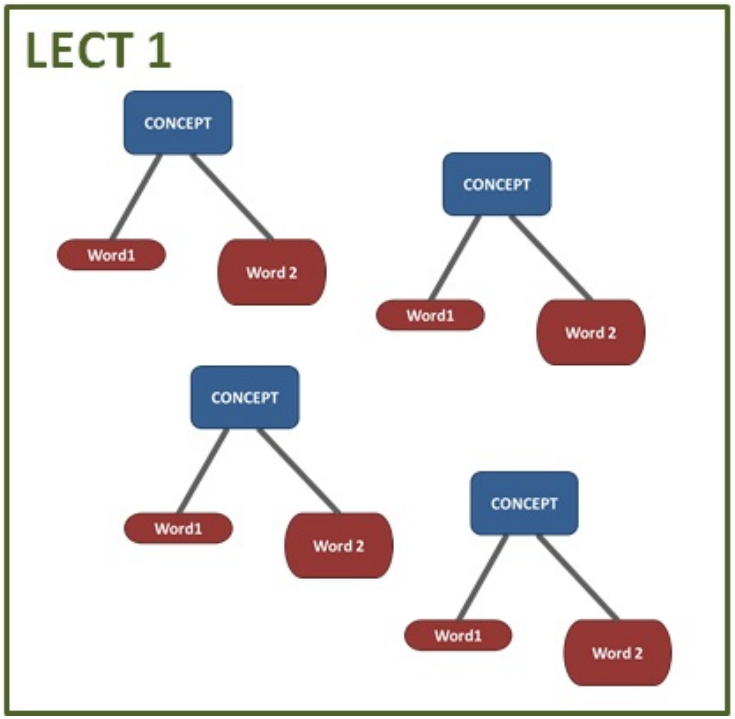
FORMAL ONOMASIOLOGICAL VARIATION

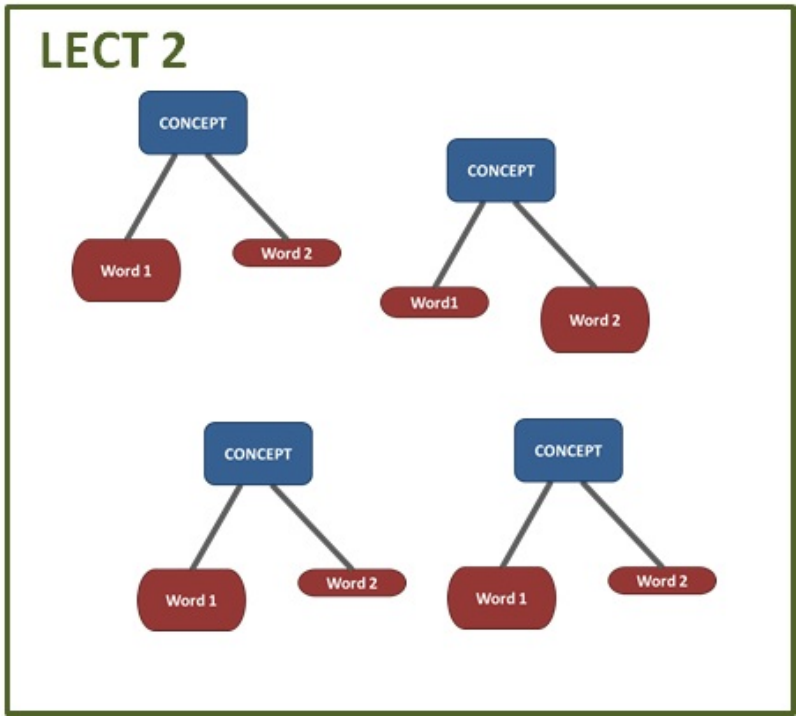




The Profile-Based Approach (1999)

AGGREGATED LEXICAL VARIATION





The Profile-Based Approach (1999)

Profile based measurement of lexical variation

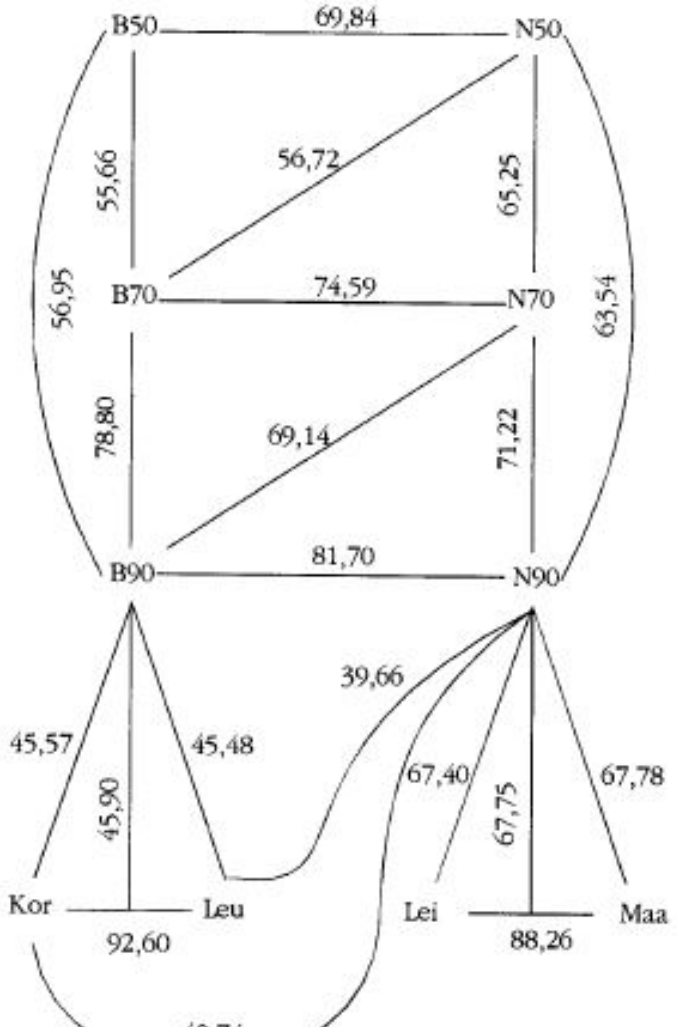
Do B and NL use the same lexemes to refer to a given concept?

- define a set of synonyms = profile
- collect all instances from 2 corpora (B vs NL) + disambiguate
- Calculate relative frequency of synonyms in 2 corpora
- Overlap in relative frequency = uniformity measure

	B	NL	overlap
jeans	85	30	30
spijkerbroek	15	70	15
			45

The Profile-Based Approach (1999)

AGGREGATED UNIFORMITY B-NL



The Profile-Based Approach (1999)

- sports and clothing terms: Geeraerts et al. 1999, 2003
- empirical, corpus-based, quantitative
- BUT time consuming, not readily scalable
  - Profiles (synonyms) manually selected
  - Occurrences of synonyms manually disambiguated

sem·metrix project (Peirsman & Heylen)

WHAT?

- semi-automatic generation of synonym sets (profiles)
- Word Sense Disambiguation for profile membership

HOW?

- Distributional semantic modelling in large corpora

Overview

Semantic Vector Spaces

Linguistic origin: Distributional Hypothesis

- "You shall know a word by the company it keeps" (Firth)
- a word's meaning can be induced from its co-occurring words
- long tradition of collocation studies in corpus linguistics

Semantic Vector Spaces in Computational Linguistics

- standard technique in statistical NLP for the large-scale automatic modeling of (lexical) semantics
- aka Vector Spaces Models, Distributional Semantic Models, Word Spaces,... (cf Turney & Pantel 2010 for overview)
- generalised, large-scale collocation analysis
- mainly used for automatic thesaurus extraction:
  $\Rightarrow$ words occurring in same contexts have similar meaning

Type-level SVS

Collect co-occurrence frequencies for a large part of the vocabulary and put them in a matrix

	transport	train	commute	ticket	scene	sugar	cream	now
subway	120	424	388	82	12	11	3	189
underground	154	401	376	99	305	20	1	123
coffee	5	8	18	4	1	72	102	152

Type-level SVS

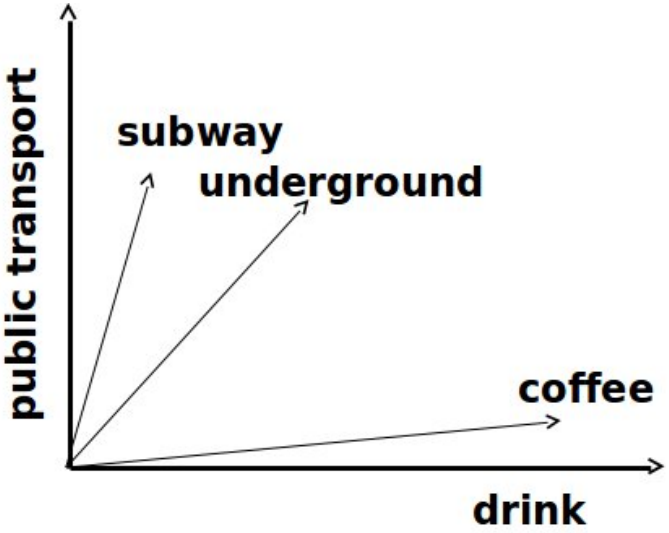
weight the raw frequencies by collocational strength (pmi)

	transport	*train*	*commute*	*ticket*	*scene*	*sugar*	*milk*	*now*
subway	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
underground	4.3	8.1	5.7	3.2	6.2	0.5	0.0	0.1
coffee	0.1	0.2	0.4	0.1	0.0	6.4	7.2	0.1

Type-level SVS

calculate word by word similarity matrix

	subway	underground	coffee
subway	1	.71	.08
underground	.71	1	.09
coffee	.08	.09	1

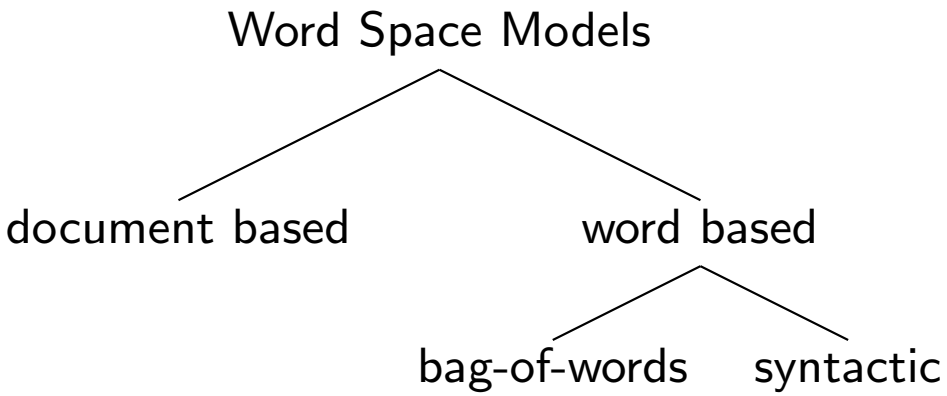


Different Context Models

Vector Space Models come in many flavours

The main difference between the models lies in how they define context

Family of models



Different Context Models

document based models

- context = stretch of text in which target word occurs
- 2 words are related when they often co-occur in text
- Landauer & Dumais 1997: Latent Semantic Analysis

word based models

- context = context words around the target word
- 2 words are related when they co-occur with the same context words, but not necessarily with each other

	DOC.1	DOC.2	DOC.3	DOC.4	DOC.5	DOC.6	DOC.7	DOC.8
ongeval	23	12	14	24	8	0	0	0
ongeluk	16	9	11	18	17	20	0	1
koffie	0	0	0	0	0	14	12	15

	auto	slachtoffer	vrachtwagen	file	gekwetst	suiker	melk	kop
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5
koffie	5	8	18	4	1	72	102	93

Different Context Models

Within word based models:

bag-of-words

- context words in window of n words left and right of target word
- a bag of unstructured context features

syntactic features

- context words in specific syntactic relation with target word
- takes clause structure into account
- Lin 1998, Padó & Lapata 2007

The wagging dog barked at the postman on the bike

	wagging	dog	bark	postman	bike
dog	1	0	1	1	1
postman	1	1	1	0	1

The wagging dog barked at the postman on the bike

	$=$subj.bark	$+$adj.wagging	$=$PC.bark.at	$+$PP.on.bike
dog	1	1	0	0
postman	0	0	1	1

Different Context Models

Within the bag-of-words models:

1st order co-occurrences

- context = words in immediate proximity to the target
- Levy & Bullinaria 2001

2nd order co-occurrences

- context = context words of context words of target
- can generalise over semantically related context words
- Schütze 1998

NB syntactic models are also 1st order models

‘s Morgens veroorzaakte een ongeval een file

‘s Morgens veroorzaakte een ongeval een file

Hij kan ‘s morgens zijn **bed** niet uit
Neem ‘s morgens **tijd** voor een **ontbijt**
‘s Morgens gaat de **wekker** altijd te **vroeg** af

Roken veroorzaakt **kanker**
CO2 veroorzaakt **klimaatopwarming**
De **sneeuw** veroorzaakt **problemen**

‘s Morgens veroorzaakte een ongeval een file

Hij kan ‘s morgens zijn **bed** niet uit
Neem ‘s morgens **tijd** voor een **ontbijt**
‘s Morgens gaat de **wekker** altijd te **vroeg** af

Roken veroorzaakt **kanker**
CO2 veroorzaakt **klimaatopwarming**
De **sneeuw** veroorzaakt **problemen**

‘s Morgens veroorzaakte een ongeval een file

Ze **stond uren** in de file
Op de **ringweg staat** een file
Met de **auto** sta je sowieso in de file

Hij kan ‘s morgens zijn **bed** niet uit
Neem ‘s morgens **tijd** voor een **ontbijt**
‘s Morgens gaat de **wekker** altijd te **vroeg** af

Roken veroorzaakt **kanker**
CO2 veroorzaakt **klimaatopwarming**
De **sneeuw** veroorzaakt **problemen**

‘s Morgens veroorzaakte een ongeval een file

Ze **stond uren** in de file
Op de **ringweg staat** een file
Met de **auto** sta je sowieso in de file

Hij kan ‘s morgens zijn **bed** niet uit
Neem ‘s morgens **tijd** voor een **ontbijt**
‘s Morgens gaat de **wekker** altijd te **vroeg** af

Roken veroorzaakt **kanker**
CO2 veroorzaakt **klimaatopwarming**
De **sneeuw** veroorzaakt **problemen**

‘s Ochtends veroorzaakte een ongeval een file

Ze **stond uren** in de file
Op de **ringweg staat** een file
Met de **auto** sta je sowieso in de file

Hij kan ‘s ochtends zijn **bed** niet uit
Neem ‘s ochtends **tijd** voor een **ontbijt**
‘s ochtends gaat de **wekker** altijd te **vroeg** af

Different Context Models

Comparing models

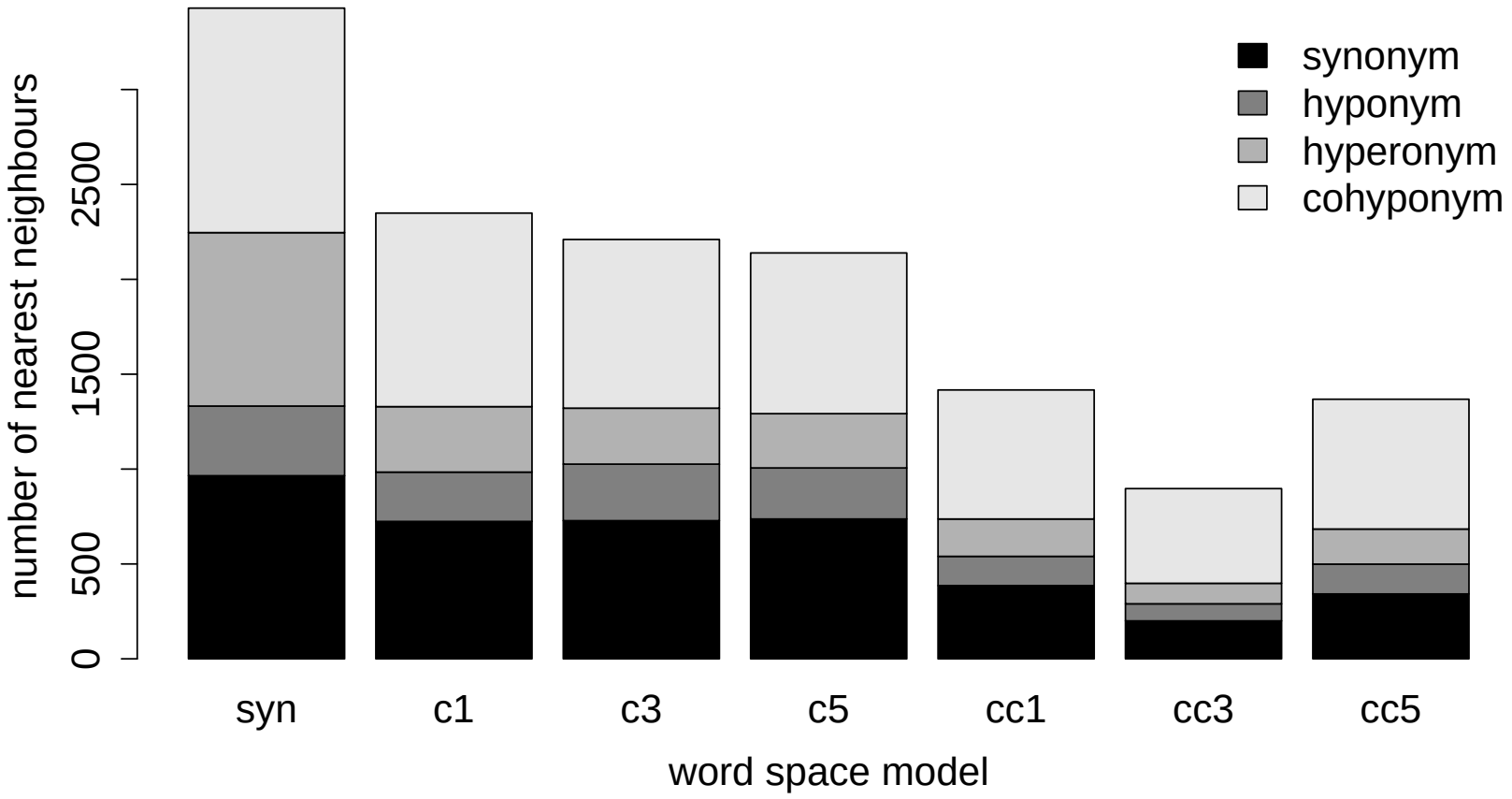
- syntactic model
- first-order bag-of-word models with context sizes 1, 3 and 5
- second-order bag-of-word models with context sizes 1, 3 and 5

