Cognitive sociolinguistics meets Culturomics:

A longitudinal distributional semantic analysis of lexical variation in immigration discourse

Kris Heylen & Dirk Speelman

KULeuven
Quantitative Lexicology and Variational Linguistics

Purpose of the talk

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

Theoretical: Advocate a Cognitive Sociolinguistic approach that naturally integrates extra- and intra-linguistic factors in the analysis of (lexical) variation

Methodological: Showcase Semantic Vector Space Models as an exploratory tool for analysing lexical semantics and variation in large corpora,

Overview

1. Background: The Rise and Fall of a political correct term
2. Cognitive Sociolinguistics
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the attestation level
9. Conclusion

Overview

1. Background: The Rise and Fall of a political correct term
2. Cognitive Sociolinguistics
3. Semantic Vector Spaces
4. Identifying alternative expressions
5. Analysing Semantic Structure
6. Measuring semantic change in registers
7. Lexical variation on the attestation level
9. Conclusion

Background: The Rise and Fall of a political correct term

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

The Fall: On September 19, 2012, Belgian left-of-centre newspaper *De Morgen* decides to ban the word *allochtoon* citing the following reasons:

- the word is vaguely defined
- a catchall for a very diverse group of people
- the word is stigmatising and discriminating

Since then, different Belgian and Dutch organisations have followed suit.

Background: The Rise and Fall of a political correct term

Research Questions:

- In what contexts is *allochtoon* exactly used? How vague is the term?
- Why did it loose its political correct status? Did the usage change since the 90s? Did it acquire negative connotations?
- Are there alternative terms? Did *allochtoon* replace another term or was is it replaced itself?
- Is the apparent negative connotation typical for high-brow newspapers ? Is the usage and meaning change the same in different registers?

Overview

2. Cognitive Sociolinguistics

Lexical alternations

- Pervasive, salient and an important variable in dialectology
- However, semantic aspects are neutralised or disregarded
- Almost neglected in current corpus-based Sociolinguistics

Why this neglect?

- sociolinguistic variable = 1 function, multiple forms
- But meaning is notoriously difficult to delineate:

> "it is inadequate at the current state of sociolinguistic research to extend to other levels of analysis of variation the notion of sociolinguistic variable originally developed on the basis of phonological data. The quantitative studies of variation which deal with morphological, syntactic, and lexical alternation suffer from the lack of an articulated theory of meanings."
>
> (Lavandera 1978: 171)

2. Cognitive Sociolinguistics

Subtle and complex lexical-semantic variation that interacts with extra-linguistic factors can be and is studied empirically in corpus data:

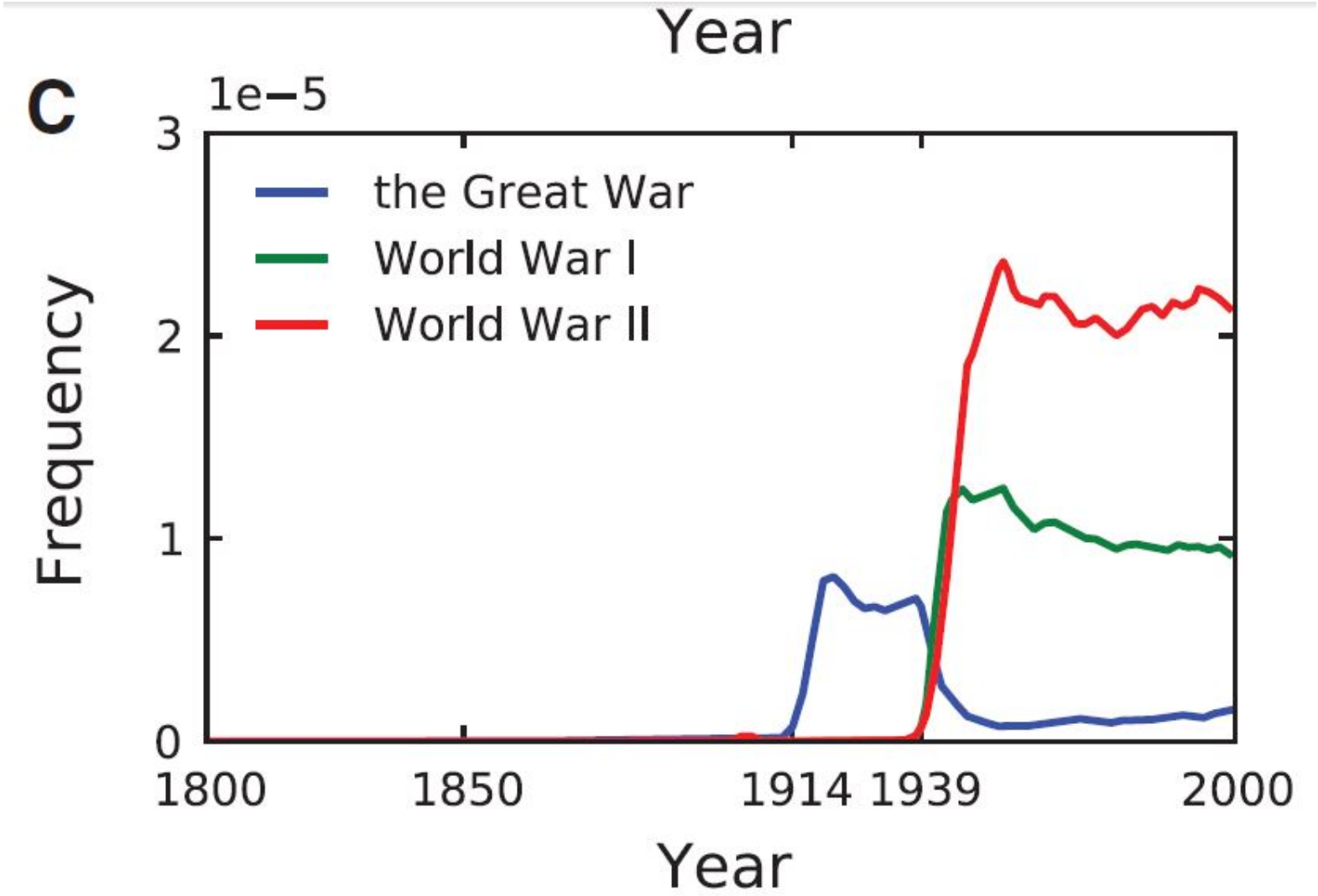
- Critical Discourse analysis in corpus data (Orpin 2005 on *Corruption*, Baker 2012, 2014 on *Muslims*
- Culturomics approach (Michel et al. 2011) to google-ngrams.

However, these studies have been criticized

- providing purely applied linguistic analyses without any theoretical underpinning
- showing severe methodological weaknesses: confirming the linguist's preconceptions based on shallow data analysis

2. Cognitive Sociolinguistics

Culturomics (Michel et al. 2011):



2. Cognitive Sociolinguistics

Within Cognitive Linguistics, recent theoretical and empirical trends have given rise to Cognitive Sociolinguistics (Kristiansen & Dirven 2008; Geeraerts, Kristiansen& Peirsman 2010) that aims to study the complex interplay between *lectal*, *formal* and *semantic* variation. It does so:

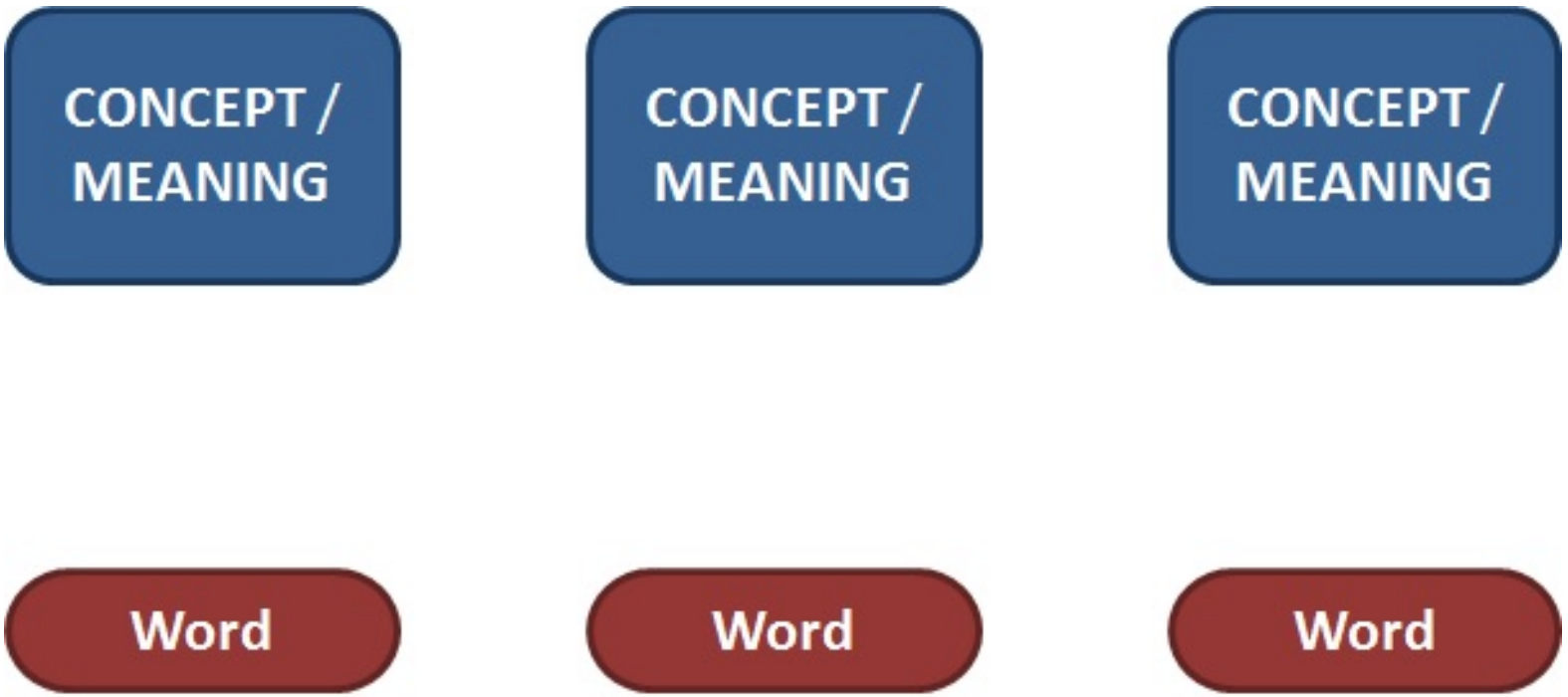
- within a meaning-centered theory of language
- taking a usage-based perspective of language
- emphasis on the a socio-cultural aspects of semantic structure
- commitment to the use of advanced quantitative methods

Previous studies within Cognitive Sociolinguistics on political and ideological discourse:

- Koller 2008 on Corporate Branding
- Peirsman, Heylen & Geeraerts 2010 on the conceptualisation of Muslims pre and post 911

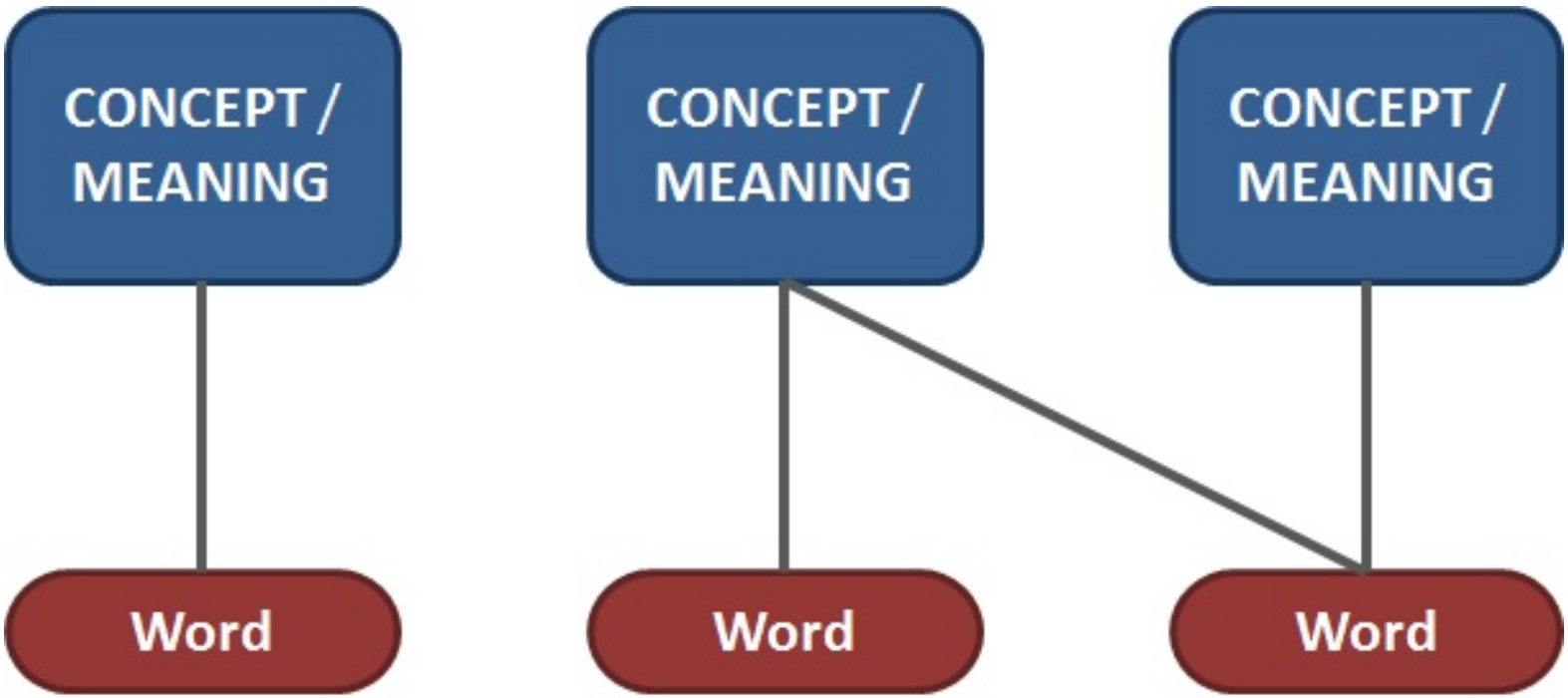
Analysing the Structure of Lexical Variation

How are concepts mapped onto lexemes?



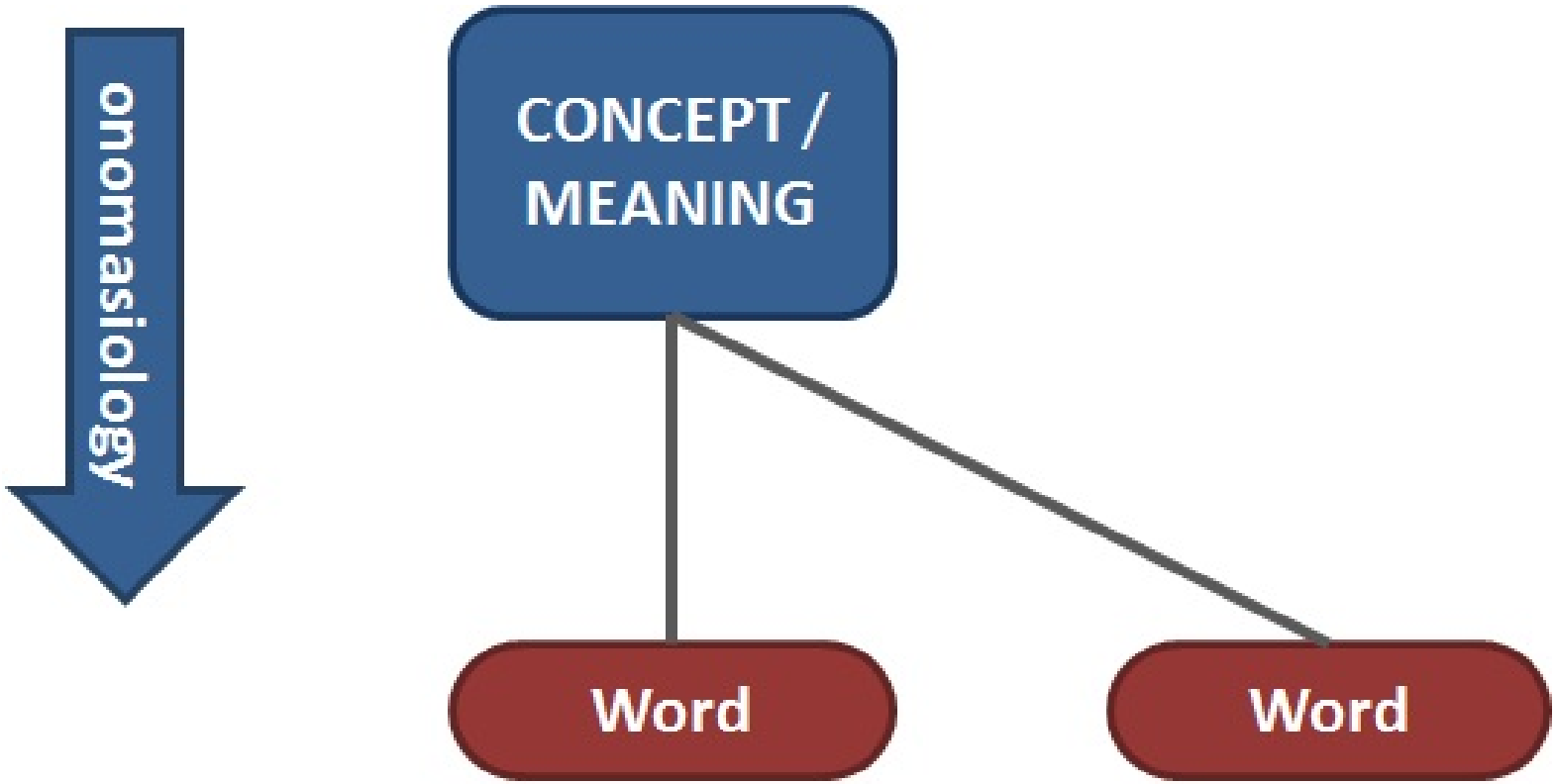
Analysing the Structure of Lexical Variation

How are concepts mapped onto lexemes?



