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

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
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 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

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
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 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

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
3. Analysing Lexical semantics in CSL

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):

[Diagram showing concepts and words related to lexical semantics]
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:

- Concept / Meaning
- Word
- Word
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:

L1 = Quality Newspapers  L2 = Popular newspapers
3. Analysing Lexical semantics in CSL

USAGE-BASED STUDY:

CORPUS

1999

CONCEPT

PERSON WITH IMMIGRATION BACKGROUND

L1 = Quality Newspapers  L2 = Popular newspapers

2005
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

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
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: ⇒ words occurring in same contexts have similar meaning
4. Semantic Vector Spaces

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

<table>
<thead>
<tr>
<th></th>
<th>work</th>
<th>foreign</th>
<th>citizenship</th>
<th>laws</th>
<th>space</th>
<th>sugar</th>
<th>cream</th>
<th>now</th>
</tr>
</thead>
<tbody>
<tr>
<td>immigrant</td>
<td>120</td>
<td>424</td>
<td>388</td>
<td>82</td>
<td>12</td>
<td>11</td>
<td>3</td>
<td>189</td>
</tr>
<tr>
<td>alien</td>
<td>154</td>
<td>401</td>
<td>376</td>
<td>99</td>
<td>305</td>
<td>20</td>
<td>1</td>
<td>123</td>
</tr>
<tr>
<td>coffee</td>
<td>5</td>
<td>8</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>72</td>
<td>102</td>
<td>152</td>
</tr>
</tbody>
</table>
4. Semantic Vector Spaces

weight the raw frequencies by collocational strength (pmi)

<table>
<thead>
<tr>
<th></th>
<th>work</th>
<th>foreign</th>
<th>citizenship</th>
<th>laws</th>
<th>space</th>
<th>sugar</th>
<th>milk</th>
<th>now</th>
</tr>
</thead>
<tbody>
<tr>
<td>immigrant</td>
<td>5.3</td>
<td>7.9</td>
<td>6.5</td>
<td>4.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>alien</td>
<td>4.3</td>
<td>8.1</td>
<td>5.7</td>
<td>3.2</td>
<td>6.2</td>
<td>0.5</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>coffee</td>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.1</td>
<td>0.0</td>
<td>6.4</td>
<td>7.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>
4. Semantic Vector Spaces

calculate word by word similarity matrix

<table>
<thead>
<tr>
<th></th>
<th>immigrant</th>
<th>alien</th>
<th>coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>immigrant</td>
<td>1</td>
<td>.71</td>
<td>.08</td>
</tr>
<tr>
<td>alien</td>
<td>.71</td>
<td>1</td>
<td>.09</td>
</tr>
<tr>
<td>coffee</td>
<td>.08</td>
<td>.09</td>
<td>1</td>
</tr>
</tbody>
</table>
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
5. Identifying alternative expressions

- calculate contextual similarity between 10K Dutch nouns
- sort by similarity to *allochtoon*
5. Identifying alternative expressions

<table>
<thead>
<tr>
<th>Alternative</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>allochtoon</td>
<td>1.0</td>
</tr>
<tr>
<td>migrant</td>
<td>0.71</td>
</tr>
<tr>
<td>vreemdeling</td>
<td>0.48</td>
</tr>
<tr>
<td>immigrant</td>
<td>0.47</td>
</tr>
<tr>
<td>buitenlander</td>
<td>0.47</td>
</tr>
<tr>
<td>nieuwkomer</td>
<td>0.32</td>
</tr>
<tr>
<td>gastarbeider</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Table alternatives to *allochtoon*
5. Identifying alternative expressions
5. Identifying alternative expressions

Concept frequency in the corpus per month

 Lexeme relative frequencies in Belgian Newspapers

- dark blue: allochtoon
- red: migrant
5. Identifying alternative expressions

Concept frequency in the corpus per month per newspaper type
5. Identifying alternative expressions

- **Concept frequency in Quality Newspapers**
  - Quality newspapers

- **Concept frequency in Popular Newspapers**
  - Popular newspapers
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

Lexeme distribution in Belgian Newspapers
5. Identifying alternative expressions

Lexeme distribution in Quality Newspapers

Lexeme distribution in Popular Newspapers
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
6. Identifying and structuring collocations
6. Identifying and structuring collocations