Parameters

- data: Twente Nieuws Corpus (400M words)
- 2000 dimensions: the most frequent features in the corpus
- weighting scheme: point-wise mutual information
- similarity calculated with cosine

Different Context Models

1 neighbour



Different Context Models

- Syntactic model > first-order bag-of-words > second-order bag-of-words
- The "stricter" the definition of context, the more semantically similar words.
- demo: https://perswww.kuleuven.be/ ~u0042527/demo.html

- AUTOMATIC GENERATION OF SYNONYM PROFILES

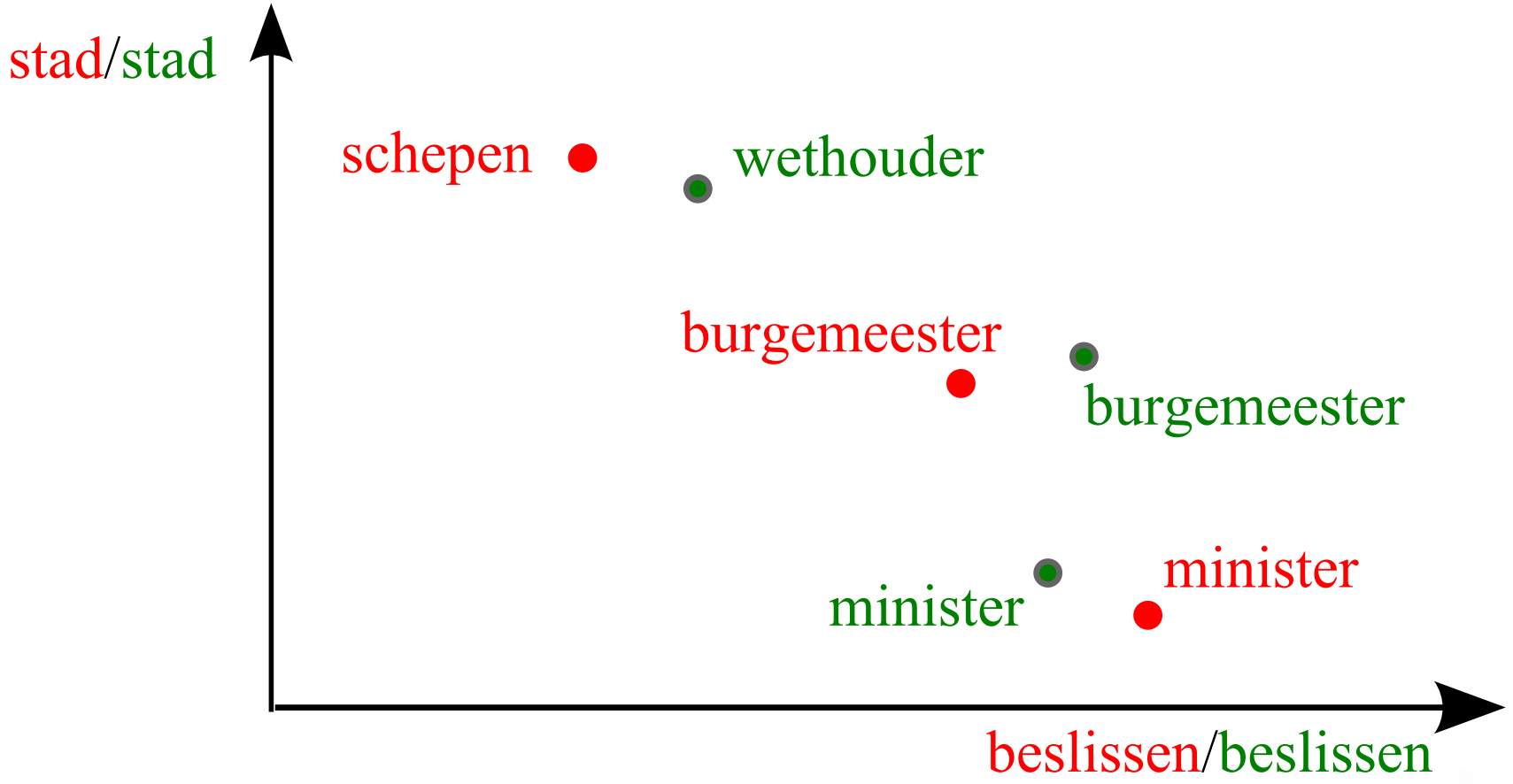
Extension: Finding Equivalents

Bilectal Word Space Models

- Extend Word Space Models from one corpus to two corpora
- Word co-occurrences are relatively constant across varieties
- Identification of synonyms in different varieties of the same language
  - BE dessert – NL toetje, UK nappy – US diaper

	road	*drive*	*engine*	*traffic*	*baby*	*clean*	*toilet*	*legs*
car (UK)	120	424	221	189	6	11	3	2
car (US)	152	389	178	167	15	9	5	1
nappy(UK)	3	12	16	9	270	189	318	167
diaper (US)	8	11	13	7	301	205	324	179

Two varieties



Lexical Sociolectometry (Ruette)

Large-scale Profile-based approach

- Measure distances betweeen language varieties based on onomasiological variation.
- Aggregated over automatically selected profiles of near-synonyms

Corpora and varieties

	Usenet	Popular news	Quality news	Legalese	Total
BE	22 million	905 million	373 million	70 million	1.4 billion
NL	26 million	126 million	161 million	115 million	428 million
Total	48 million	1 billion	499 million	185 million	1.8 billion

Lexical Sociolectometry (Ruette)

Clustering by Committee

1. All pairwise similarities between 10K nouns in the corpus are calculated.
2. A cluster algorithm (Pantel et al., 2002) finds sets of similar nouns (profile).
3. A trained linguist goes through these sets and filters the sets that contain synonyms.

For now, semi-automatic variable set generation

- The SVS approach ensures automatic, bottom-up generation of candidate variables, with high recall, but low precision.
- The manual correction of the trained linguist ensures the quality of the variable set.

Lexical Sociolectometry

Concept	Items
Manner	wijze, manier
Genocide	volk_moord, genocide
Poll	peiling, opiniepeiling
Marihuana	cannabis, marihuana
Putsch	staatsgreep, coup
Meningitis	hersenvliesontsteking, meningitis
Demonstrator	demonstrant, betoger
Airport	vliegveld, luchthaven
Coldness	koude, kou
Torture	marteling, foltering
Victory	zege, overwinning
Homosexual	homo, homoseksueel
Saxophone	sax, saxofoon
Internetprovider	provider, internetprovider, internetaanbieder
Airconditioning	airconditioning, airco
Religion	religie, godsdienst
The other side	overkant, overzijde
Explosion	explosie, ontploffing
Restroom	toilet, wc

Lexical Sociolectometry (Ruette)

Distance metric

- "different lexemes refering to the same concept"
- concept naming: if two subcorpora use the same word to refer to a certain concept, they are considered to be similar
- similarity between two subcorpora is measured by summing the City-block distance for every concept

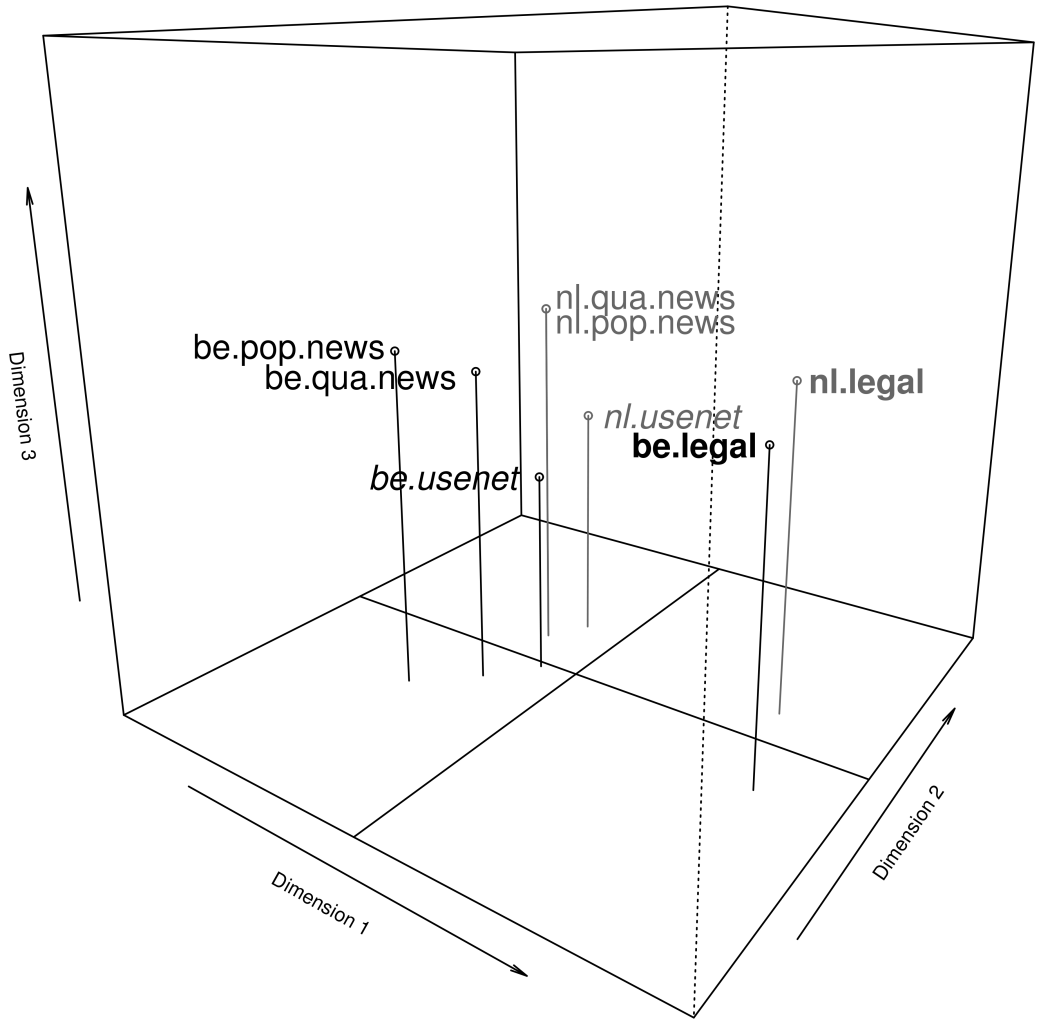
Distance metric

$$D_{CB,L}(V_1, V_2) = \frac{1}{2}\sum_{i=1}^{n} |R_{V_1,L}(x_i) - R_{V_2,L}(x_i)| \quad (1)$$

$$D_{CB}(V_1, V_2) = \sum_{i=1}^{m} (D_{L_i}(V_1, V_2) W(L_i)) \quad (2)$$

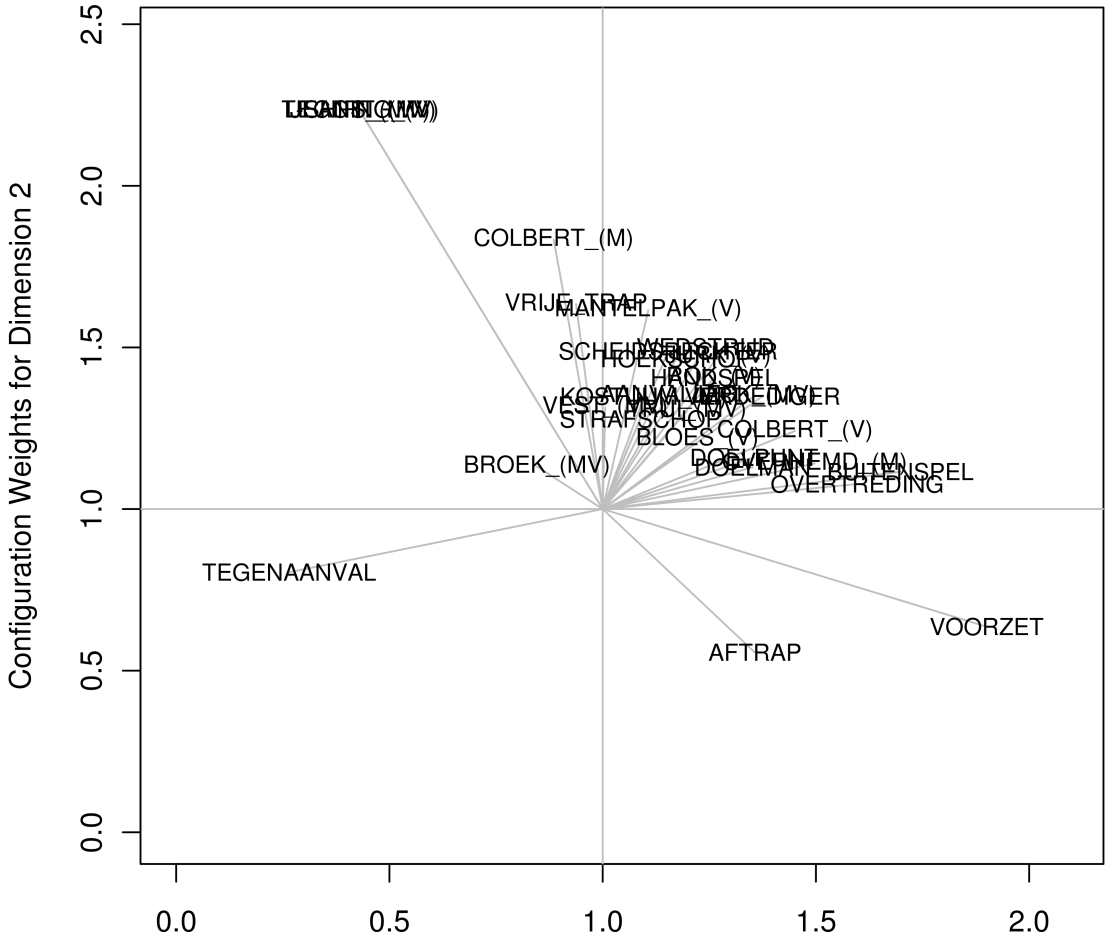
Lexical Sociolectometry (Ruette)

Non-parametric MDS



Lexical Sociolectometry (Ruette)