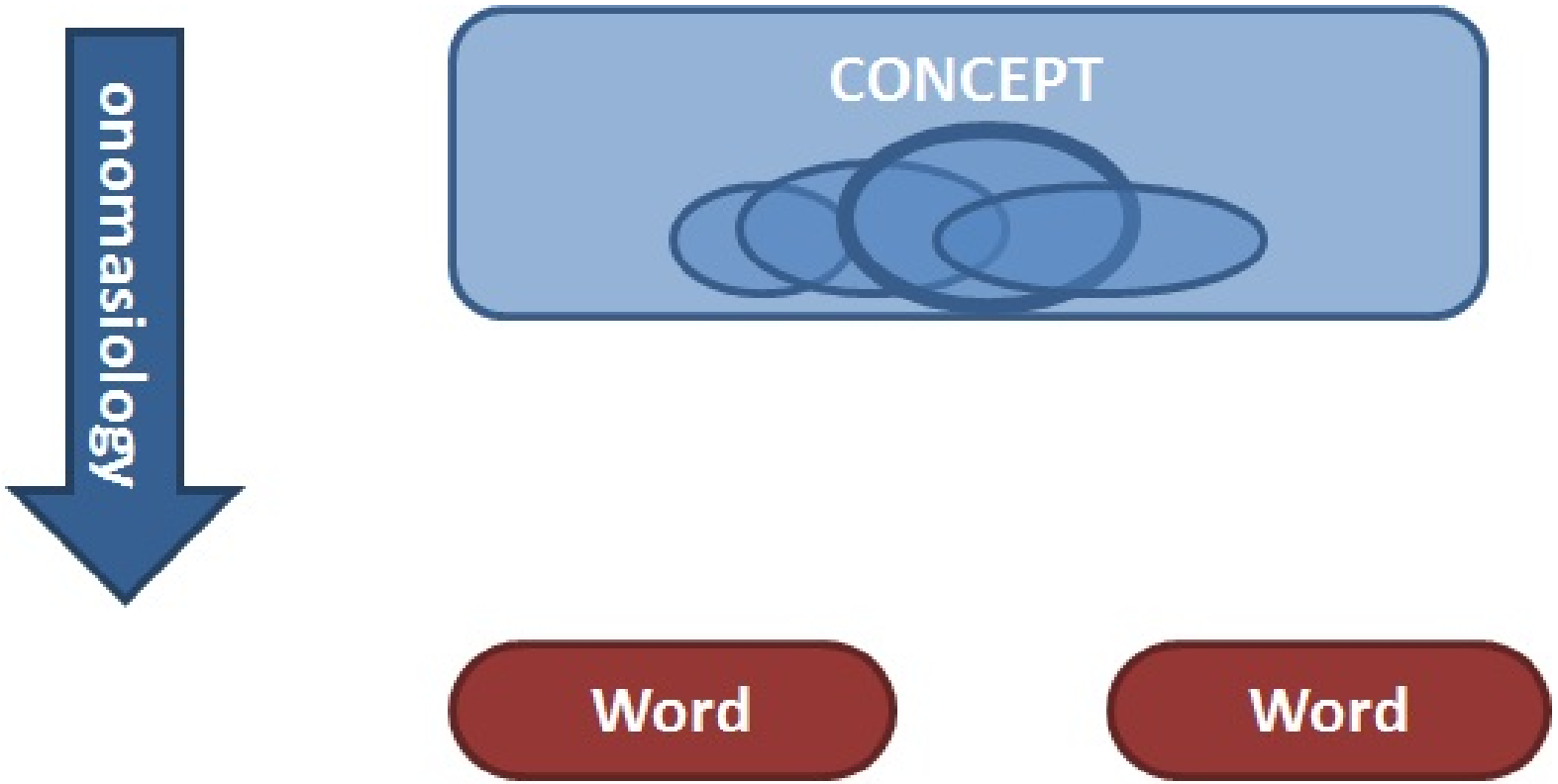
Analysing the Structure of Lexical Variation

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



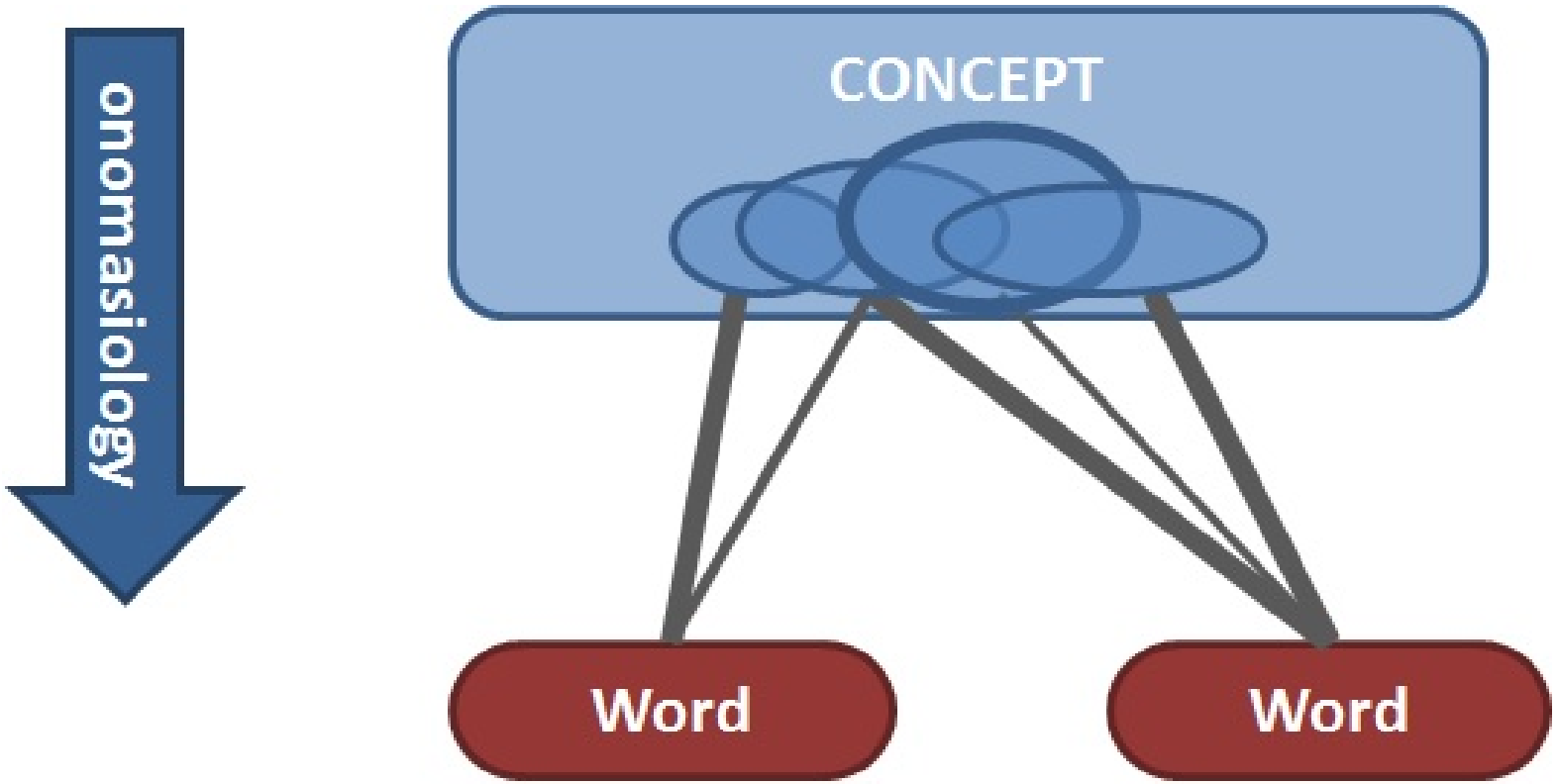
Analysing the Structure of Lexical Variation

A concept has a complex internal structure:
PROTOTYPE STRUCTURE:



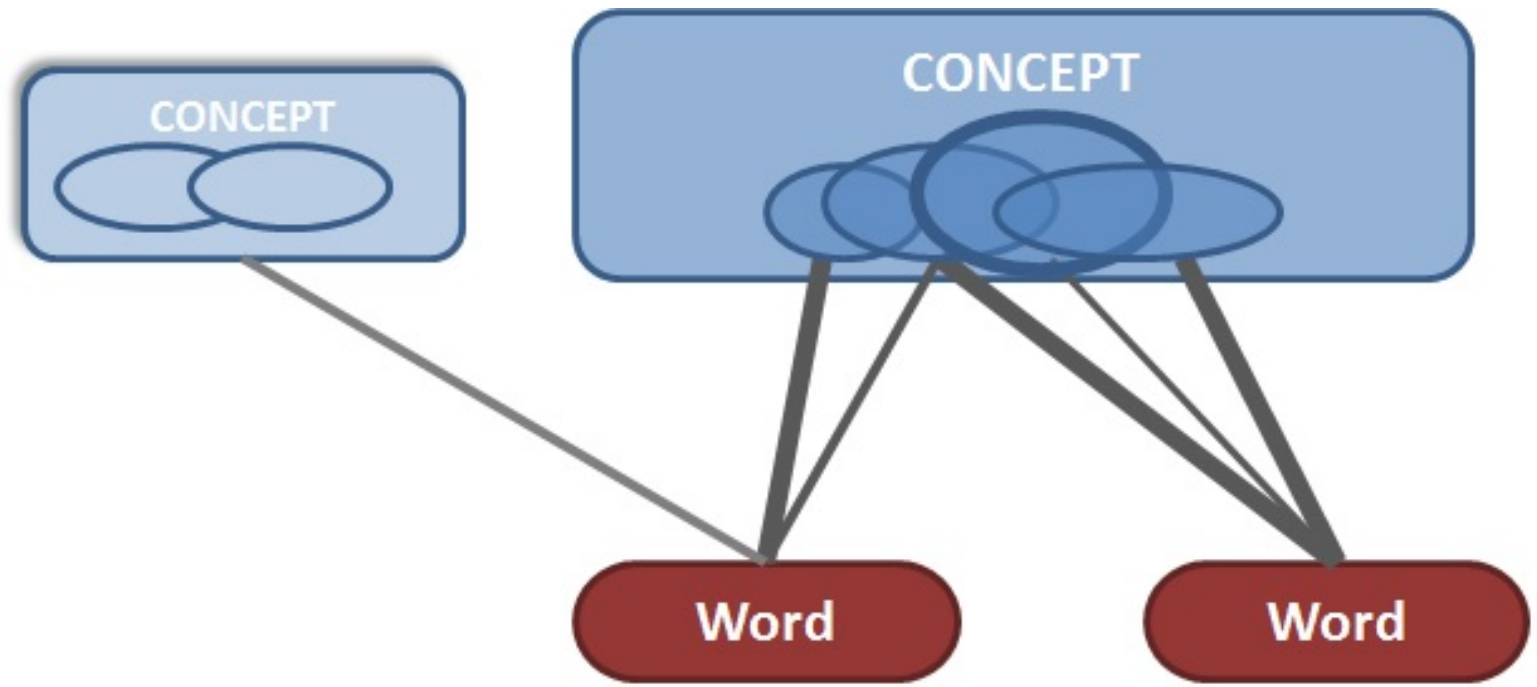
Analysing the Structure of Lexical Variation

Semantic features have different weight in lexemes:
PROTOTYPE STRUCTURE:



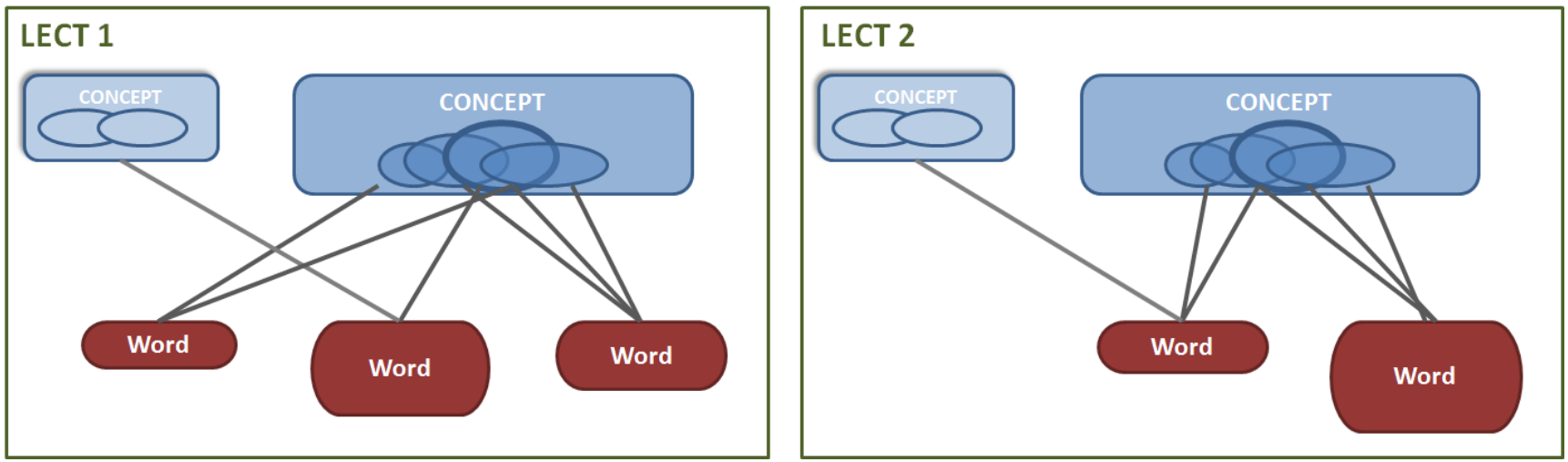
Analysing the Structure of Lexical Variation

