Mapping semantic space in comparable corpora:

Semantic Vector Spaces as an analysis tool for lexical variation

Thomas Wielfaert, Kris Heylen & Dirk Speelman

KULeuven
Quantitative Lexicology and Variational Linguistics

Purpose of the talk

Theoretical: Study the Structure of Lexical Variation and show how a concept is mapped differently onto lexical alternatives in different varieties of the same language

Methodological: Use Semantic Vector Space Models as an exploratory tool for analysing lexical semantics in large comparable corpora

Descriptive: A short term diachronic analysis of the lexicalisation of the politically loaded concept IMMIGRANTS in Belgian Dutch, stratified by register

Overview

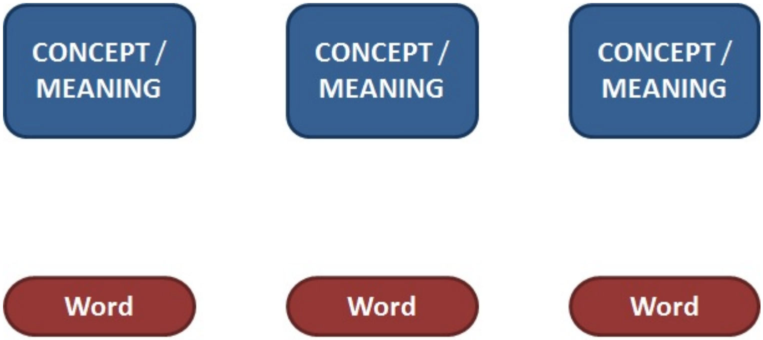
1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

Overview

1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

Analysing the Structure of Lexical Variation

How are concepts mapped onto lexemes?



CONCEPT / MEANING

CONCEPT / MEANING

CONCEPT / MEANING

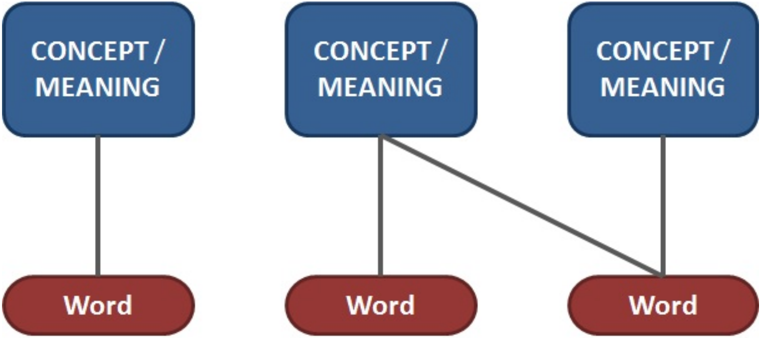
Word

Word

Word

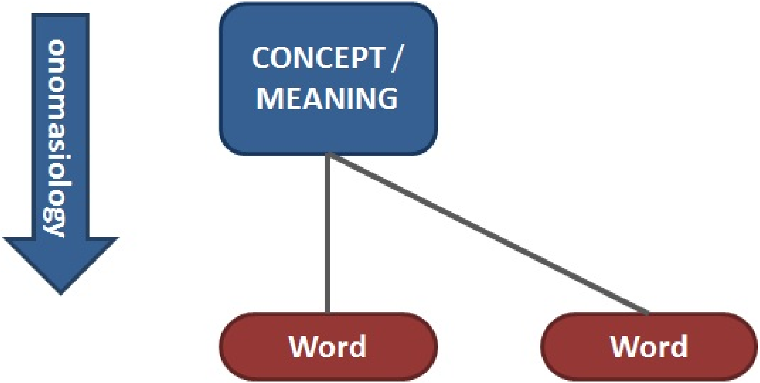
Analysing the Structure of Lexical Variation

How are concepts mapped onto lexemes?



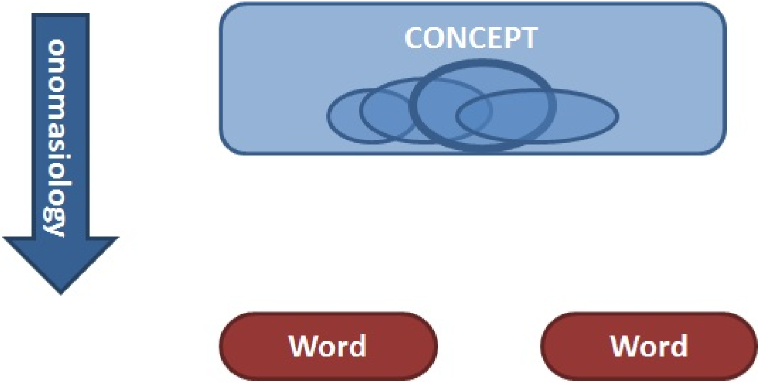
Analysing the Structure of Lexical Variation

Taking the perspective of the concept:
Which lexemes are available to express a given concept?