Extract strongest concept collocations from matrix

<table>
<thead>
<tr>
<th></th>
<th>jobs</th>
<th>racisme</th>
<th>integratie</th>
<th>misdaad</th>
<th>stemrecht</th>
<th>suiker</th>
<th>zon</th>
<th>hond</th>
</tr>
</thead>
<tbody>
<tr>
<td>allochtoon</td>
<td>5.3</td>
<td>7.9</td>
<td>6.5</td>
<td>4.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>migrant</td>
<td>4.3</td>
<td>8.1</td>
<td>5.7</td>
<td>3.2</td>
<td>6.2</td>
<td>0.5</td>
<td>0.0</td>
<td>0.1</td>
</tr>
</tbody>
</table>
6. Identifying and structuring collocations

Make collocation-profile matrix for collocations

<table>
<thead>
<tr>
<th></th>
<th>jobs</th>
<th>racisme</th>
<th>integratie</th>
<th>misdaad</th>
<th>stemrecht</th>
<th>suiker</th>
<th>zon</th>
<th>hond</th>
</tr>
</thead>
<tbody>
<tr>
<td>jobs</td>
<td>5.3</td>
<td>7.9</td>
<td>6.5</td>
<td>4.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>racisme</td>
<td>4.3</td>
<td>8.1</td>
<td>5.7</td>
<td>3.2</td>
<td>6.2</td>
<td>0.5</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>integratie</td>
<td>5.3</td>
<td>7.9</td>
<td>6.5</td>
<td>6.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>misdaad</td>
<td>4.3</td>
<td>8.1</td>
<td>5.7</td>
<td>2.2</td>
<td>6.2</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>stemrecht</td>
<td>5.3</td>
<td>7.9</td>
<td>6.5</td>
<td>8.0</td>
<td>0.8</td>
<td>0.9</td>
<td>0.3</td>
<td>0.0</td>
</tr>
</tbody>
</table>
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

illegal immigration

newcomers
6. Identifying and structuring collocations

![Diagram showing collocations related to crime, 2002 riots, and education.]
6. Identifying and structuring collocations

![Diagram showing collocations between emancipation and discrimination]
6. Identifying and structuring collocations
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
7. Measuring lectal and diachronic distances

Newspaper type
L1 = Quality
L2 = Popular
7. Measuring lectal and diachronic distances

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

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

<table>
<thead>
<tr>
<th></th>
<th>jobs</th>
<th>racisme</th>
<th>integratie</th>
<th>misdaad</th>
<th>stemrecht</th>
<th>suiker</th>
<th>zon</th>
</tr>
</thead>
<tbody>
<tr>
<td>jobs</td>
<td>5.3</td>
<td>7.9</td>
<td>6.5</td>
<td>4.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>werk</td>
<td>4.3</td>
<td>8.1</td>
<td>5.7</td>
<td>3.2</td>
<td>6.2</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>arbeidsmarkt</td>
<td>5.3</td>
<td>7.9</td>
<td>6.5</td>
<td>6.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>LABOURMARKET</td>
<td>5.3</td>
<td>7.1</td>
<td>7.7</td>
<td>2.2</td>
<td>6.2</td>
<td>0.4</td>
<td>0.0</td>
</tr>
</tbody>
</table>
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

![Graphs showing changes over time]
7. Measuring lectal and diachronic distances

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT

QUALITY NP

POPULAR NP

MUSLIMS
7. Measuring lectal and diachronic distances

ALLOCHTOON TAKES OVER CONTEXTS FROM MIGRANT
7. Measuring lectal and diachronic distances

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON
7. Measuring lectal and diachronic distances

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON
7. Measuring lectal and diachronic distances

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON
7. Measuring lectal and diachronic distances

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT
7. Measuring lectal and diachronic distances

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

QUALITY NP

POPULAR NP

LABOUR MARKET
7. Measuring lectal and diachronic distances

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT
7. Measuring lectal and diachronic distances
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
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

<table>
<thead>
<tr>
<th>arbeidsmarkt</th>
<th>allochtonen</th>
<th>jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>7.1</td>
<td>5.2</td>
</tr>
<tr>
<td>5.1</td>
<td>0.1</td>
<td>3.1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>3.1</td>
<td>4.1</td>
<td>3.7</td>
</tr>
<tr>
<td>4.7</td>
<td>3.1</td>
<td>3.9</td>
</tr>
<tr>
<td>2.2</td>
<td>3.8</td>
<td>2.9</td>
</tr>
</tbody>
</table>

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
8. Structuring concordances

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

Descriptive: *allochtoon* vs. *migrant*

- *allochtoon* replaces *migrant* in frequency
- Immigration discussions seem to have strong ’seasonal peaks, especially in high-brow newspapers
- *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

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.wiefaert@arts.kuleuven.be
dirk.speelman@arts.kuleuven.be
References I


References II


References III