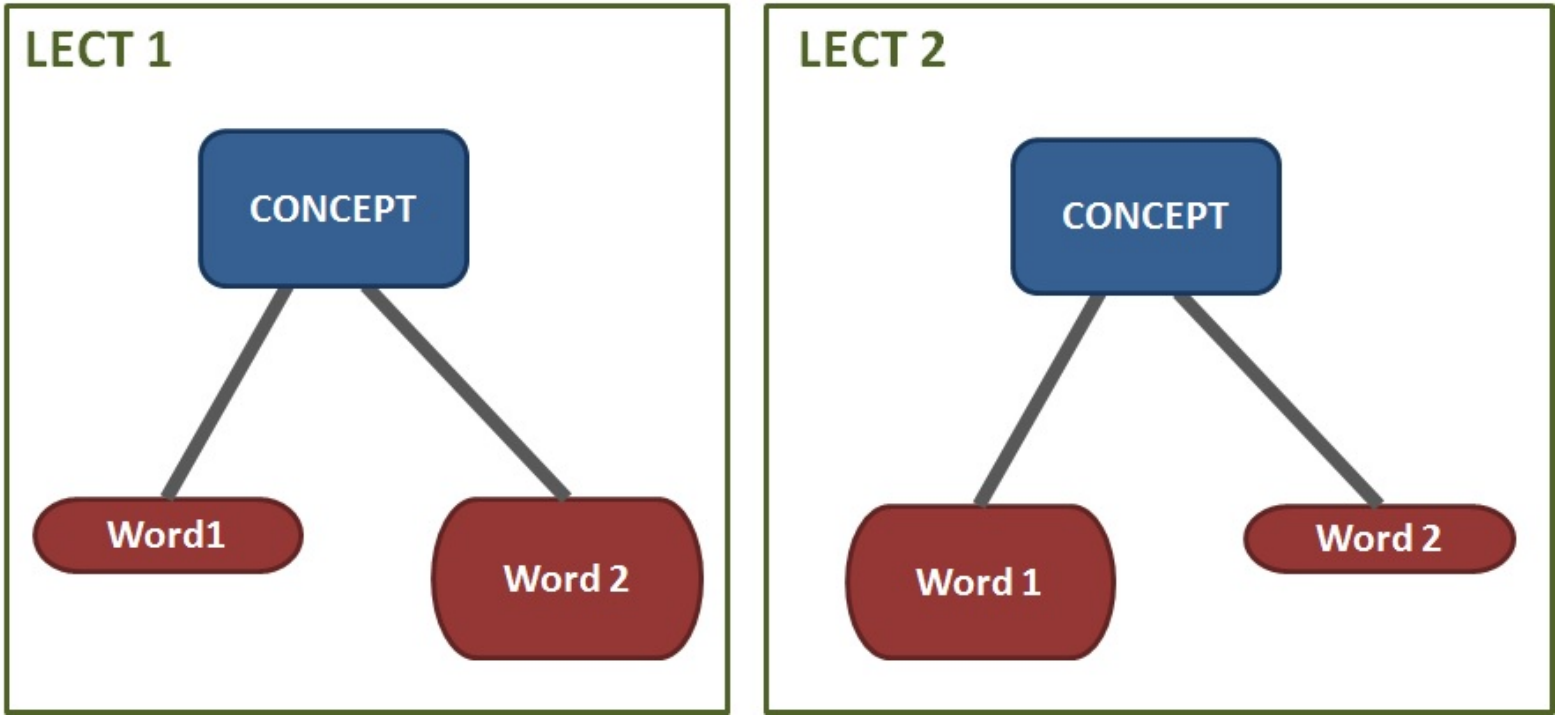
Individual difference scaling



Overview

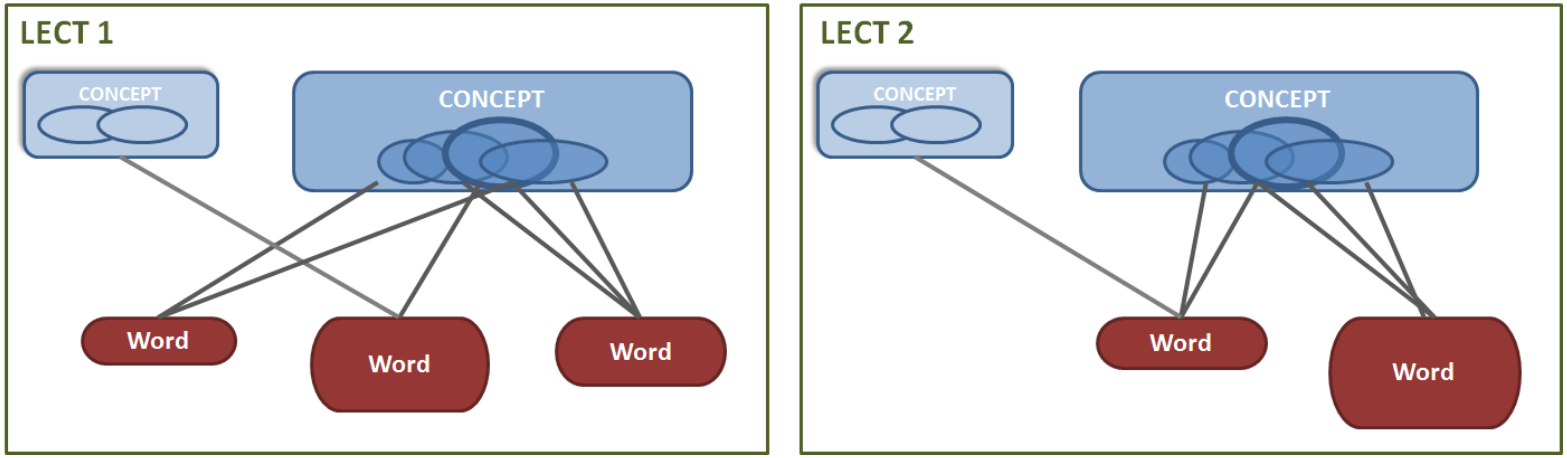
Semasiological Variation

Profile-Based approach



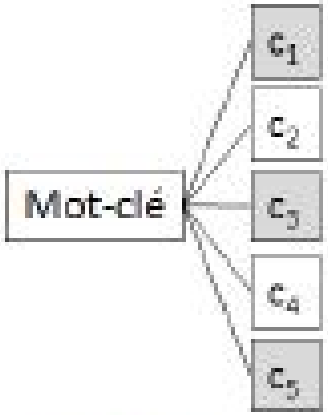
Semasiological Variation

Polysemy:



Measuring Polysemy (Bertels)

- Univocity hypothesis in Wuesterian terminology: Technical terms are monosemous
- Bertels: correlation between keyness and monosemy
- Monosemy-measure: overlap of second-order co-occurrences

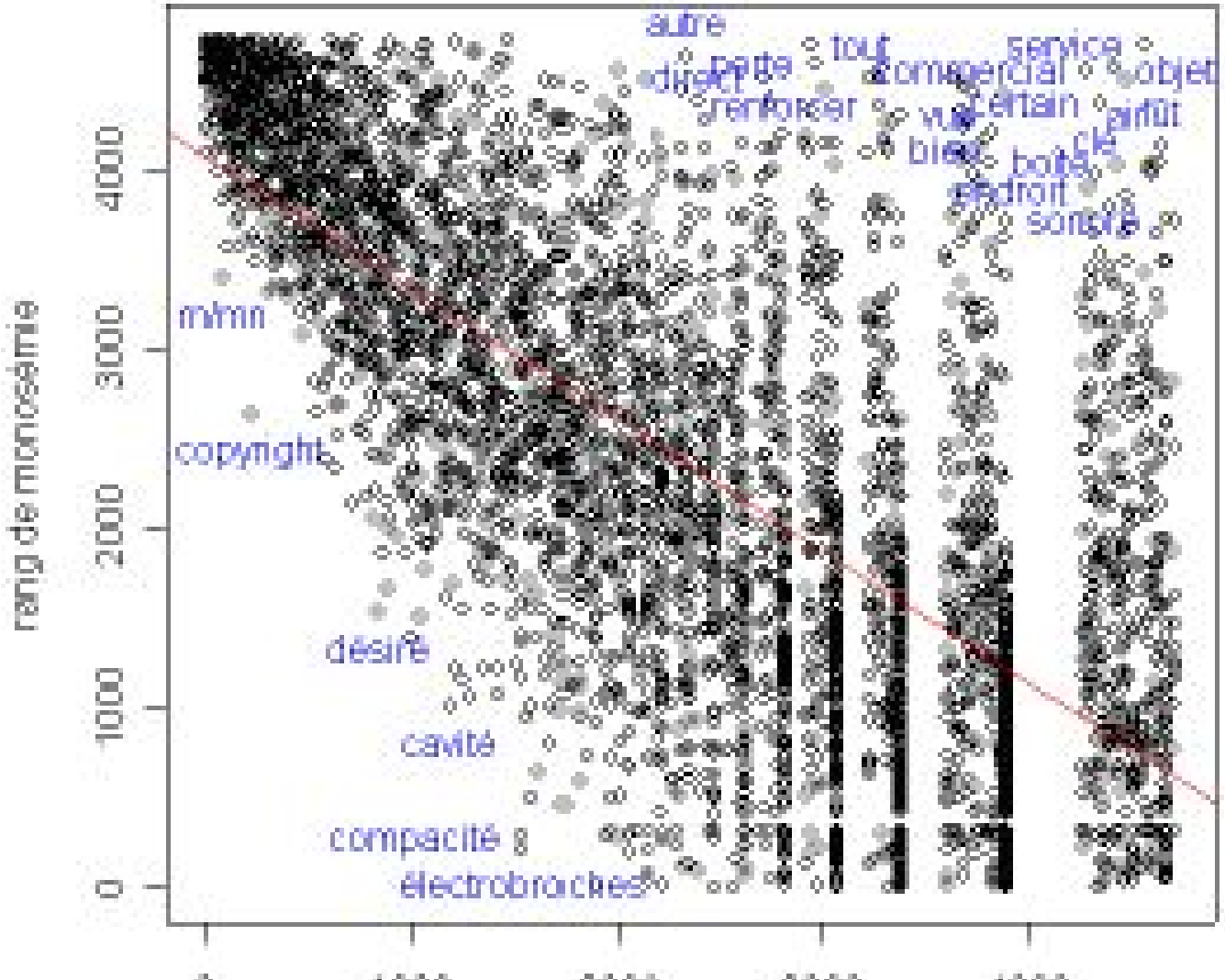


cc_1	cc_2	cc_3	cc_4	cc_5	cc_6	cc_7	cc_8	cc_9	cc_{10}
1	0	1	0	0	0	1	1	0	1
1	0	1	0	1	0	1	0	1	1
0	0	1	0	1	1	0	1	0	0
1	1	1	0	0	0	1	1	0	1
1	0	1	1	0	0	1	1	0	0

FIGURE 1 – Schéma : mot-clé + cooccurrents de premier et de deuxième ordre.

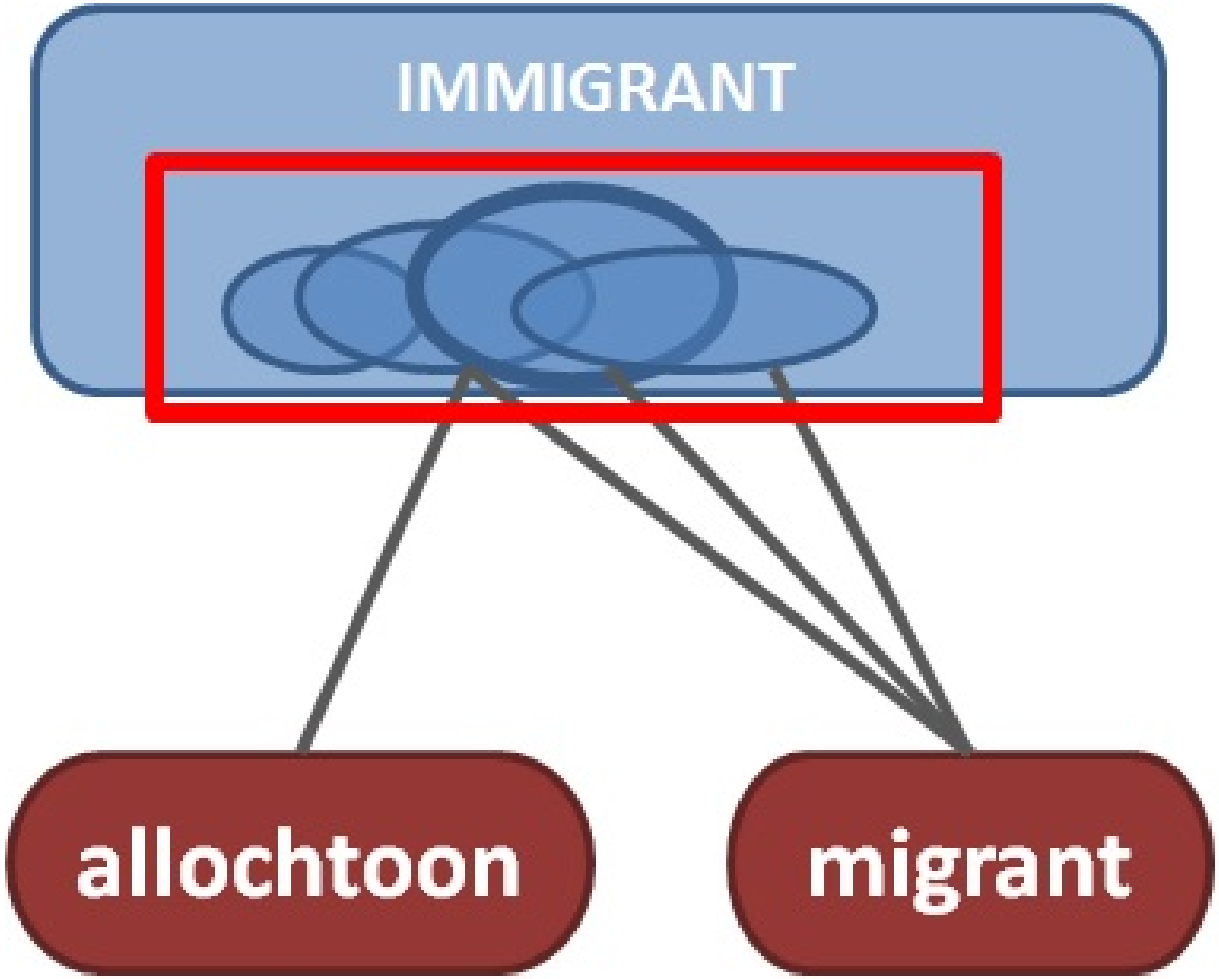
Measuring Polysemy (Bertels)

Inverse relation between keyness and monosemy!



Analysing Semantic Structure

Which semantic features constitute the prototypical structure of the concept?



Analysing Semantic Structure

Extract strongest concept collocations from matrix

	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*	*hond*
allochtoon	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
migrant	4.3	8.1	5.7	3.2	6.2	0.5	0.0	0.1

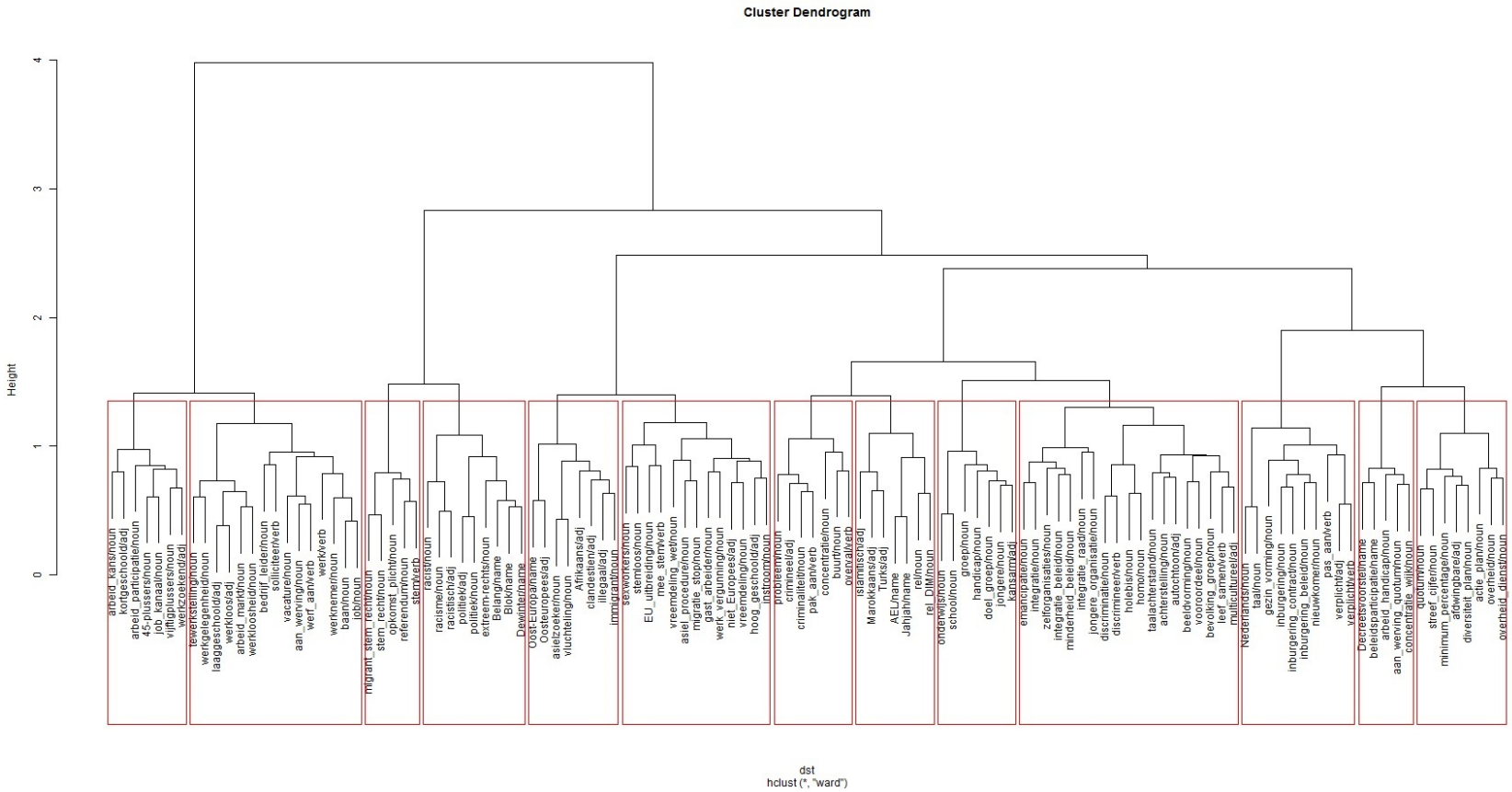
Analysing Semantic Structure

Make weighted co-occurrence matrix for these collocations

	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*	*hond*
jobs	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
racisme	4.3	8.1	5.7	3.2	6.2	0.5	0.0	0.1
integratie	5.3	7.9	6.5	6.0	0.8	0.6	0.1	0.0
misdaad	4.3	8.1	5.7	2.2	6.2	0.4	0.0	0.1
stemrecht	5.3	7.9	6.5	8.0	0.8	0.9	0.3	0.0