Some lexemes can have an additional meaning:
Polysemy/homonymy



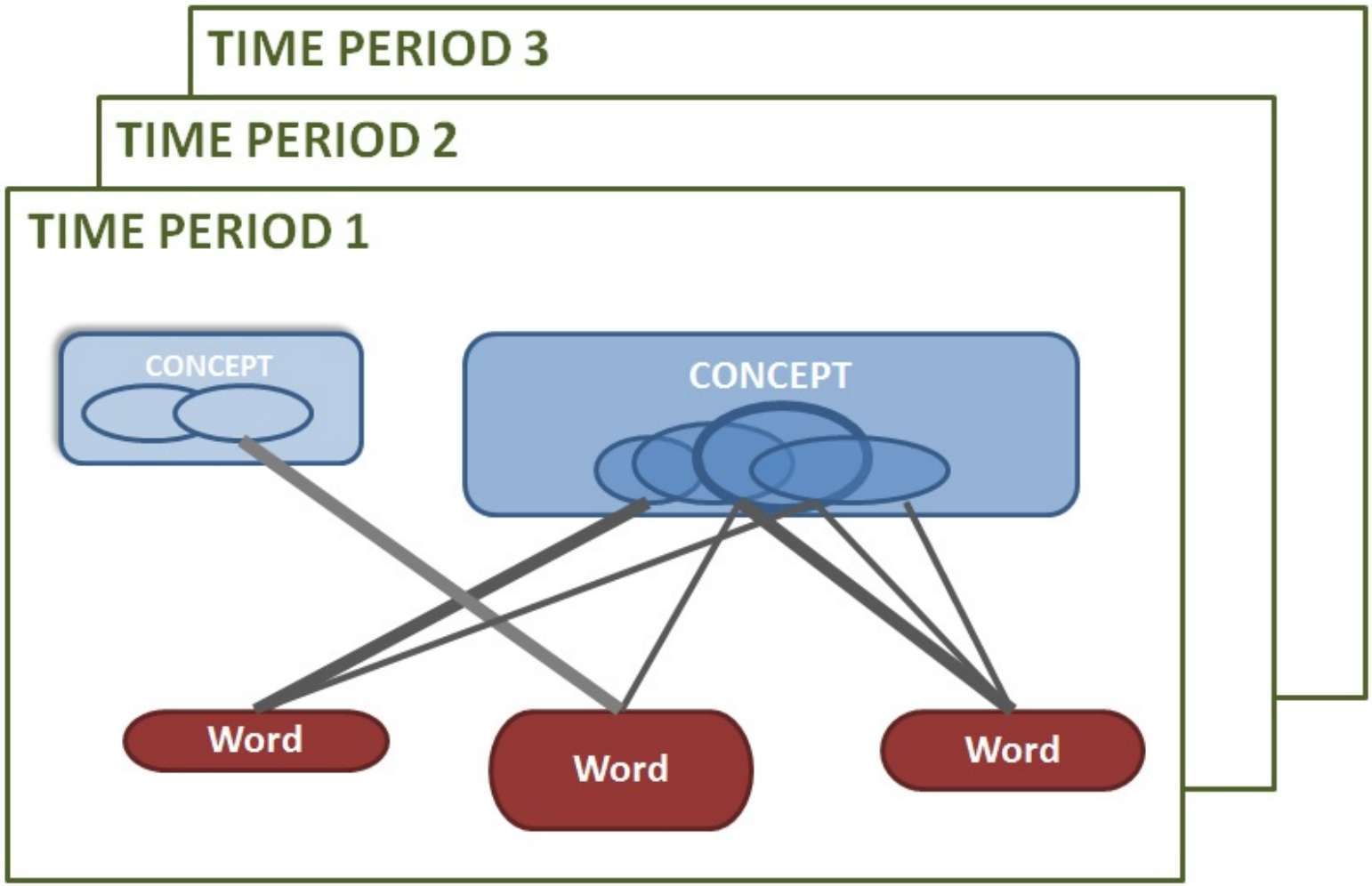
Analysing the Structure of Lexical Variation

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



Analysing the Structure of Lexical Variation

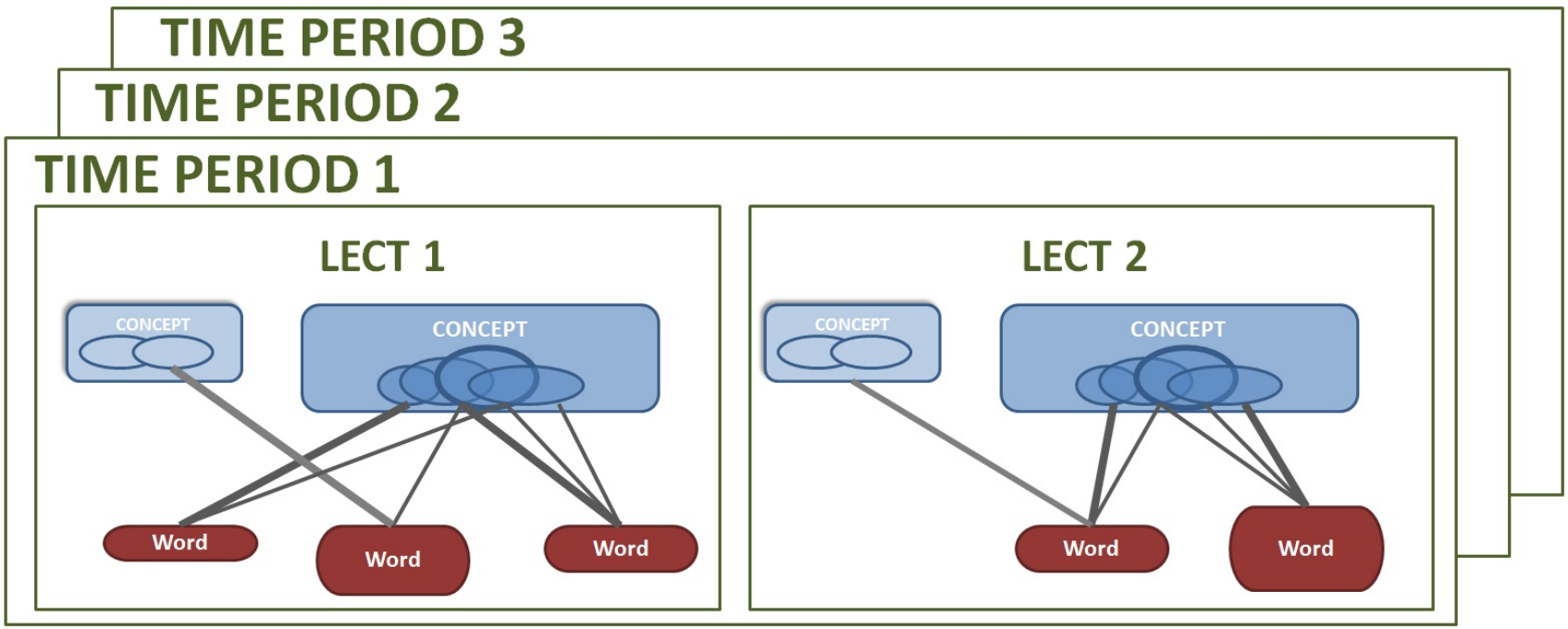
Mapping between concept and lexemes can change over time:
DIACHRONIC VARIATION:



Analysing the Structure of Lexical Variation

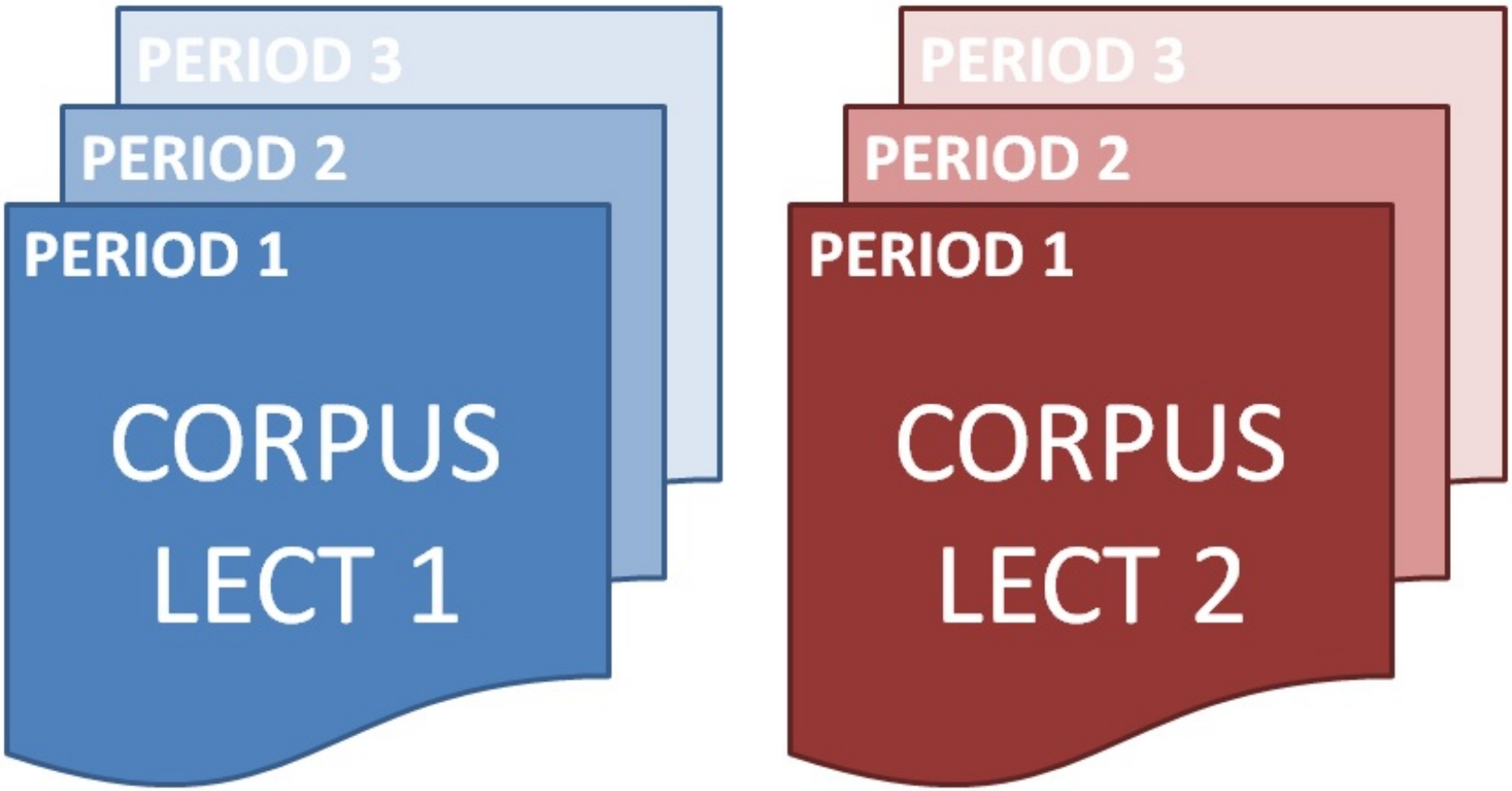
How do all these different factors interact?

STRUCTURE OF LEXICAL VARIATION (Geeraerts et al. 1994)



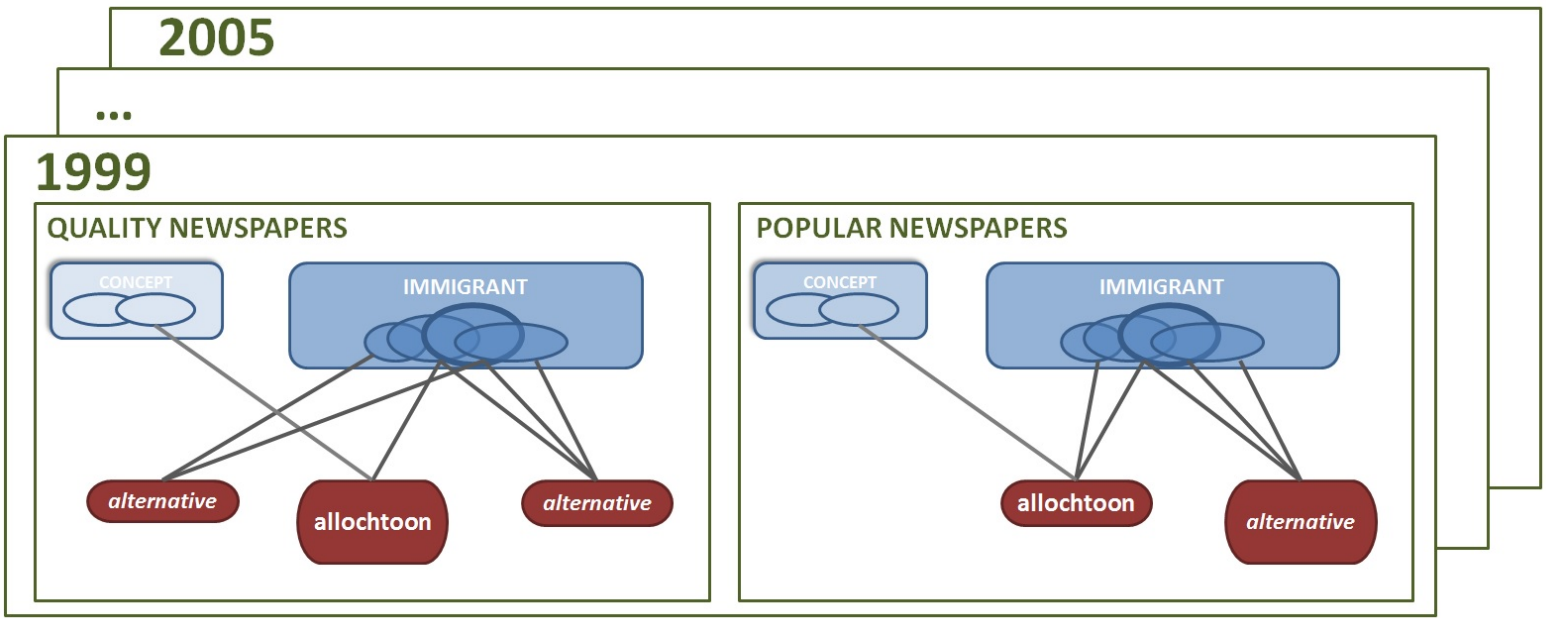
Analysing the Structure of Lexical Variation

Usage based analysis:
STRATIFIED CORPORA



Case study: IMMIGRANTS

PERSON WITH IMMIGRATION BACKGROUND:



Case study: IMMIGRANTS

STRATIFIED CORPORA OF BELGIAN DUTCH (1.3G words)

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

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

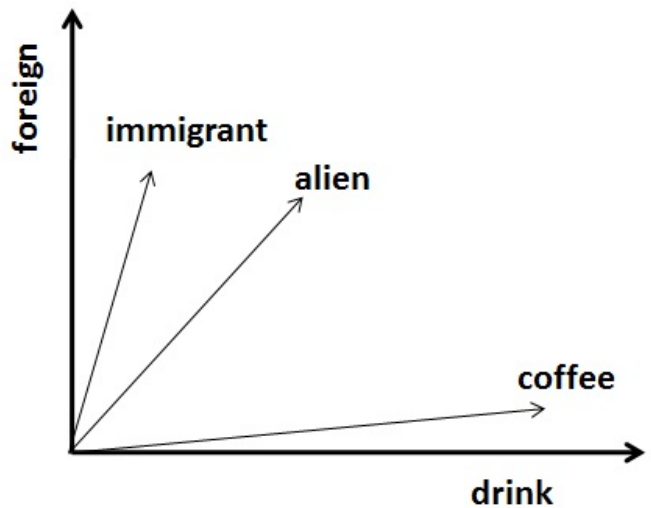
weight the raw frequencies by collocational strength (pmi)

	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

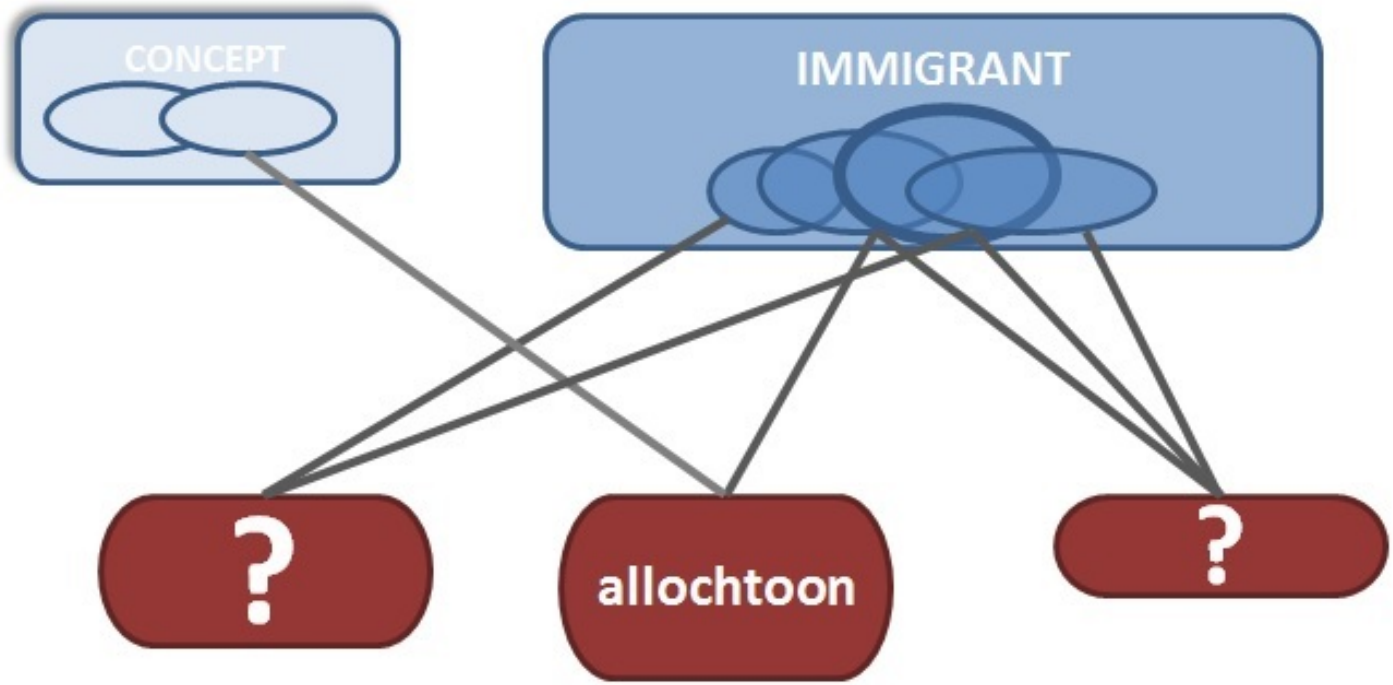
calculate word by word similarity matrix

