Tracking Immigration Discourse through Time:

A Semantic Vector Space Approach to Discourse Analysis

Kris Heylen, Thomas Wielfaert & 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: Integrate typical research questions from Critical Discourse Analysis into the usage-based and lectally enriched framework of Cognitive Sociolinguistics

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

Overview

1. Background: The Rise and Fall of a political correct term
2. CDA meets CSL
3. Analysing Lexical semantics in CSL
4. Semantic Vector Spaces
5. Identifying alternative expressions
6. Identifying and structuring collocations
7. Measuring lectal and diachronic distances
8. Structuring concordances
9. Conclusion

Overview

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

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. CDA meets CSL

Politically and ideologically loaded concepts have been studied by Critical Discourse Analysis (CDA) since the 1970s(Fairclough, Wodak). However, CDA has been criticized for:

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

However, recent rapprochements with:

- Cognitive Linguistics, providing a stronger theoretical basis (see Dirven, Polzenhagen & Wolf 2007 for an overview; Hart 2011 on *Immigration*)
- Corpus Linguistics for a usage-based methodology (Orpin 2005 on *Corruption*, Baker 2012 on *Muslims*

2. CDA meets CSL

These theoretical and empirical trends link up CDA with Cognitive Sociolinguistics (Kristiansen & Dirven 2008; Geeraerts, Kristiansen& Peirsman 2010) and allow to study political discourse:

- 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

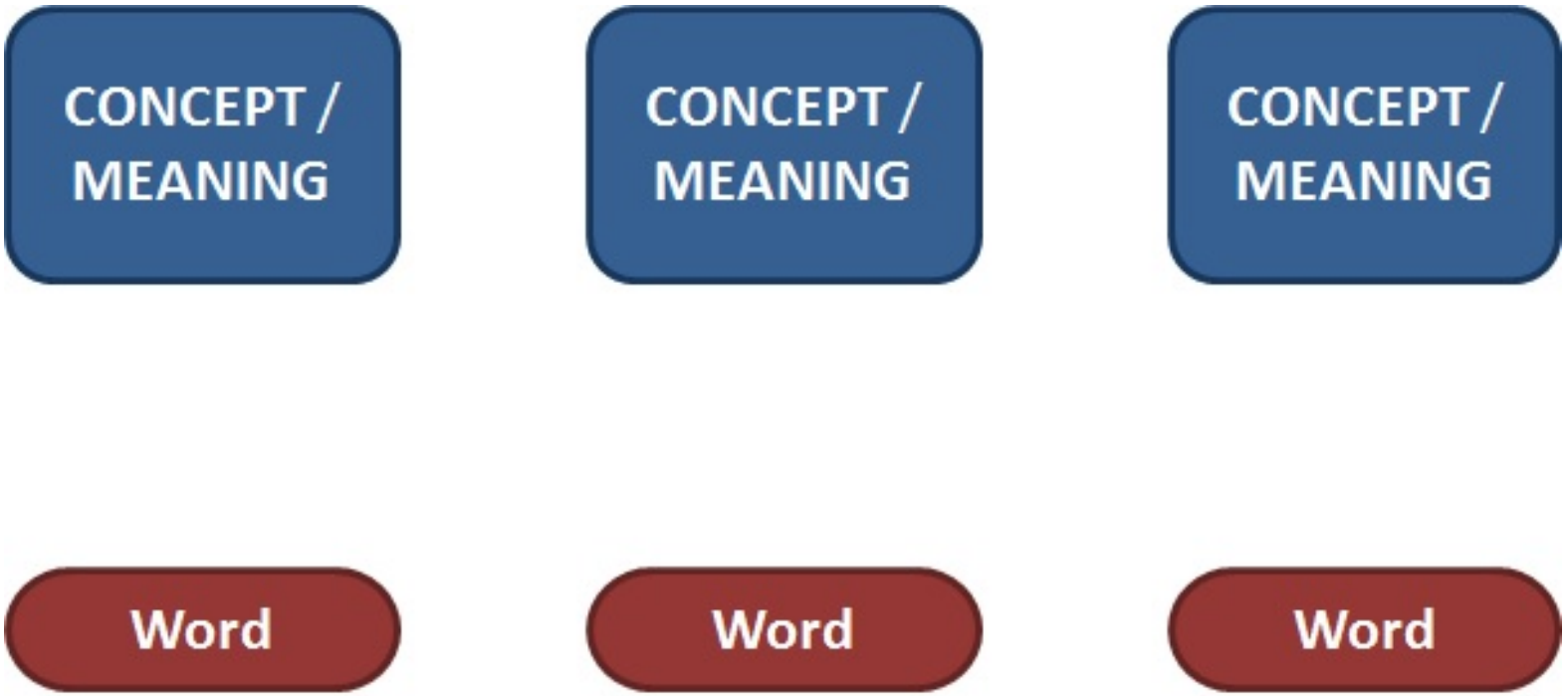
Previous studies within Cognitive Sociolinguistics on political and ideological discourse:

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

Overview

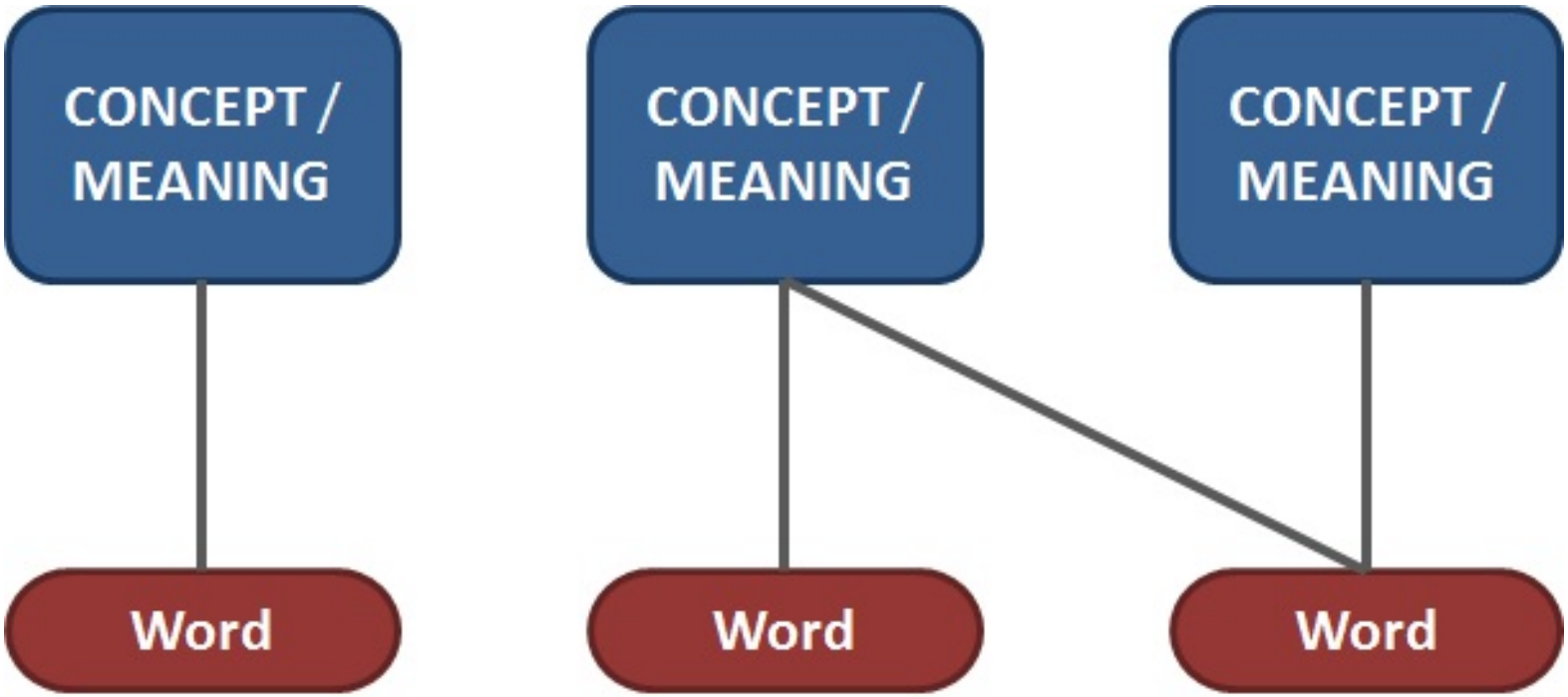
3. Analysing Lexical semantics in CSL

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):



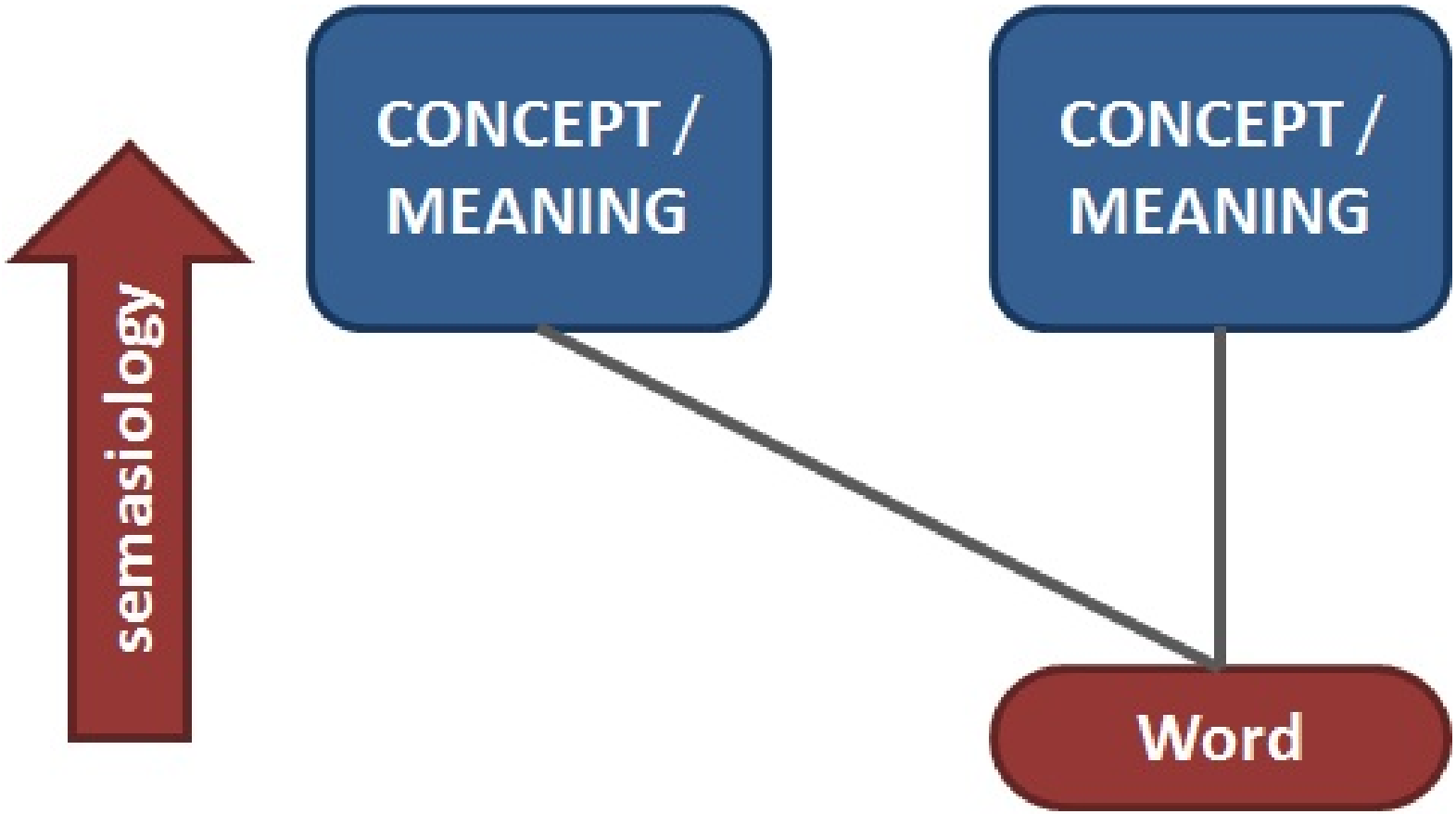
3. Analysing Lexical semantics in CSL

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):



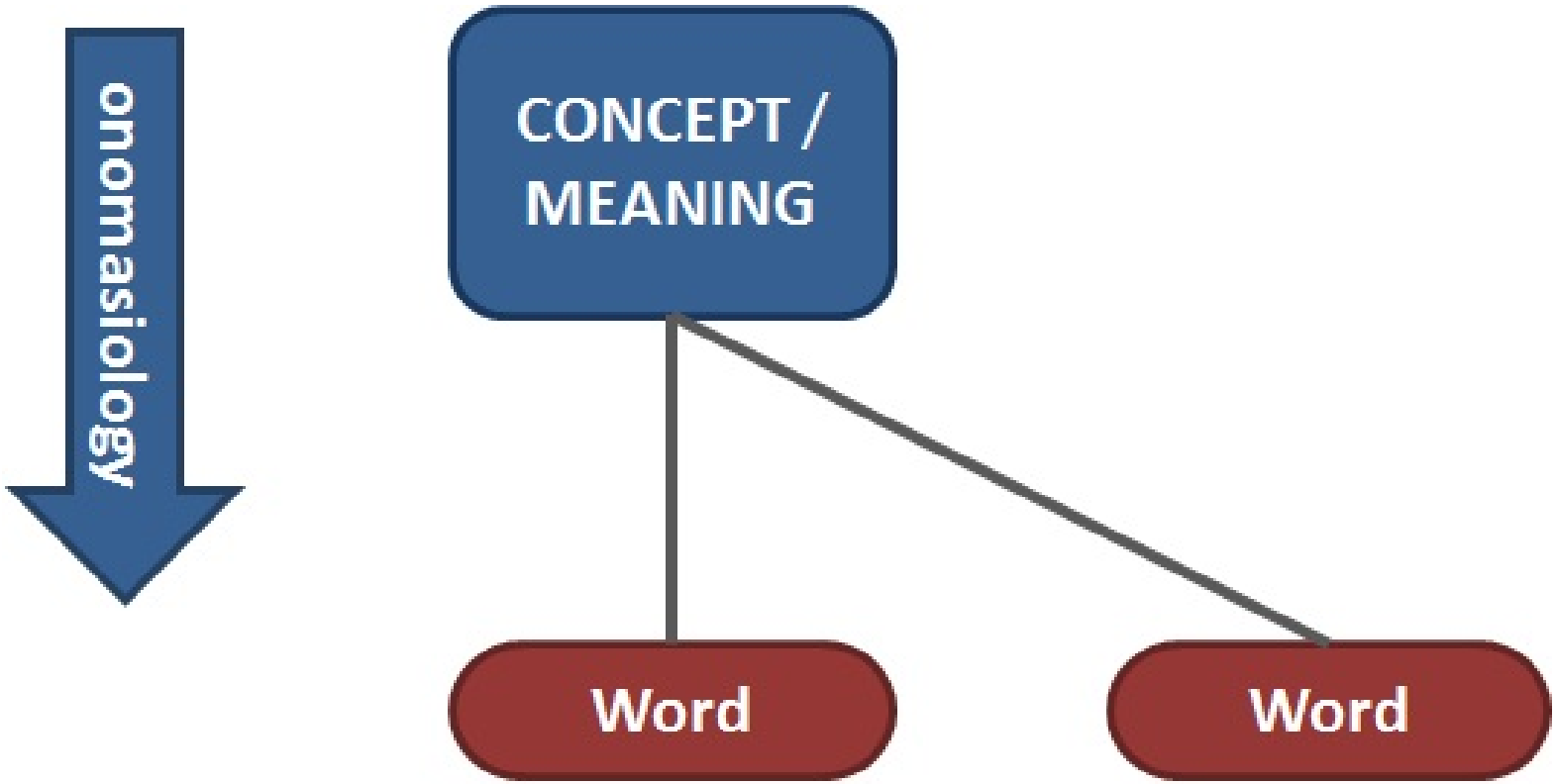
3. Analysing Lexical semantics in CSL

SEMASIOLOGY:



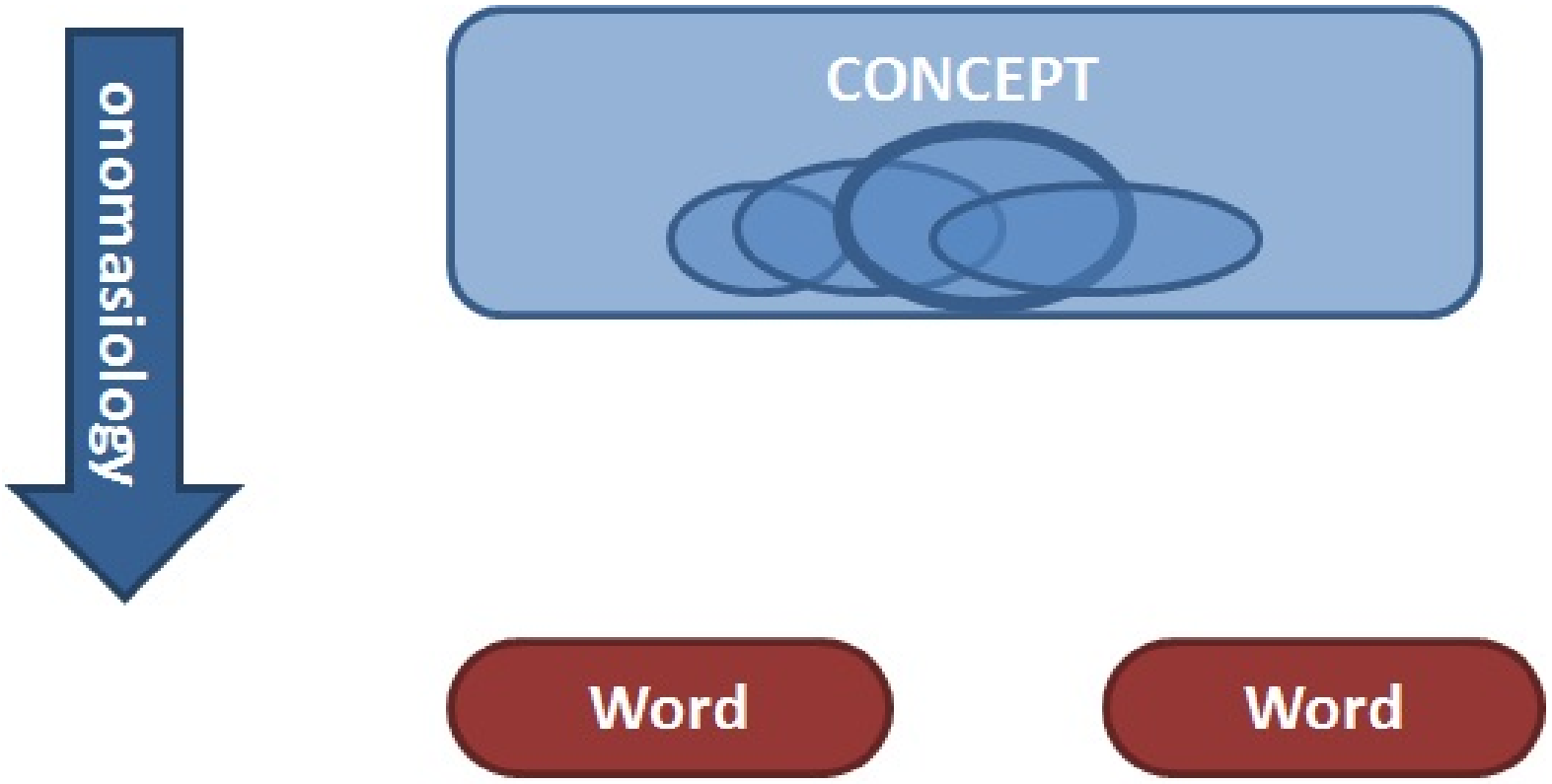
3. Analysing Lexical semantics in CSL

ONOMASIOLOGY:



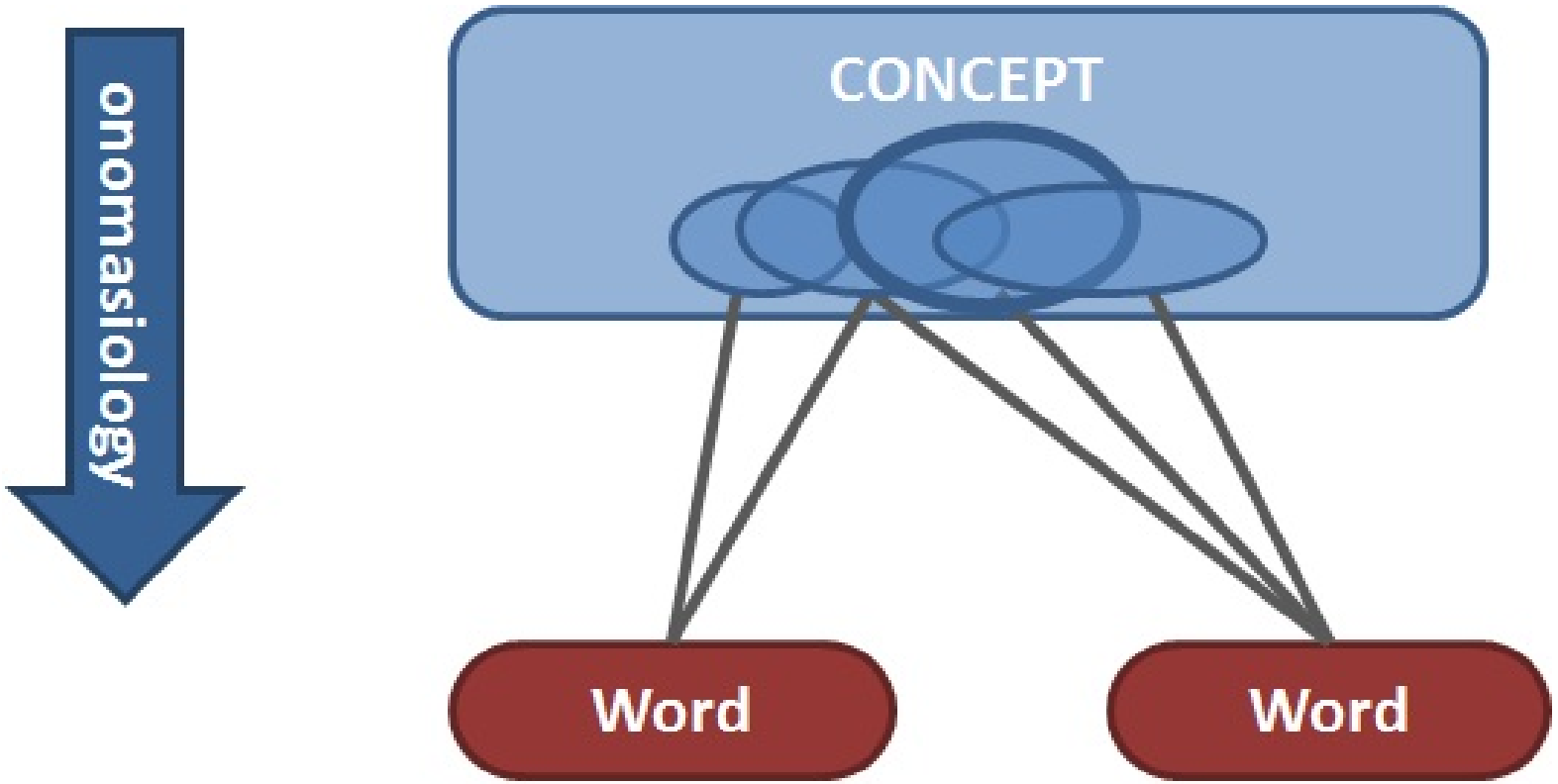
3. Analysing Lexical semantics in CSL

PROTOTYPE STRUCTURE:



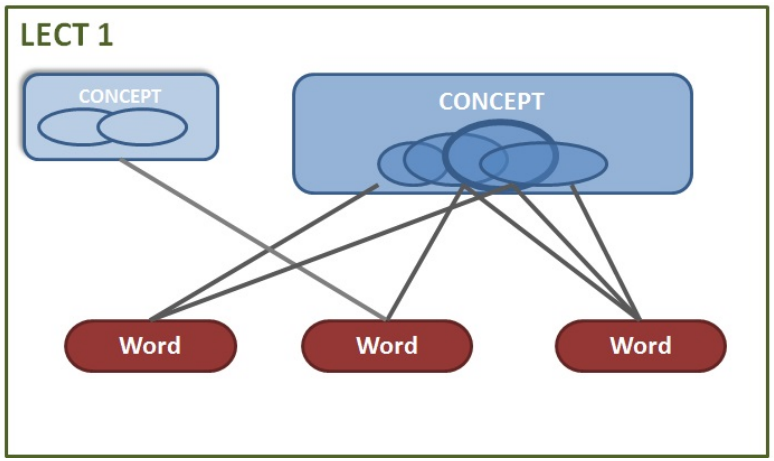
3. Analysing Lexical semantics in CSL

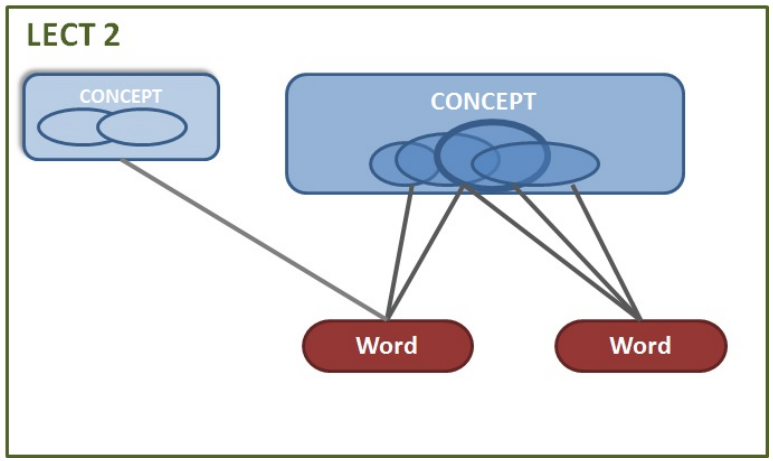
PROTOTYPE STRUCTURE:



3. Analysing Lexical semantics in CSL

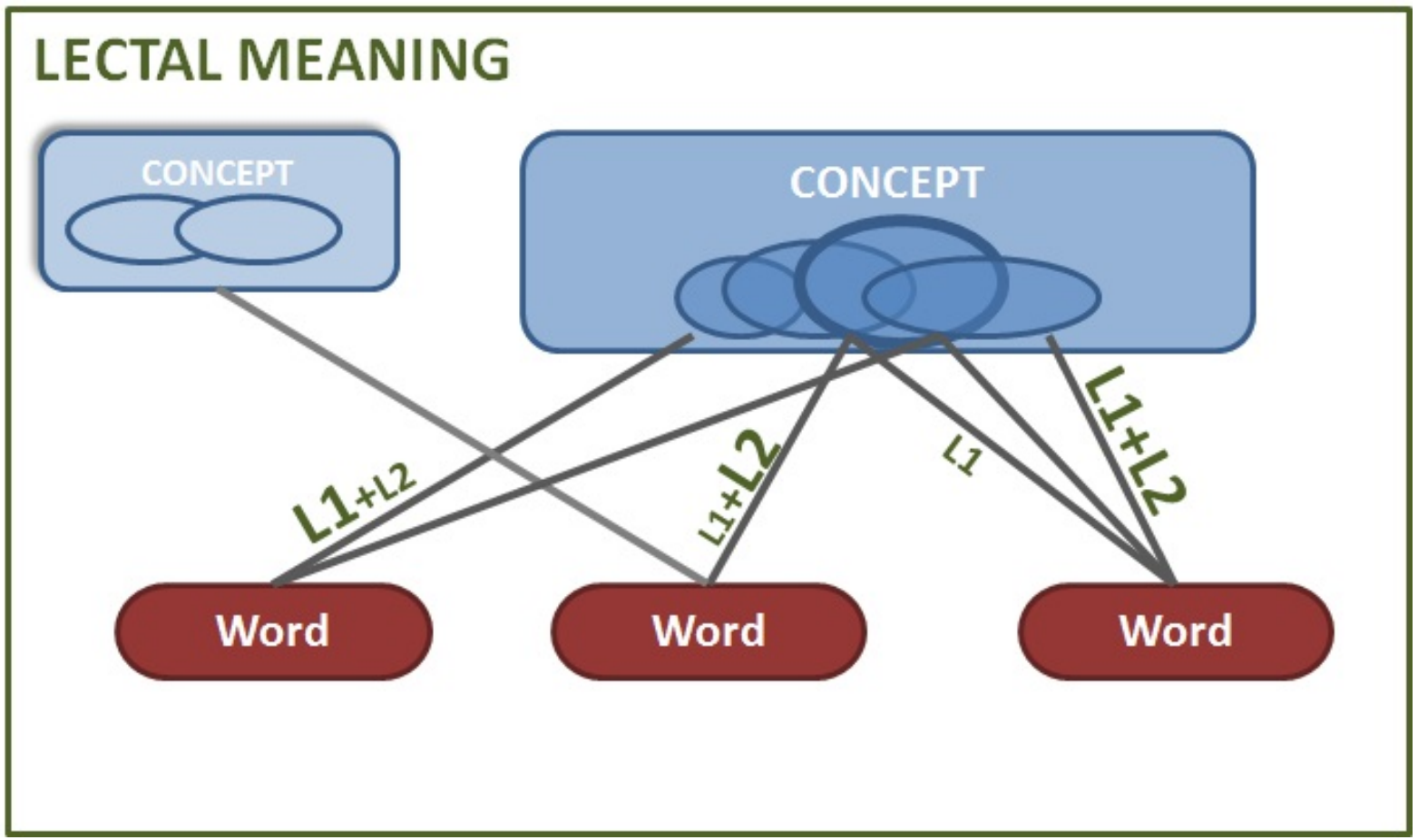
LECTAL VARIATION:





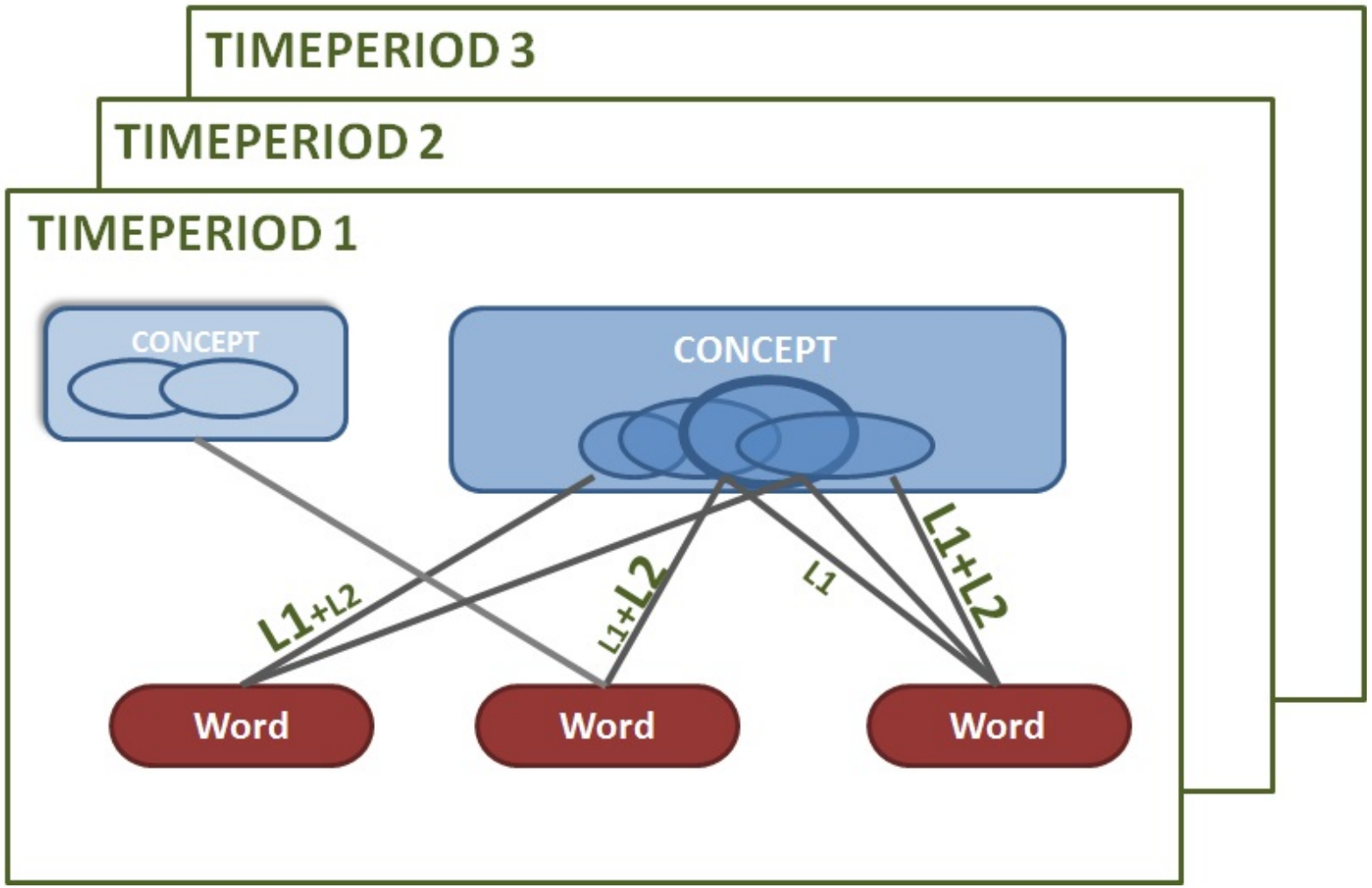
3. Analysing Lexical semantics in CSL

LECTAL MEANING:



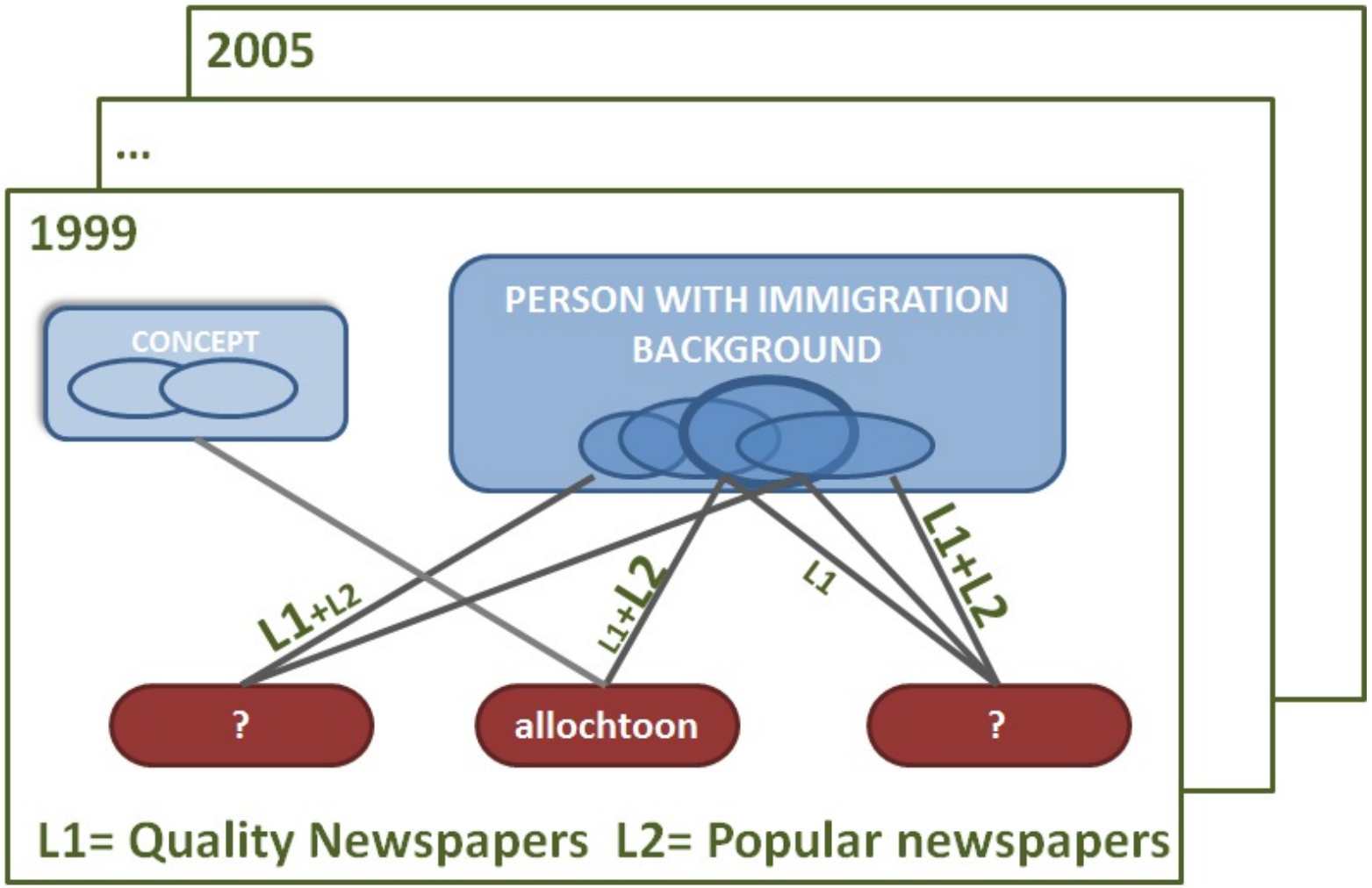
3. Analysing Lexical semantics in CSL

DIACHRONIC VARIATION:



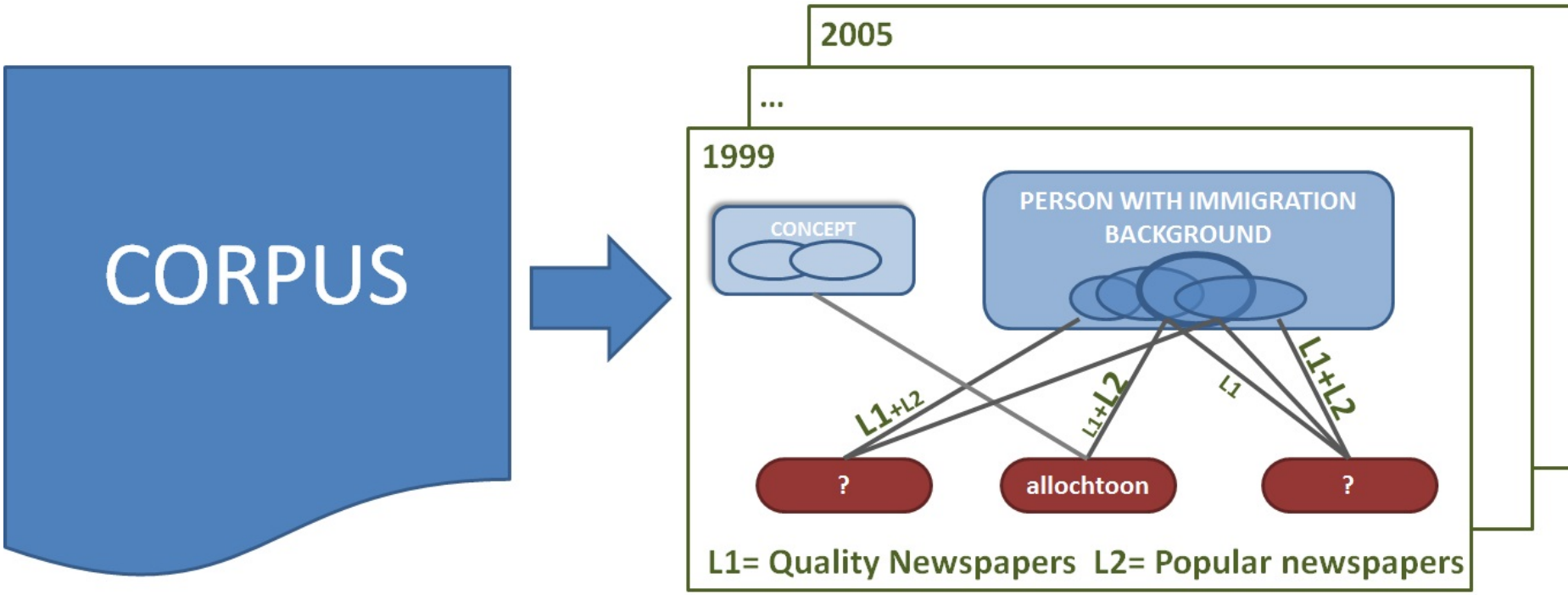
3. Analysing Lexical semantics in CSL

PERSON WITH IMMIGRATION BACKGROUND:



3. Analysing Lexical semantics in CSL

USAGE-BASED STUDY:



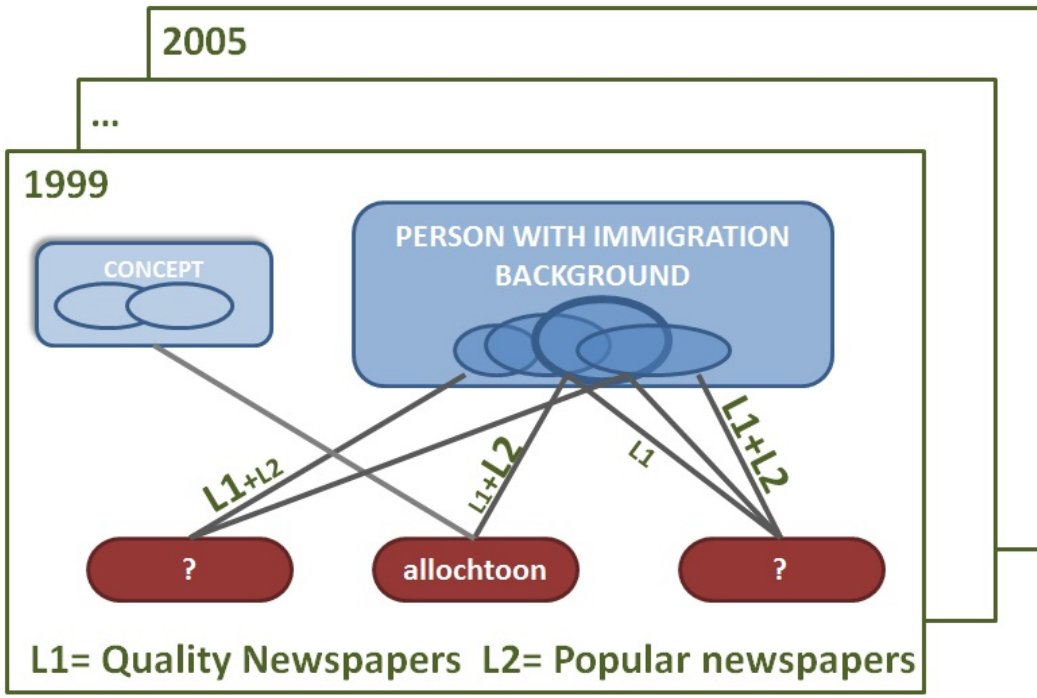
3. Analysing Lexical semantics in CSL

HOW TO FIND PATTERNS IN LARGE AMOUNTS OF DATA?

CORPUS

- Belgian Dutch **1.3 billion** words
- 1999-2005
- 3 quality newspapers (Standaard, De Morgen, De Tijd)
- 3 popular newspapers (Laatste Nieuws, Nieuwsblad, Belang v Limburg)
- 22,306 occ. of *allochtoon*

BIG DATA



Overview

4. 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

4. Semantic Vector Spaces

Collect co-occurrence frequencies for a large part of the vocabulary and p ut 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

4. Semantic Vector Spaces

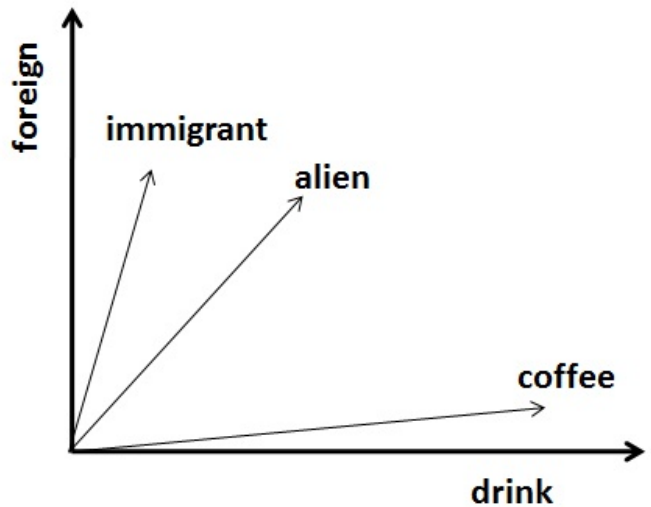
weight the raw frequencies by collocational strength (pmi)

	work	*foreign*	*citizenship*	*laws*	*space*	*sugar*	*milk*	*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

4. Semantic Vector Spaces

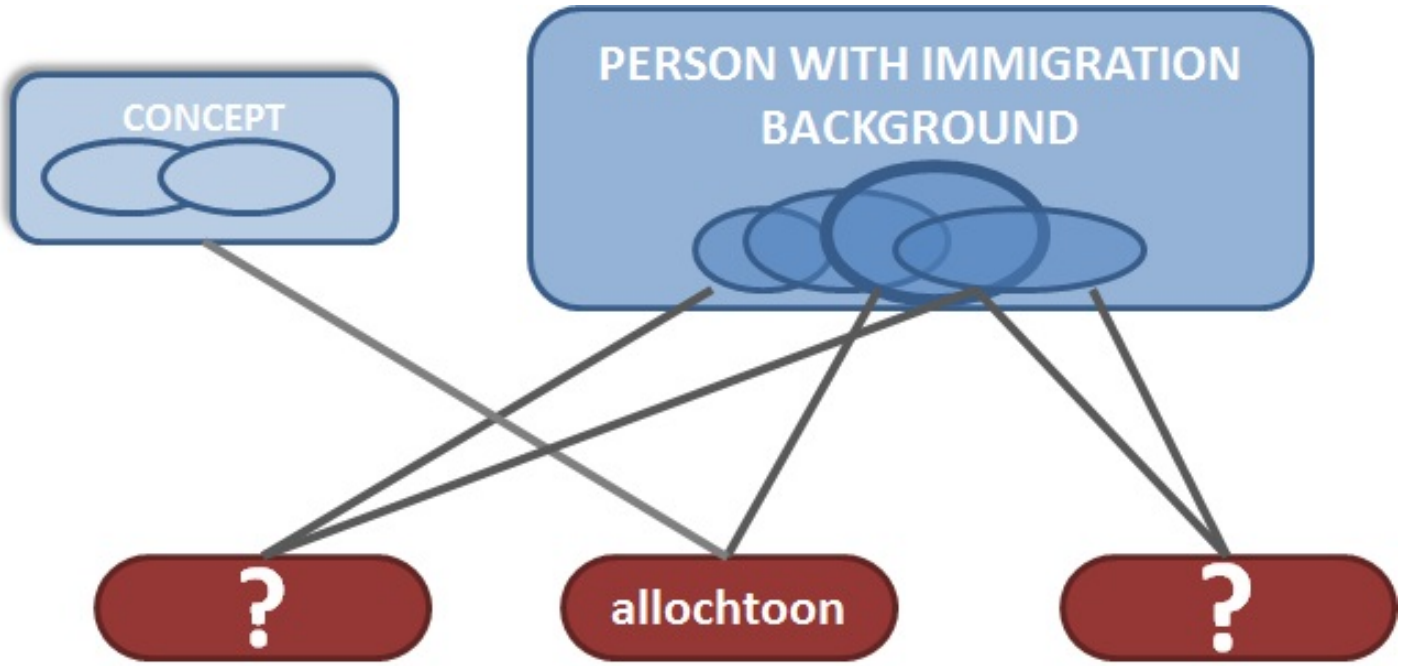
calculate word by word similarity matrix

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



Overview

5. Identifying alternative expressions



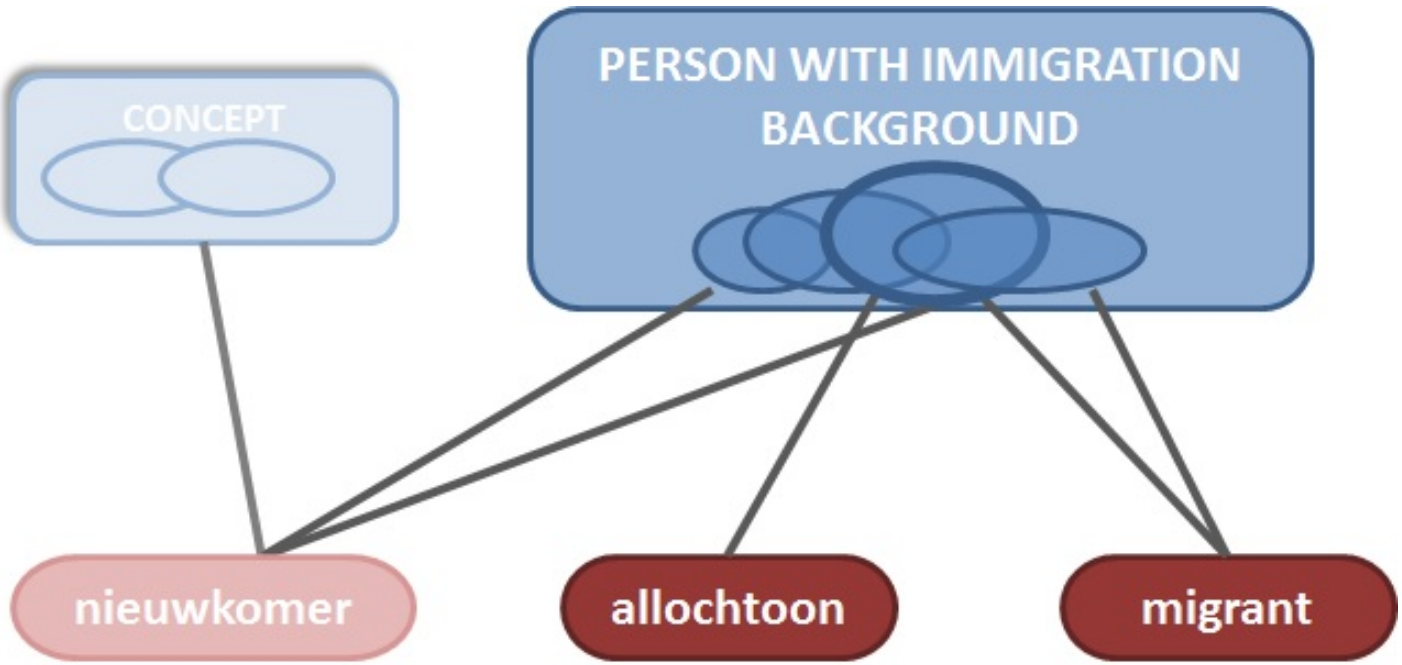
- calculate contextual similarity between 10K Dutch nouns
- sort by similarity to *allochtoon*