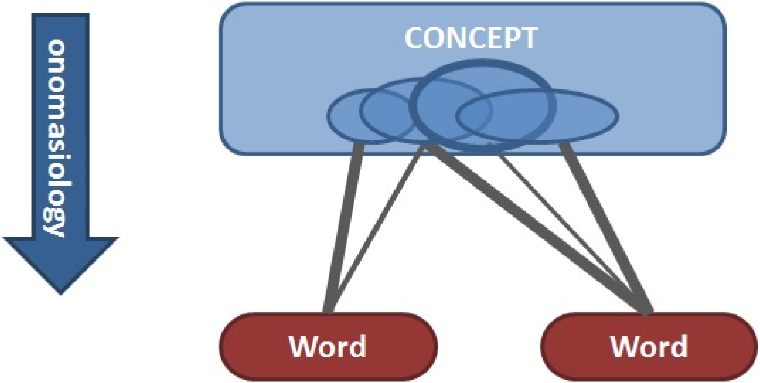
Analysing the Structure of Lexical Variation

A concept has a complex internal structure:
PROTOTYPE STRUCTURE:



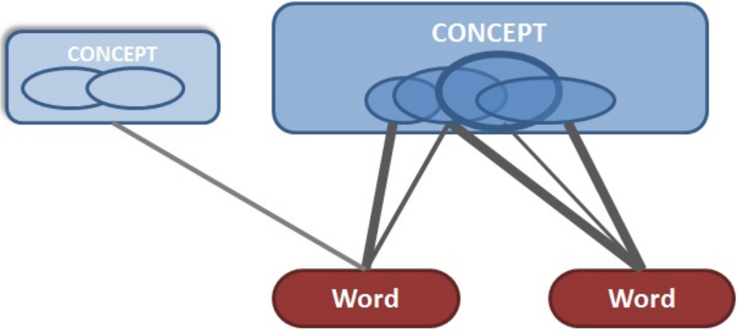
Analysing the Structure of Lexical Variation

Semantic features have different weight in lexemes:
PROTOTYPE STRUCTURE:



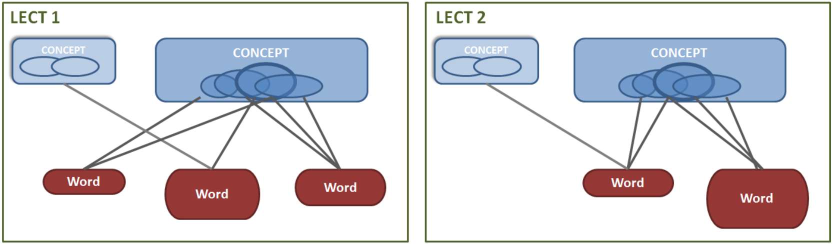
Analysing the Structure of Lexical Variation

Some lexemes can have an additional meaning:
Polysemy/homonymy



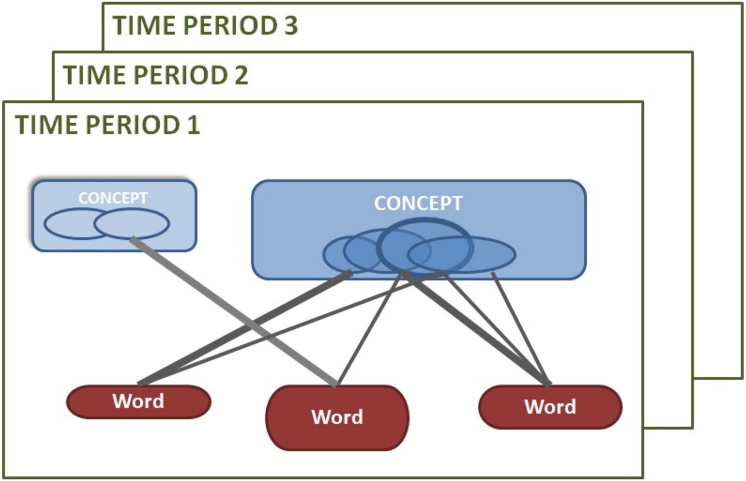
Analysing the Structure of Lexical Variation

Mapping can be different in different *lects* (regiolects, registers,...)
LECTAL VARIATION



Analysing the Structure of Lexical Variation

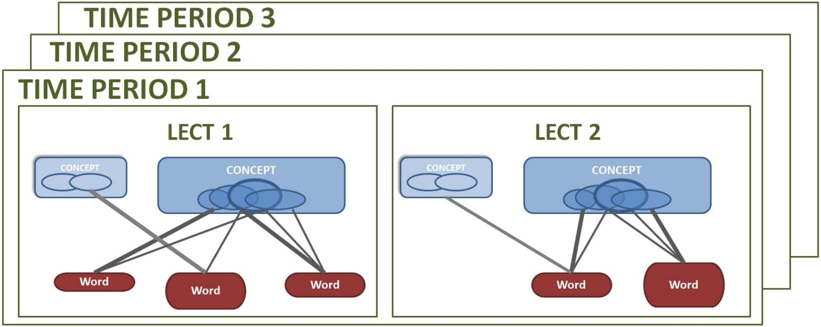
Mapping between concept and lexemes can change over time: DIACHRONIC VARIATION:



Analysing the Structure of Lexical Variation

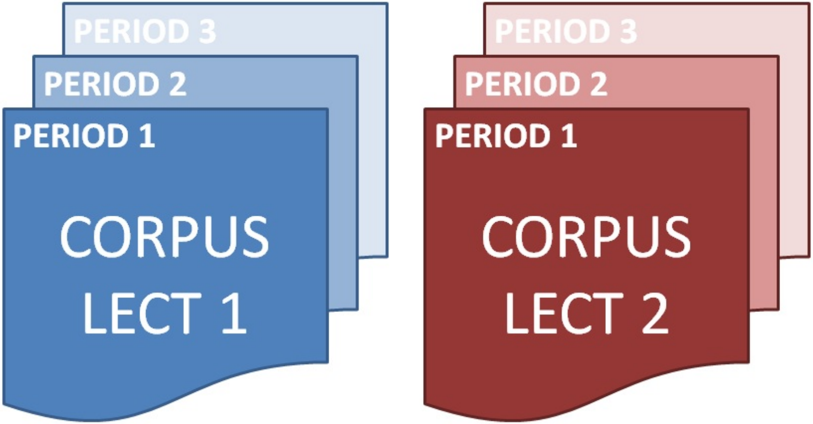
How do all these different factors interact?