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



Overview

Identifying alternative expressions



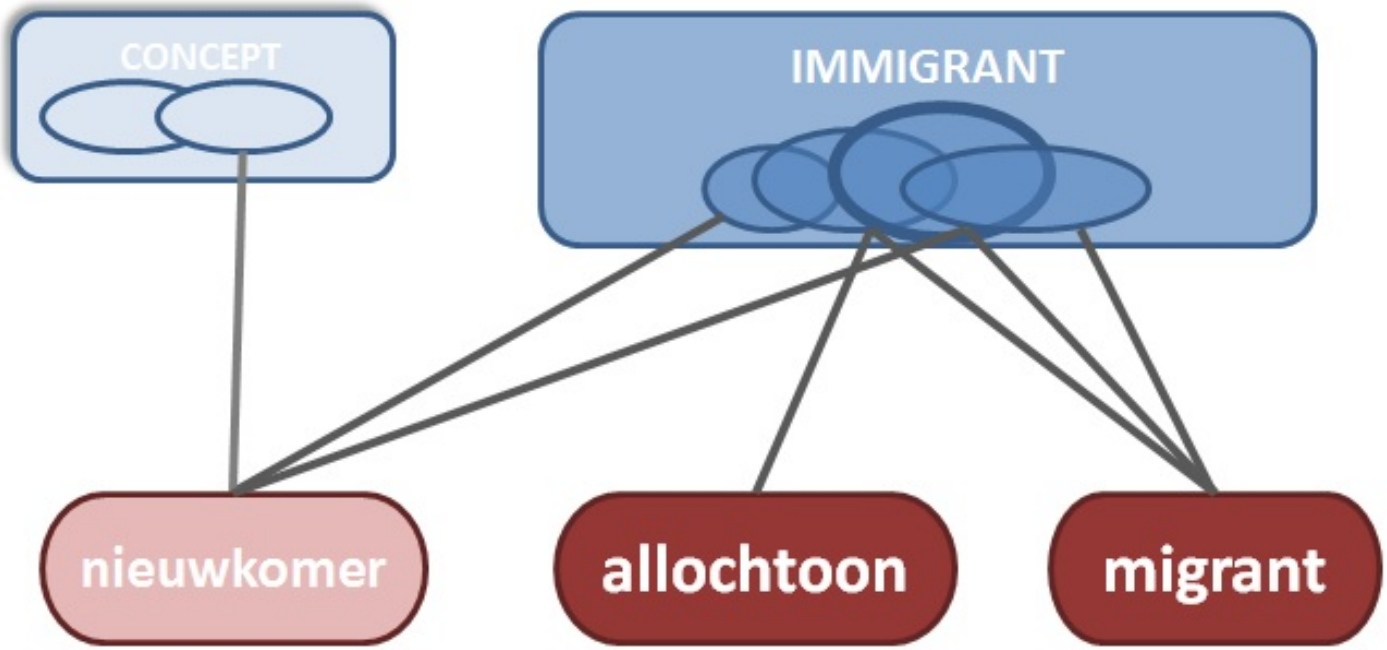
- calculate contextual similarity between 10K Dutch nouns
- sort by similarity to *allochtoon*

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*

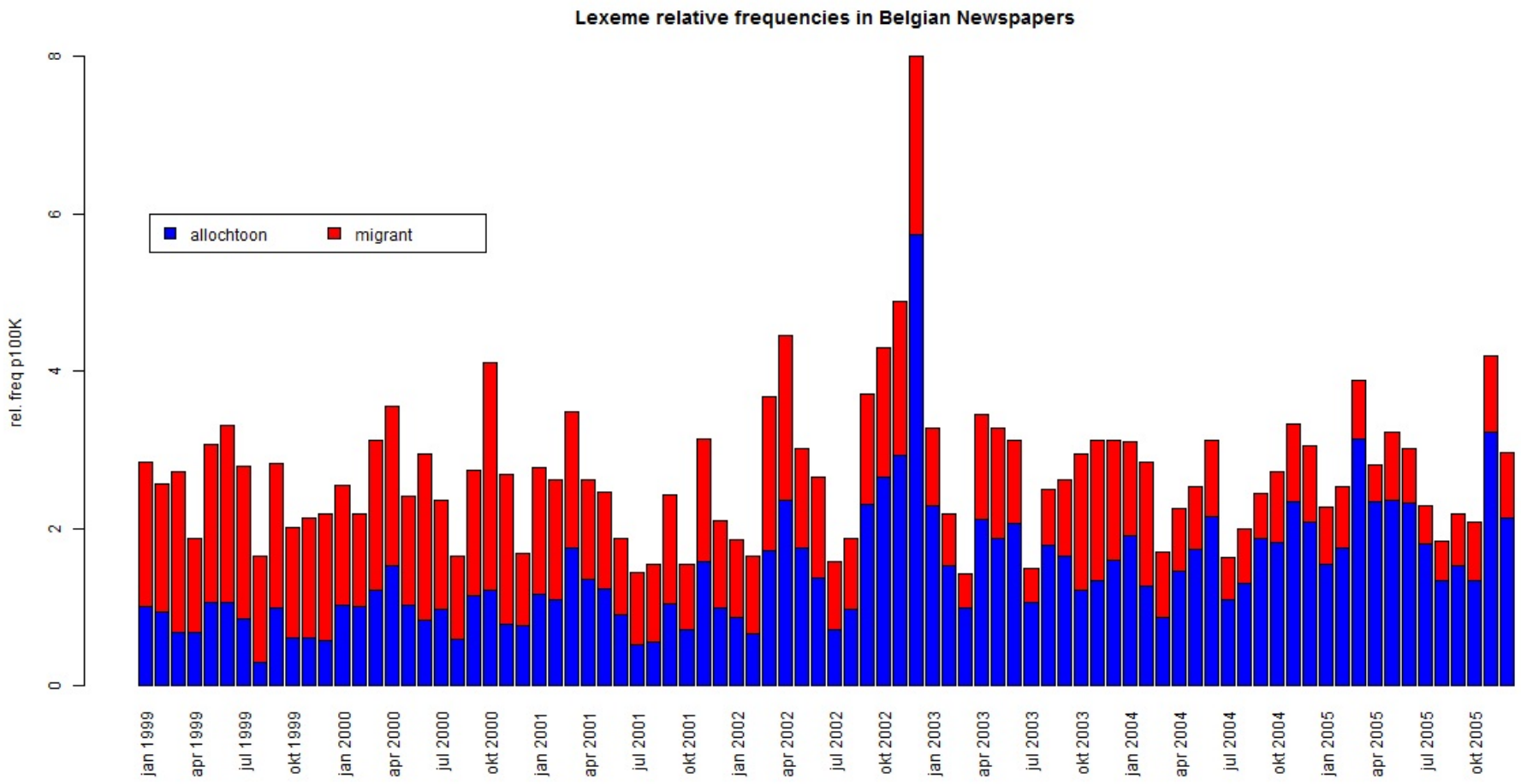
Identifying alternative expressions



Identifying alternative expressions

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

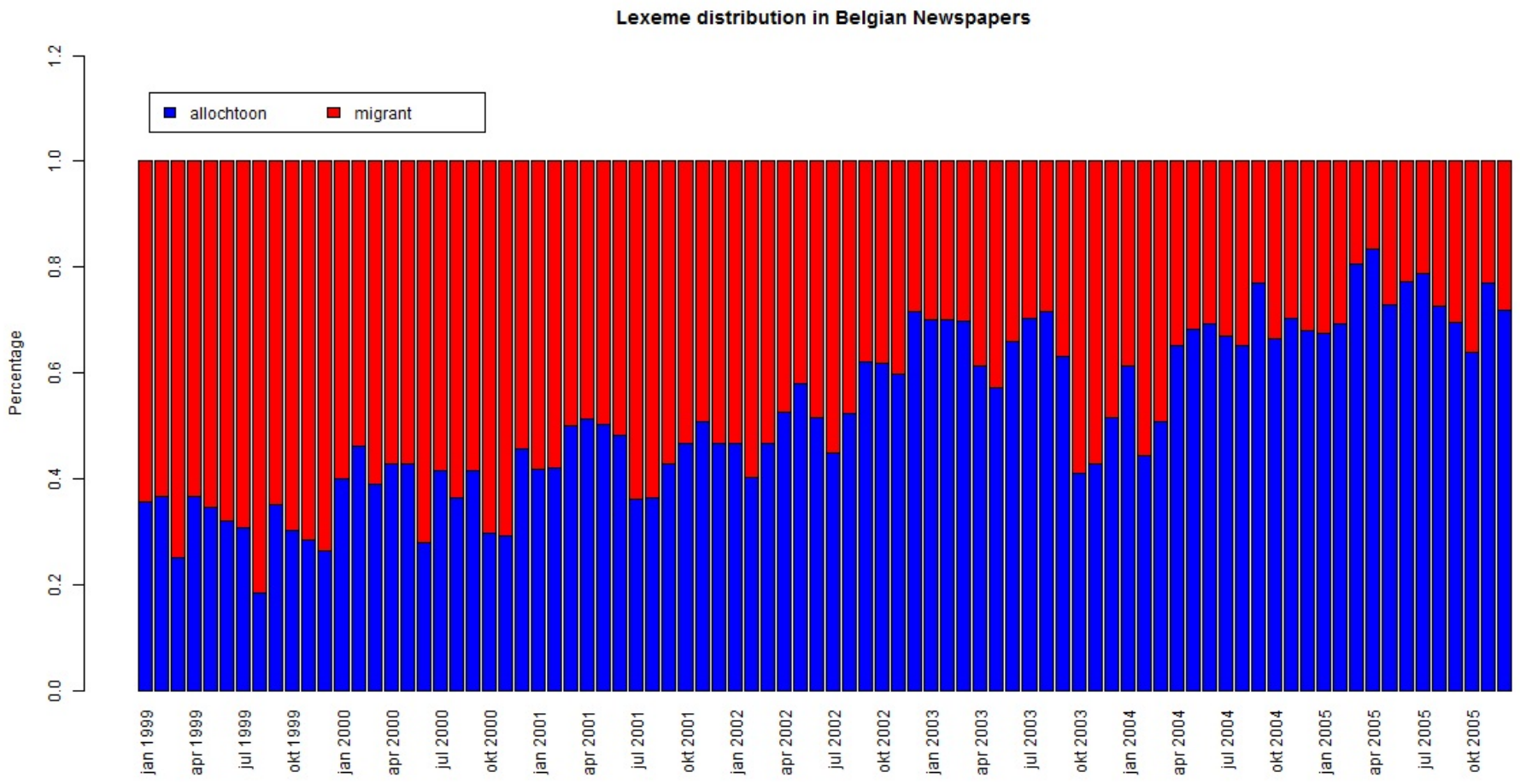
immigrant-talk seems to be a seasonal phenomenon



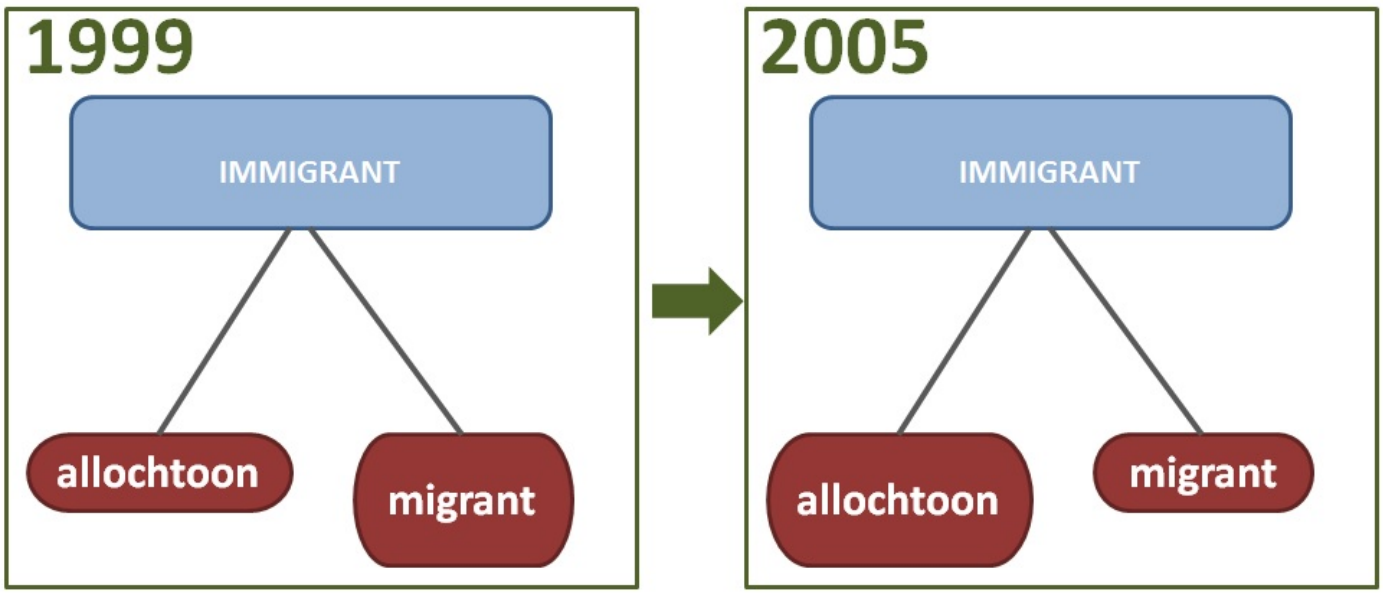
Identifying alternative expressions

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

allochtoon becomes more frequent than *migrant*



Identifying alternative expressions

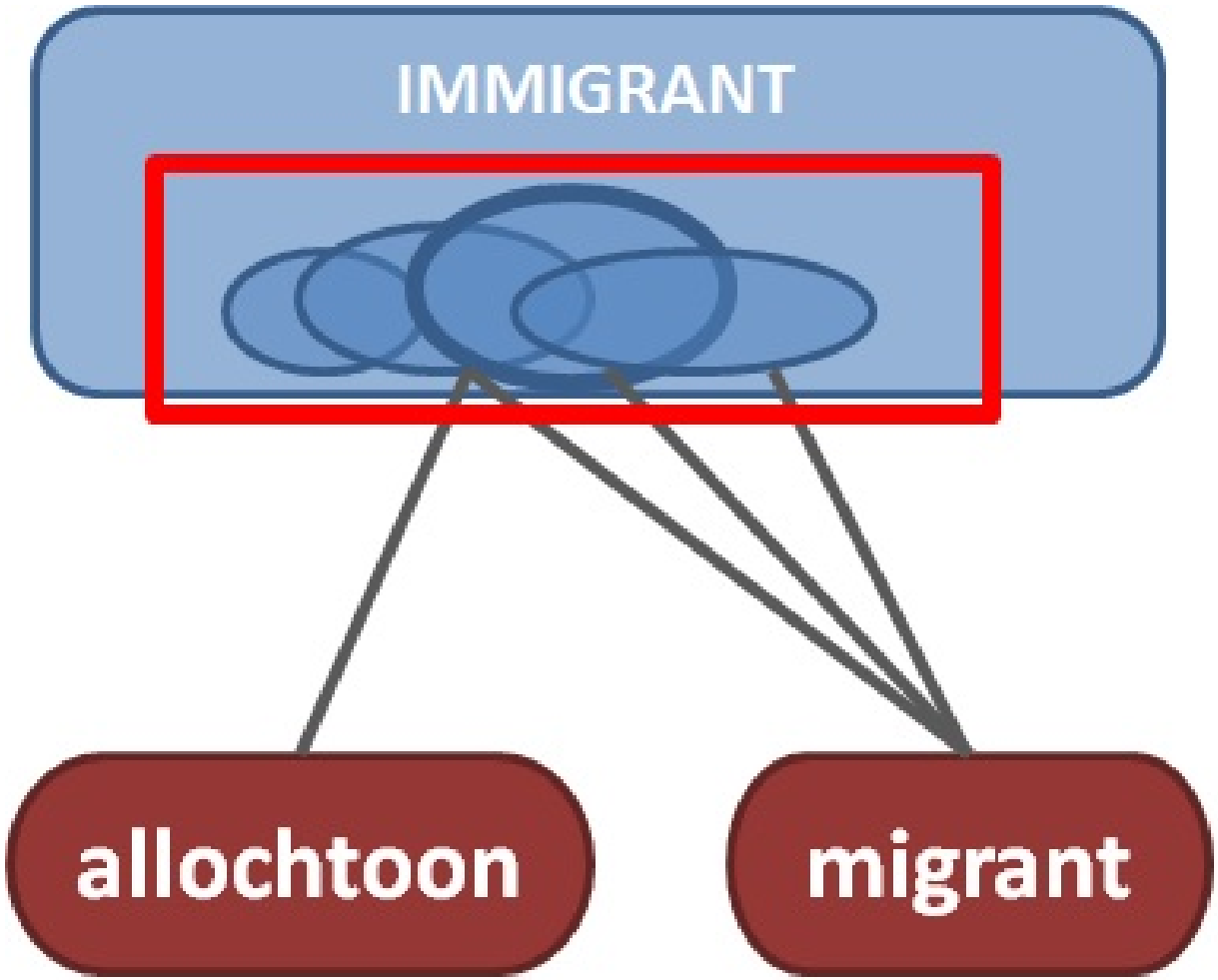


Is this change in frequency also indicative of semantic change?