5. Identifying alternative expressions

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

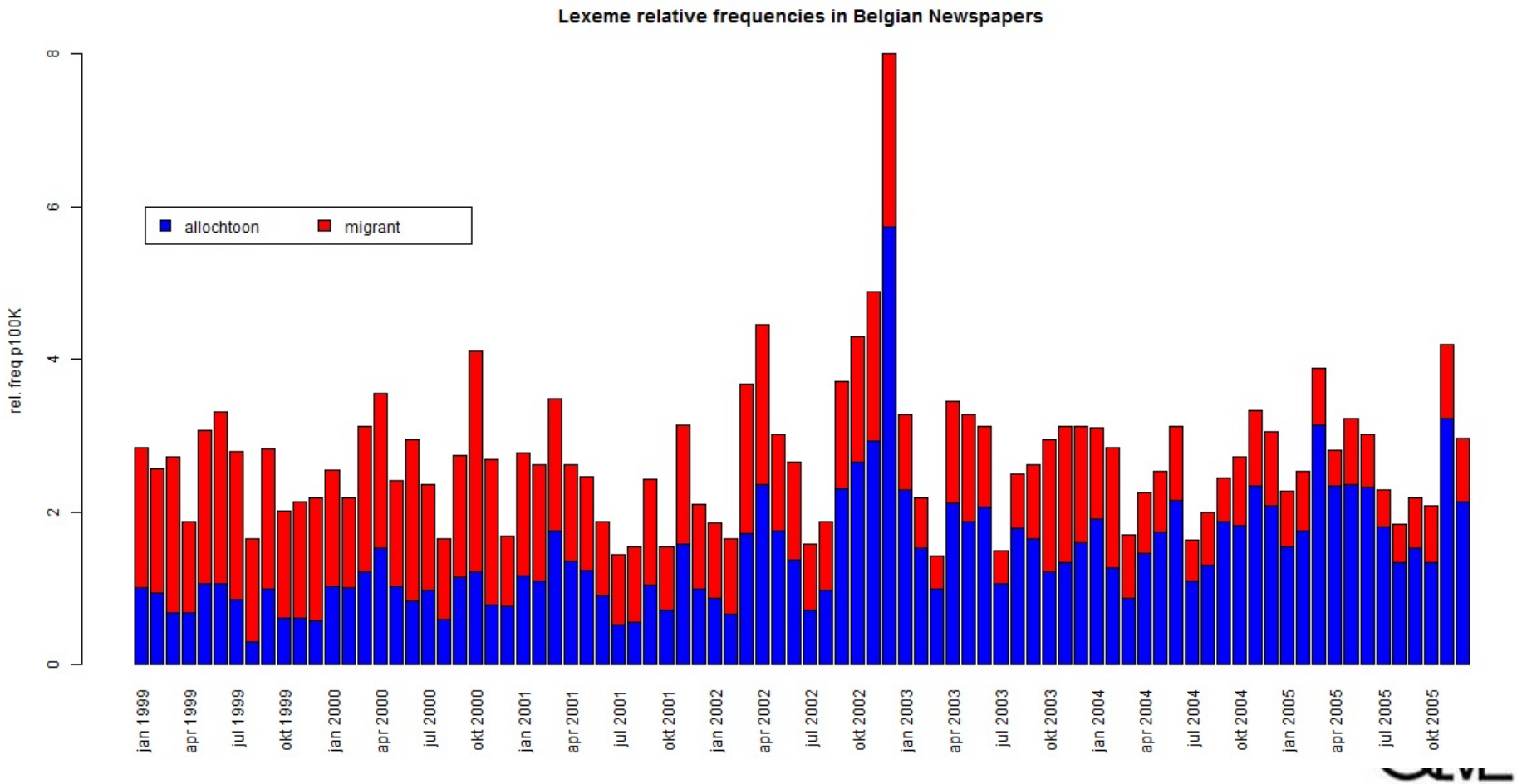
Table alternatives to *allochtoon*

5. Identifying alternative expressions



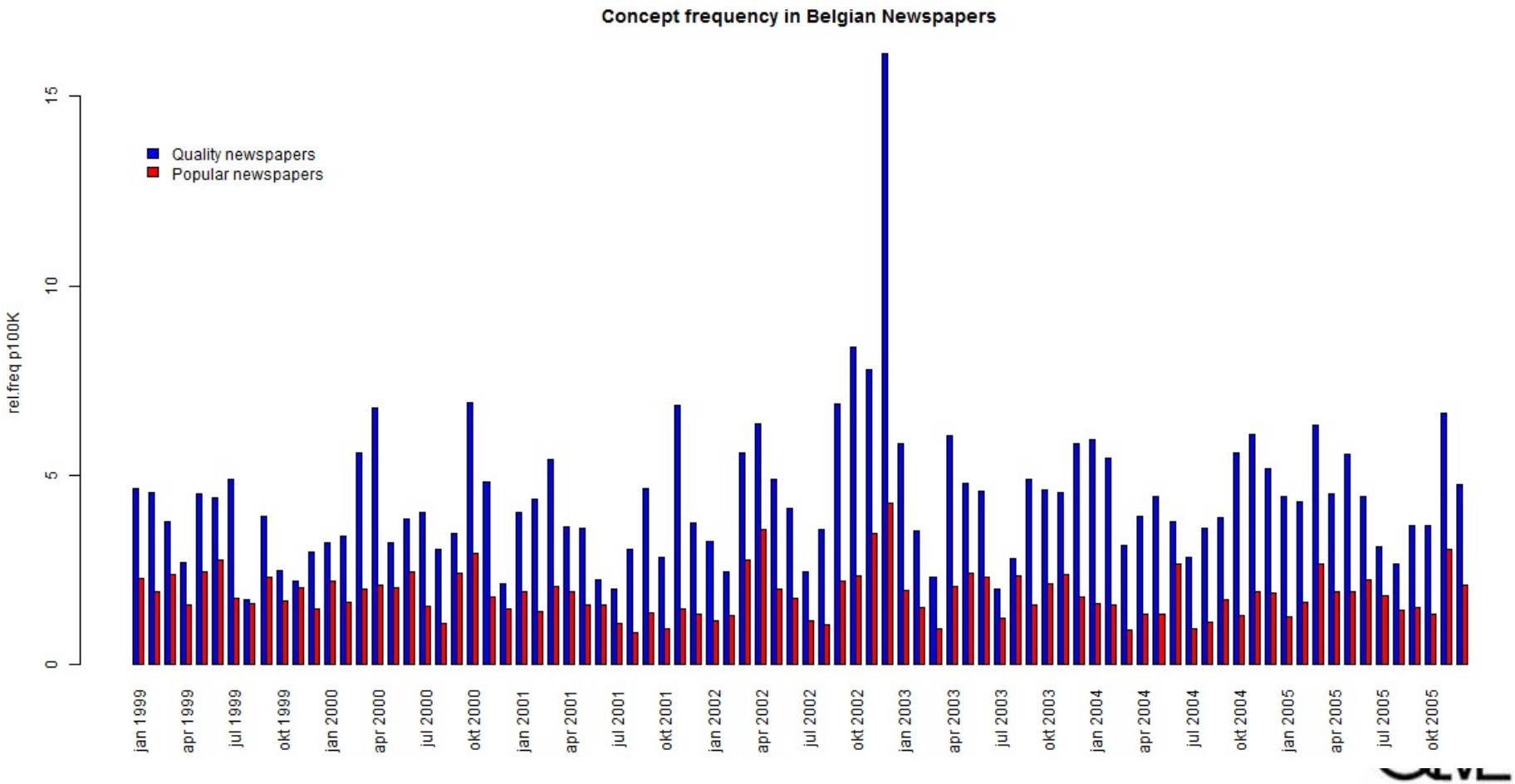
5. Identifying alternative expressions

Concept frequency in the corpus per month

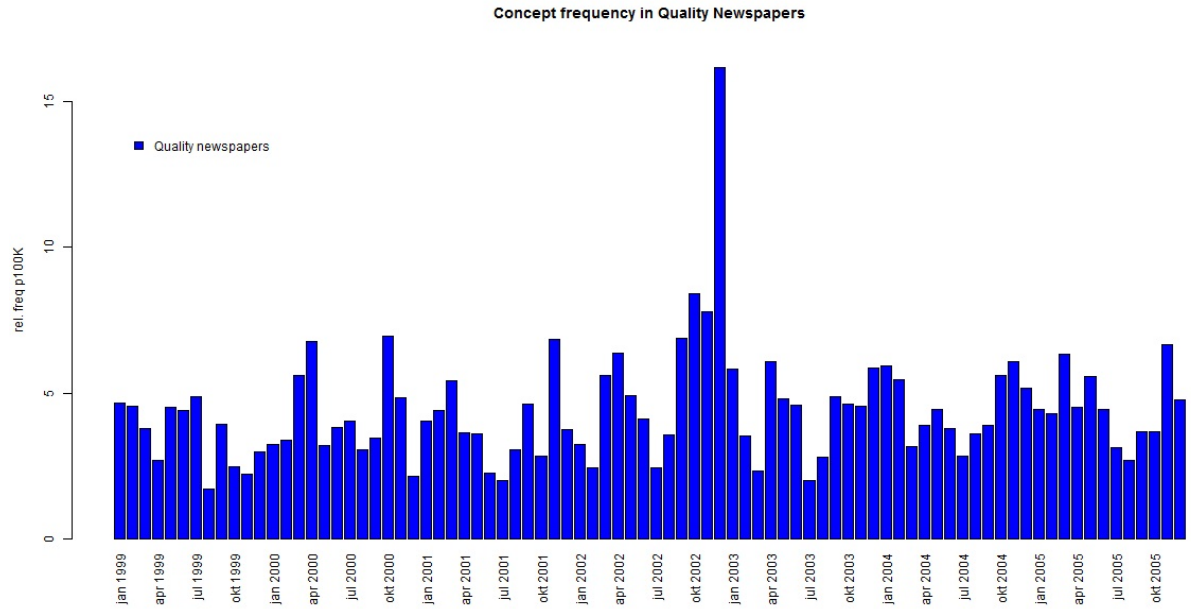


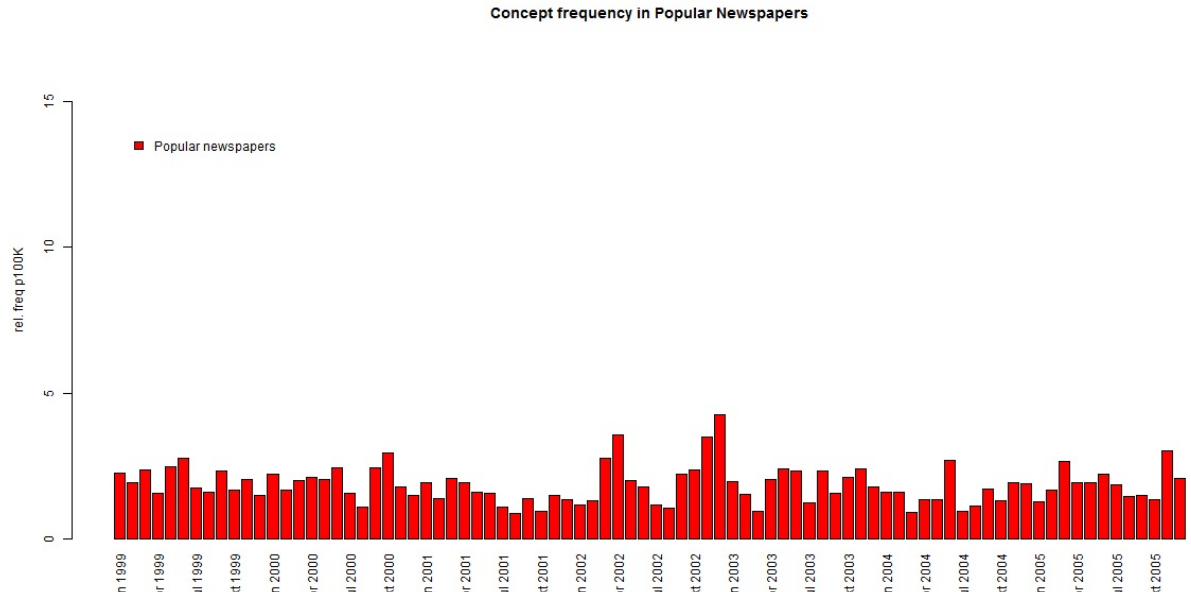
5. Identifying alternative expressions

Concept frequency in the corpus per month per newspaper type



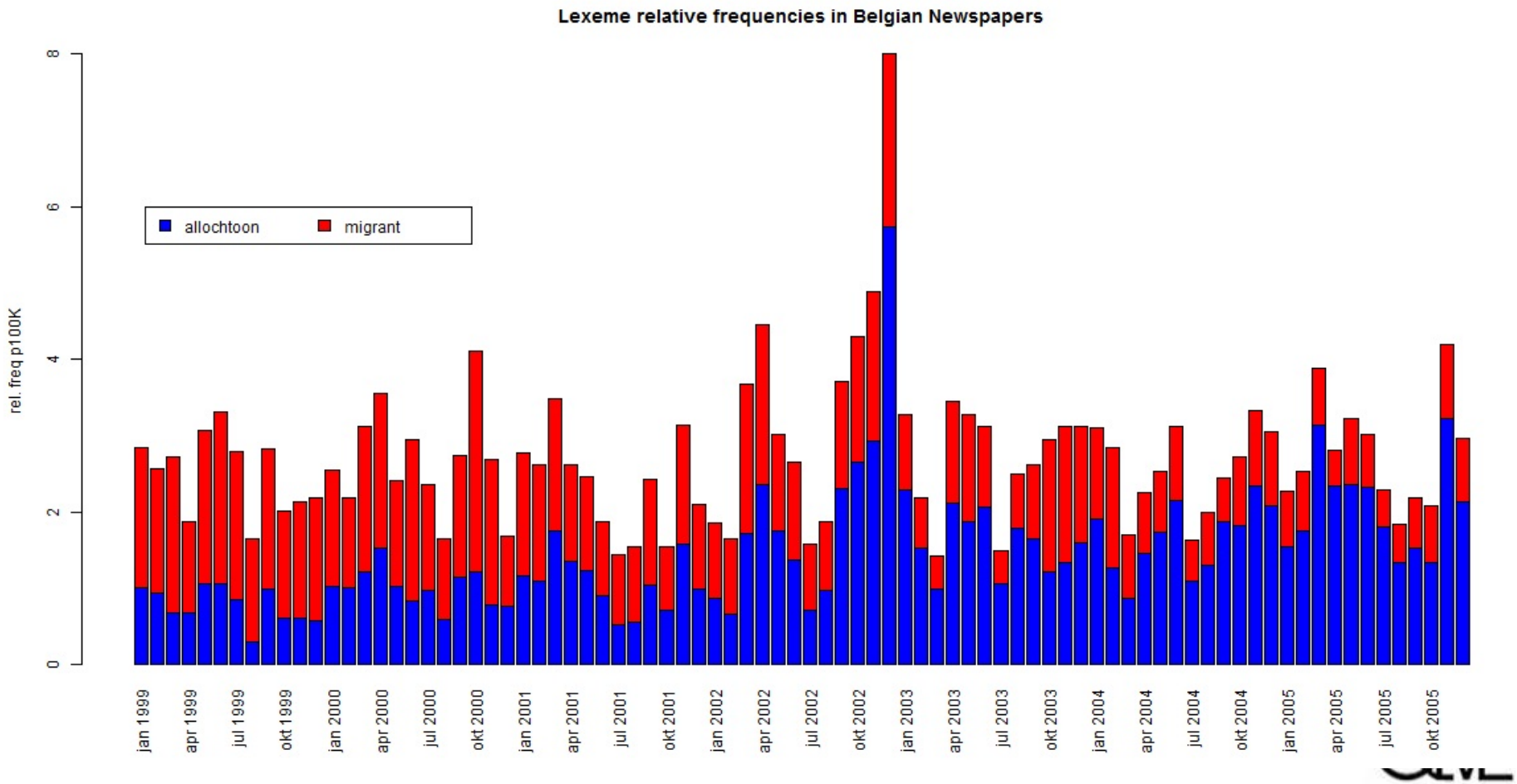
5. Identifying alternative expressions





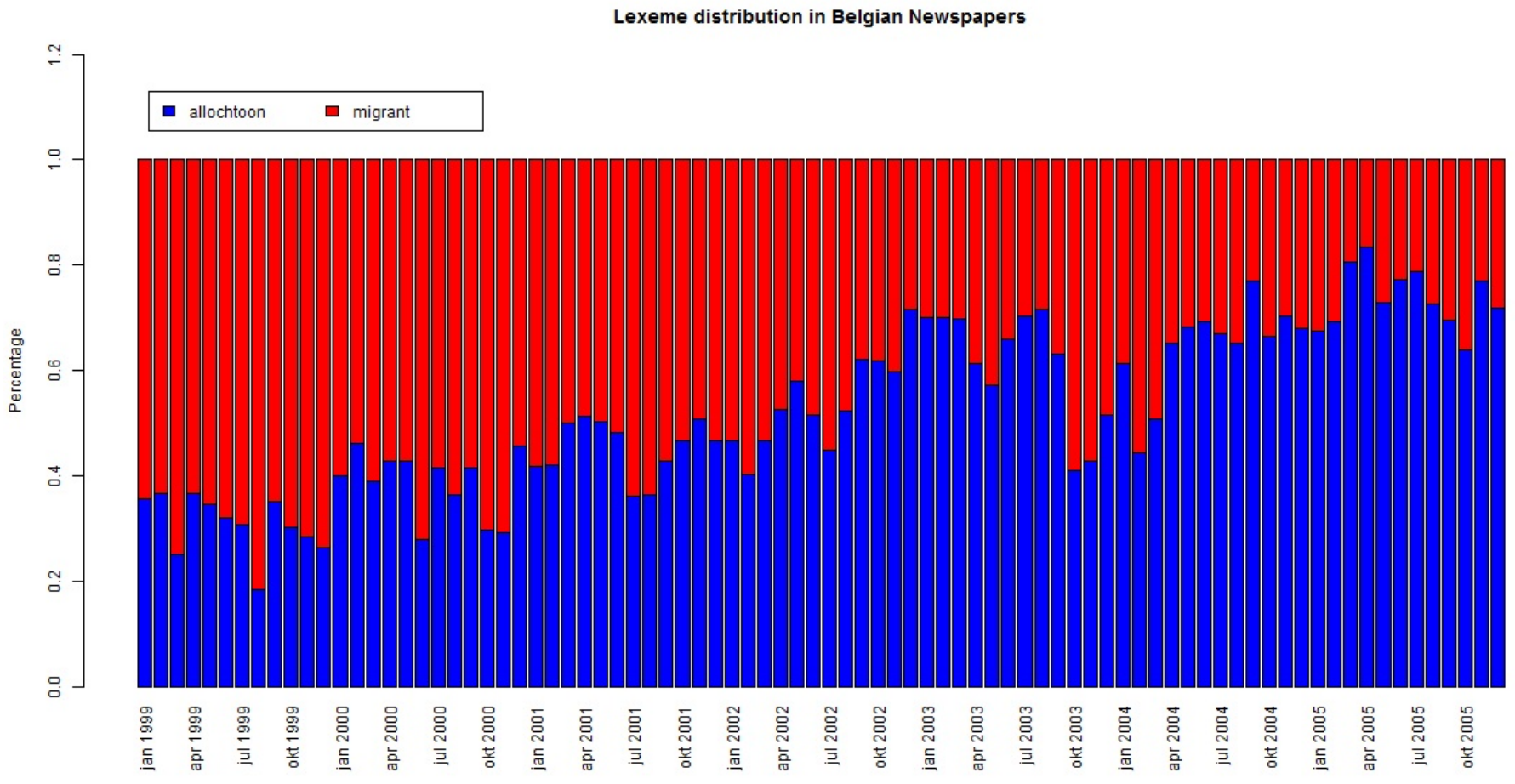
5. Identifying alternative expressions

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

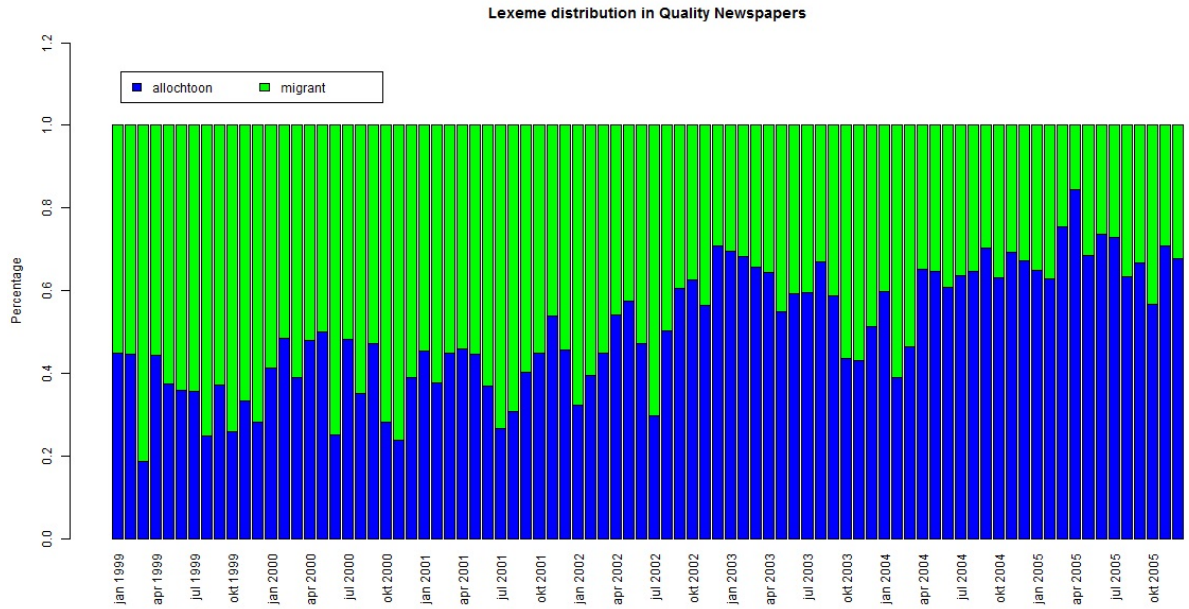


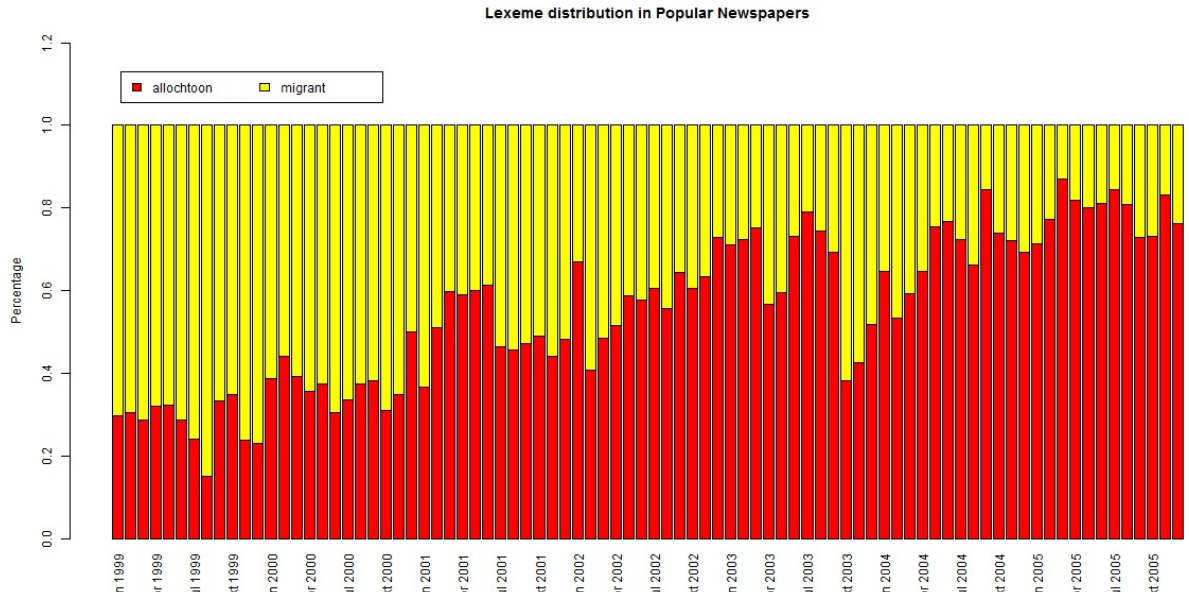