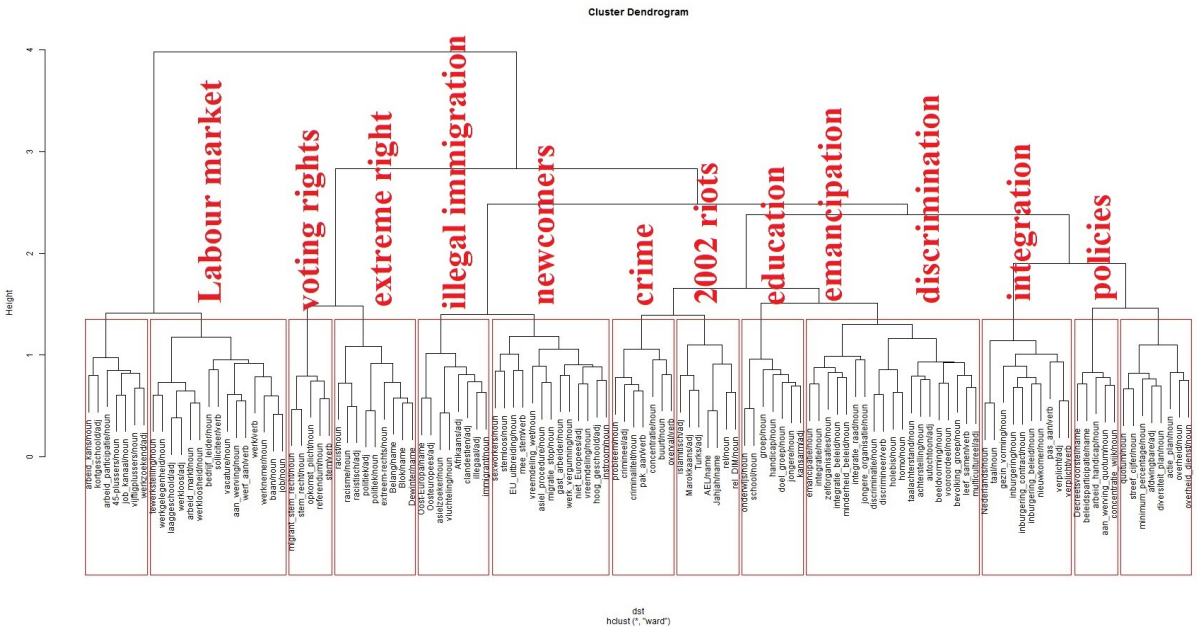
Analysing Semantic Structure

Calculate similarity between collocations and feed to it a (hierarchical) cluster analysis

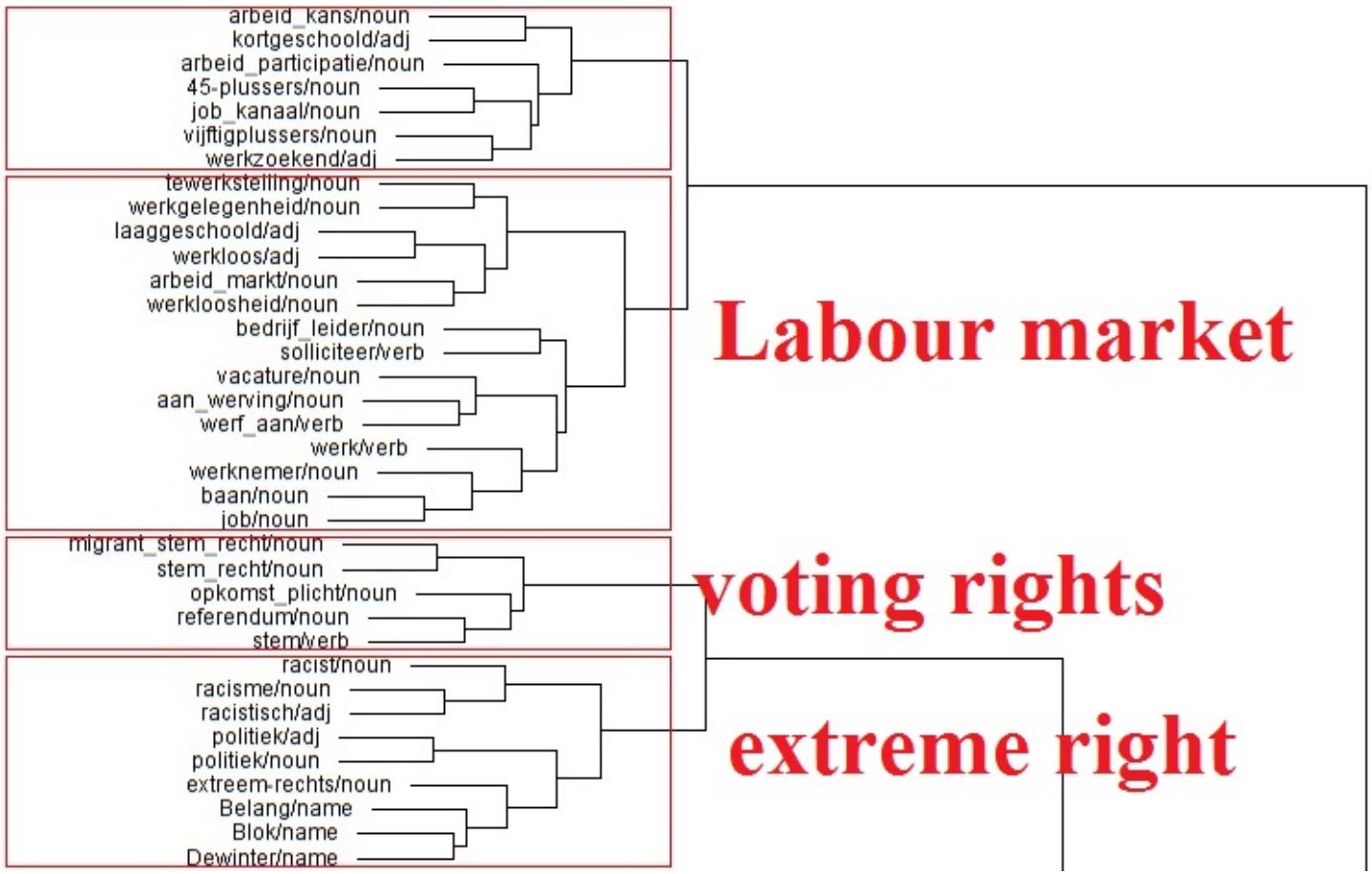


Analysing Semantic Structure

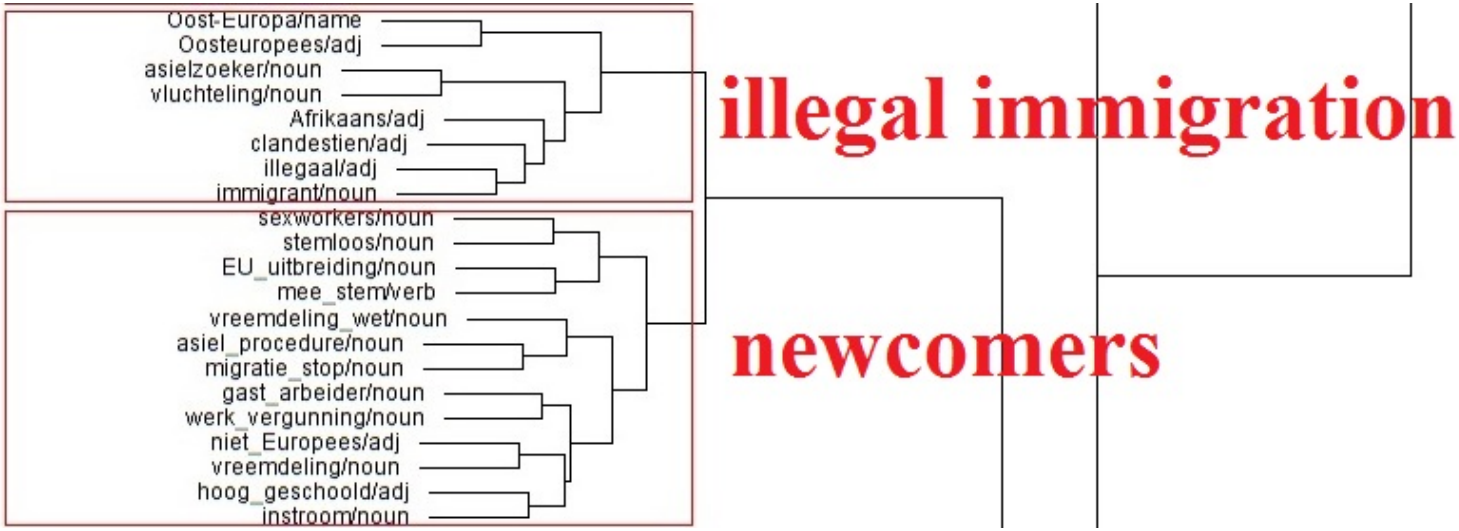
Clusters of contextually related collocations ≈ semantic features
Clusters can be labeled manually