Overview

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

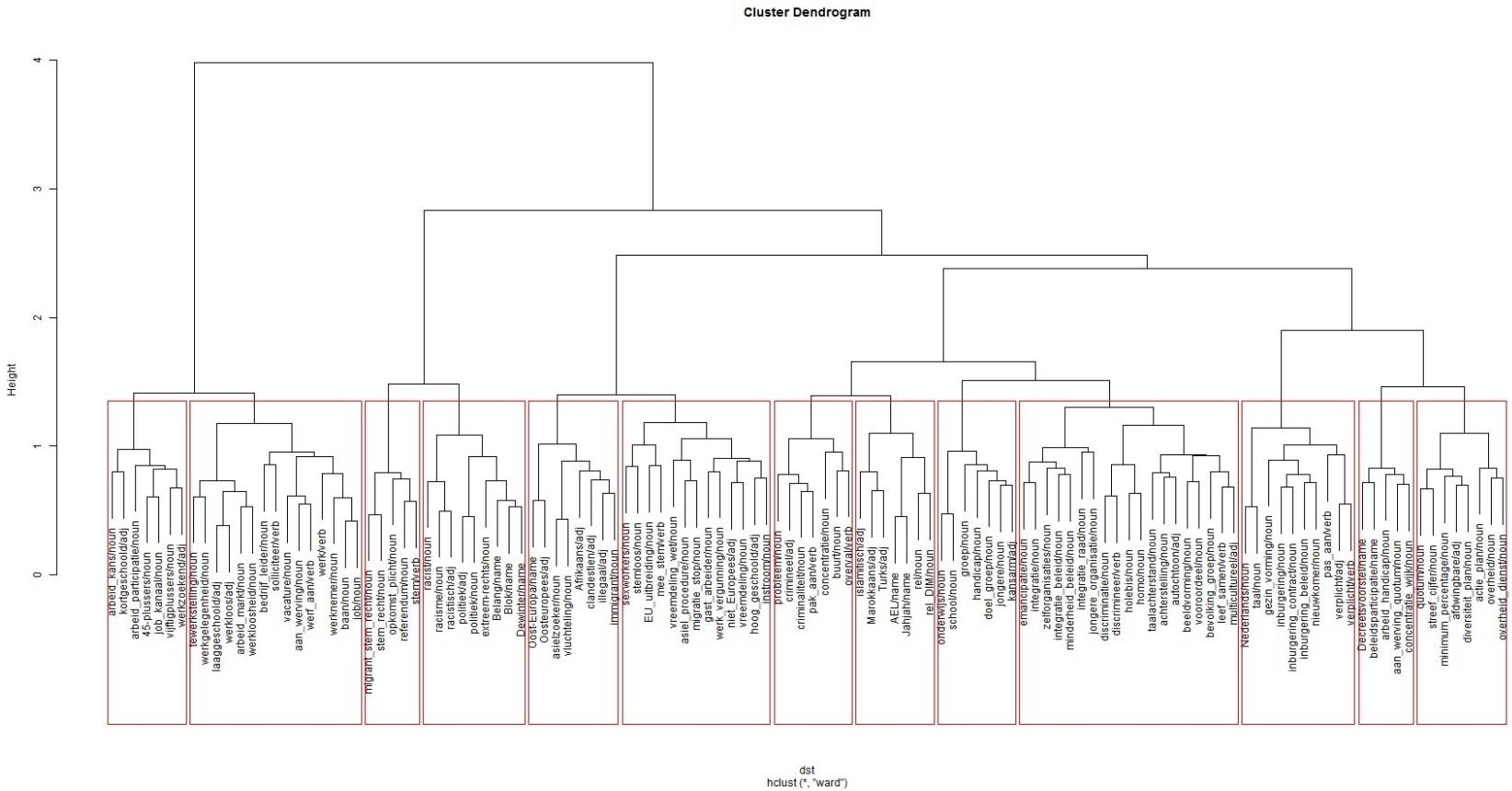
Analysing Semantic Structure

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

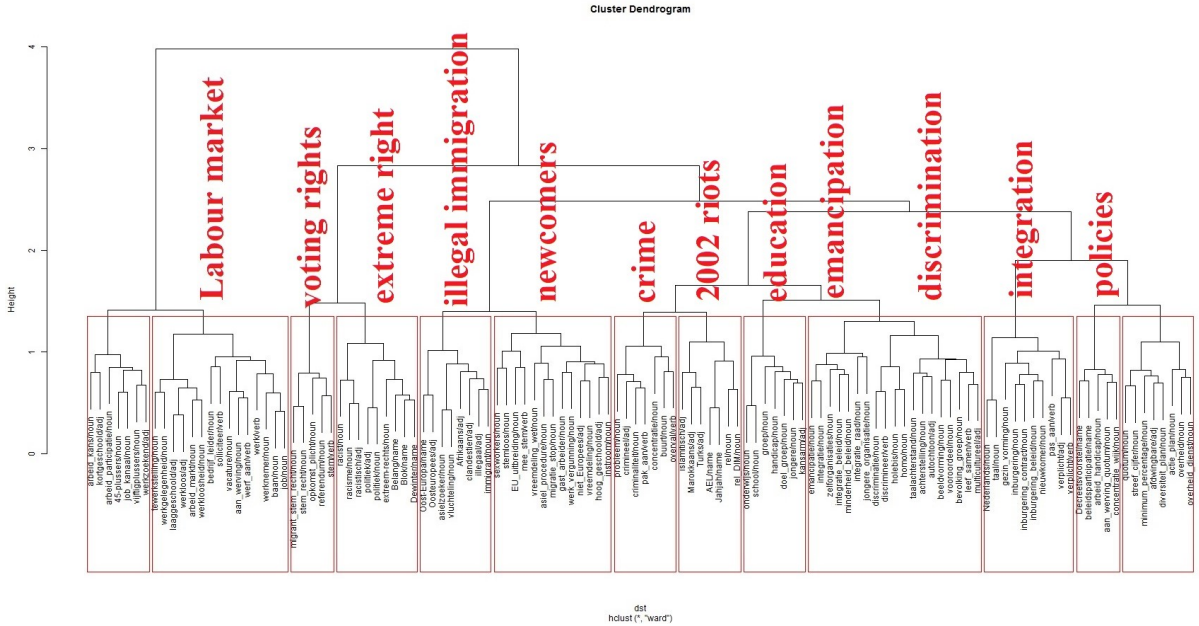
Analysing Semantic Structure

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

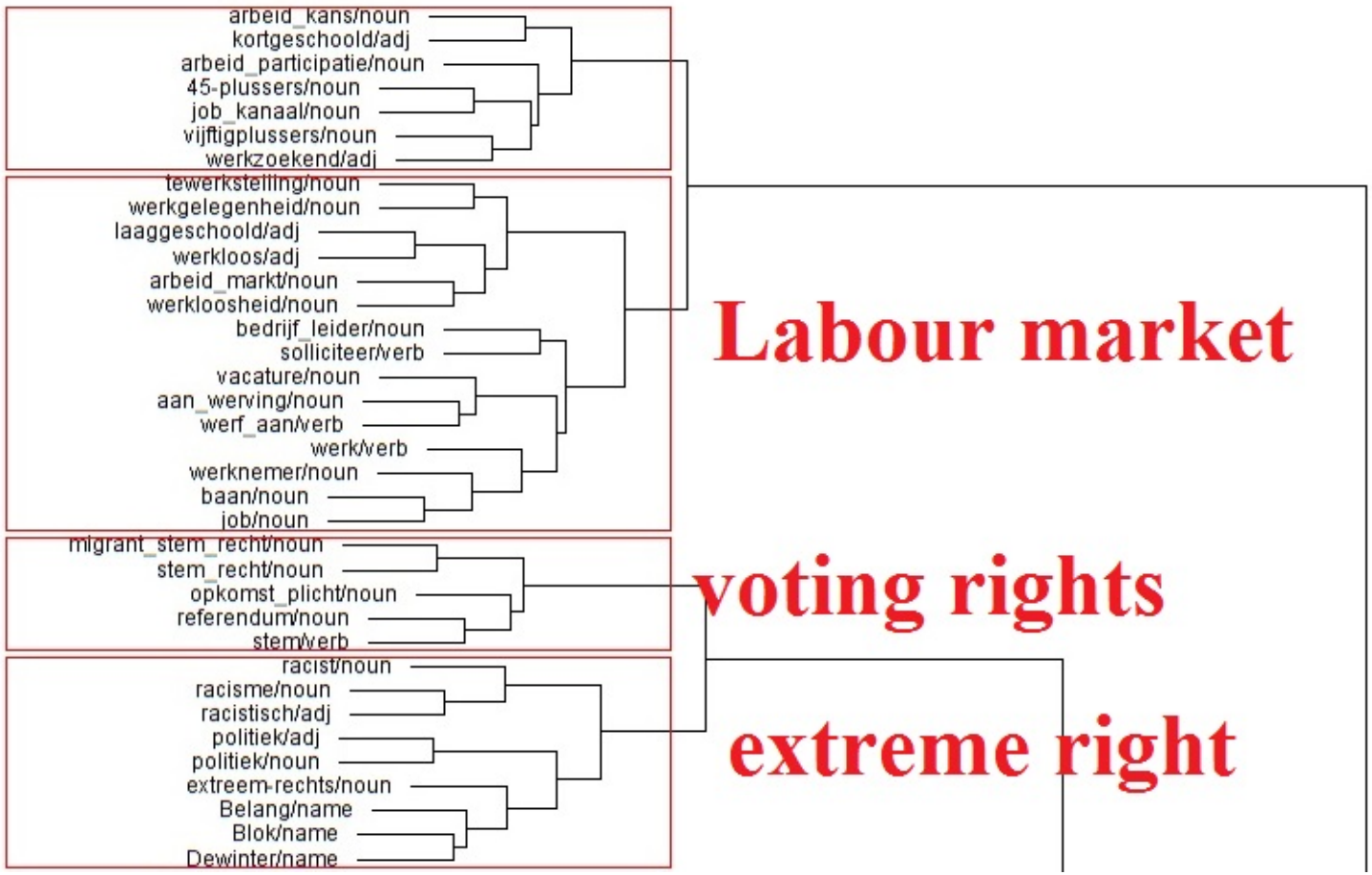


Analysing Semantic Structure

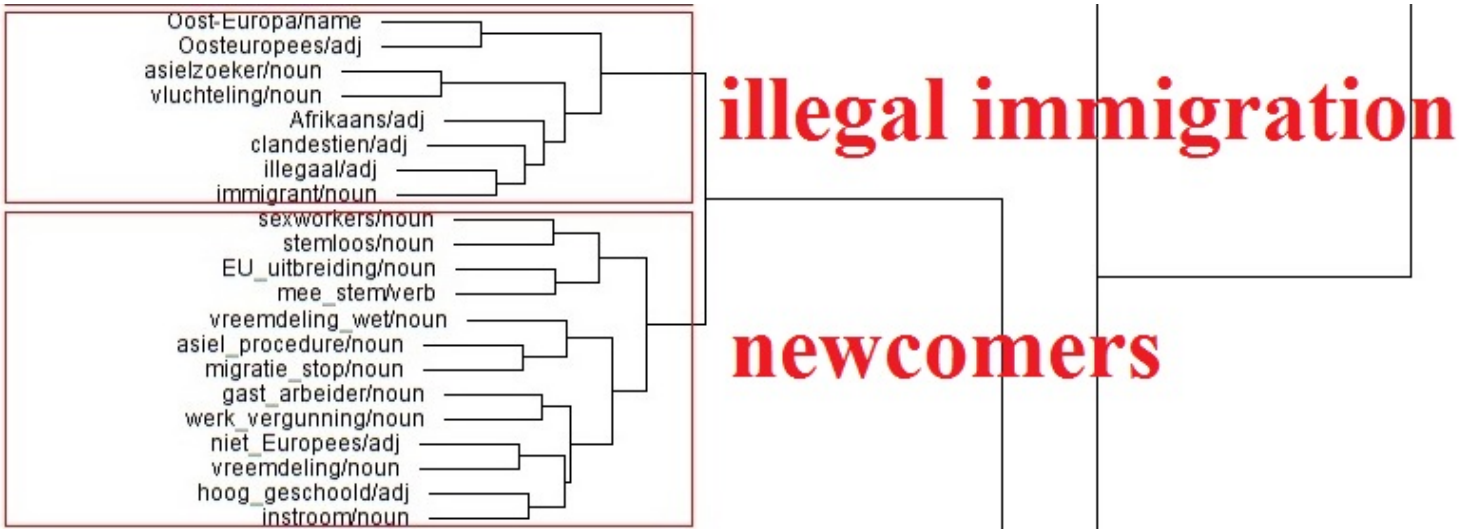
Clusters of contextually related collocations ≈ semantic features
Clusters can be labeled manually