5. Identifying alternative expressions

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



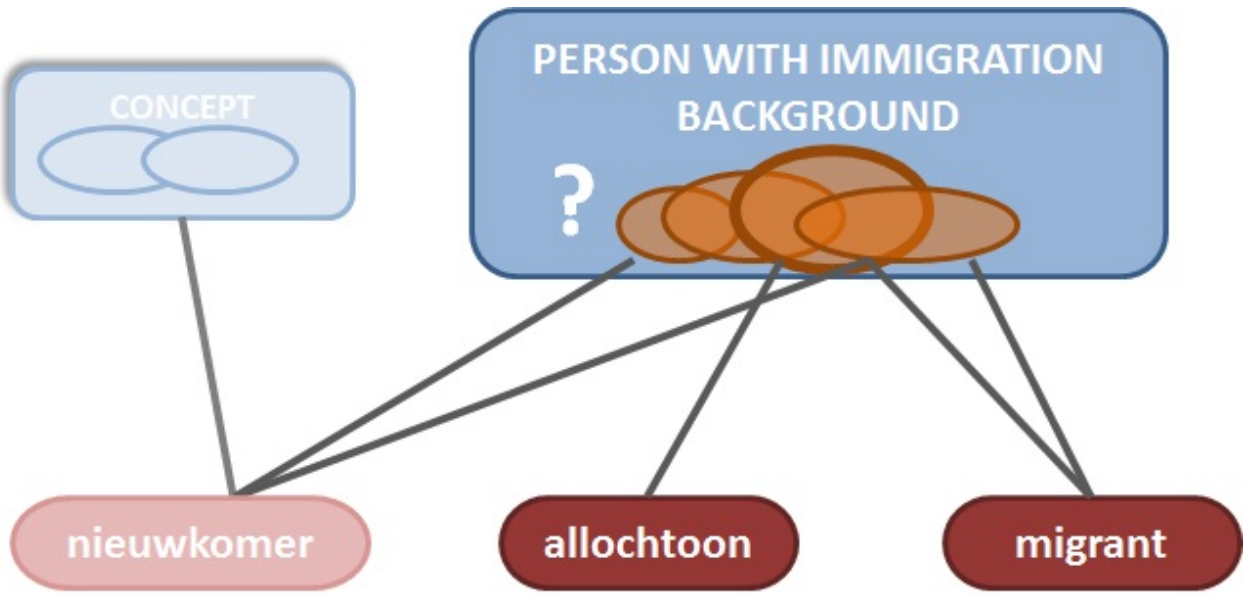
5. Identifying alternative expressions





Overview

6. Identifying and structuring collocations



6. Identifying and structuring collocations

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

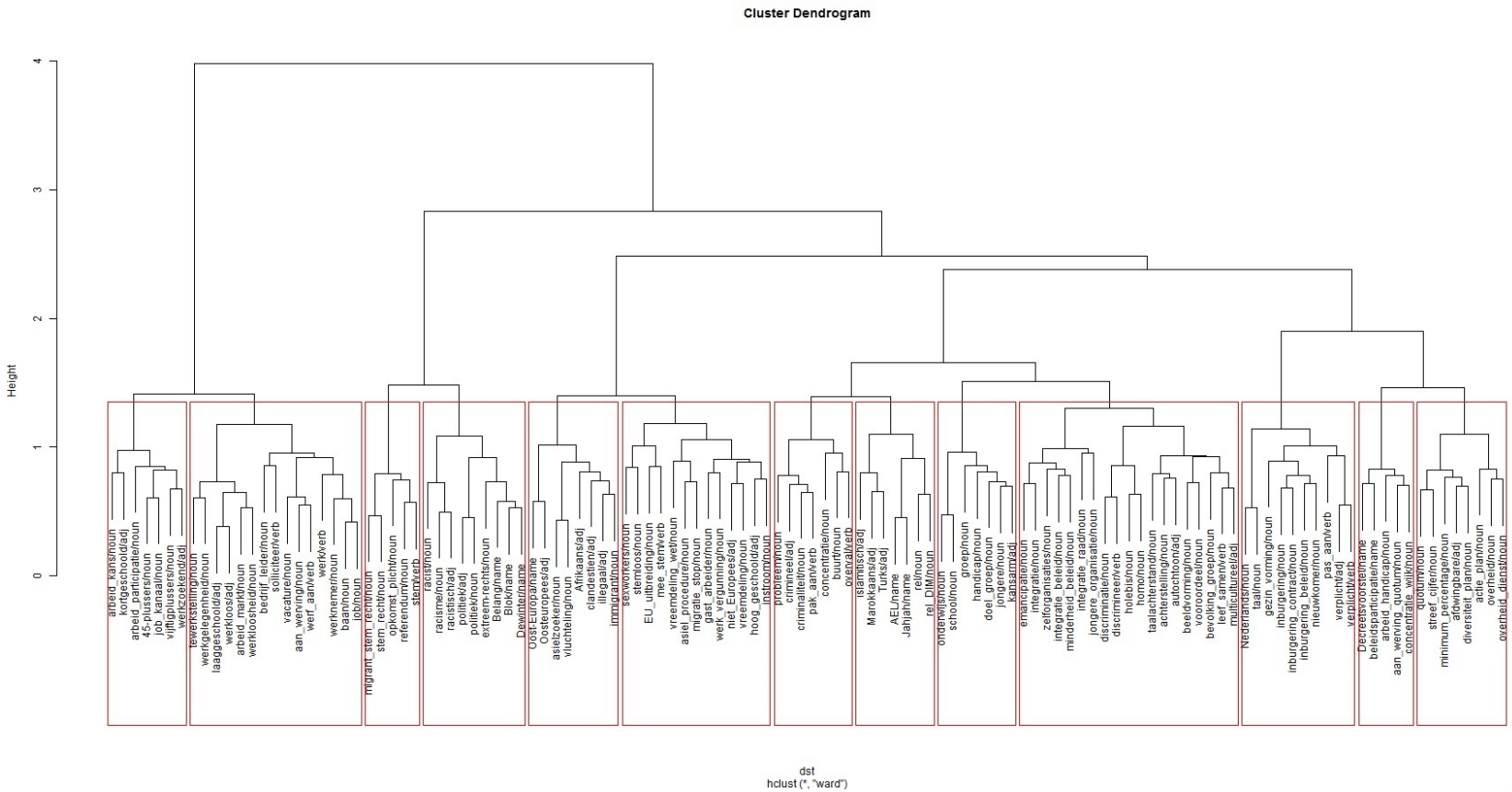
6. Identifying and structuring collocations

Make collocation-profile matrix for collocations

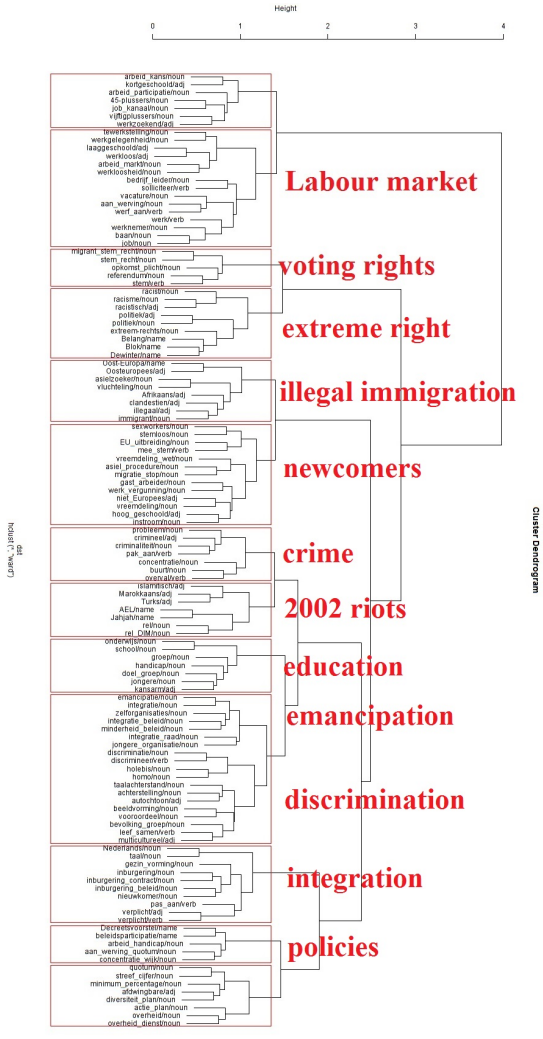
	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*	*hond*
jobs	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
racisme	4.3	8.1	5.7	3.2	6.2	0.5	0.0	0.1
integratie	5.3	7.9	6.5	6.0	0.8	0.6	0.1	0.0
misdaad	4.3	8.1	5.7	2.2	6.2	0.4	0.0	0.1
stemrecht	5.3	7.9	6.5	8.0	0.8	0.9	0.3	0.0