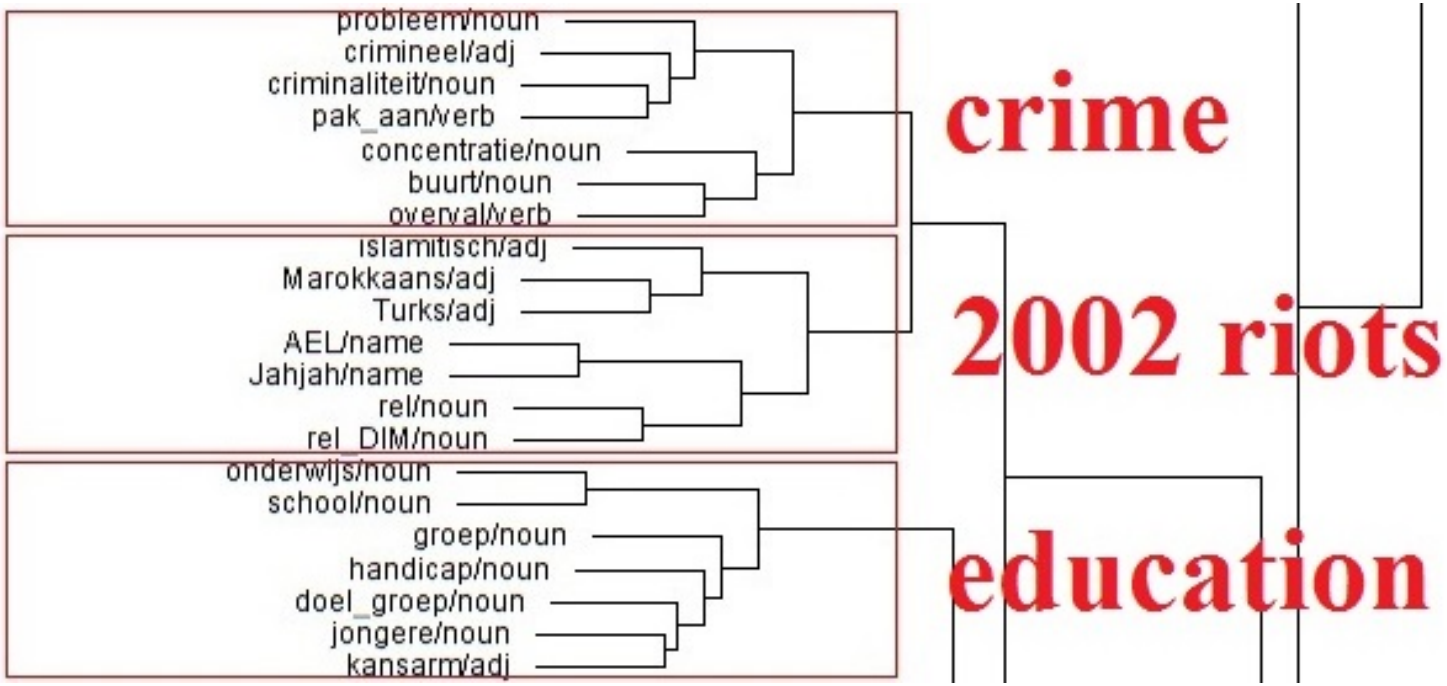
Analysing Semantic Structure



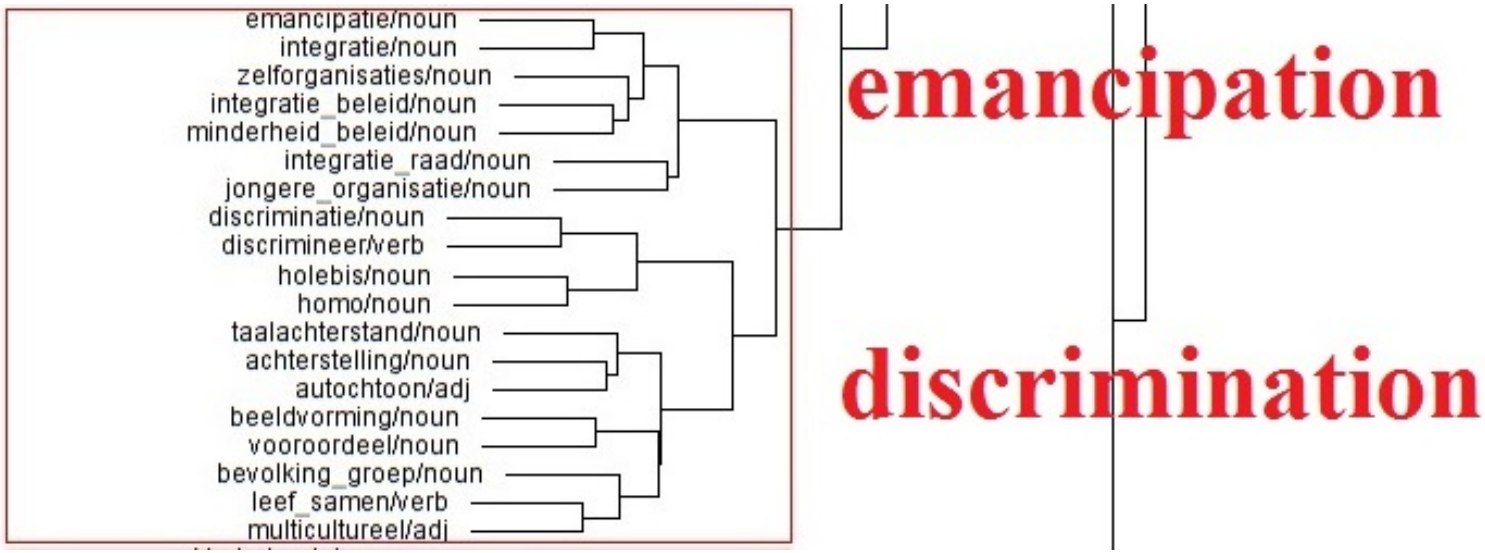
Analysing Semantic Structure



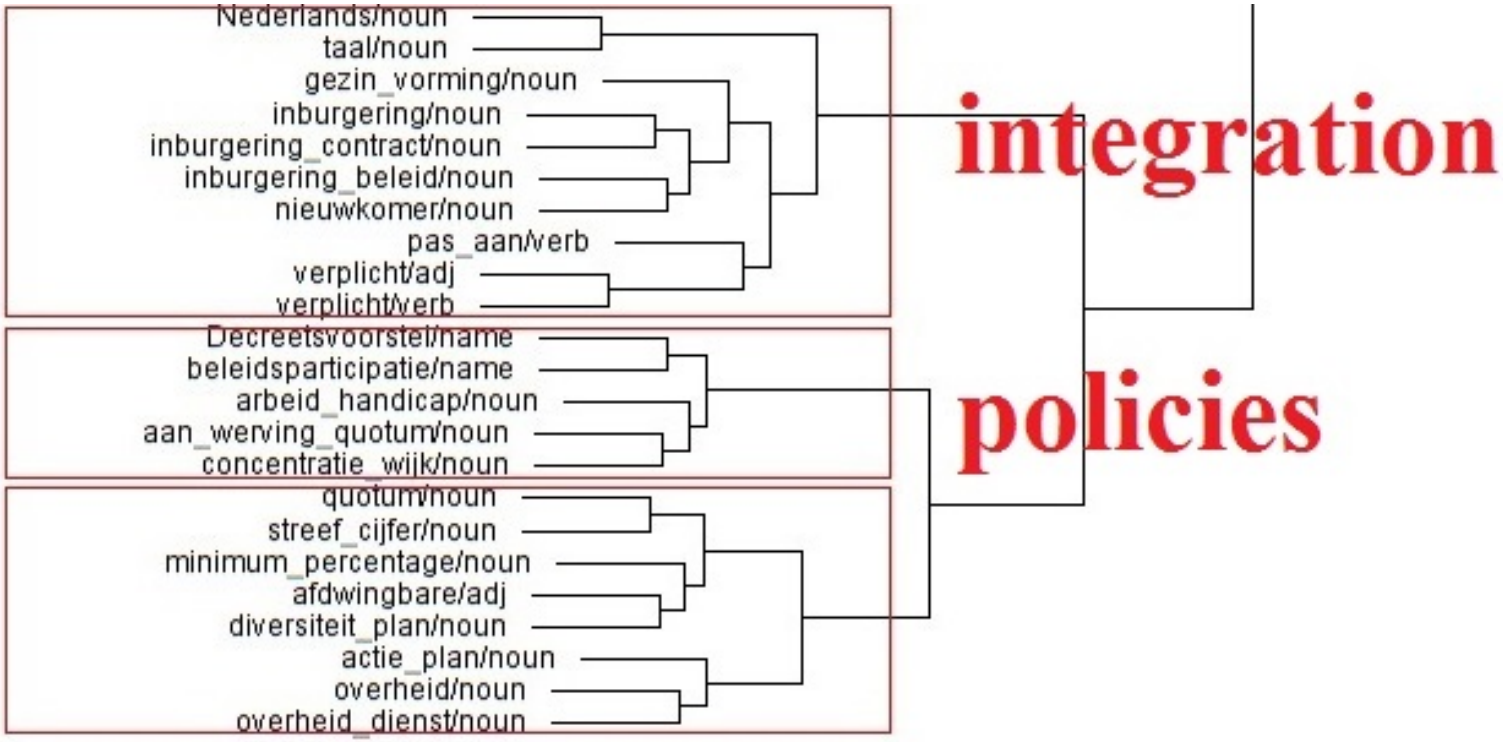
Analysing Semantic Structure



Analysing Semantic Structure

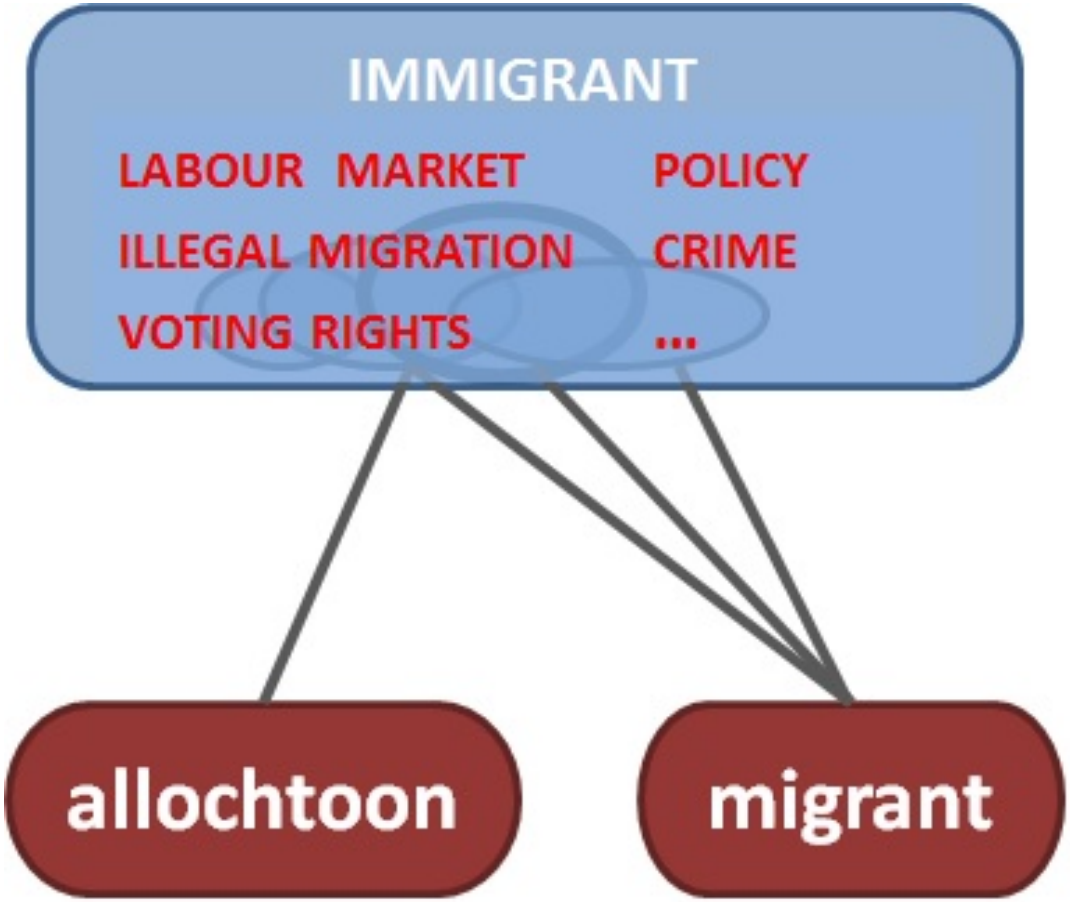


Analysing Semantic Structure



Analysing Semantic Structure

Contextually defined "semantic features" that constitute the prototypical structure of the concept

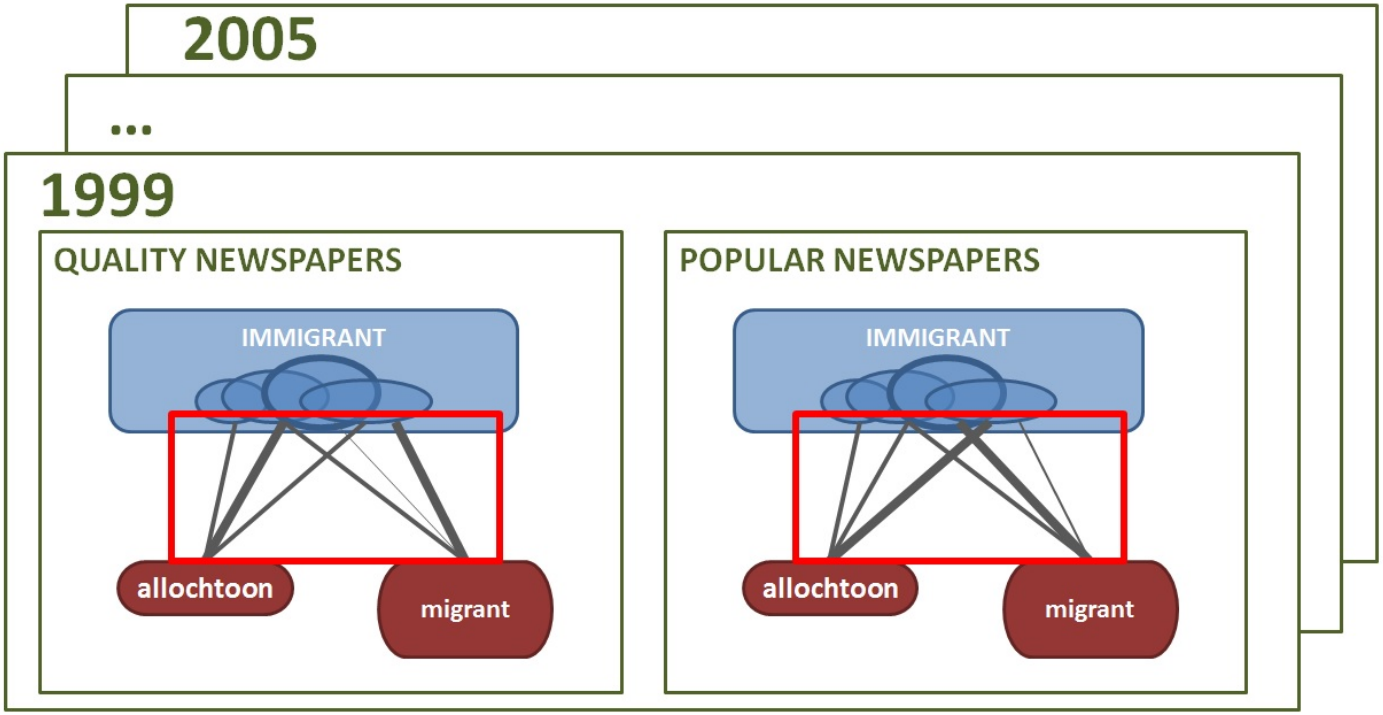


Measuring semantic change in registers

- How strong are *allochtoon* and *migrant* associated with the different context cluster/semantic features
- Is the strength of association the same in quality and popular newspapers?
- Does the strength of association change over time?

Measuring semantic change in registers

What is association strength between semantic features and lexemes in different registers and periods?



Measuring semantic change in registers

STEP 1

Make separate vectors per variant, per year, and per newspaper type

	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*
allochtoon/1999pop	5.3	7.9	6.5	4.0	0.8	0.6	0.0
migrant/1999pop	4.3	8.1	5.7	3.2	6.2	0.5	0.0
allochtoon/1999qual	4.3	2.9	7.5	8.1	0.3	1.6	0.3
migrant/1999qual	4.3	4.2	5.7	3.2	6.2	0.5	0.0
allochtoon/2000pop	5.8	3.5	6.5	5.1	1.3	0.0	0.1
migrant/2000pop	2.9	2.4	4.7	2.2	4.2	0.3	0.7

Measuring semantic change in registers

STEP 2

Make vector per context cluster through aggregation

	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*
jobs	5.3	7.9	6.5	4.0	0.8	0.6	0.0
werk	4.3	8.1	5.7	3.2	6.2	0.5	0.0
arbeidsmarkt	5.3	7.9	6.5	6.0	0.8	0.6	0.1
LABOURMARKET	5.3	7.1	7.7	2.2	6.2	0.4	0.0

Measuring semantic change in registers

STEP 3

Combine variant/year/type vectors and context cluster vectors in 1 matrix

	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*
allochtoon/1999pop	5.3	7.9	6.5	4.0	0.8	0.6	0.0
migrant/1999pop	4.3	8.1	5.7	3.2	6.2	0.5	0.0
allochtoon/1999qual	4.3	2.9	7.5	8.1	0.3	1.6	0.3
migrant/1999qual	4.3	4.2	5.7	3.2	6.2	0.5	0.0
allochtoon/2000pop	5.8	3.5	6.5	5.1	1.3	0.0	0.1
...	...	...	...	...	...	...	...
LABOURMARKET	5.3	7.1	7.7	2.2	6.2	0.4	0.0
...	...	...	...	...	...	...	...

Measuring semantic change in registers

STEP 4

Calculate the cosine similarity (≈ association strength) of each variant/year/type vector to each context cluster vector

	LABOUR	ILLEGAL	EXTREME	POLICY	CRIME	VOTING	RACISM
allochtoon/1999pop	0.3	0.9	0.5	0.0	0.8	0.6	0.0
migrant/1999pop	0.3	0.1	0.7	0.2	0.2	0.5	0.0
allochtoon/1999qual	0.3	0.9	0.5	0.1	0.3	0.6	0.3
migrant/1999qual	0.3	0.2	0.7	0.2	0.2	0.5	0.0
allochtoon/2000pop	0.8	0.5	0.5	0.1	0.3	0.0	0.1
migrant/2000pop	0.9	0.4	0.7	0.2	0.2	0.3	0.7

Measuring semantic change in registers

STEP 5

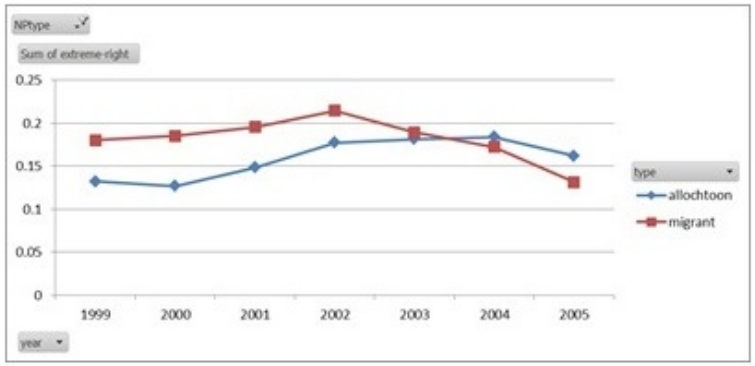
Plot the change of association strength per context cluster and newspaper type

Measuring semantic change in registers

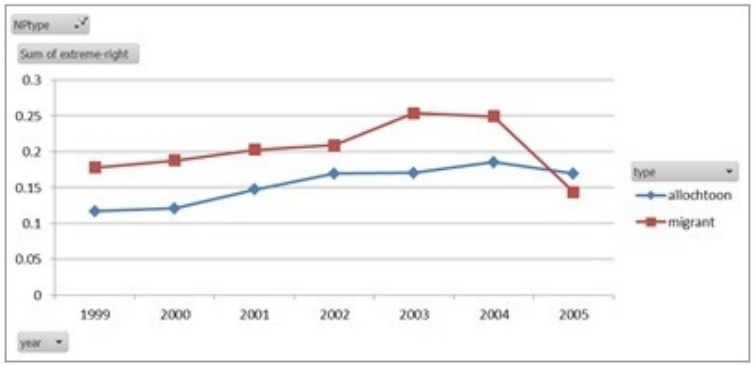
ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

EXT-RIGHT

QUALITY



POPULAR

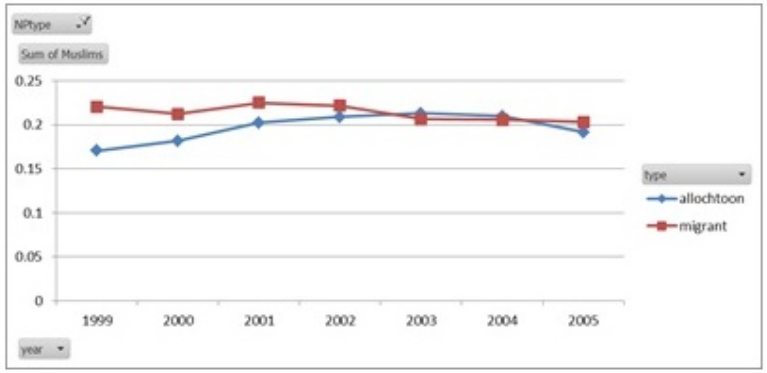


Measuring semantic change in registers

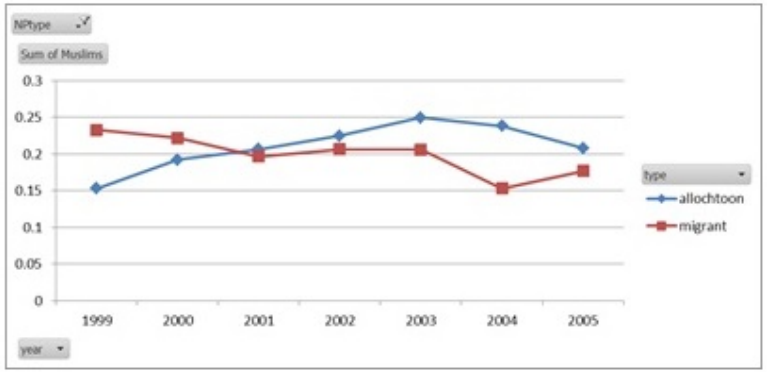
ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

MUSLIMS

QUALITY NP



POPULAR NP

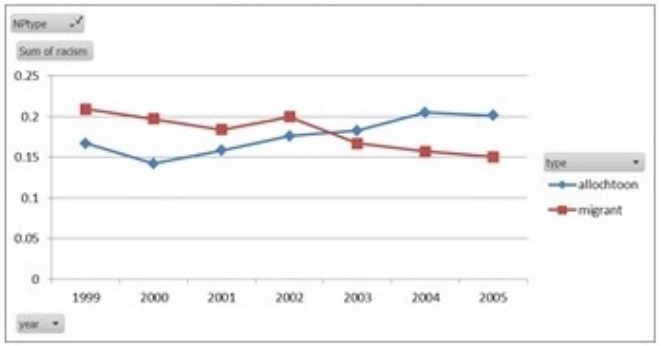


Measuring semantic change in registers

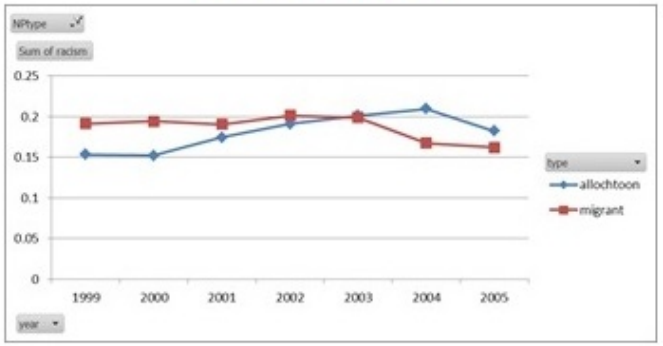
ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