STRUCTURE OF LEXICAL VARIATION (Geeraerts et al. 1994)



Analysing the Structure of Lexical Variation

Usage based analysis:
COMPARABLE CORPORA



Overview

1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

Case study: IMMIGRANTS

Allochtoon: Dutch, < Greek allos (other) + chthon (soil), *Person with an immigration background*, in use since early 1990s

Public discussion: *allochtoon* has become a politically incorrect term and is banned by:

- Sept. 2012,*De Morgen* (Belgian left-of-centre, high-brow newspaper)
- Feb. 2013, City of Ghent
- Feb. 2013 City of Amsterdam

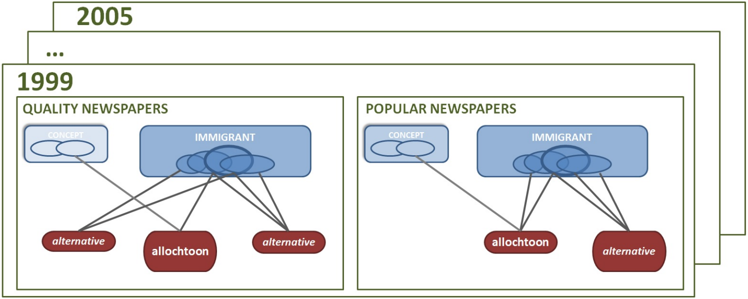
Case study: IMMIGRANTS

Research Questions:

- In what contexts is *allochtoon* exactly used? How negative are the contexts?
- Did the usage change since the 90s? Did it acquire more negative connotations?
- Are there alternative terms? Did *allochtoon* replace another term or was is it replaced itself?
- Is the apparent negative connotation typical for intellectual communities and high-brow newspapers? Is the usage and meaning change the same in different registers?

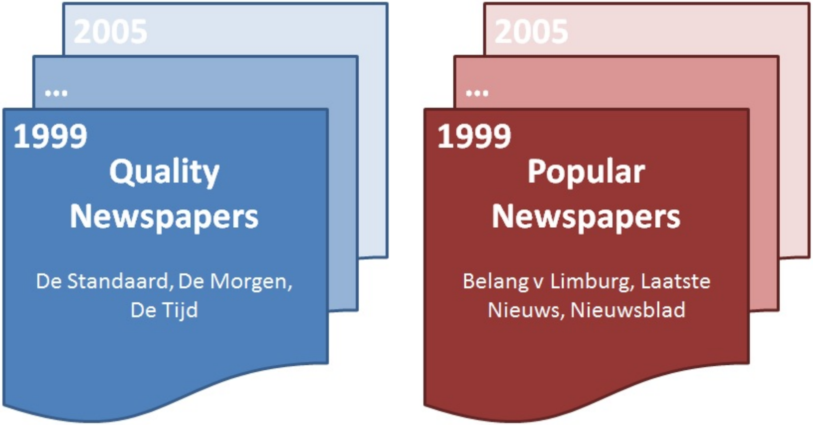
Case study: IMMIGRANTS

PERSON WITH IMMIGRATION BACKGROUND:



Case study: IMMIGRANTS

COMPARABLE CORPORA OF BELGIAN DUTCH (1.3G words)



2005

...

1999

Quality Newspapers

De Standaard, De Morgen, De Tijd

2005

...

1999

Popular Newspapers

Belang v Limburg, Laatste Nieuws, Nieuwsblad

Overview

1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

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

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
- words occurring in same contexts have similar meaning

Semantic Vector Spaces

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

	work	foreign	citizenship	laws	space	sugar	cream	now
immigrant	120	424	388	82	12	11	3	189
alien	154	401	376	99	305	20	1	123
coffee	5	8	18	4	1	72	102	152

Semantic Vector Spaces

Similar co-occurrence pattern indicates usage in similar contexts and hence semantic similarity

	work	foreign	citizenship	laws	space	sugar	cream	now
immigrant	120	424	388	82	12	11	3	189
alien	154	401	376	99	305	20	1	123
coffee	5	8	18	4	1	72	102	152

Semantic Vector Spaces

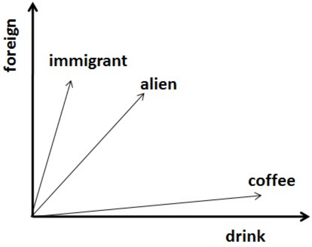
weight the raw frequencies by collocational strength (pmi)

	work	foreign	citizenship	laws	space	sugar	cream	now
immigrant	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
alien	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

Semantic Vector Spaces

calculate word by word similarity matrix

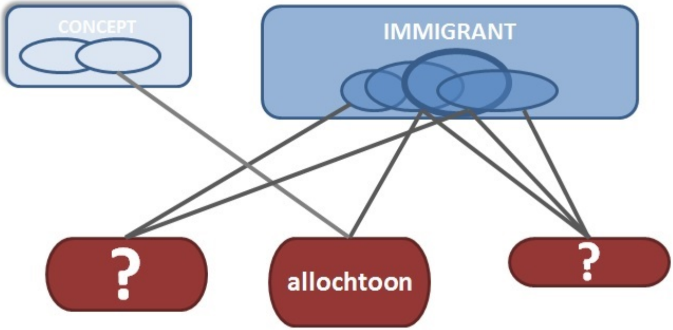
	immigrant	alien	coffee
immigrant	1	.71	.08
alien	.71	1	.09
coffee	.08	.09	1