6. Identifying and structuring collocations

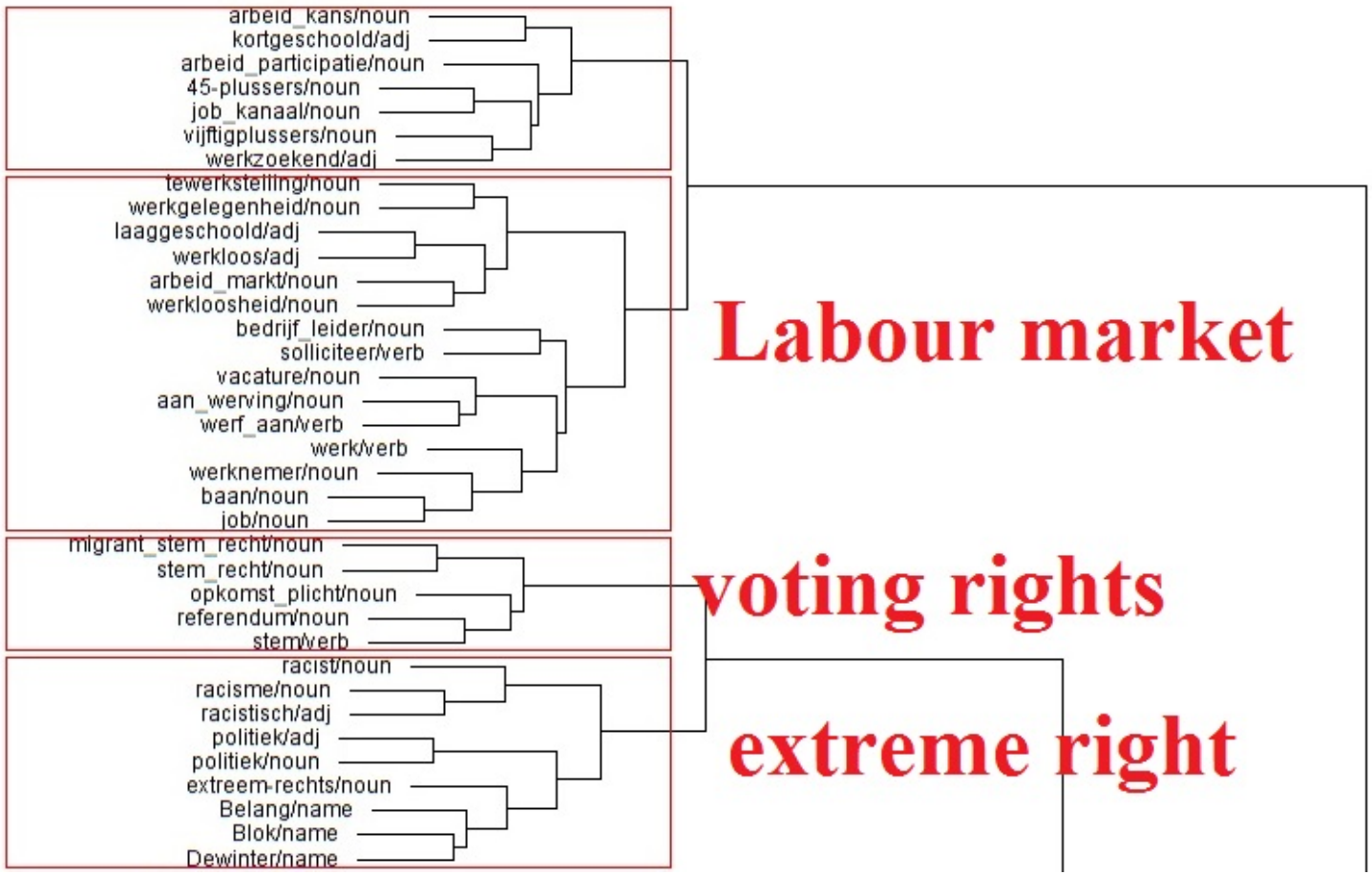
Calculate similarity between collocations and feed to a cluster analysis



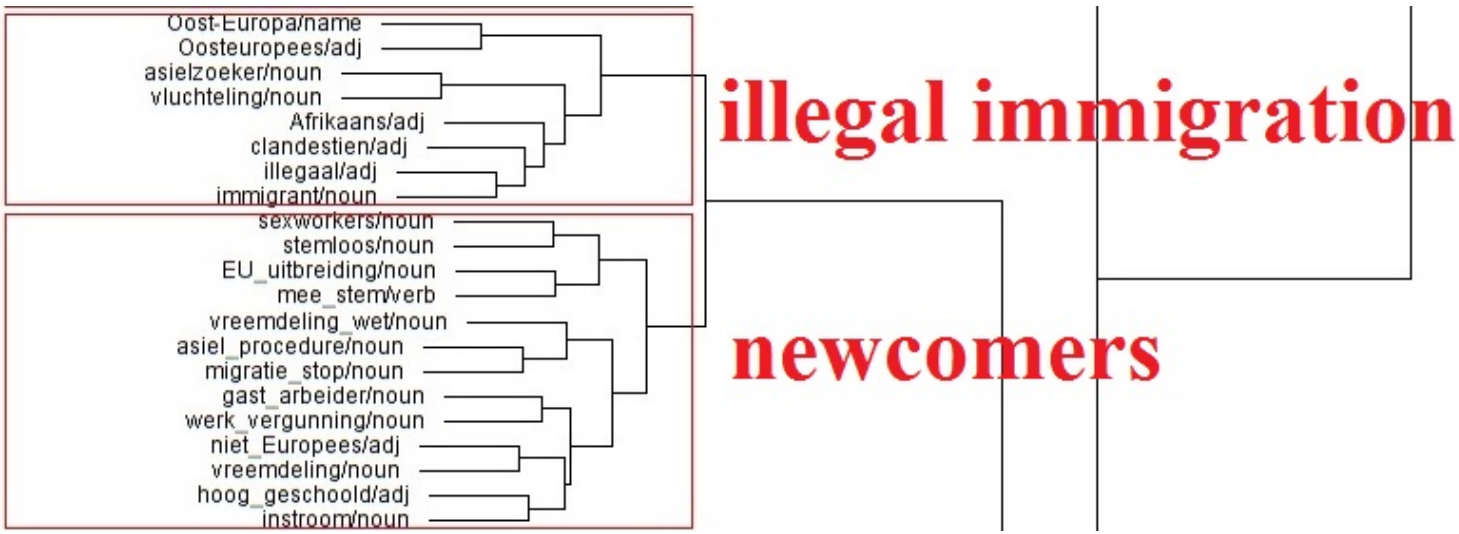
6. Identifying and structuring collocations



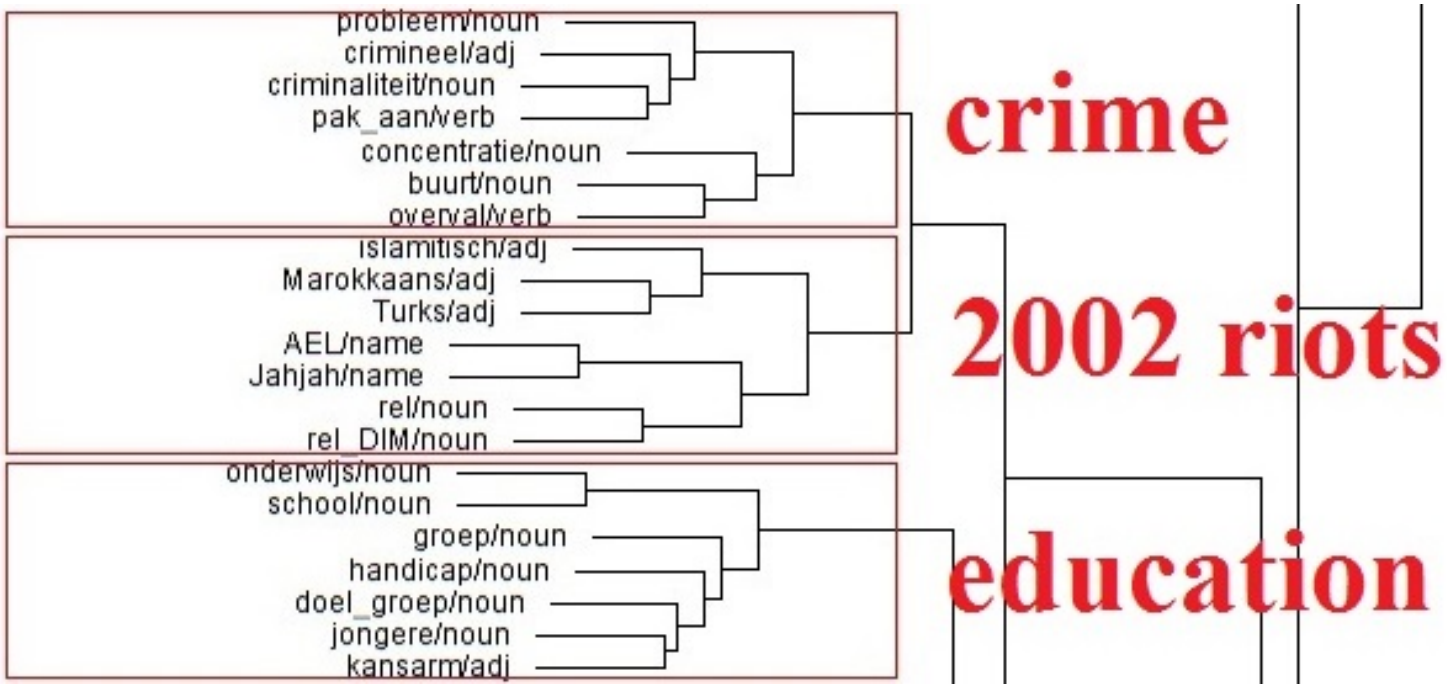
6. Identifying and structuring collocations



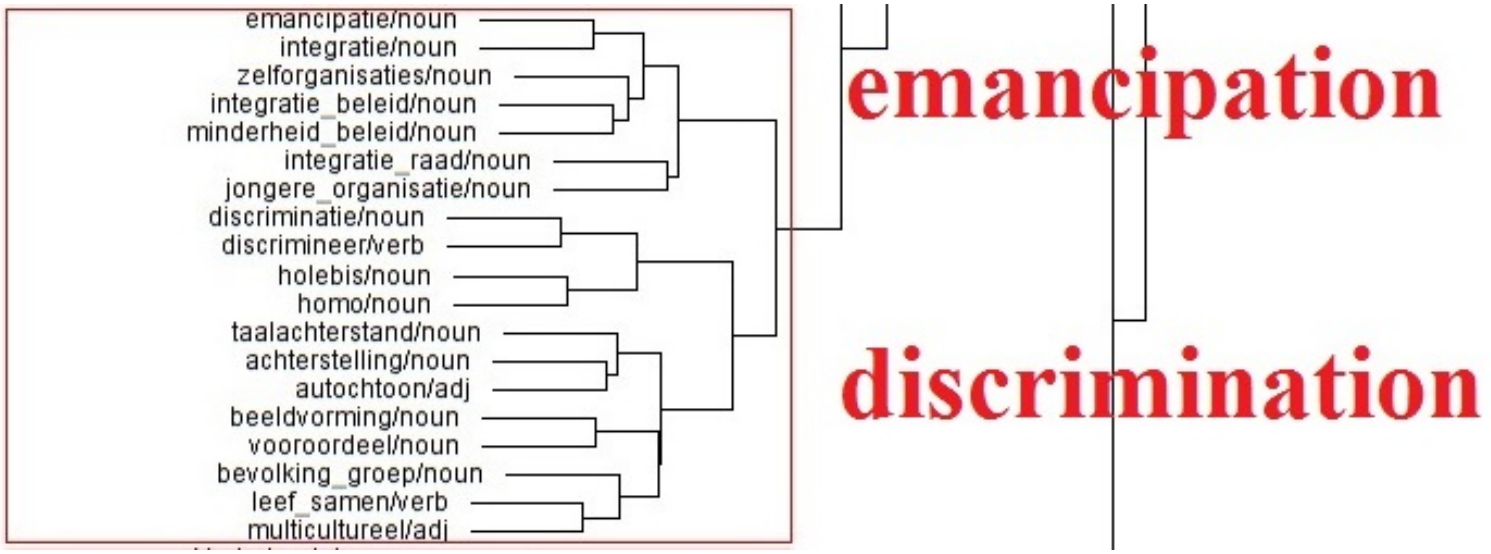
6. Identifying and structuring collocations



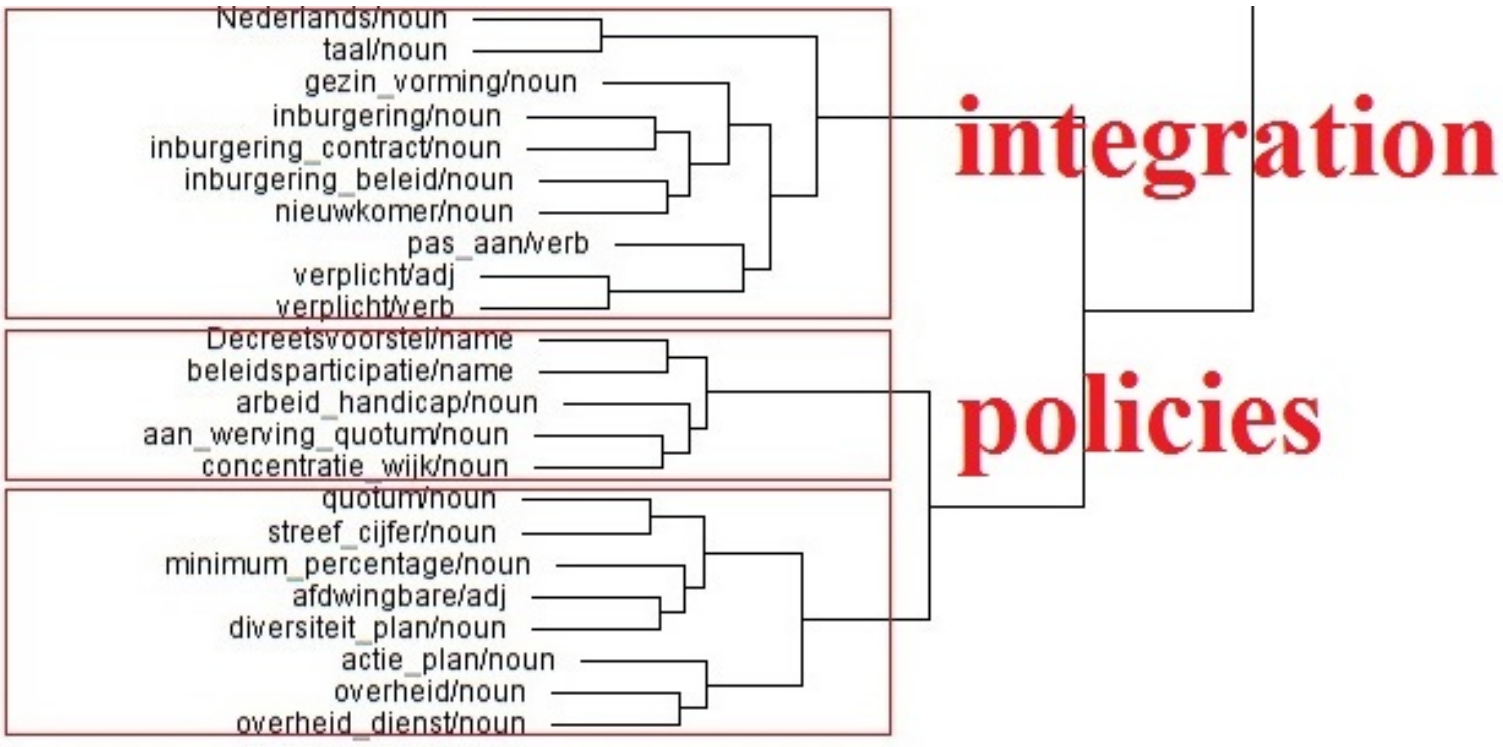
6. Identifying and structuring collocations



