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?
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

LECT 1

LECT 2
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 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 CORPORAS OF BELGIAN DUTCH (1.3G words)
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.

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

<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>
### 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>cream</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>
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>

Diagram showing the relationship between words in a vector space with axes for "foreign" and "drink".
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

<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*
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

<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>
### Analysing Semantic Structure

Make weighted co-occurrence matrix for these 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>
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

Labour market

voting rights

extreme right
Analysing Semantic Structure

illegal immigration

newcomers
Analysing Semantic Structure
Analysing Semantic Structure

discrimination

emancipation
Analysing Semantic Structure
Analysing Semantic Structure

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

Diagram:
- IMMIGRANT
  - LABOUR
  - MARKET
  - POLICY
  - ILLEGAL MIGRATION
  - CRIME
  - VOTING RIGHTS
  - ...
- allochtoon
- migrant
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

<table>
<thead>
<tr>
<th>Type</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>
Measuring semantic change in registers

**STEP 2**
Make vector per context 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>
Measuring semantic change in registers

**STEP 3**
Combine variant/year/type vectors and context cluster vectors in 1 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>
</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>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>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

...
Measuring semantic change in registers

STEP 4
Calculate the cosine similarity (≈ association strength) of each variant/year/type vector to each context cluster vector

<table>
<thead>
<tr>
<th></th>
<th>LABOUR</th>
<th>ILLEGAL</th>
<th>EXTREME</th>
<th>POLICY</th>
<th>CRIME</th>
<th>VOTING</th>
<th>RACISM</th>
</tr>
</thead>
<tbody>
<tr>
<td>allochtoon/1999pop</td>
<td>0.3</td>
<td>0.9</td>
<td>0.5</td>
<td>0.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>migrant/1999pop</td>
<td>0.3</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>allochtoon/1999qual</td>
<td>0.3</td>
<td>0.9</td>
<td>0.5</td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>migrant/1999qual</td>
<td>0.3</td>
<td>0.2</td>
<td>0.7</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>allochtoon/2000pop</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1</td>
<td>0.3</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>migrant/2000pop</td>
<td>0.9</td>
<td>0.4</td>
<td>0.7</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.7</td>
</tr>
</tbody>
</table>
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

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
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

QUALITY

POPULAR
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

[Graphs showing changes in voting rights over time for Quality and Popular categories]
Measuring semantic change in registers

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT
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

**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

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

<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>
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

<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><em>allochtoon</em>$_1$</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><em>allochtoon</em>$_2$</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><em>allochtoon</em>$_3$</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><em>migrant</em>$_1$</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><em>migrant</em>$_2$</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><em>migrant</em>$_3$</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>
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
Lexical variation on the exemplar level

STEP 5: use MDS to plot similarly 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