Overview

1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

Identifying alternative expressions



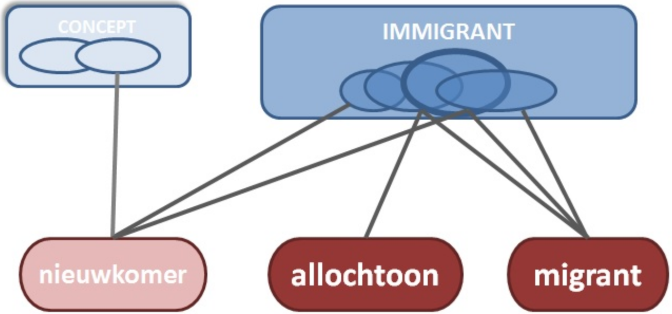
- calculate contextual similarity between 10K Dutch nouns
- sort by similarity to *allochtoon*

Identifying alternative expressions

allochtoon	1.0
migrant	0.71
vreemdeling	0.48
immigrant	0.47
buitenlander	0.47
nieuwkomer	0.32
gastarbeider	0.29

Table alternatives to *allochtoon*

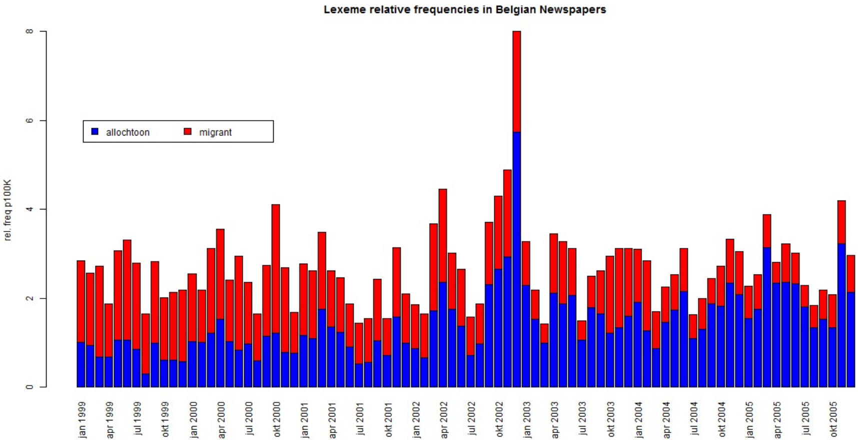
Identifying alternative expressions



Identifying alternative expressions

Normalised frequency of *allochtoon* and *migrant* per month

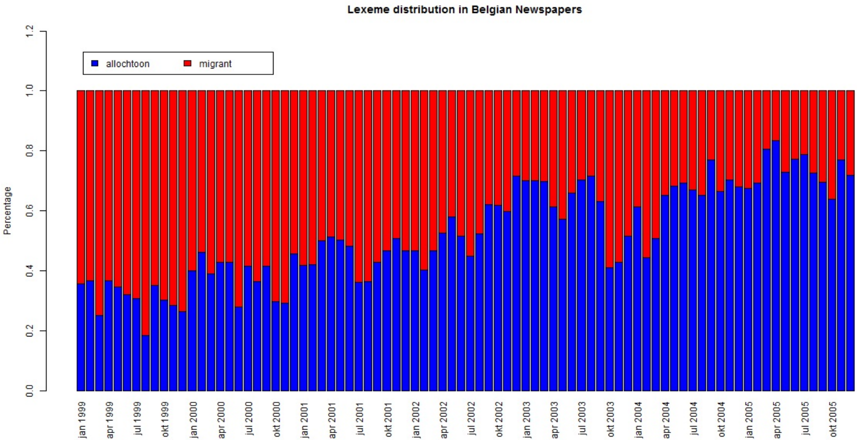
immigrant-talk seems to be a seasonal phenomenon



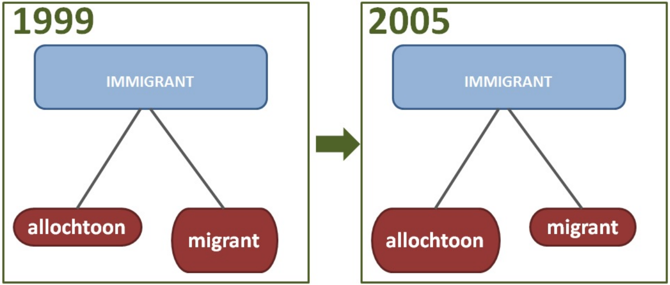
Identifying alternative expressions

Proportion of *allochtoon* and *migrant* in the corpus per month

allochtoon becomes more frequent than *migrant*



Identifying alternative expressions



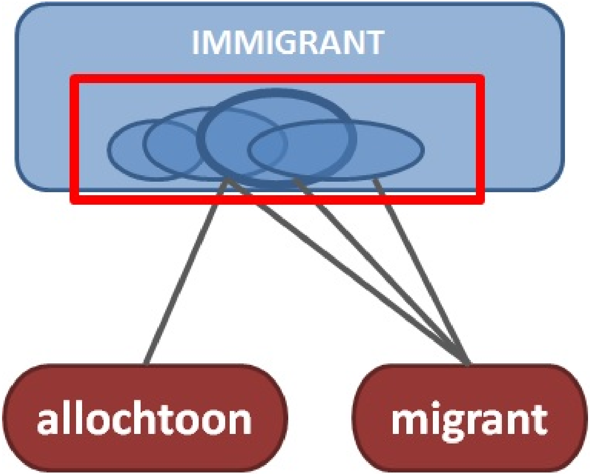
Is this change in frequency also indicative of semantic change?