RACISM

QUALITY



POPULAR

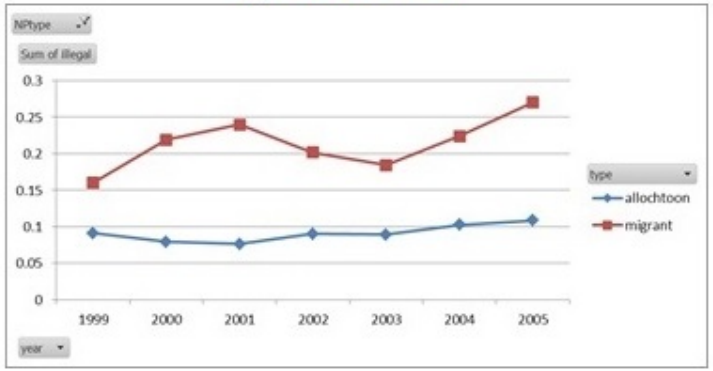


Measuring semantic change in registers

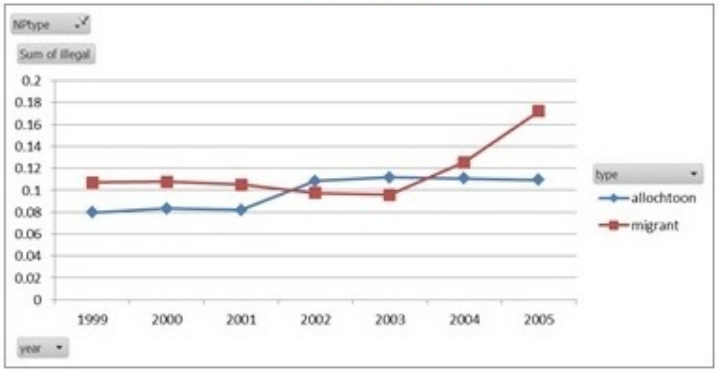
MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

ILLEGAL

QUALITY



POPULAR

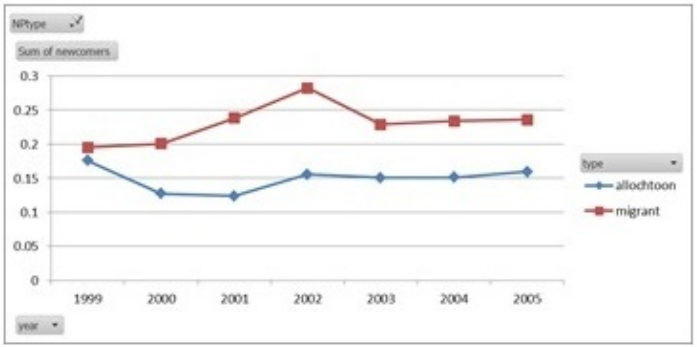


Measuring semantic change in registers

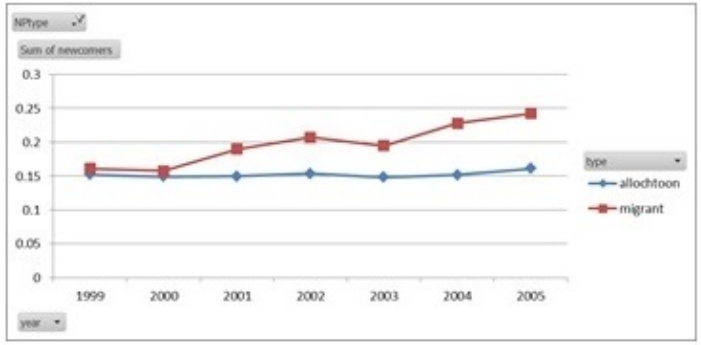
MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

NEW-COMERS

QUALITY



POPULAR

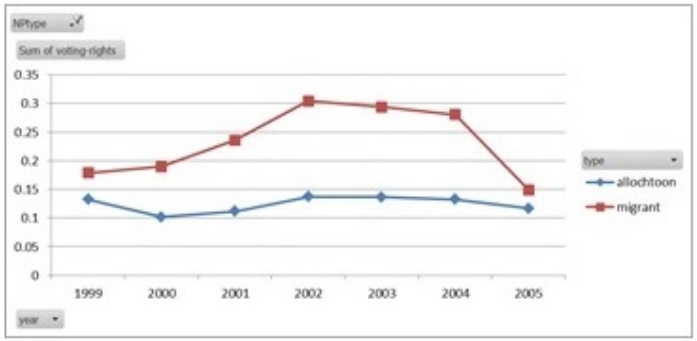


Measuring semantic change in registers

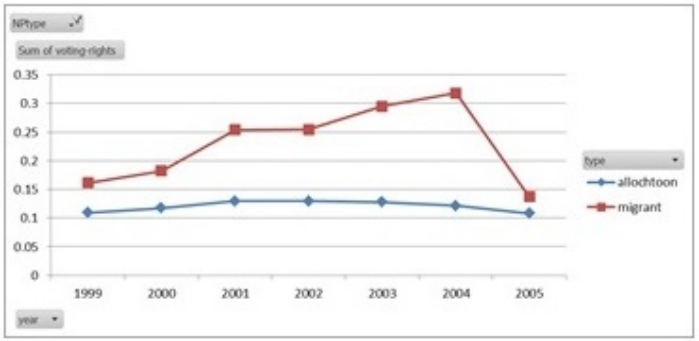
MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

VOTING RIGHTS

QUALITY



POPULAR

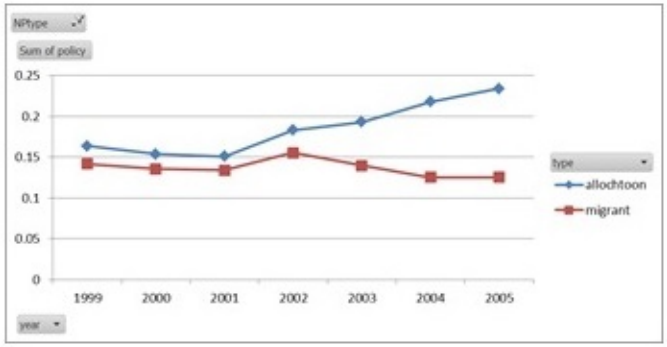


Measuring semantic change in registers

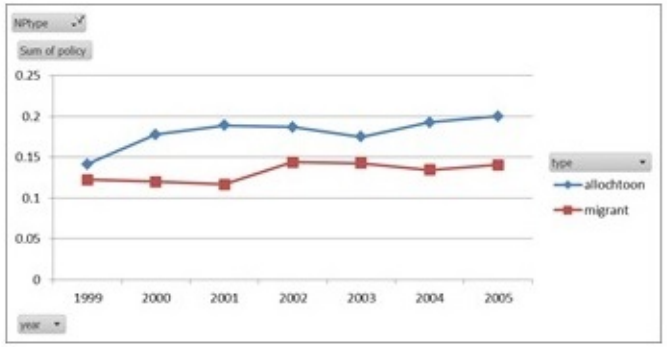
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

POLICY

QUALITY



POPULAR

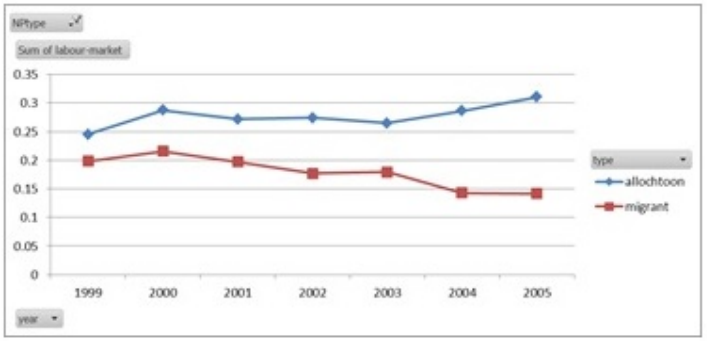


Measuring semantic change in registers

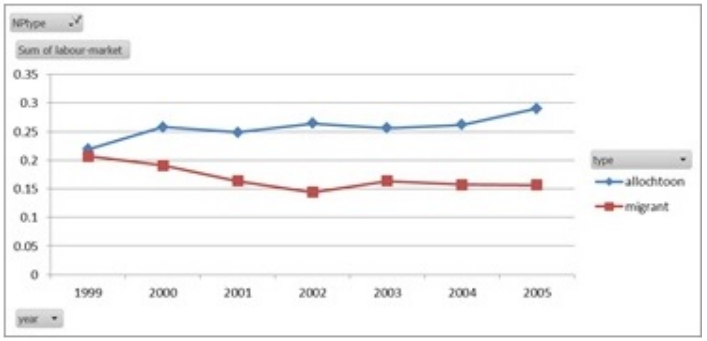
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

LABOUR MARKET

QUALITY NP



POPULAR NP

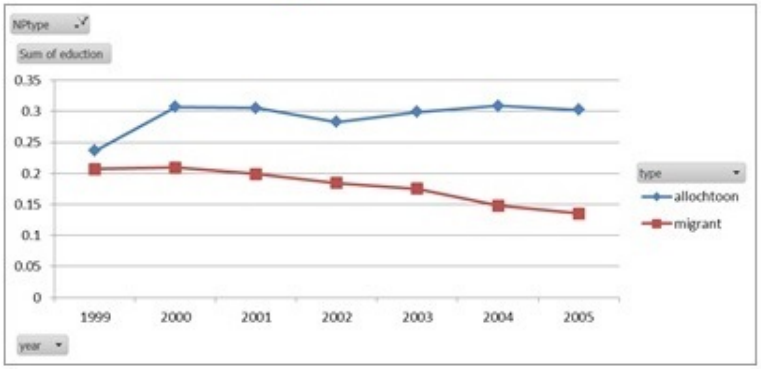


Measuring semantic change in registers

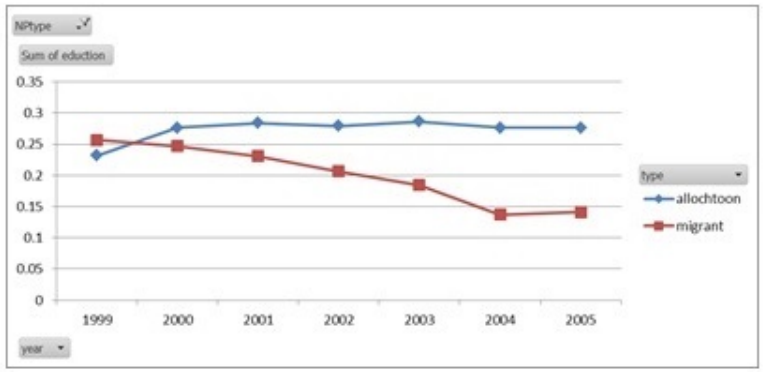
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

EDUCATION

QUALITY



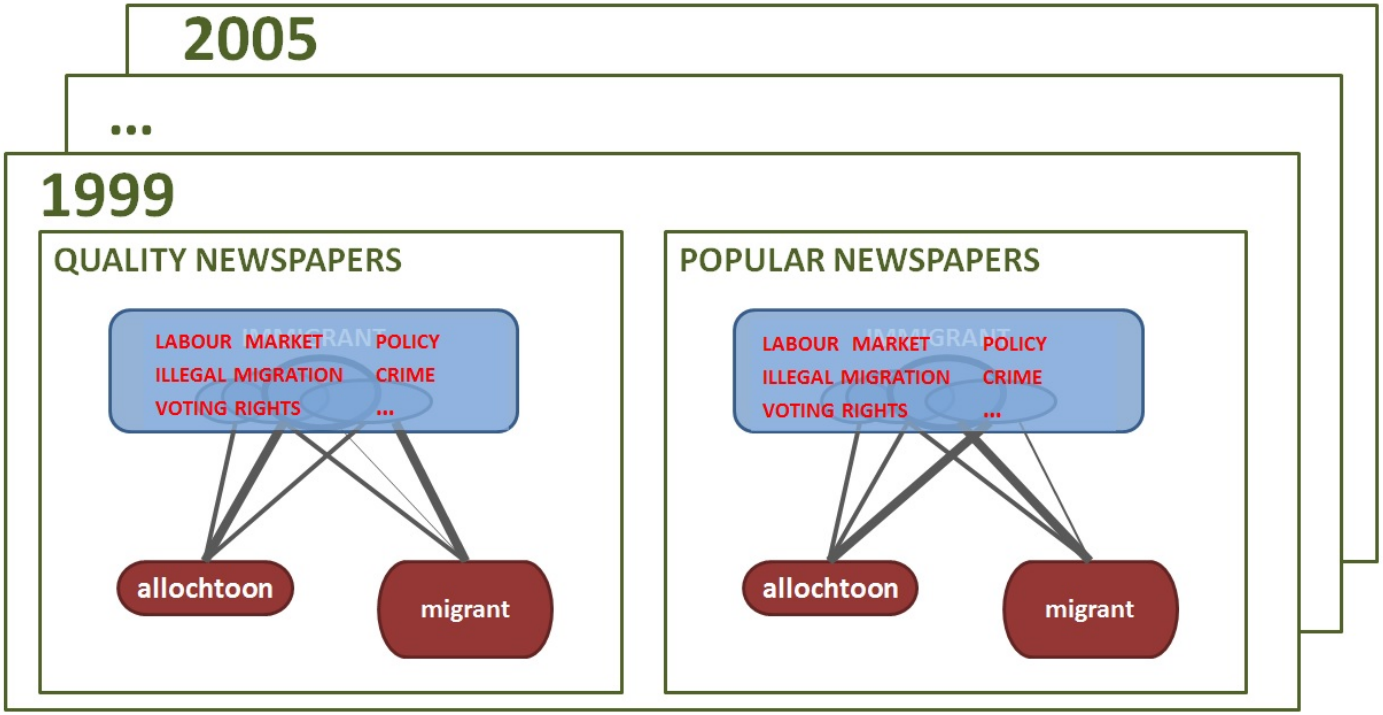
POPULAR



Measuring semantic change in registers

Measuring semantic change in registers

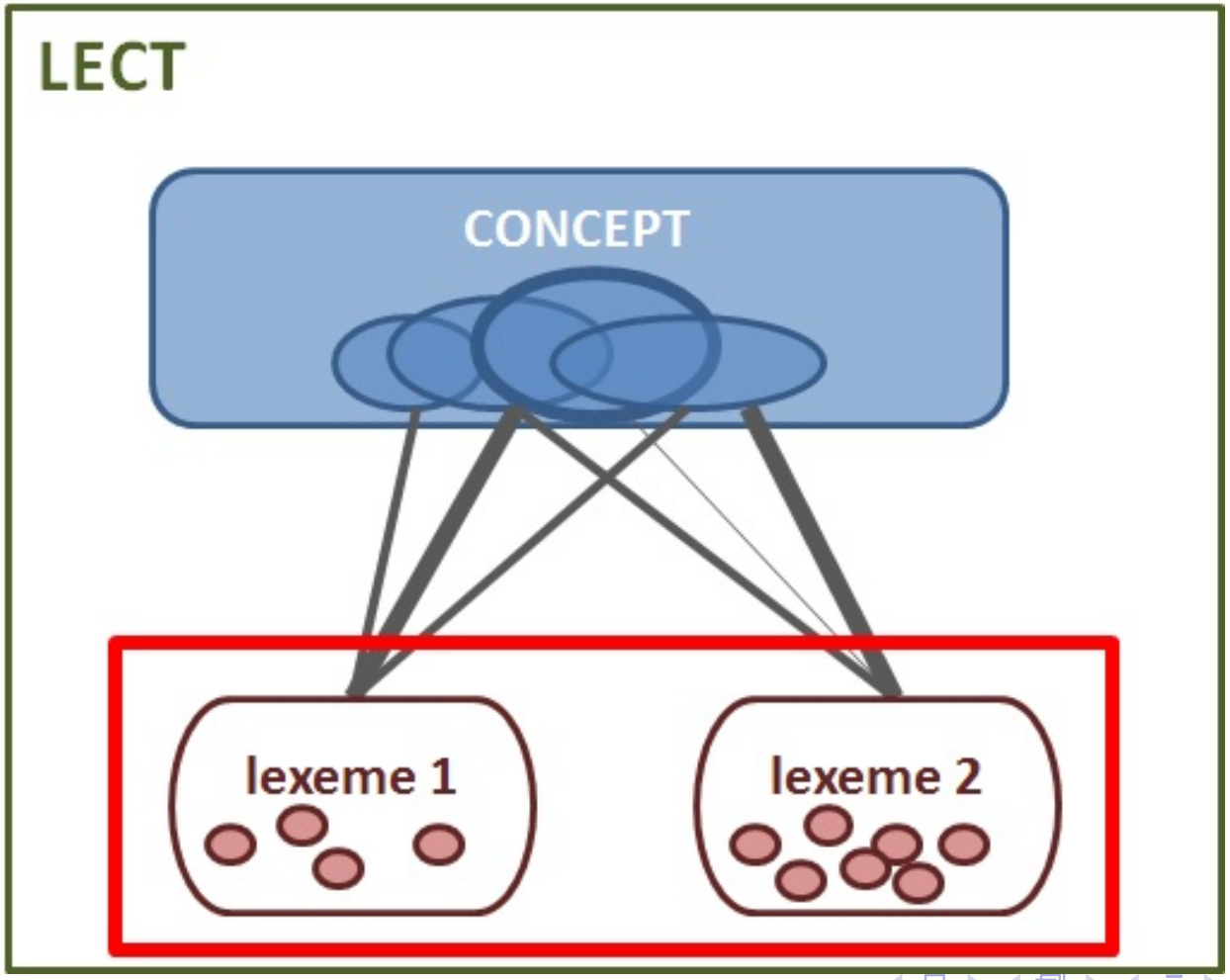
Association strength between semantic features and lexemes differ between registers and changes over time.