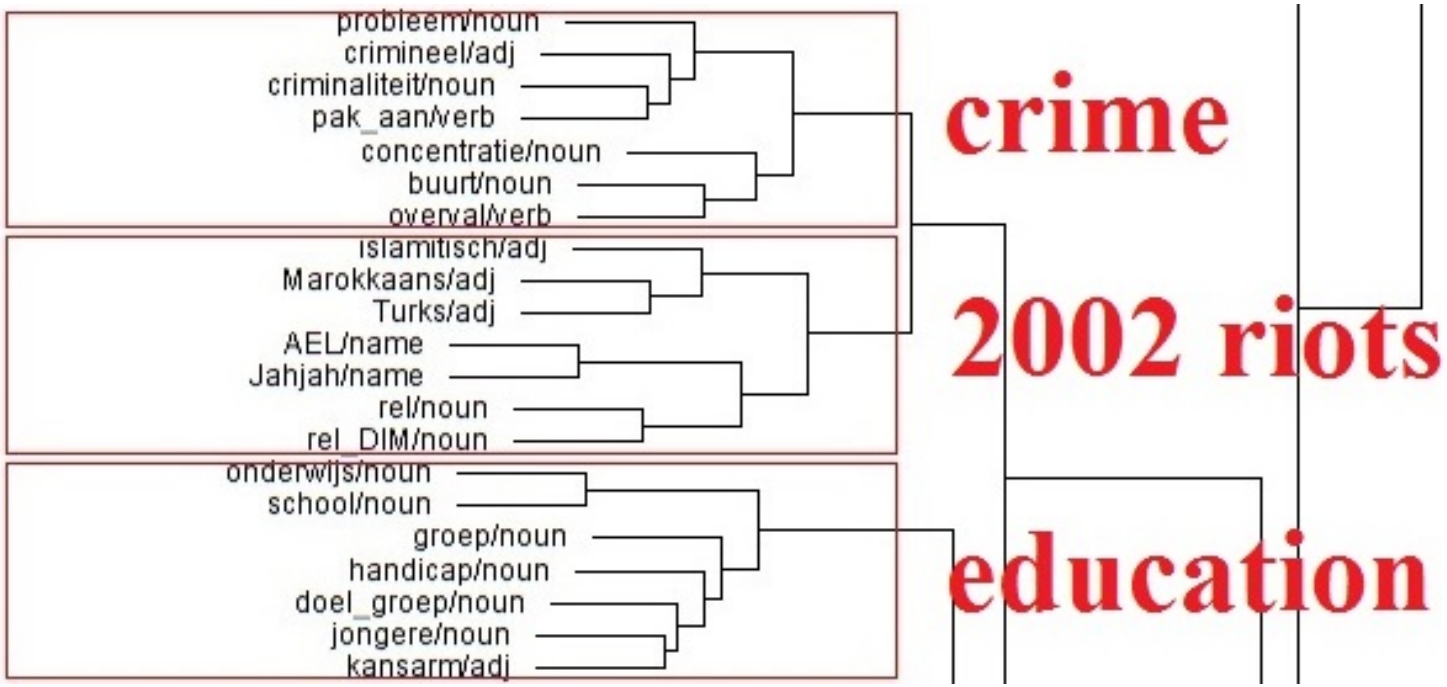
Analysing Semantic Structure



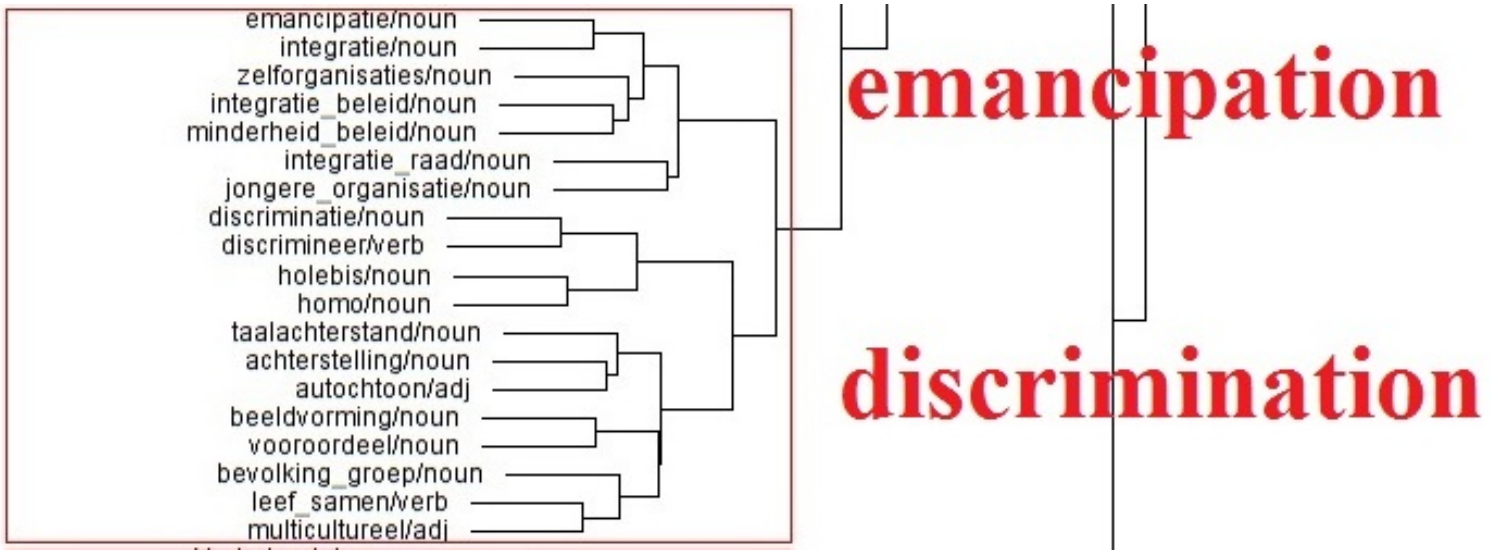
Analysing Semantic Structure



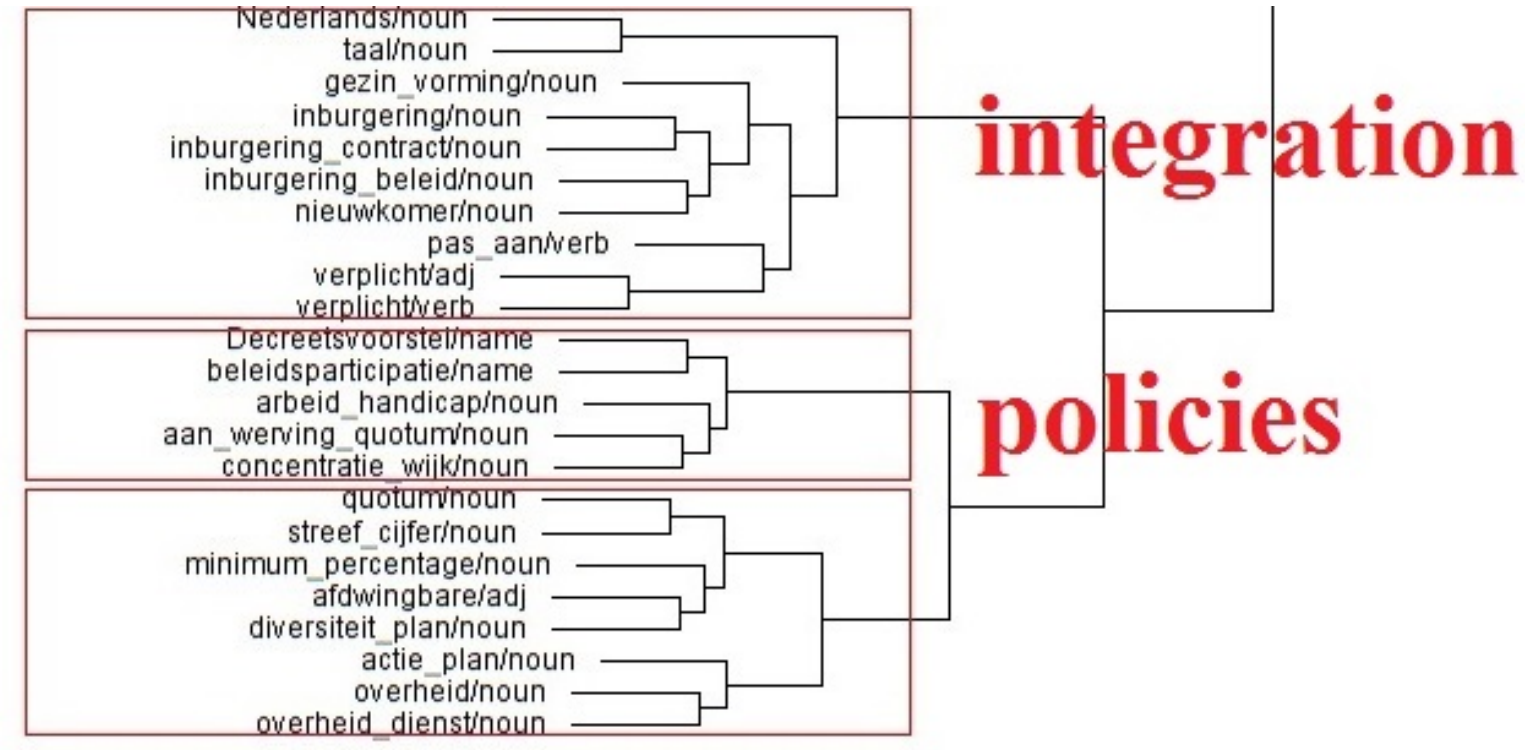
Analysing Semantic Structure



Analysing Semantic Structure

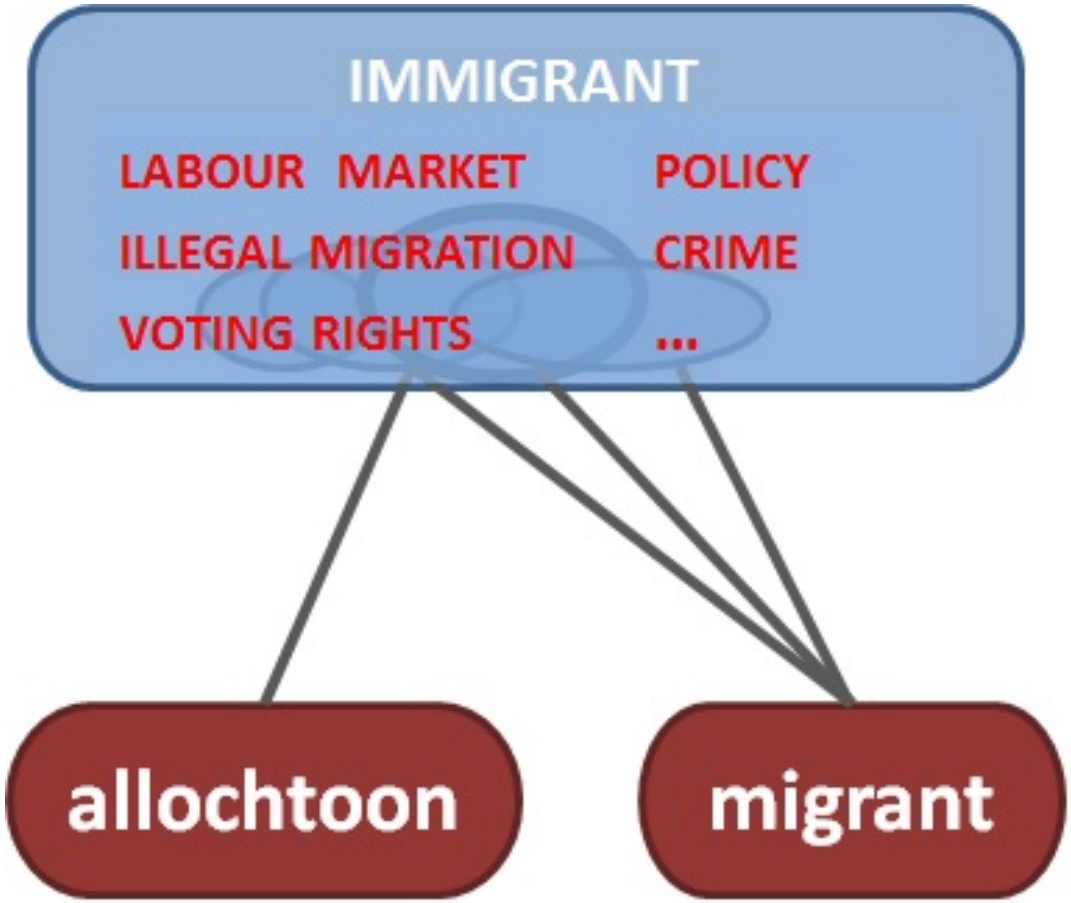


Analysing Semantic Structure



Analysing Semantic Structure

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



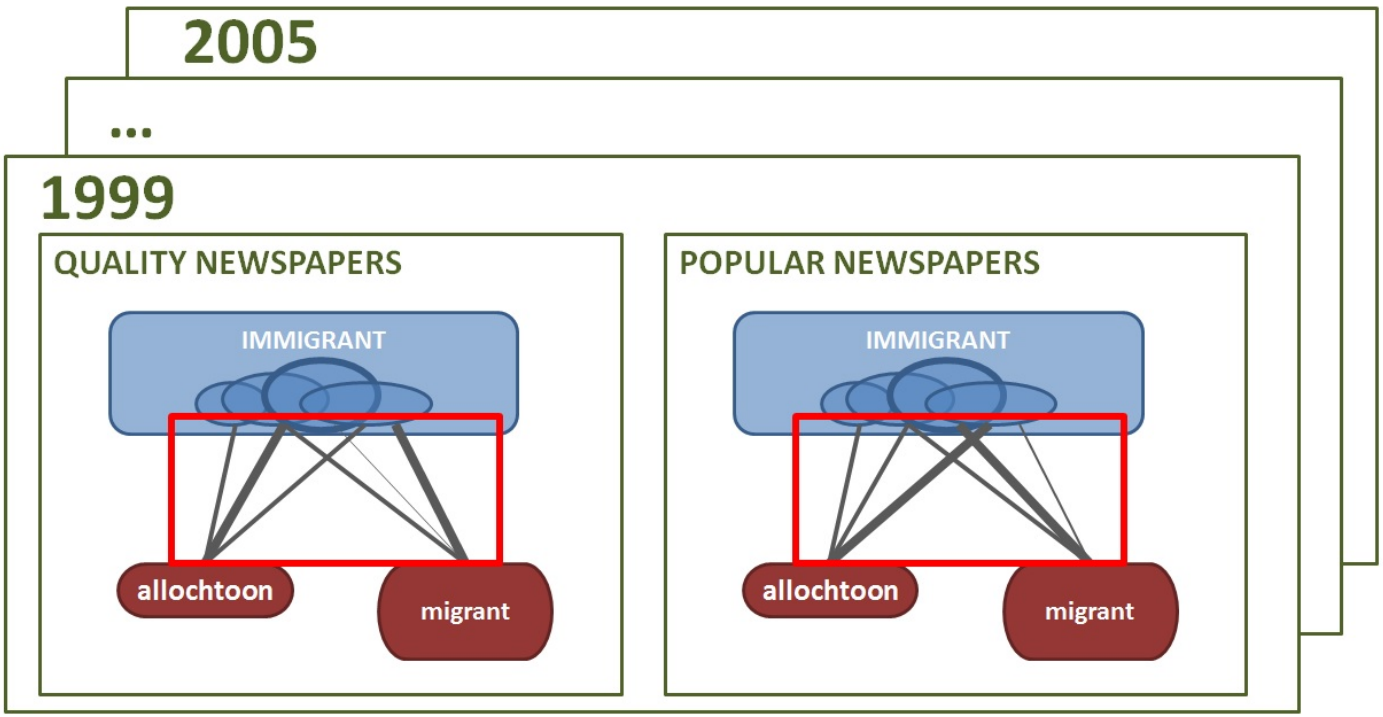
Overview

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

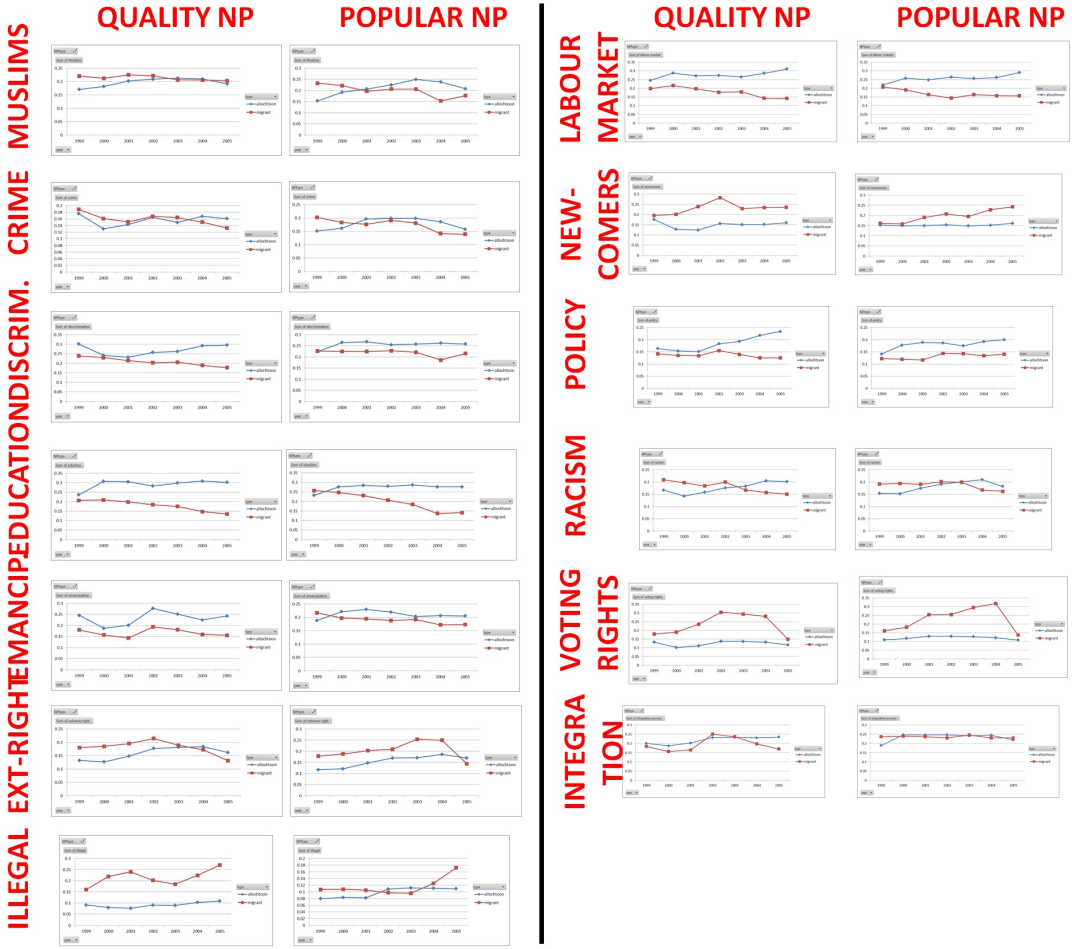
Calculate the cosine similarity ($\approx$ association strength) of each variant/year/type vector to each context cluster vector