Overview

1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

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

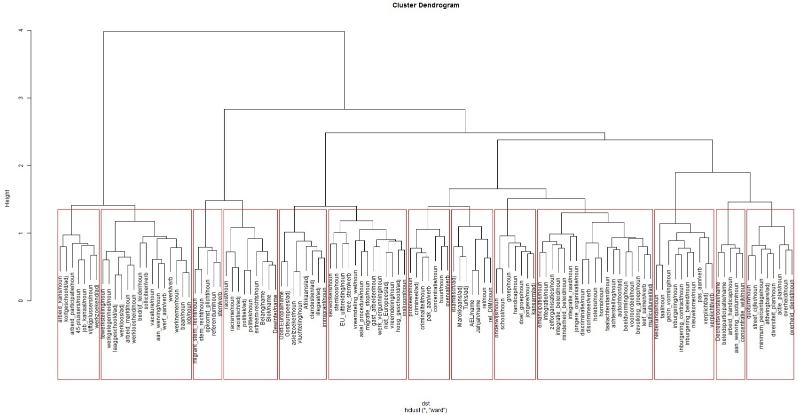
Analysing Semantic Structure

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

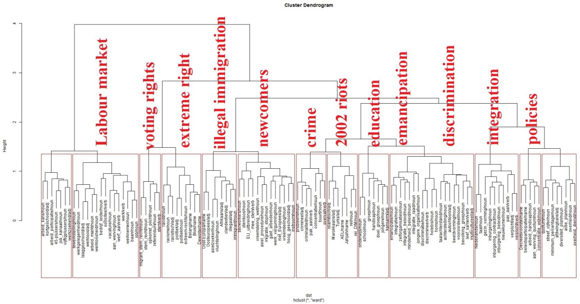
Analysing Semantic Structure

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

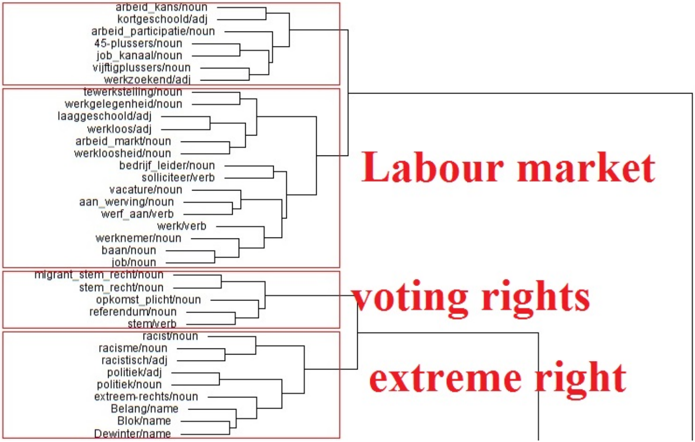


Analysing Semantic Structure

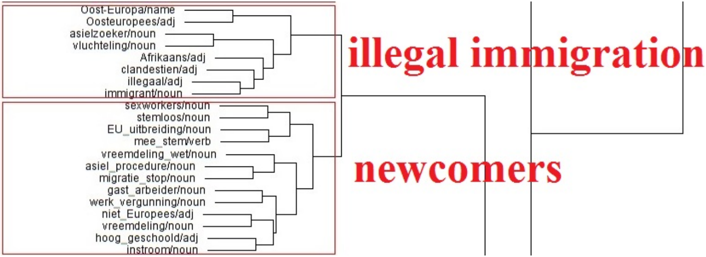
Clusters of contextually related collocations ≈ semantic features
Clusters can be labeled manually