6. Identifying and structuring collocations

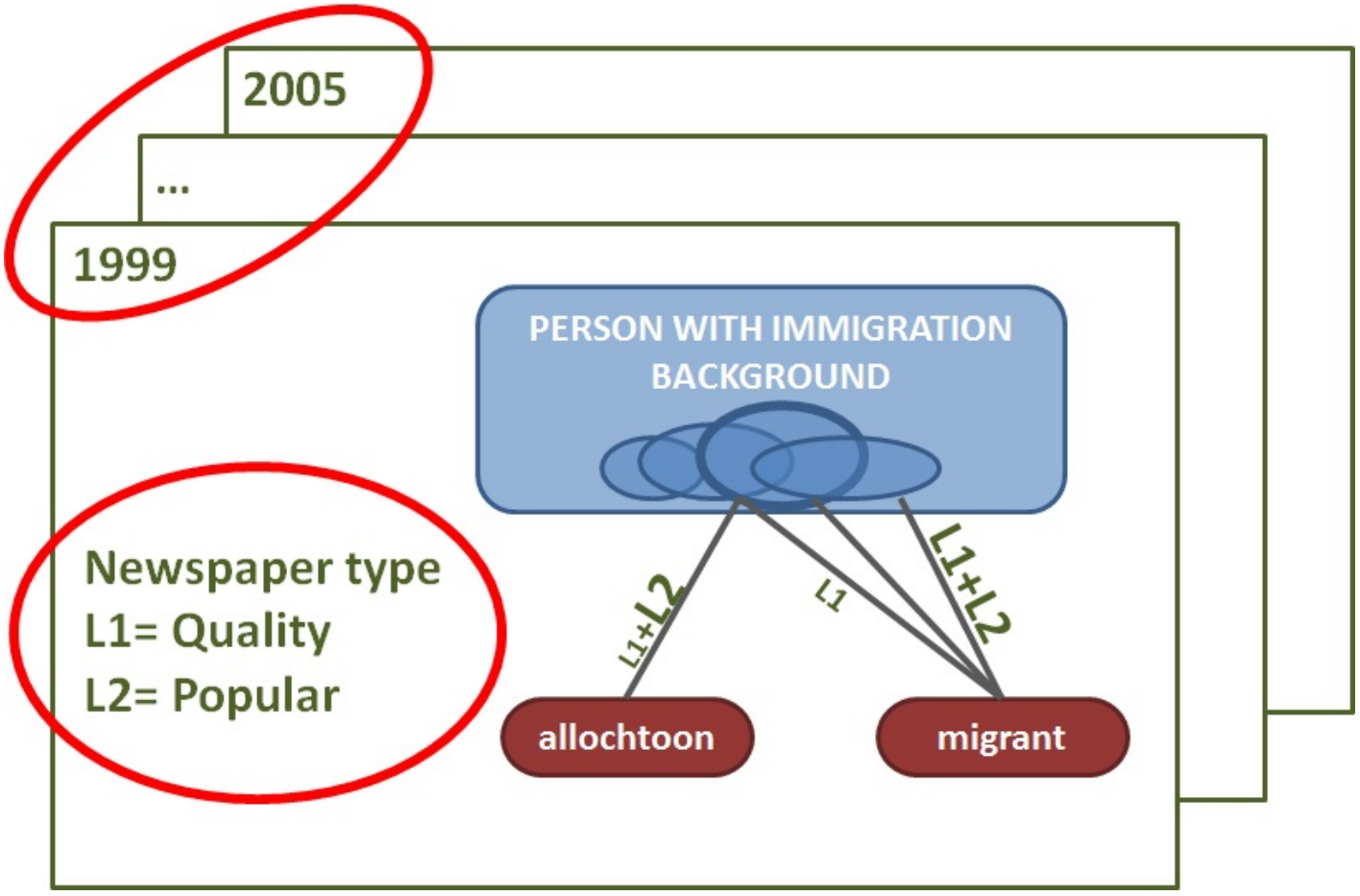


6. Identifying and structuring collocations



Overview

7. Measuring lectal and diachronic distances



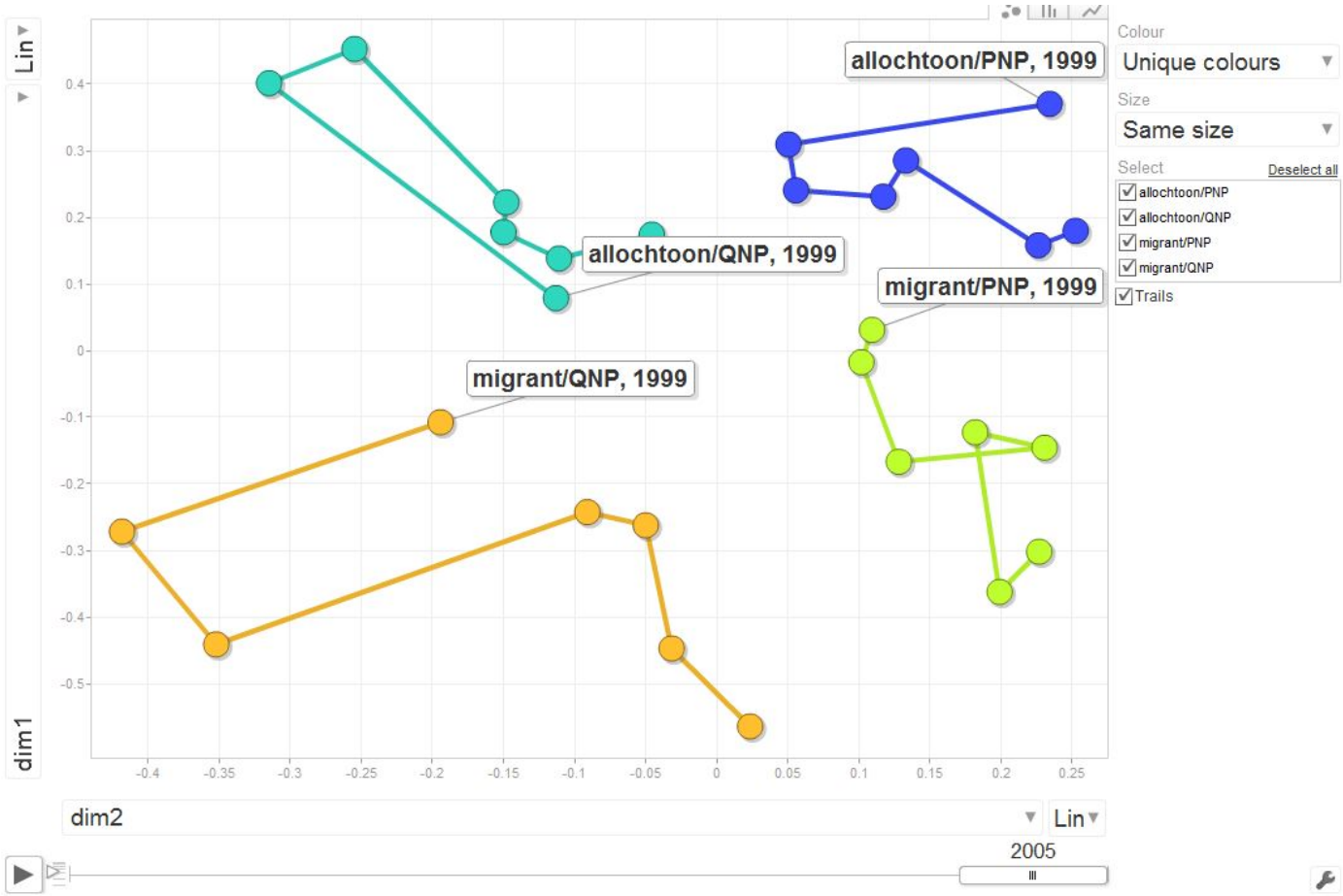
7. Measuring lectal and diachronic distances

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

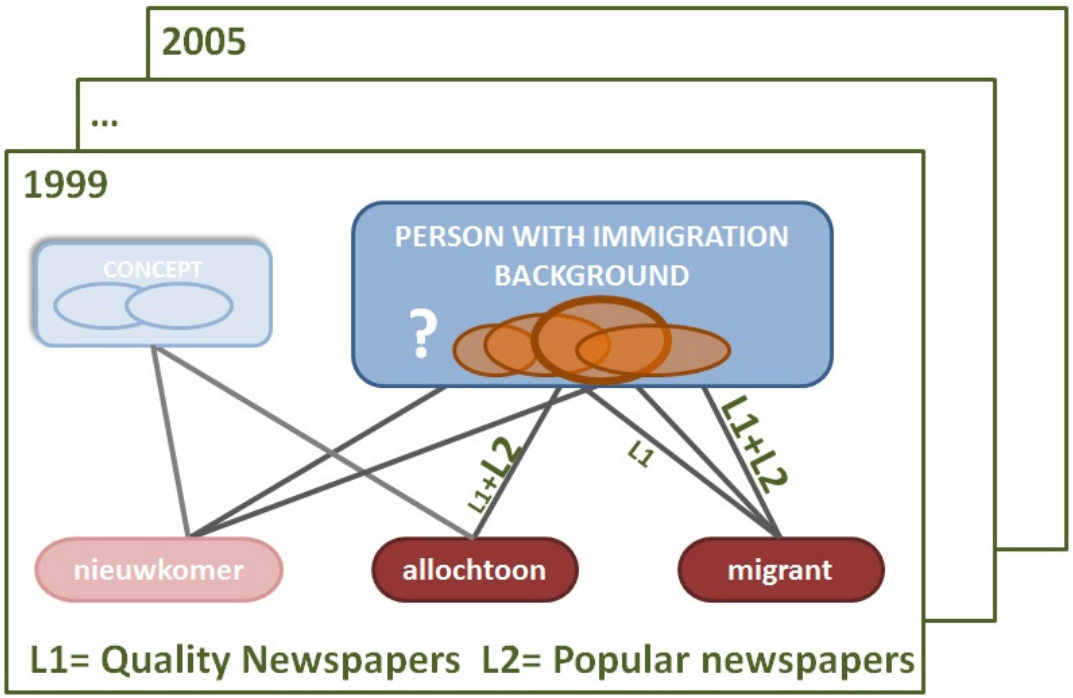
	jobs	*racisme*	*integratie*	*misdaad*	*stemrecht*	*suiker*	*zon*
allochtoon/1999pop	5.3	7.9	6.5	4.0	0.8	0.6	0.0
migrant/1999pop	4.3	8.1	5.7	3.2	6.2	0.5	0.0
allochtoon/1999qual	4.3	2.9	7.5	8.1	0.3	1.6	0.3
migrant/1999qual	4.3	4.2	5.7	3.2	6.2	0.5	0.0
allochtoon/2000pop	5.8	3.5	6.5	5.1	1.3	0.0	0.1
migrant/2000pop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

7. Measuring lectal and diachronic distances

Calculate similarity matrix and use MDS to plot in 2D
Visualise convergence/divergence with Motion Chart



7. Measuring lectal and diachronic distances



6. Identifying and structuring collocations

Make vector per collocation 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

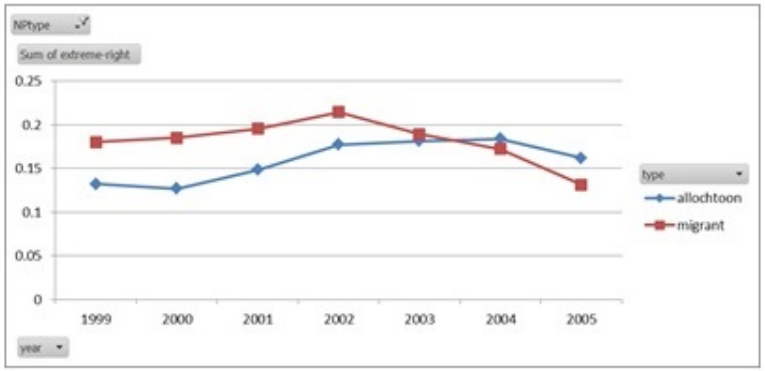
7. Measuring lectal and diachronic distances

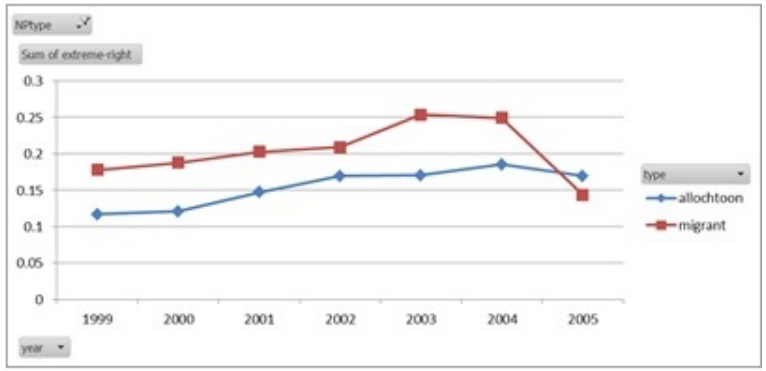
Calculate similarity of each cluster vector to each year/np-vector of *allochtoon* and *migrant*

7. Measuring lectal and diachronic distances

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

EXT-RIGHT



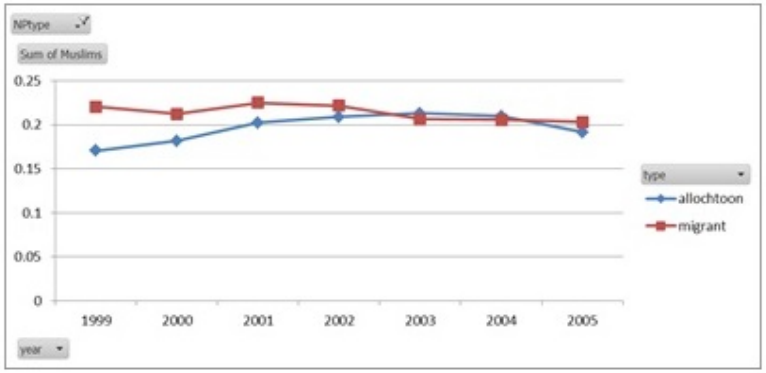


7. Measuring lectal and diachronic distances

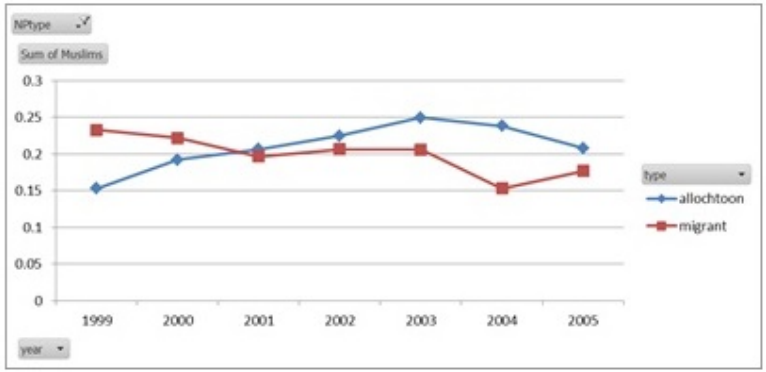
ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