Token clouds

How are the individual tokens/exemplars structured?

≈ Word Sense Disambiguation



Token-level SVS

Make a vector for each occurrence of the variants

the teacher saw the dog chasing the cat

Token-level SVS

Make a vector for each occurrence of the variants

	3.2	4.3			0.8		7.1
	5.1	2.2			3.7		0.1
	0.2	3.5			2.3		0.3
	3.1	1.9			2.9		4.1
	4.7	0.2			1.3		3.1
	2.2	3.1			4.1		3.8
the	teacher	saw	the	dog	chasing	the	cat

Token-level SVS

Make a vector for each occurrence of the variants

					AVERAGE
3.2	4.3		0.8	7.1	3.9
5.1	2.2		3.7	0.2	2.8
0.2	3.5		2.3	0.3	1.6
3.1	1.9		2.9	4.1	3.0
4.7	0.2		1.4	3.1	2.3
2.2	3.1		4.1	3.8	3.3
teacher	saw	dog	chasing	cat	

Token-level SVS

Weighting

	3.2	4.3		0.8	7.1
	5.1	2.2		3.7	0.1
	0.2	3.5		2.3	0.3
	3.1	1.9		2.9	4.1
	4.7	0.2		1.3	3.1
	2.2	3.1		4.1	3.8
	teacher	saw	dog	chasing	cat
PMI weights	0.4	0.8		2.1	1.5

Context words are not equally informative for the meaning of dog.

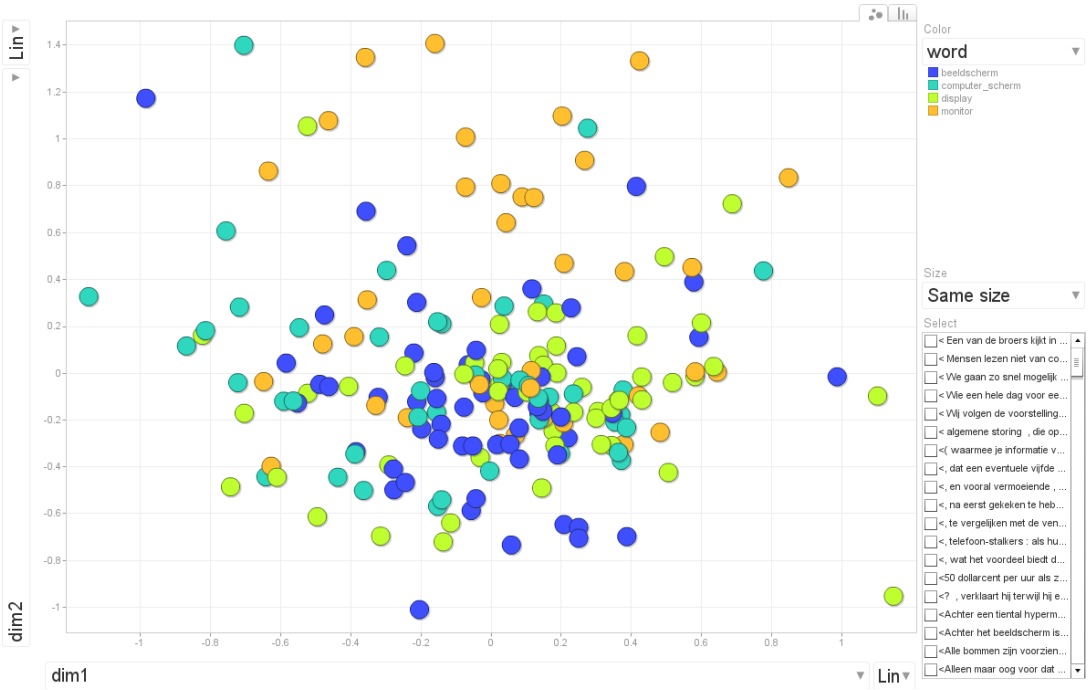
Token-level SVS

Weighted vectors

					WEIGHTED AVERAGE
3.2x0.4	4.3x0.8		0.8x2.1	7.1x1.5	4.3
5.1x0.4	2.2x0.8		3.7x2.1	0.2x1.5	3.0
0.2x0.4	3.5x0.8		2.3x2.1	0.3x1.5	2
3.1x0.4	1.9x0.8		2.9x2.1	4.1x1.5	3.8
4.7x0.4	0.2x0.8		1.4x2.1	3.1x1.5	2.4
2.2x0.4	3.1x0.8		4.1x2.1	3.8x1.5	4.4
teacher	saw	dog	chasing	cat	

Token clouds

Calculate similarity between all tokens
use MDS and googlevis to plot in 2D
Demo: https://perswww.kuleuven.be/~u0038536/googleVis



Calibration

Semantic Vector Spaces, and especially token-level SVSs are parameter-rich.

Examples of parameters

- Bag-of-Words $\leftrightarrow$ Dependency Models
- Size of the context window for co-occurrences
- Size of the context window for weights
- Weighting scheme:
  Pointwise Mutual Information $\leftrightarrow$ Log-Likelihood Ratio
- Include $\leftrightarrow$ exclude highly-frequent (function words) words

Token Clouds

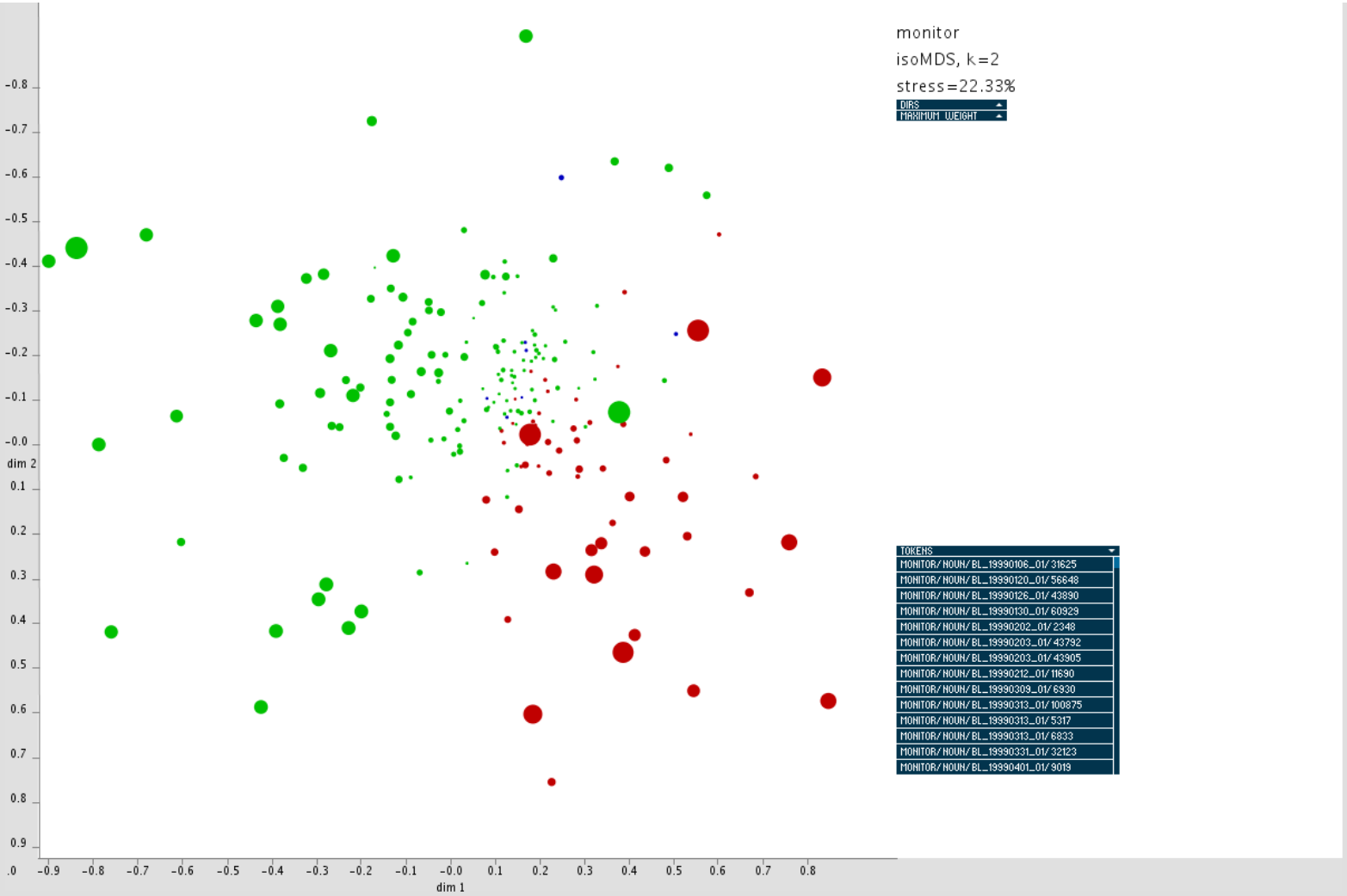
- Calibration could benefit from visual analytics of the different solutions.
- Using manually disambiguated data facilitates the visual evaluation as we can color–code the tokens for their different meanings.
- Misclassified tokens are quickly identified.
- We built our own, customisable tool to explore these token clouds.

Token Clouds

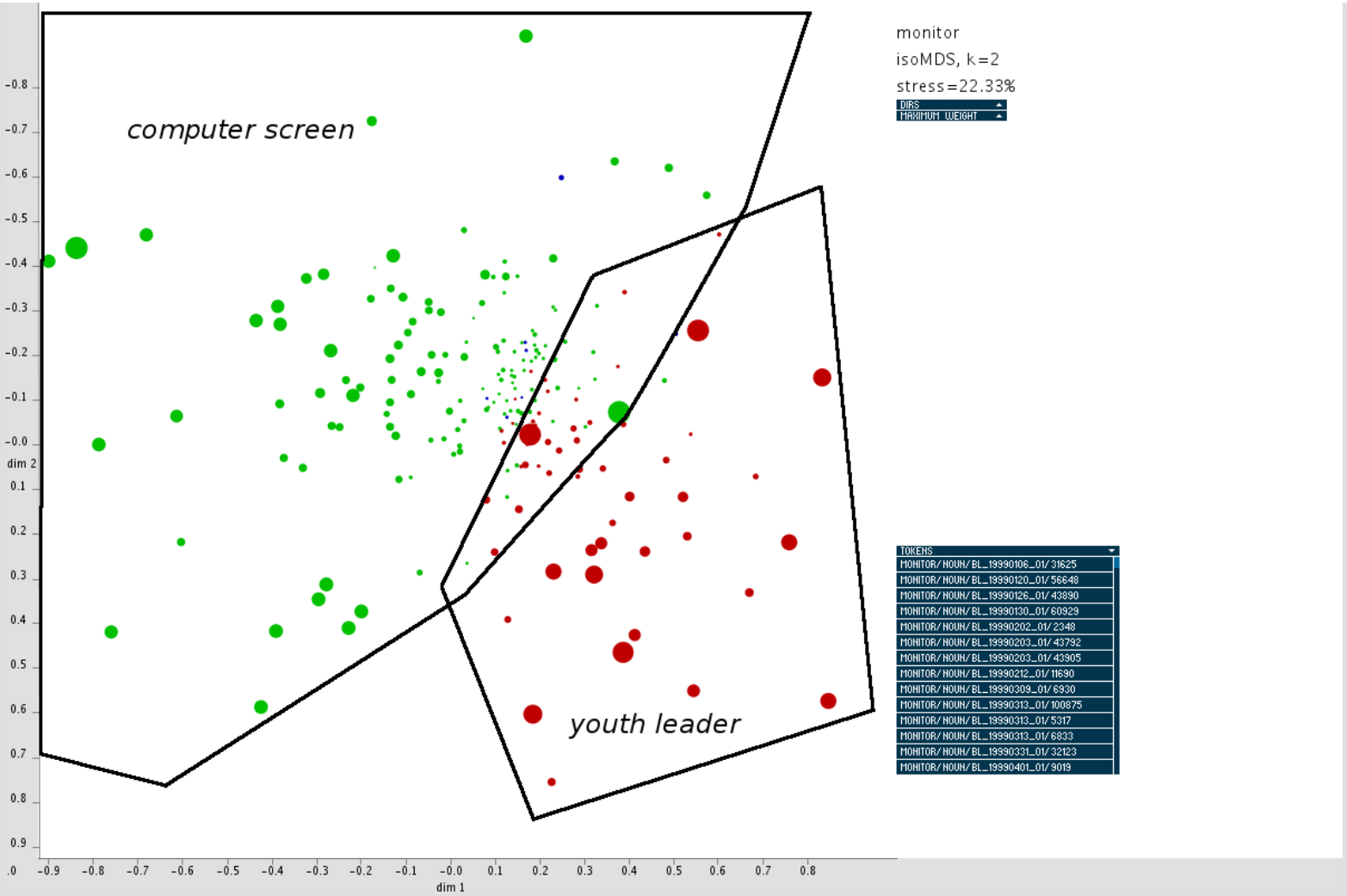
Dutch noun monitor

- Manually disambiguted data for the concept of BEELDSCHERM (display)
- Semasiological study of a polysemous word with lectal variation (geographical and register)