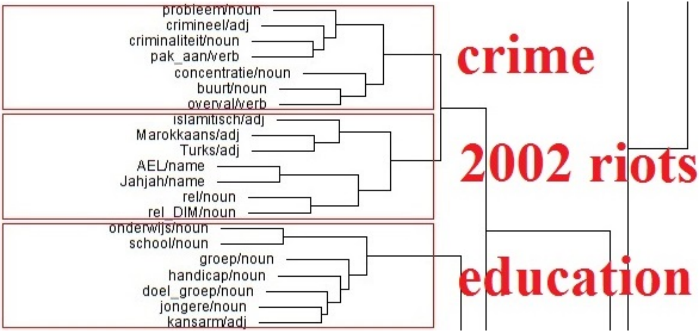
Analysing Semantic Structure



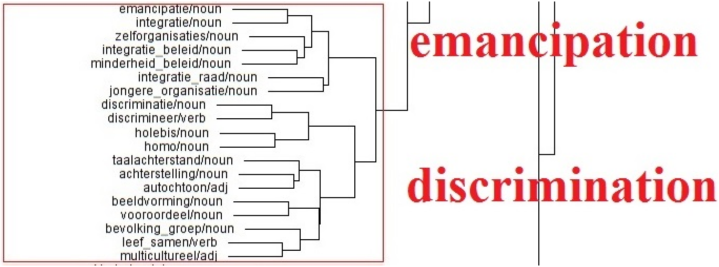
Analysing Semantic Structure



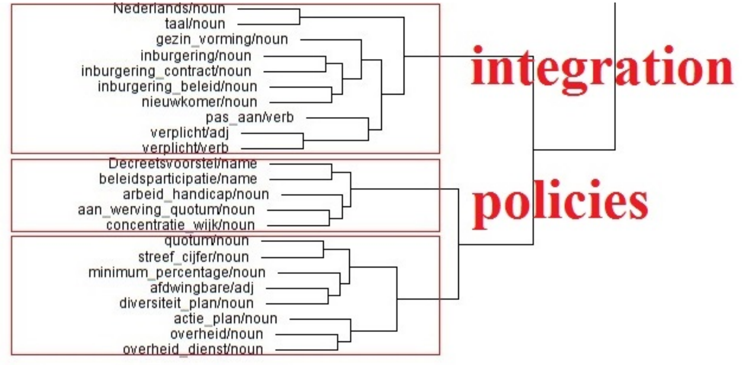
Analysing Semantic Structure



Analysing Semantic Structure

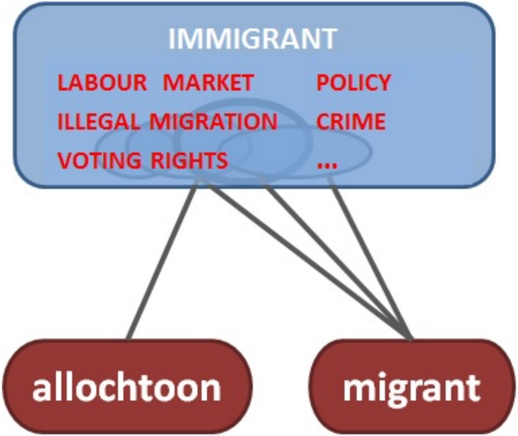


Analysing Semantic Structure



Analysing Semantic Structure

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



Overview

1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

Measuring semantic change in registers

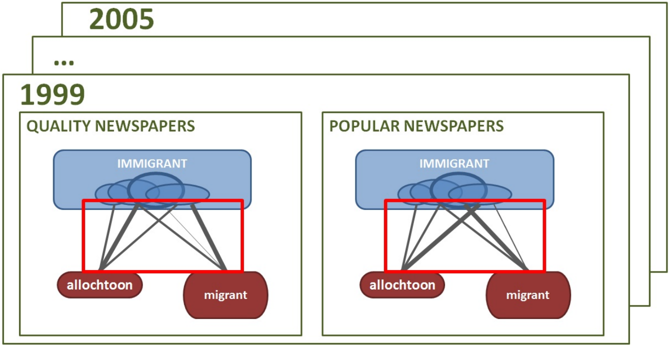
- How strong are *allochtoon* and *migrant* associated with the different context cluster/semantic features

- 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

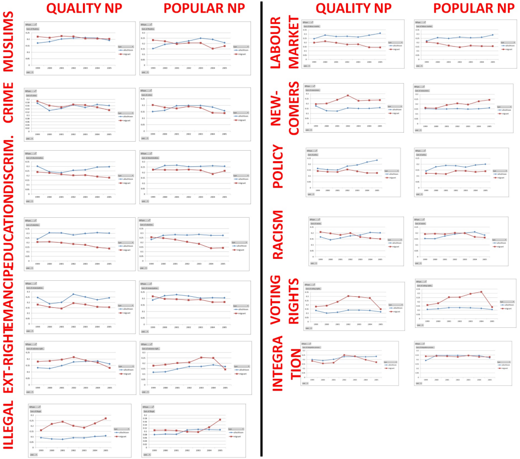
Calculate the cosine similarity (≈ association strength) of each variant/year/type vector to each context cluster vector

	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

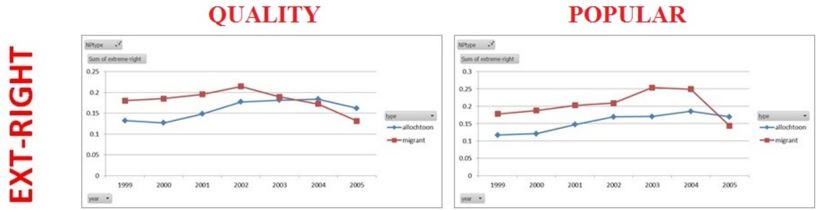
STEP 5

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



Measuring semantic change in registers

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

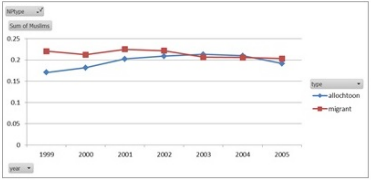


Measuring semantic change in registers

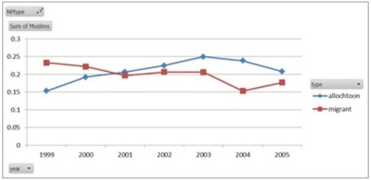
ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