MUSLIMS

QUALITY NP



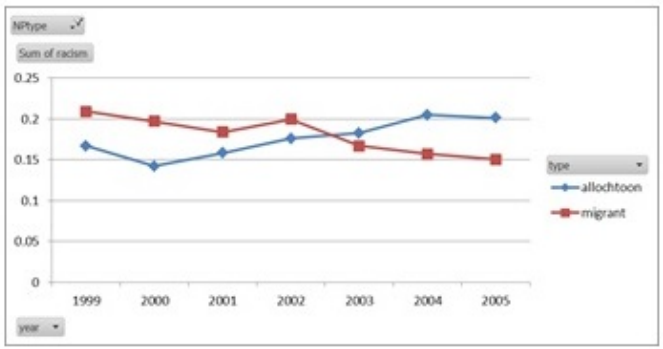
POPULAR NP

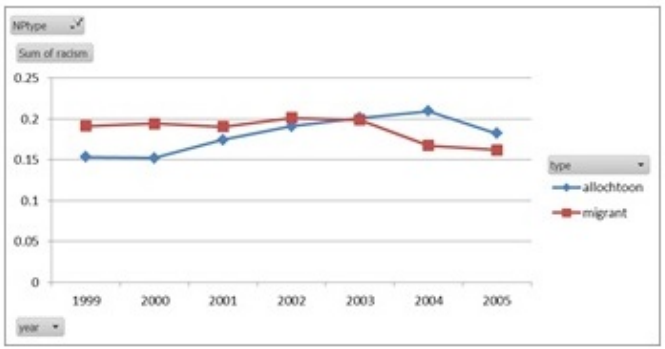


7. Measuring lectal and diachronic distances

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

RACISM

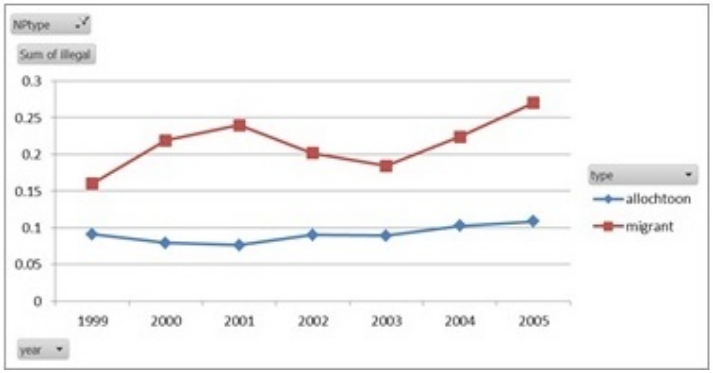


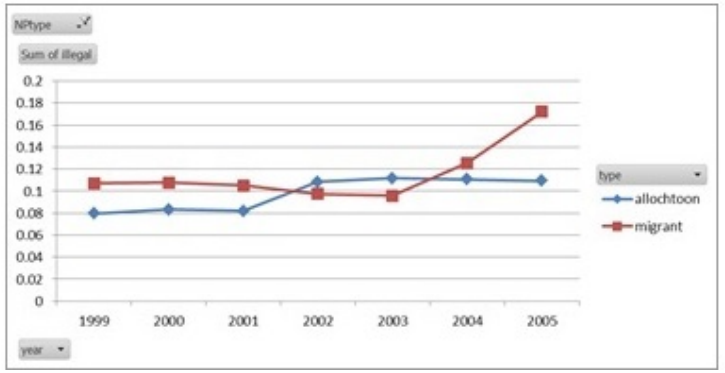


7. Measuring lectal and diachronic distances

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

ILLEGAL

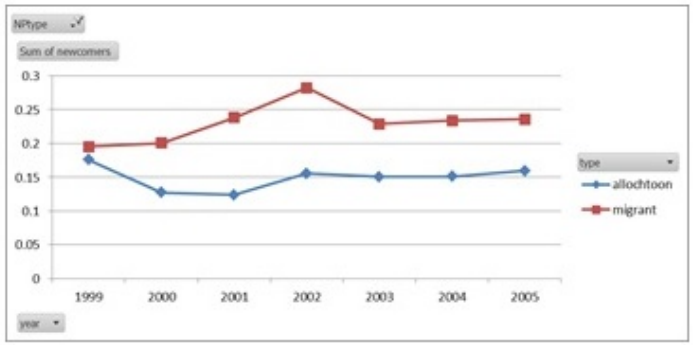


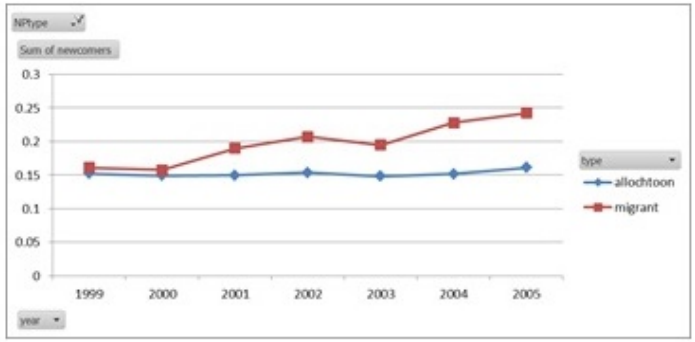


7. Measuring lectal and diachronic distances

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

NEW-COMERS

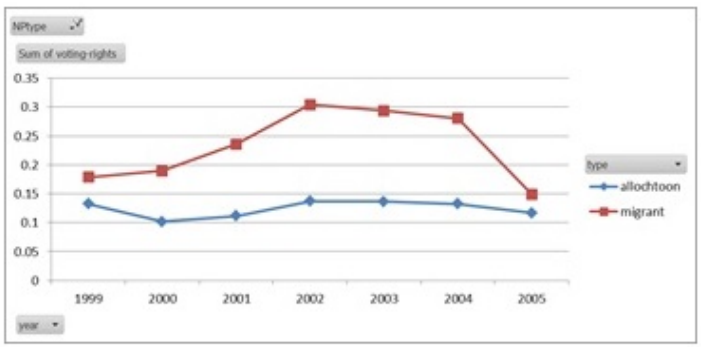


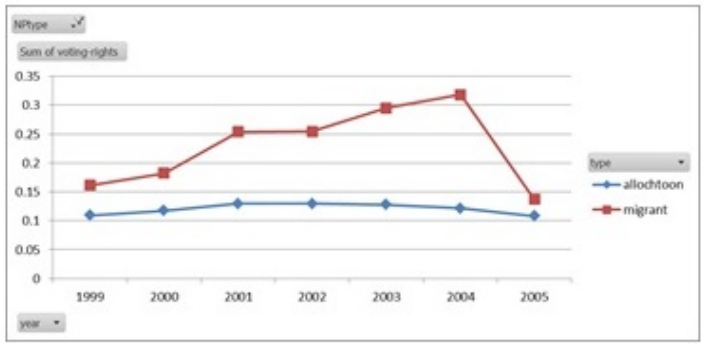


7. Measuring lectal and diachronic distances

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

VOTING RIGHTS

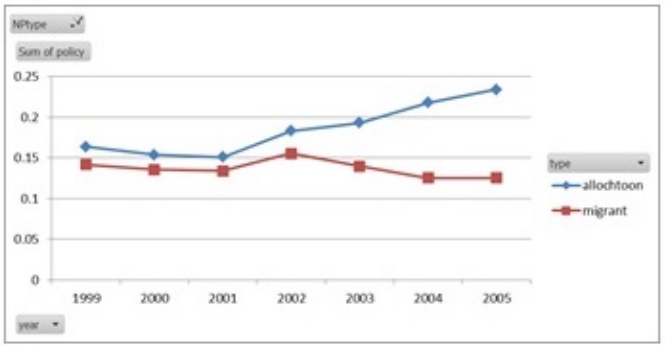


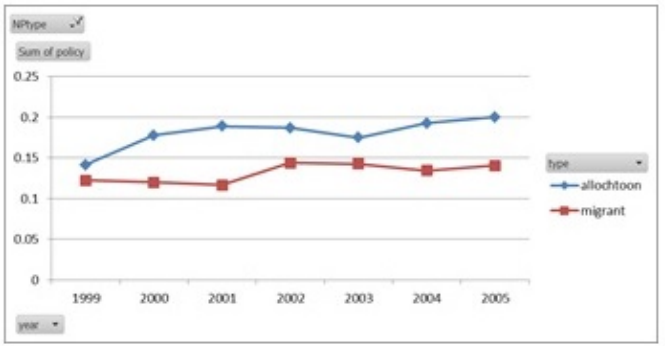


7. Measuring lectal and diachronic distances

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

POLICY



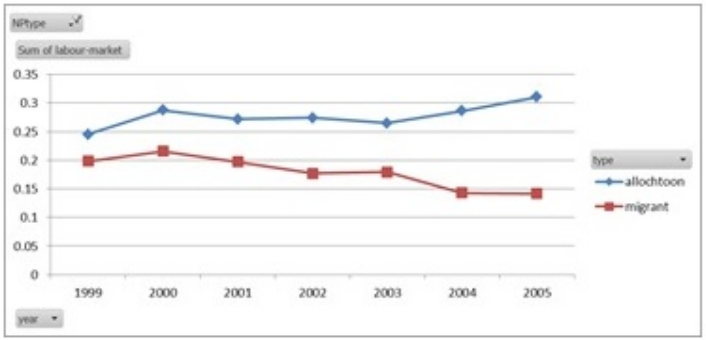


7. Measuring lectal and diachronic distances

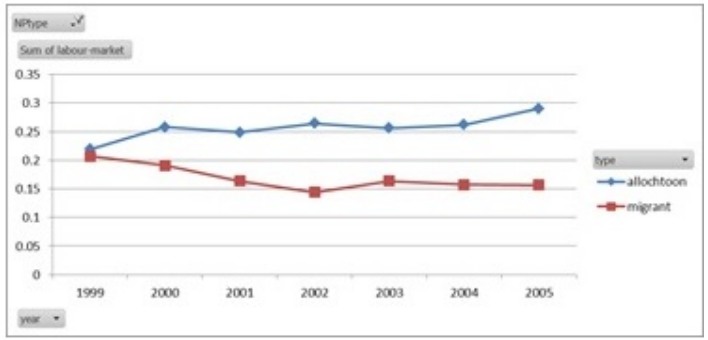
ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

LABOUR MARKET

QUALITY NP



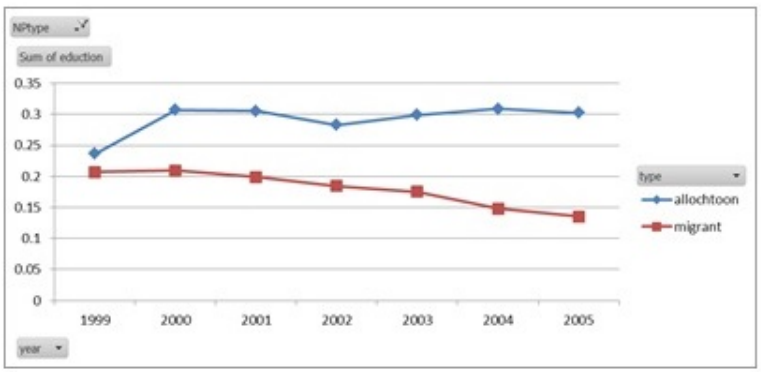
POPULAR NP

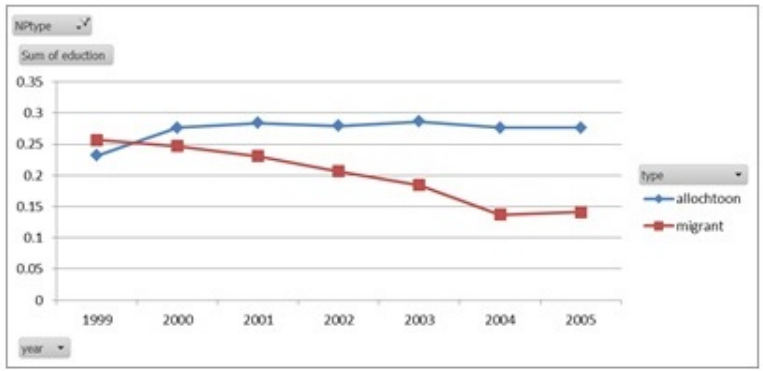


7. Measuring lectal and diachronic distances

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

EDUCATION

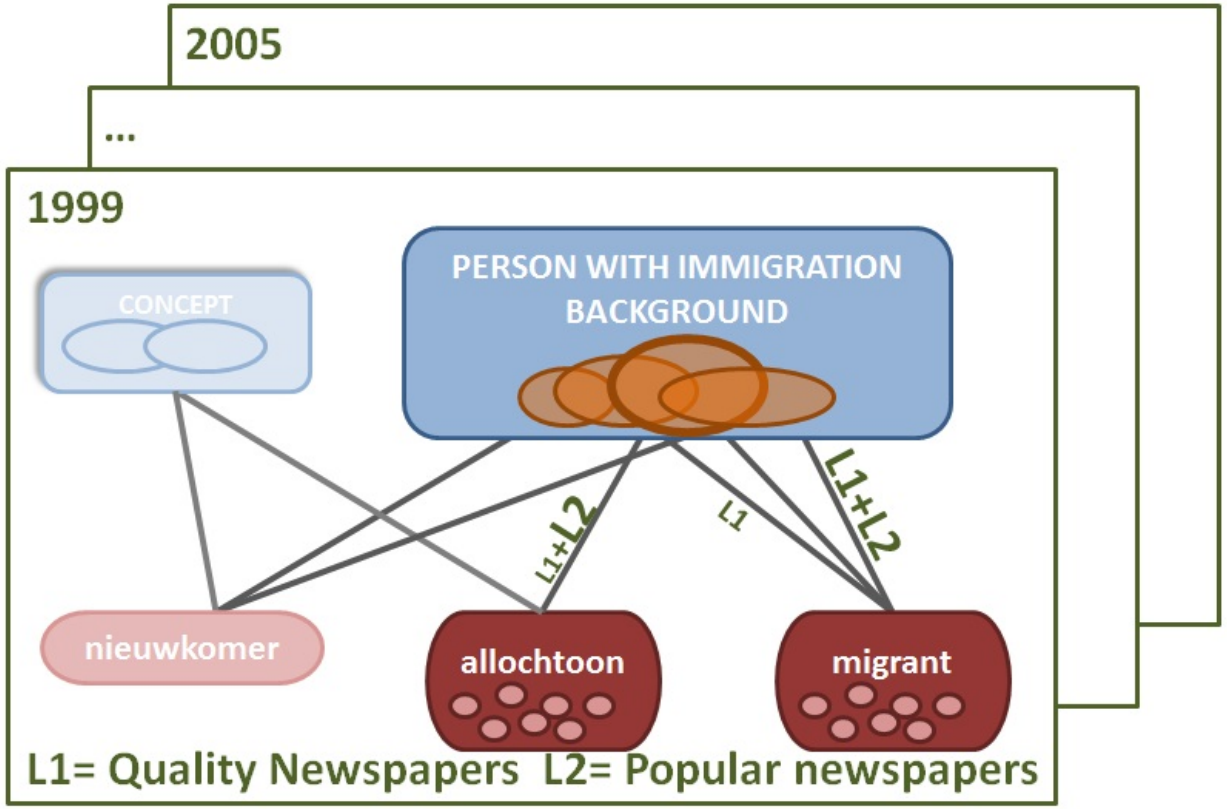




7. Measuring lectal and diachronic distances

Overview

8. Structuring concordances



- Study how individual occurrences are structured by context types

8. Structuring concordances

Make a vector for each occurrence of the variants

op de arbeidsmarkt zijn er voor allochtonen nauwelijks jobs

8. Structuring concordances

Make a vector for each occurrence of the variants

	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

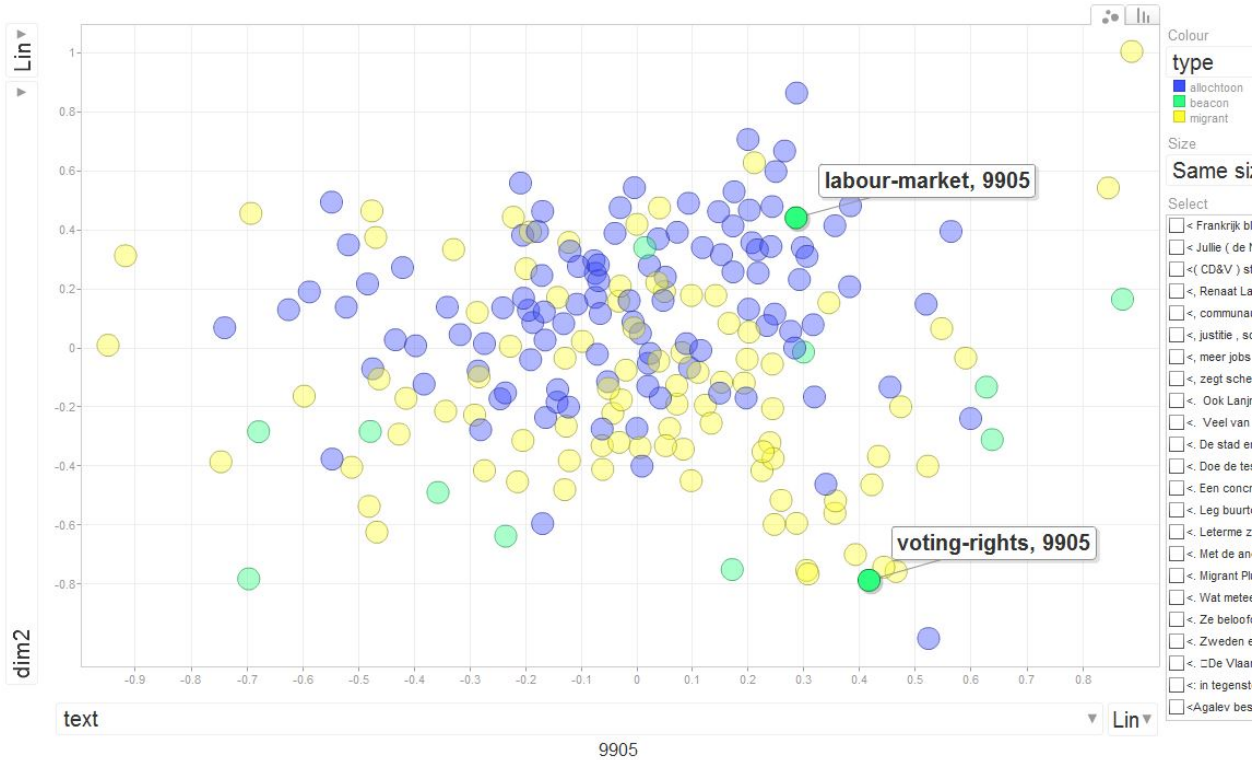
8. Structuring concordances

Make a vector for each occurrence of the variants

			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	

8. Structuring concordances

Calculate similarity between all tokens
use MDS and googlevis to plot in 2D



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

CDA research in CSL:

- political and ideological discourse can be studied empirically in large datasets
- diachronic and lectal variation need to be taken into account

For more information:

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

kris.heylen@arts.kuleuven.be

thomas.wielfaert@arts.kuleuven.be

dirk.speelman@arts.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.