	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

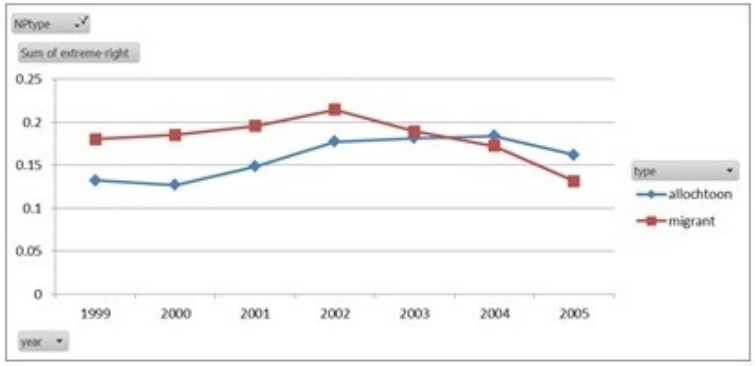


Measuring semantic change in registers

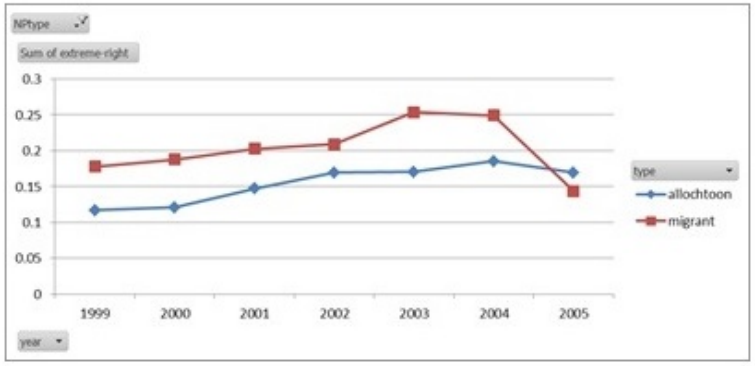
ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

EXT-RIGHT

QUALITY

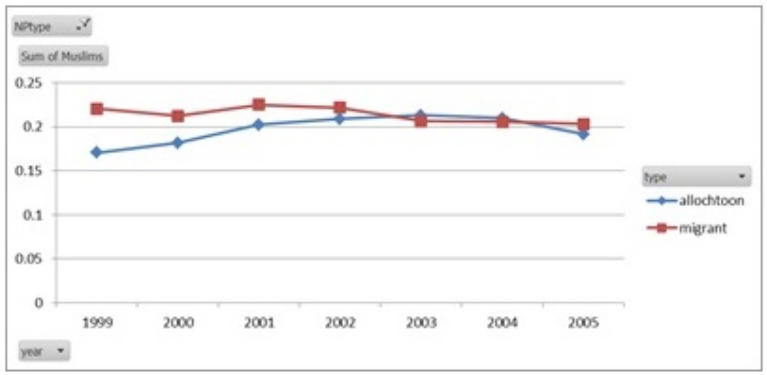
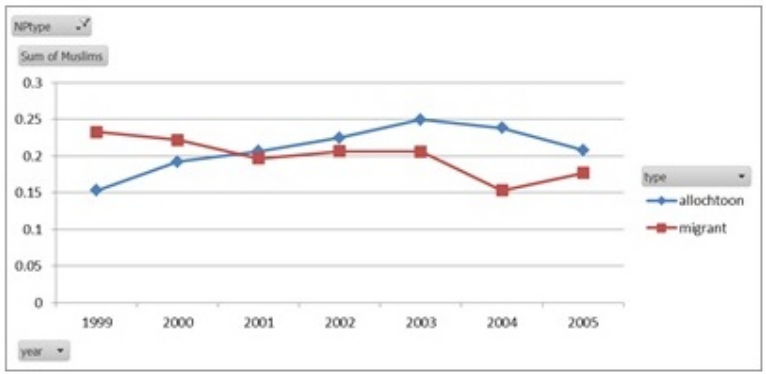


POPULAR



Measuring semantic change in registers

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

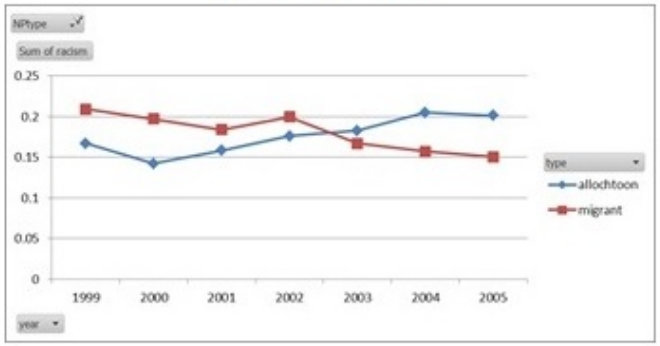
	QUALITY NP	POPULAR NP
MUSLIMS		

Measuring semantic change in registers

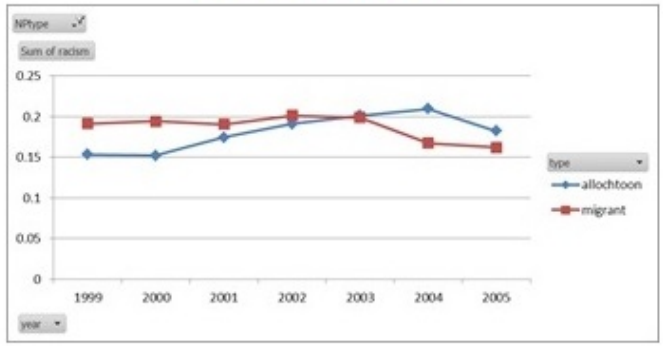
ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

RACISM

QUALITY



POPULAR

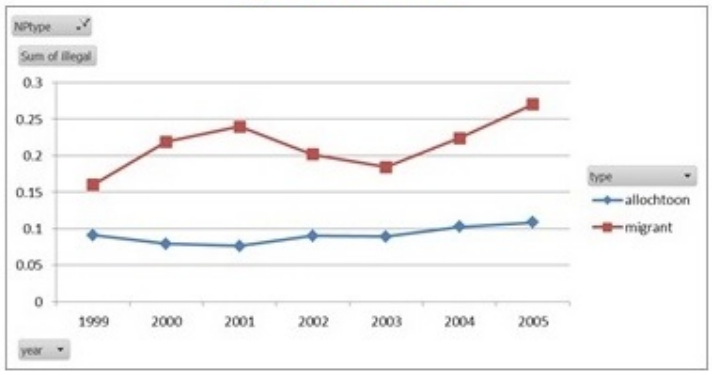


Measuring semantic change in registers

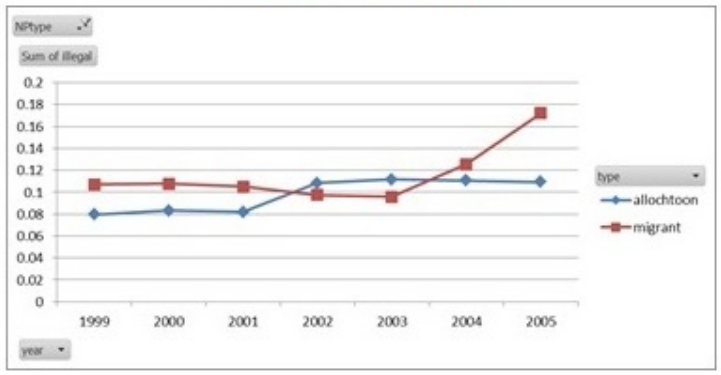
MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

ILLEGAL

QUALITY



POPULAR

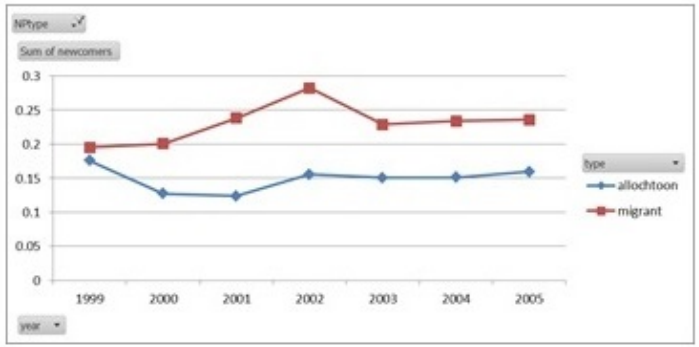


Measuring semantic change in registers

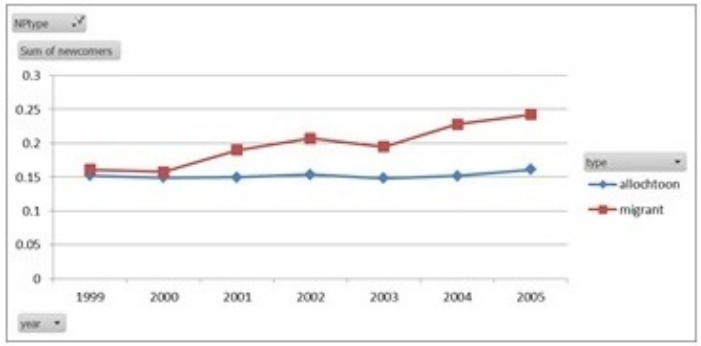
MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

NEW-COMERS

QUALITY



POPULAR

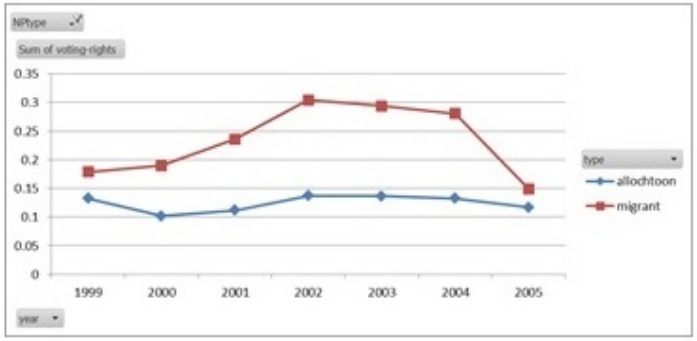


Measuring semantic change in registers

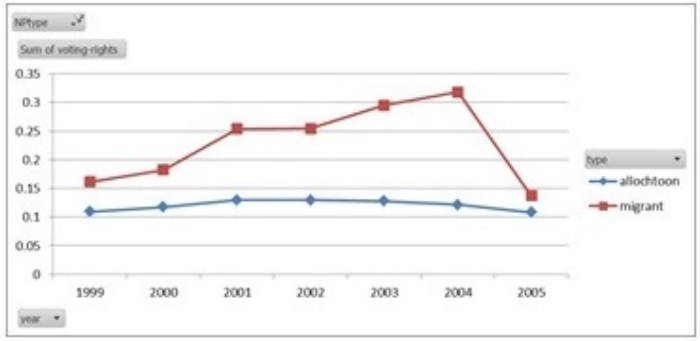
MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

VOTING RIGHTS

QUALITY



POPULAR

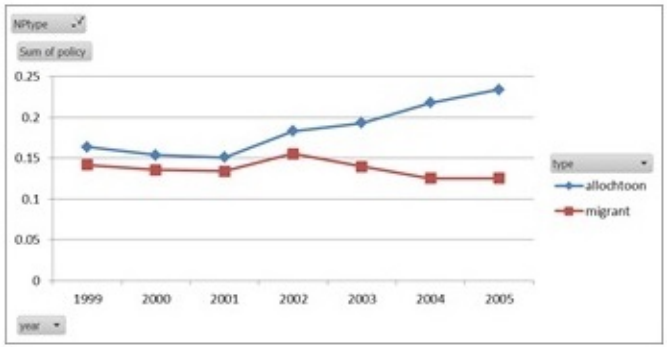


Measuring semantic change in registers

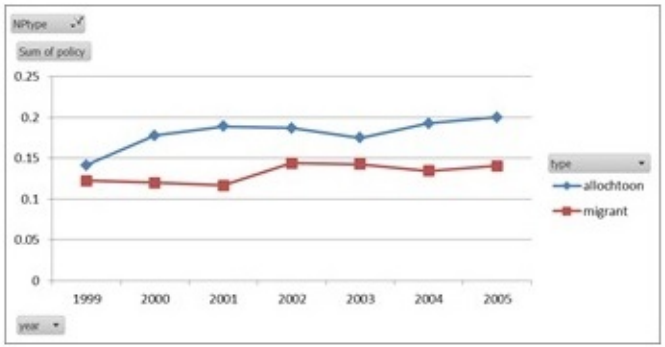
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

POLICY

QUALITY

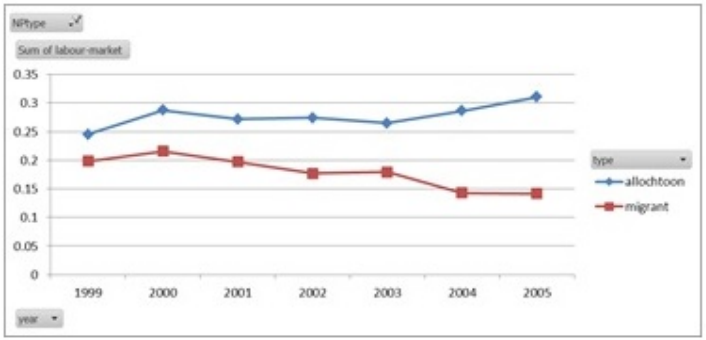
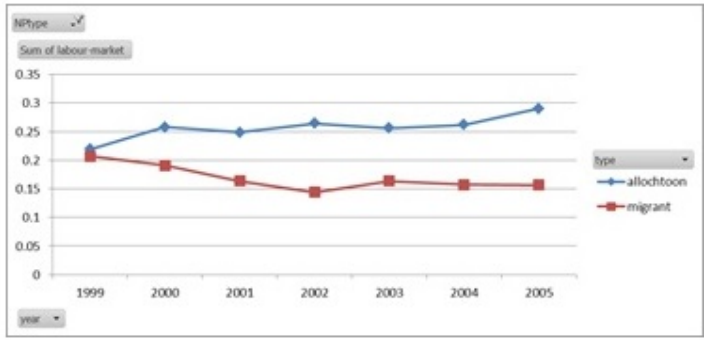


POPULAR



Measuring semantic change in registers

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

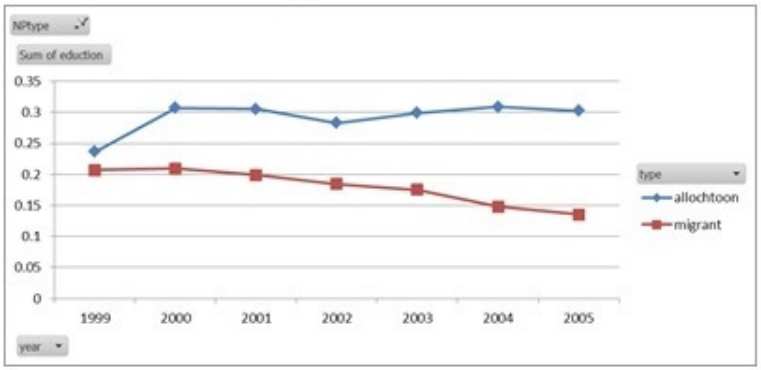
	QUALITY NP	POPULAR NP
LABOUR MARKET		

Measuring semantic change in registers

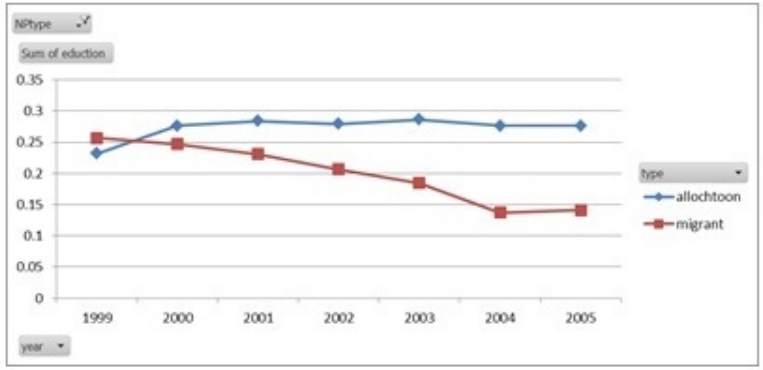
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