MUSLIMS

QUALITY NP

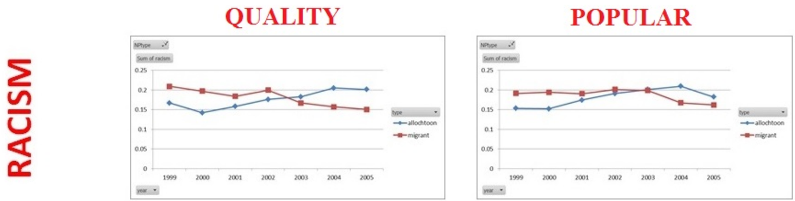


POPULAR NP



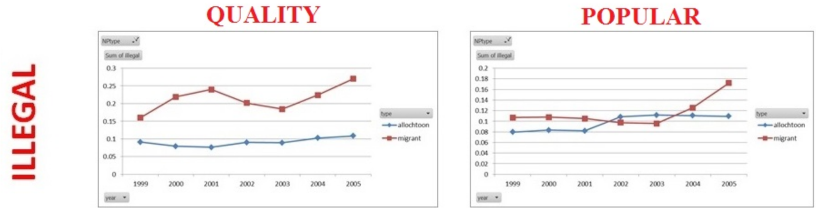
Measuring semantic change in registers

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT



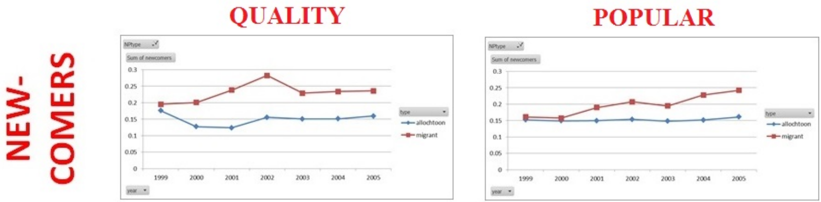
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON



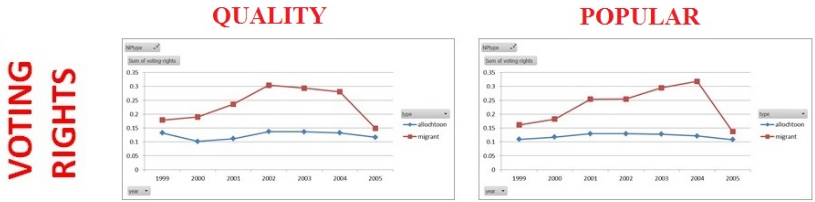
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON



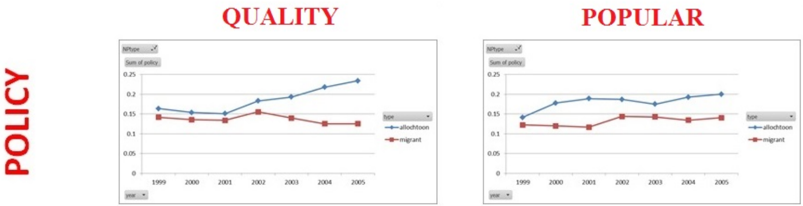
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON