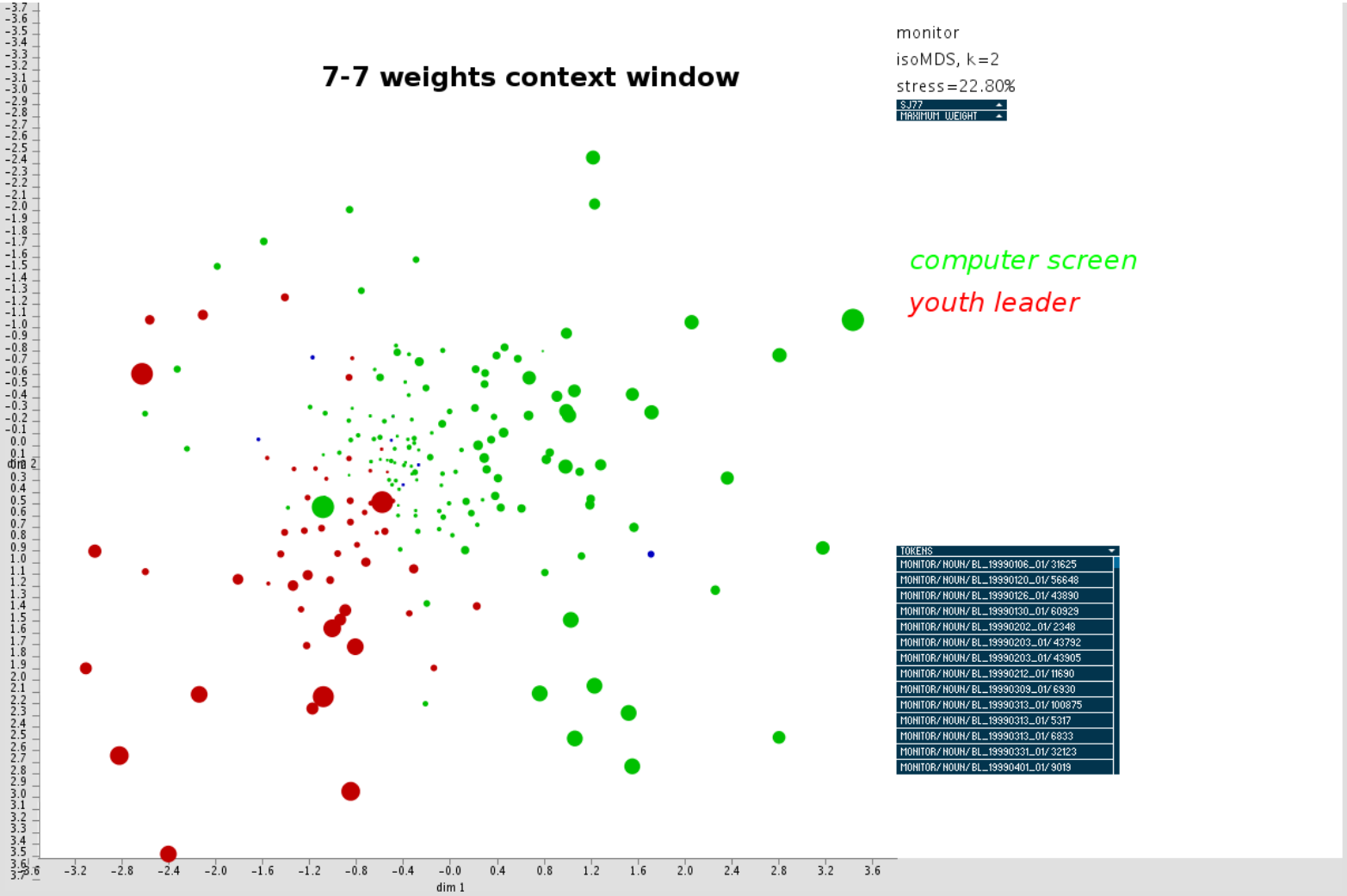
4. Visual Analytics



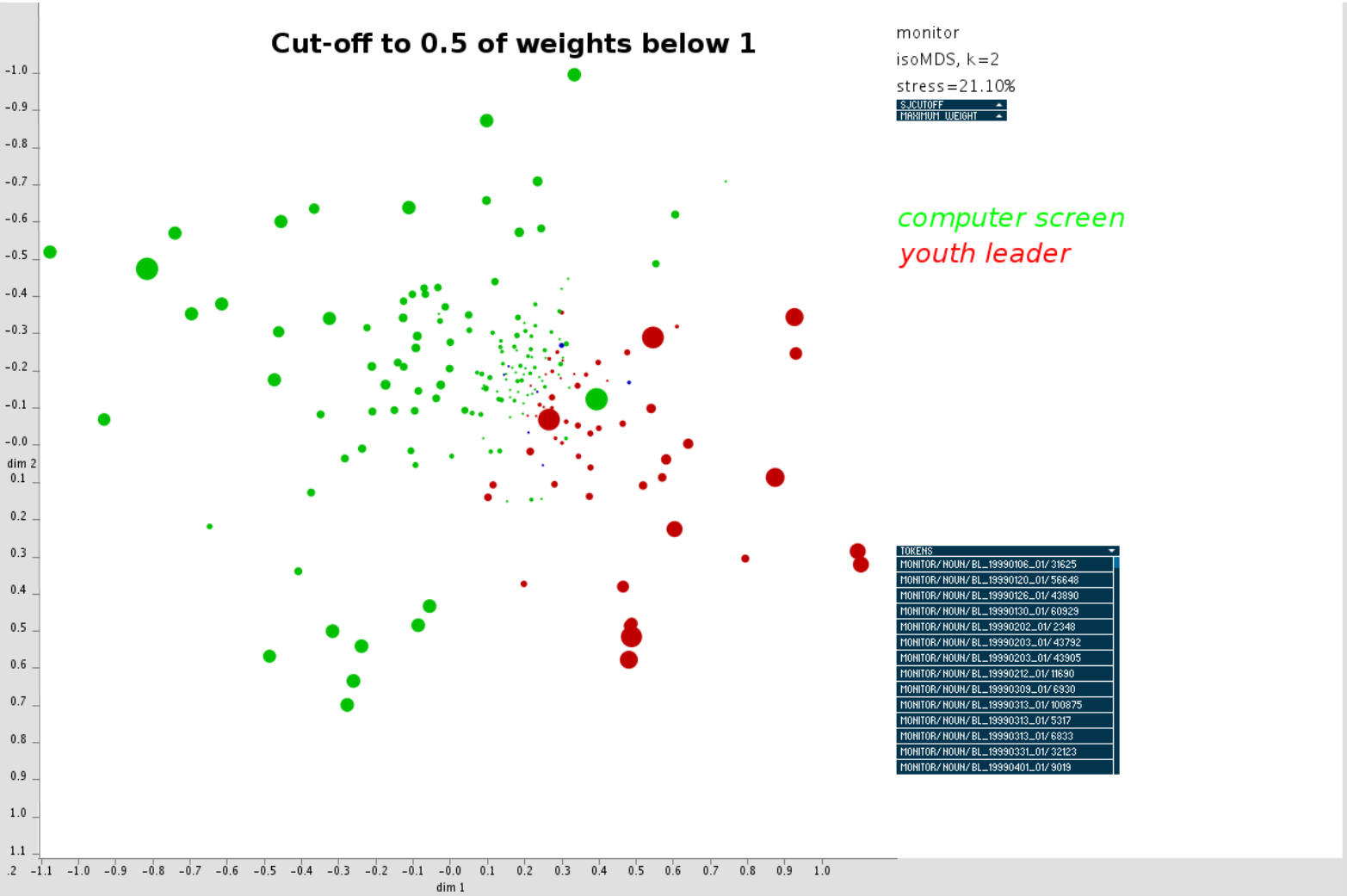
4. Visual Analytics



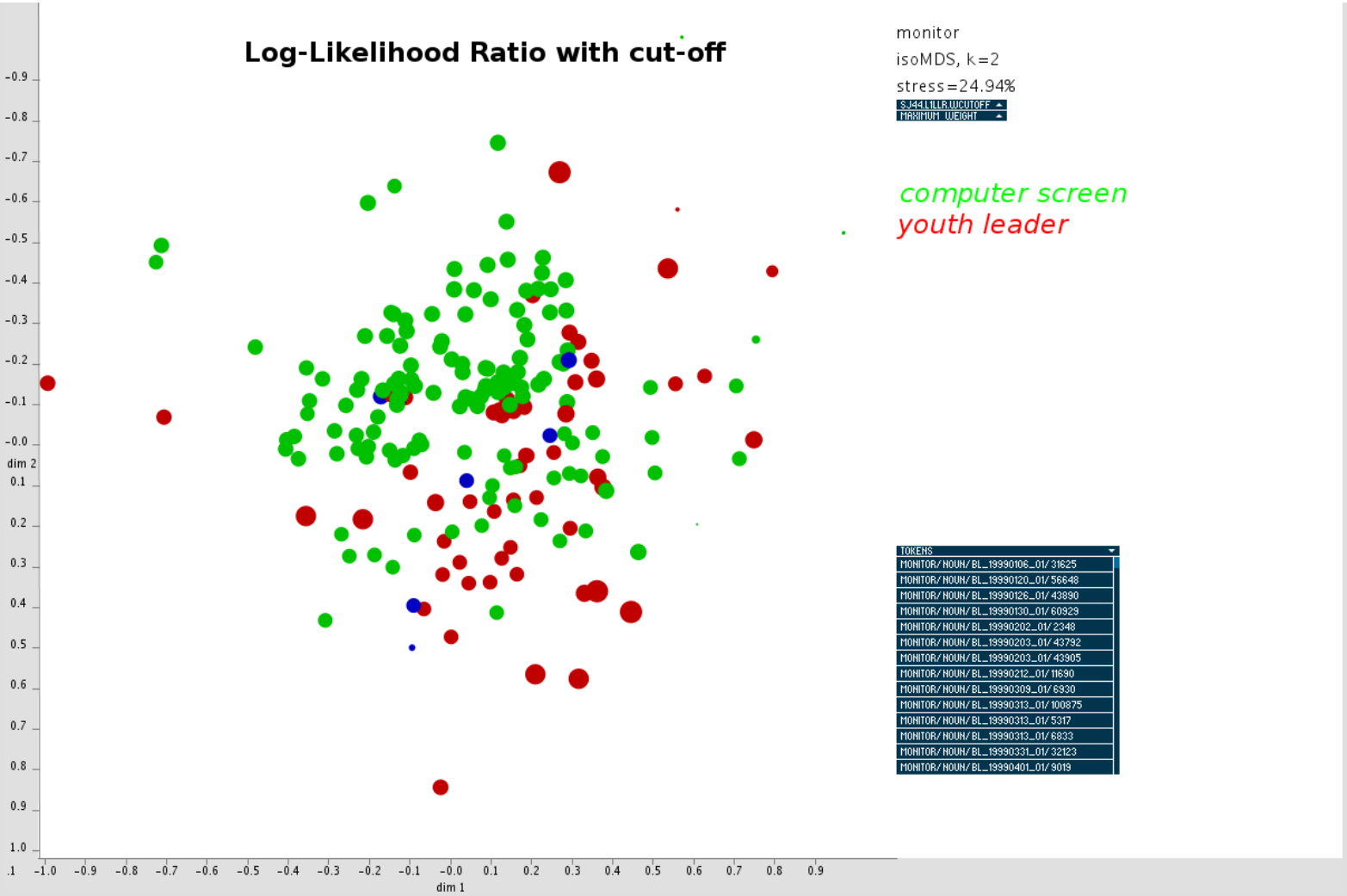
4. Visual Analytics



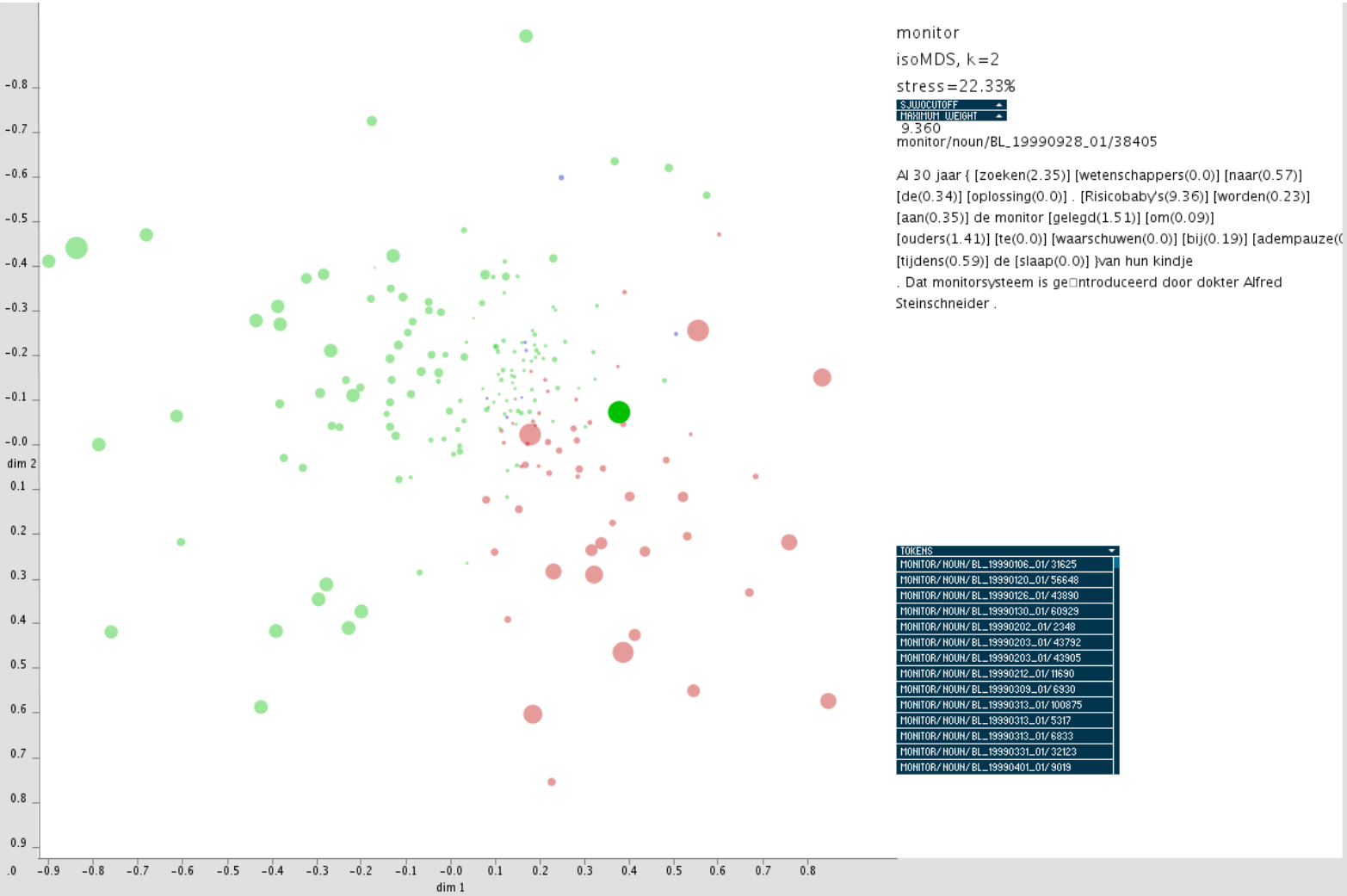
4. Visual Analytics



4. Visual Analytics



4. Visual Analytics



Overview

Conclusion

Lexicology in Cognitive Sociolinguistics

- Variant level: Structure of Lexical Variation
- Variety level: Lexical Sociolectometry

Distributional Semantic Modelling

- to find synonym profiles
- measure polysemy
- structure collocations
- measure semantic change
- examplar level modelling

For more information:

http://wwwling.arts.kuleuven.be/qlvl
kris.heylen@arts.kuleuven.be
dirk.geeraerts@arts.kuleuven.be
dirk.speelman@arts.kuleuven.be

Bertels, Ann, De Hertog, Dirk, & Heylen, Kris. 2012. Etude sémantique des mots-clés et des marqueurs lexicaux stables dans un corpus technique. *Pages 239–252 of: Actes de la conférence conjointe JEP-TALN-RECITAL 2012, Grenoble, volume 2: TALN*.

Geeraerts, Dirk, Grondelaers, Stefan, & Bakema, Peter. 1994. *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin/New York, Mouton de Gruyter.

Geeraerts, Dirk, Grondelaers, Stefan, & Speelman, Dirk. 1999. *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kledings- en voetbaltermen*. Amsterdam: Meertens Instituut.

Geeraerts, Dirk, Kristiansen, Gitte, & Peirsman, Yves (eds). 2010. *Advances in cognitive sociolinguistics*. Berlin/New York: Mouton de Gruyter.

Heylen, Kris, & Ruette, Tom. 2013. Degrees of semantic control in measuring aggregated lexical distances. *Pages 353–374 of:*

Borin, Lars, & Saxena, Anju (eds), *Approaches to Measuring Linguistic Differences*. Trends in Linguistics. Studies and Monographs. Berlin: Mouton de Gruyter.

Heylen, Kris, Peirsman, Yves, & Geeraerts, Dirk. 2008a. Automatic synonymy extraction. *Pages 101–116 of:* Verberne, Suzan, van Halteren, Hans, & Coppen, Peter-Arno (eds), *Computational linguistics in the Netherlands 2007*. Amsterdam: Rodopi.

Heylen, Kris, Peirsman, Yves, Geeraerts, Dirk, & Speelman, Dirk. 2008b. Modelling word similarity: An evaluation of automatic synonymy extraction algorithms. *In:* Calzolari, Nicoletta, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, Piperidis, Stelios, & Tapias, Daniel (eds), *Proceedings of the Sixth International Language Resources and Evaluation*. Marrakech: European Language Resources Association.

Heylen, Kris, Speelman, Dirk, & Geeraerts, Dirk. 2012. Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. *Pages 16–24 of:* Butt, Miriam,

Carpendale, Sheelagh, Penn, Gerald, Prokic, Jelena, & Cysouw, Michael (eds), *Proceedings of the EACL-2012 joint workshop of LINGVIS & UNCLH: Visualization of Language Patters and Uncovering Language History from Multilingual Resources*. Stroudsburg: Association for Computational Linguistics.

Kristiansen, Gitte, & Dirven, Rene (eds). 2008. *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Cognitive Linguistics Research. Mouton De Gruyter, Berlin.

Landauer, Thomas K, & Dumais, Susan T. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, **104**(2), 240–411.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. *Pages 768–774 of: Proceedings of the 17th international conference on Computational linguistics-Volume 2*. Association for Computational Linguistics.

Padó, Sebastian, & Lapata, Mirella. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, **33**(2), 161–199.

Pantel, Patrick, & Lin, Dekang. 2002. Document clustering with committees. *Pages 199–206 of: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '02. New York, NY, USA: ACM.

Peirsman, Yves, Heylen, Kris, & Speelman, Dirk. 2007. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. *Pages 9–16 of:* Baroni, Marco, Lenci, Alessandro, & Sahlgren, Magnus (eds), *Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models: Beyond Words and Documents*.

Peirsman, Yves, Heylen, Kris, & Speelman, Dirk. 2008a. Putting things in order. First and second order contexts models for the calculation of semantic similarity. *Pages 907–916 of:* Heiden,

Serge, & Pincemin, Bénédicte (eds), *JADT 2008 : Actes des 9es Journées internationales dAnalyse statistique des Données Textuelles*. Lyon: Presses universitaires de Lyon.

Peirsman, Yves, Heylen, Kris, & Geeraerts, Dirk. 2008b. Size matters: tight and loose context definitions in English word space models. *Pages 34–41 of:* Baroni, Marco, Evert, Stefan, & Alessandro, Lenci (eds), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*.

Peirsman, Yves, Deyne, Simon De, Heylen, Kris, & Geeraerts, Dirk. 2008c. The Construction and Evaluation of Word Space Models. *In: Proceedings of the Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco.

Ruette, Tom. 2012. *Aggregating Lexical Variation. Towards large-scale lexical lectometry*. dissertation, KU Leuven.

Ruette, Tom, & Speelman, Dirk. 2012. Applying individual differences scaling to measurements of lexical convergence between Netherlandic and Belgian Dutch. *Pages 883–895 of:*

11th International Conference on Statistical Analysis of Textual Data (JADT 2012).

Ruette, Tom, Speelman, Dirk, & Geeraerts, Dirk. 2011. Measuring the lexical distance between registers in national varieties of Dutch. *Pages 541–554 of:* Soares da Silva, Augusto, Torres, Amadeu, & Gonçalves, Miguel (eds), *Línguas Pluricêntricas. Variação Linguística e Dimensóes Sociocognitivas*. Braga: Publicações da Faculdade de Filosofia, Universidade Católica Portuguesa.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, **24**(1), 97–124.

Turney, Peter D., & Pantel, Patrick. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, **37**(1), 141–188.

Wielfaert, Thomas, Heylen, Kris, & Speelman, Dirk. 2013. Visualisations interactives des espaces vectoriels sémantiques

pour lanalyse lexicologique. *Pages 154–166 of: Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*.