EDUCATION

QUALITY



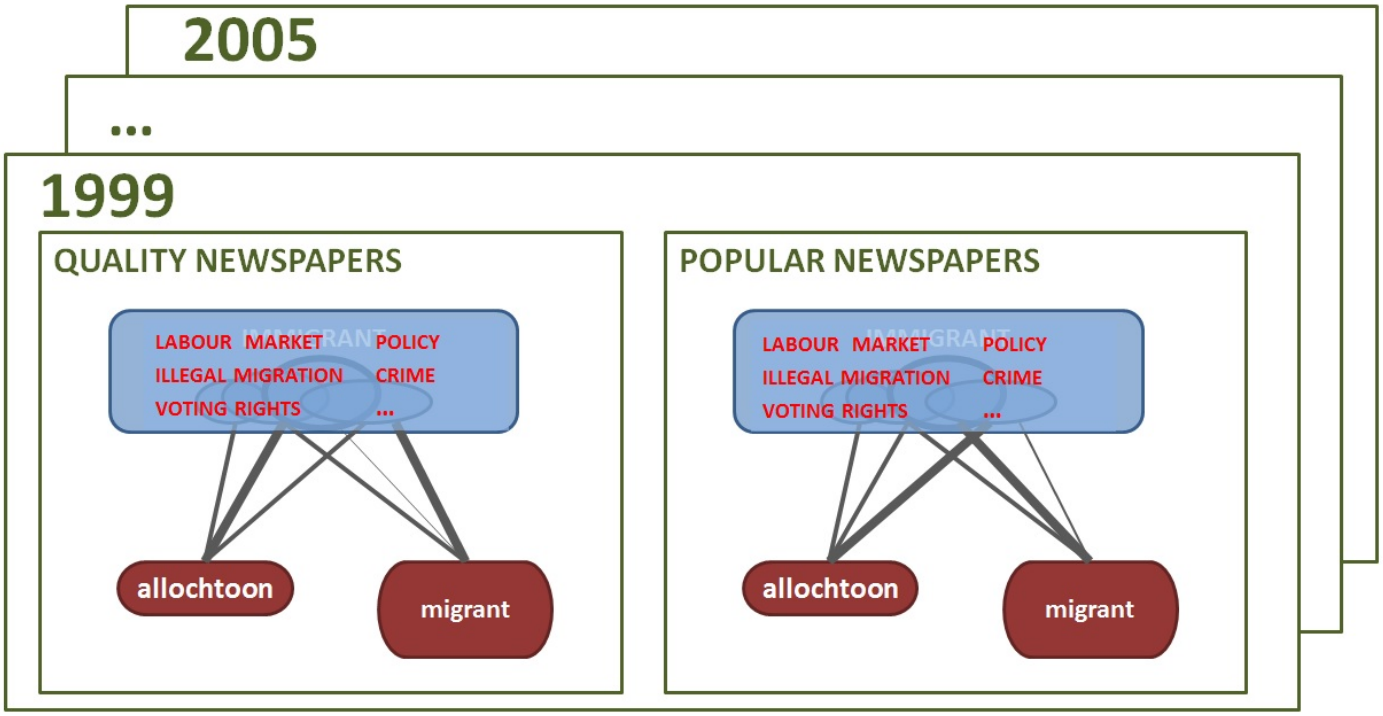
POPULAR



Measuring semantic change in registers

Measuring semantic change in registers

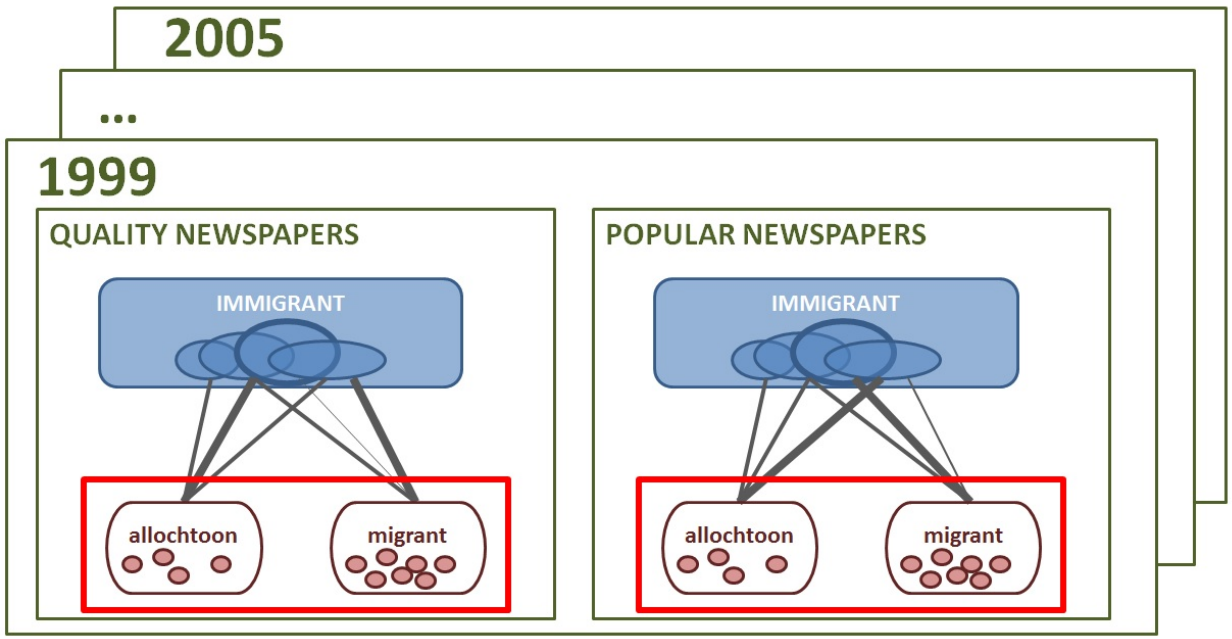
Association strength between semantic features and lexemes differ between registers and changes over time.



Overview

Lexical variation on the attestation level

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



Lexical variation on the attestation level

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

op de arbeidsmarkt zijn er voor allochtonen nauwelijks jobs

Lexical variation on the attestation level

Make a vector for each attestation 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 attestation level

Make a vector for each attestation 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 attestation level

Make a vector for each attestation 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

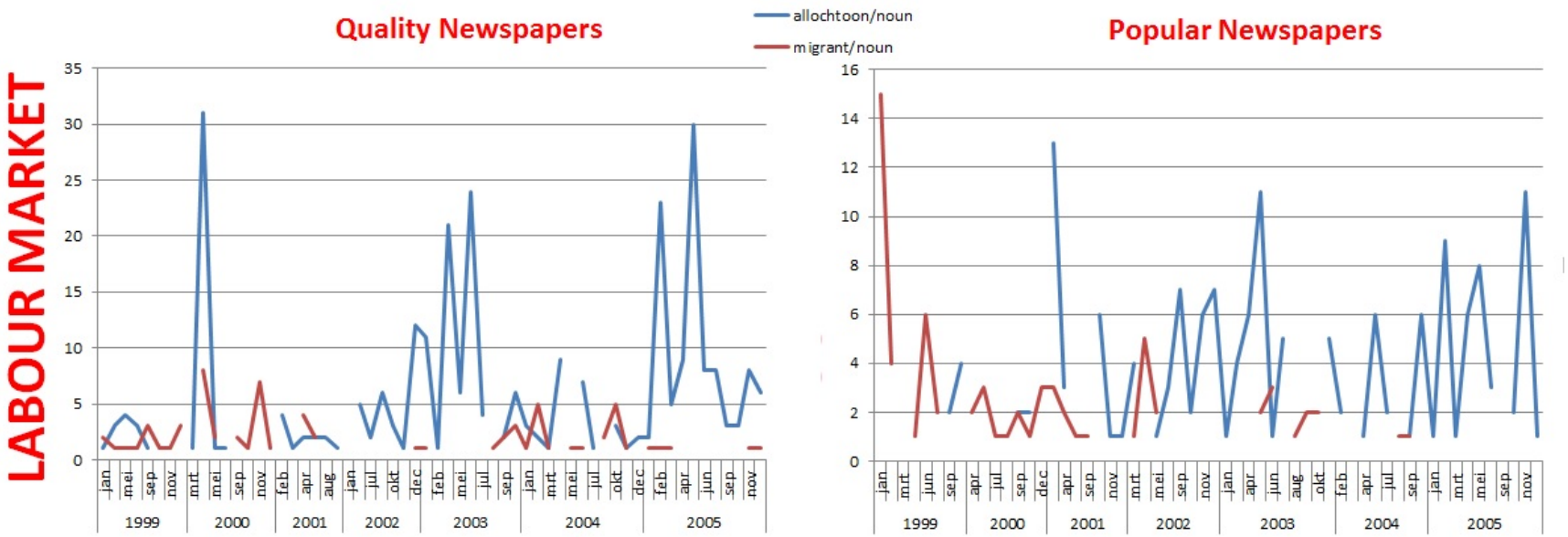
Lexical variation on the attestation level

Make a vector for each exemplar of *allochtoon* and *migrant*
STEP 4: calculate similarity matrix between attestation and cluster vectors

	LABOUR	*ILLEGAL*	*EXTREME*	*POLICY*	*CRIME*	*VOTING*	*RACISM*
$allochtoon_1$	0.1	0.9	0.5	0.4	0.8	0.6	...
$allochtoon_2$	0.4	0.3	0.7	0.2	0.2	0.5	...
$allochtoon_3$	0.3	0.9	0.4	0.3	0.3	0.6	...
$migrant_1$	0.3	0.2	0.7	0.3	0.2	0.4	...
$migrant_2$	0.8	0.5	0.5	0.1	0.1	0.0	...
$migrant_3$	0.9	0.4	0.7	0.2	0.2	0.7	...

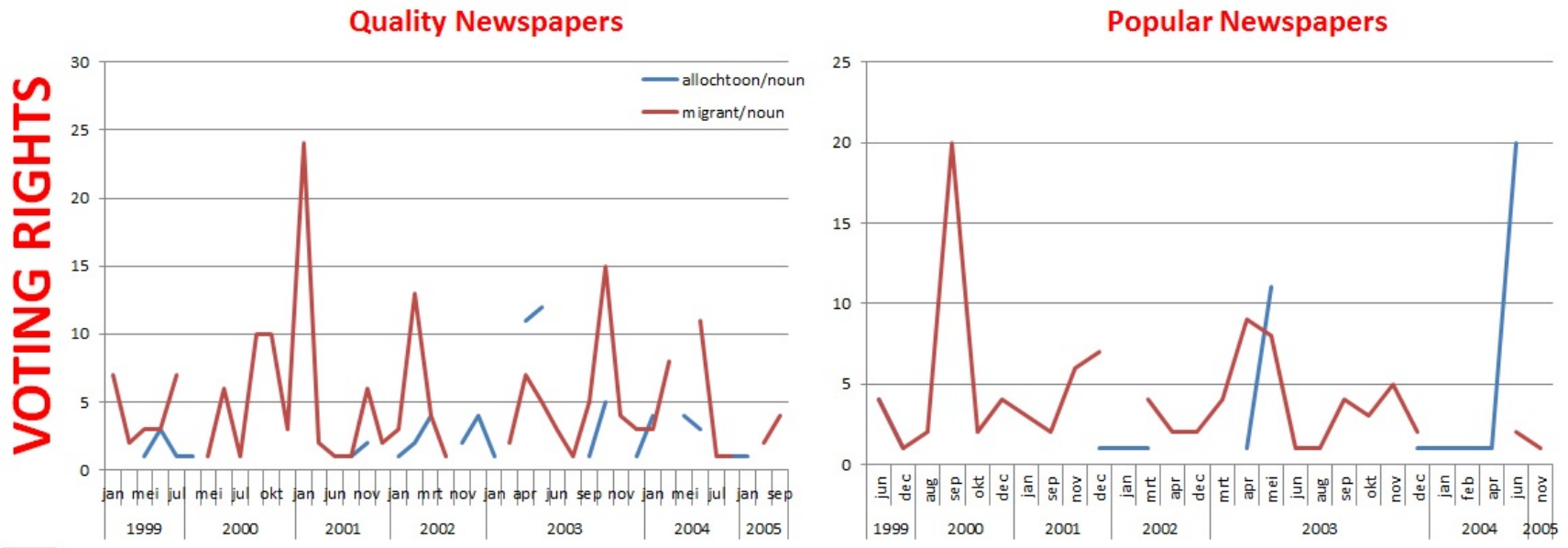
Lexical variation on the attestation level

Same evolution as on aggregated type-level, but with peaks visible



Lexical variation on the attestation level

Same evolution as on aggregated type-level, but with peaks visible



Lexical variation on the attestation level

Semantic cluster per attestation can be combined with extra-linguistic predictors in inferential modelling

```
Call:
glm(formula = variant ~ newspaper + monthnumber + resid + CL01 +
    CL02 + CL03 + CL04 + CL06 + CL07 + CL07 + CL08 + CL09 + CL10 +
    CL11 + CL12 + CL13, family = binomial, data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9664  -0.9103  -0.5307   0.9945   2.7569

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.6583084  0.2791552   2.358 0.018363 *
newspaperDM  0.2260101  0.1717969   1.316 0.188320
newspaperDS  0.1815149  0.1765265   1.028 0.303828
newspaperLN  0.0272115  0.1727756   0.157 0.874854
monthnumber -0.0294094  0.0017122 -17.177  < 2e-16 ***
resid       -0.0002842  0.0001802  -1.577 0.114732
CL01        -5.5730571  3.0764818  -1.812 0.070063 .
CL02         0.0935768  0.3008047   0.311 0.755734
CL03         4.2707470  0.4471896   9.550  < 2e-16 ***
CL04         0.8441630  0.2667436   3.165 0.001552 **
CL06        -0.0607867  0.4939538  -0.123 0.902058
CL07         0.7259322  0.3184058   2.280 0.022614 *
CL08        -0.6191612  0.2849283  -2.173 0.029777 *
CL09        -1.0547586  0.3425114  -3.079 0.002074 **
CL10        -2.1430753  0.5510571  -3.889 0.000101 ***
CL11         0.3620700  0.2924706   1.238 0.215727
```

Overview

9. Conclusion

Descriptive: *allochtoon* vs. *migrant*

- *allochtoon* replaces *migrant* in frequency
- immigration discussions seems to have strong 'seasonal peaks, especially in high-brow newspapers
- *allochtoon* gradually monopolizes socio-political contexts (labour market, eduction, policy)
- *migrant* had a flirt with 'voting rights' and specializes for 'new and 'illegal immigration.
- tendencies are stronger in quality than popular newspapers

Contra DM: Is *allochtoon* vaguely defined? No.

- *allochtoon* seem to become more and more specialized
- identifies a group that is the target of specific socio-political government policies

Methodological conclusions

Semantic Vector Spaces as large-scale, generalized collocation analysis 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 individual occurrences of lexemes and be an input to inferential models

Cognitive Sociolinguistics:

- semantically complex variation can be studied empirically in large datasets
- lectal and semantic variation can and need to be taken into account simultaneously

For more information:

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

kris.heylen@kuleuven.be dirk.speelman@kuleuven.be

References I

Baker, Paul. 2012. Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, **9**(3), 247-256.

Dirven, R., Polzenhagen, F., & Wolf, H.-G. 2007. Cognitive Linguistics, Ideology, and Critical Discourse Analysis. Pages 1222–1241 In: D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* .

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

Geeraerts, Dirk, Kristiansen, Gitte & Peirsman, Yves (Eds). 2010. *Advances in cognitive sociolinguistics.* Mouton De Gruyter, Berlin.

References II

Hart, Christopher. 2011. Moving beyond metaphor in the Cognitive Linguistics approach to CDA: Construal operations in immigration discourse. Pages 171–192 in: C. Hart (Ed.), *Critical discourse studies in context and cognition*.

Heylen, Kris, Speelman, Dirk, & Geeraerts, Dirk. 2012. Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. Pages 16–24 in: *Proceedings of the EACL-2012 joint workshop of LINGVIS & UNCLH: Visualization of Language Patters and Uncovering Language History from Multilingual Resources*.

Hilpert, Martin. 2011. Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, **16**(4), 435–461.

References III

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

Koller, Veronika. 2008. Corporate brands as socio-cognitive representations. Pages 389–418 In: G. Kristiansen & R. Dirven (Eds.) , *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*.

Orpin, Debbie. 2005. Corpus Linguistics and Critical Discourse Analysis: Examining the ideology of sleaze. *International Journal of Corpus Linguistics*, **10**(1), 37-61.

Peirsman, Yves, Heylen, Kris & Geeraerts, Dirk. 2010. Applying word space models to sociolinguistics. Religion names before and after 9/11.. Pages 111–137 In: D. Geeraerts, G. Kristiansen & Y. Peirsman (Eds.) *Advances in Cognitive Sociolinguistics*.

References IV

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