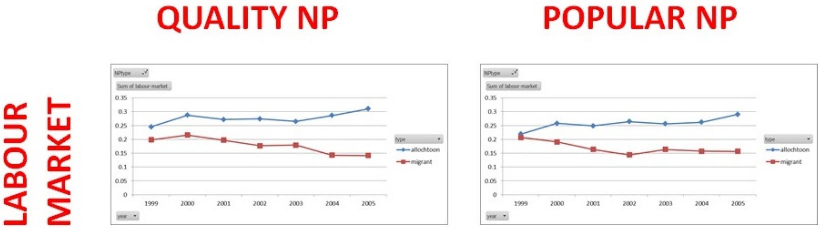
Measuring semantic change in registers

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT



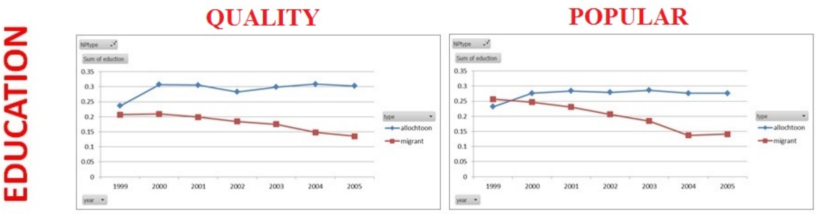
Measuring semantic change in registers

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

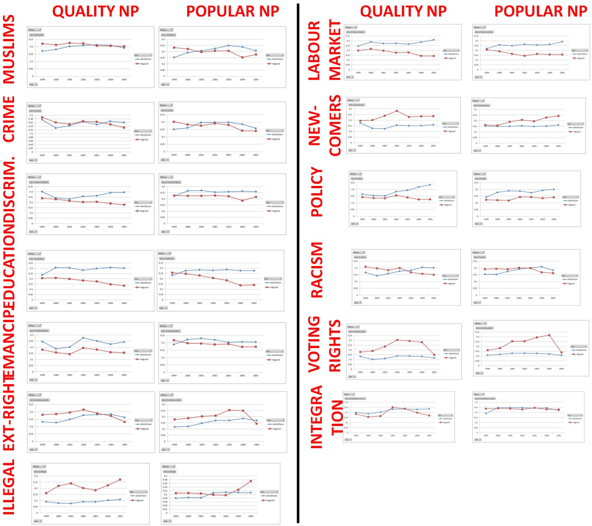


Measuring semantic change in registers

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

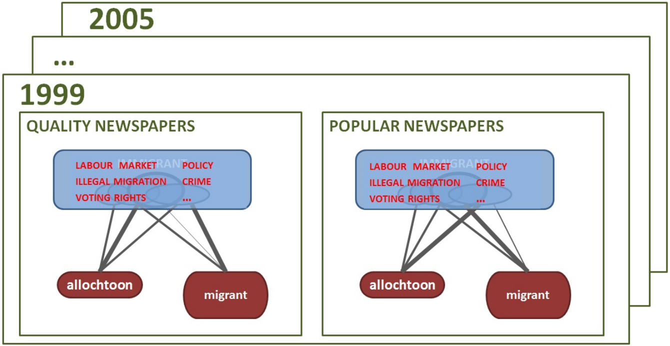


Measuring semantic change in registers



Measuring semantic change in registers

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

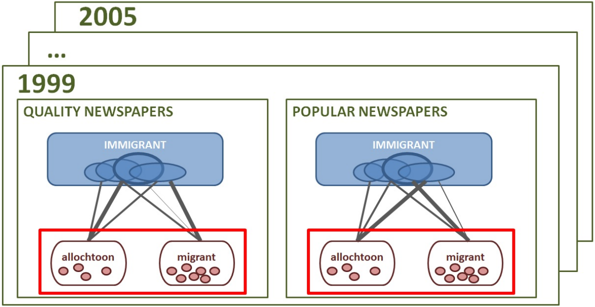


Overview

1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

Lexical variation on the exemplar level

How are the individual exemplars of *allochtoon* and *migrant* structured in context clusters?



Lexical variation on the exemplar level

Make a vector for each exemplar of *allochtoon* and *migrant*

op de arbeidsmarkt zijn er voor allochtonen nauwelijks jobs

Lexical variation on the exemplar level

Make a vector for each exemplar of *allochtoon* and *migrant*
STEP 1: retrieve the type vectors for each informative context word

	3.2				7.1
	5.1				0.1
	0.2				0.3
	3.1				4.1
	4.7				3.1
	2.2				3.8
op de	arbeidsmarkt	zijn er voor	allochtonen	nauwelijks	jobs

Lexical variation on the exemplar level

Make a vector for each exemplar of *allochtoon* and *migrant*
STEP 2: average over the type vectors of context words

			AVERAGE
3.2		7.1	5.2
5.1		0.1	3.1
0.2		0.3	0.2
3.1		4.1	3.7
4.7		3.1	3.9
2.2		3.8	2.9
arbeidsmarkt	allochtonen	jobs	

Lexical variation on the exemplar level

Make a vector for each exemplar of *allochtoon* and *migrant*
STEP 3: matrix of exemplar vector with *2nd order* co-occurrences

	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*
allochtoon$_1$	5.3	7.9	6.5	4.0	0.8	0.6	0.0
allochtoon$_2$	4.3	8.1	5.7	3.2	6.2	0.5	0.0
allochtoon$_3$	4.3	2.9	7.5	8.1	0.3	1.6	0.3
migrant$_1$	4.3	4.2	5.7	3.2	6.2	0.5	0.0
migrant$_2$	5.8	3.5	6.5	5.1	1.3	0.0	0.1
migrant$_3$	2.9	2.4	4.7	2.2	4.2	0.3	0.7

Lexical variation on the exemplar level

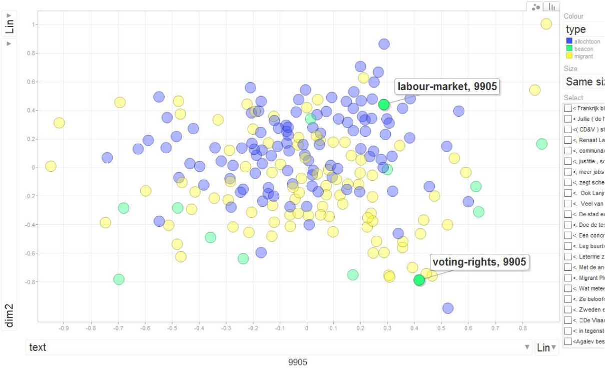
Make a vector for each exemplar of *allochtoon* and *migrant*
STEP 4: calculate similarity matrix between (sample of) exemplars

	allochtoon$_1$	*allochtoon*$_2$	*allochtoon*$_3$	*migrant*$_1$	*migrant*$_2$	*migrant*$_3$	...
allochtoon$_1$	1	0.9	0.5	0.4	0.8	0.6	...
allochtoon$_2$	0.4	1	0.7	0.2	0.2	0.5	...
allochtoon$_3$	0.3	0.9	1	0.1	0.3	0.6	...
migrant$_1$	0.3	0.2	0.7	1	0.2	0.5	...
migrant$_2$	0.8	0.5	0.5	0.1	1	0.0	...
migrant$_3$	0.9	0.4	0.7	0.2	0.2	1	...

Lexical variation on the exemplar level

STEP 5: use MDS to plot similariy matrix in 2D

STEP 6: use googleVis to make an interactive visualisation



Overview

1. Analysing the Structure of Lexical Variation
2. Case study: IMMIGRANTS
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the exemplar level
8. Conclusion

Conclusion

Descriptive: *allochtoon* vs. *migrant*

- *allochtoon* replaces *migrant* in frequency
- *allochtoon* gradually monopolizes socio-political contexts (labour market, education, policy)
- *migrant* had a flirt with 'voting rights' and specializes for 'new' and 'illegal immigration'.
- tendencies are stronger in quality than popular newspapers

Methodological conclusions

Semantic Vector Spaces can be applied to large comparable corpora in order to:

- find alternative expressions for a concept of interest
- structure the collocations into clusters of typical contexts
- quantify shifts in contextual usage and lectal differences
- structure exemplars of competing lexemes

Theoretical conclusions:

- semantic structure emerges from actual usage
- implies diachronic and lectal variation
- data shows complex concept to lexemes mapping

For more information:

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

kris.heylen@arts.kuleuven.be
thomas.wielfaert@arts.kuleuven.be
dirk.speelman@arts.kuleuven.be