Distributional Semantic Modelling in Cognitive Sociolinguistics: QLVL probes Semantic Space

Kris Heylen, Dirk Geeraerts & Dirk Speelman

KU Leuven
Quantitative Lexicology and Variational Linguistics
Purpose of the talk

**Theoretical:** Lexicology in the usage-based and lexically enriched framework of **Cognitive Sociolinguistics**

**Methodological:** Summarize QLVL’s use of **Distributional Semantic Models** in last 7 years for analysing lexical semantics in large corpora,

**Descriptive:** Charting **Lexical variation** on two levels:

1. **Varieties:** Measuring differences in word choice on an aggregate level
2. **Variants:** Word choices for individual concepts / polysemy of lexemes
Overview

1. Cognitive Sociolinguistics
   Structure of Lexical Variation
   The Profile-Based Approach

2. Onomasiological Variation
   Semantic Vector Spaces
   Different Context Models
   Lexical Sociolectometry

3. Including Semasiological Variation
   Measuring Polysemy
   Analysing Semantic Structure
   Measuring semantic change
   Token clouds

4. Conclusion
Overview

1. Cognitive Sociolinguistics
   Structure of Lexical Variation
   The Profile-Based Approach

2. Onomasiological Variation
   Semantic Vector Spaces
   Different Context Models
   Lexical Sociolectometry

3. Including Semasiological Variation
   Measuring Polysemy
   Analysing Semantic Structure
   Measuring semantic change
   Token clouds

4. Conclusion
1. Cognitive Sociolinguistics

QLVL’s research is situated in what has become known as Cognitive Sociolinguistics (Kristiansen & Dirven 2008; Geeraerts, Kristiansen & Peirsman 2010):

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

QLVL has been developing this line of research since the 1990s:

- Structure of Lexical Variation (1994)
- Profile-based approach (1999)
Structure of Lexical Variation (1994)

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):
Structure of Lexical Variation (1994)

LEXICOLOGY (Geeraerts, Grondelaers & Bakema 1994):
Structure of Lexical Variation (1994)

SEMASIOLOGY:

CONCEPT / MEANING

CONCEPT / MEANING

Word

underground

REBELS  SUBTERRANEAN PUBLIC TRANSPORT
Structure of Lexical Variation (1994)

ONOMASIOLOGY:

- **Word**: tube
- **Word**: underground
- **Word**: subway

SUBTERRANEAN PUBLIC TRANSPORT

CONCEPT / MEANING
Structure of Lexical Variation (1994)

PROTOTYPE STRUCTURE:
Structure of Lexical Variation (1994)

PROTOTYPE STRUCTURE:
Structure of Lexical Variation (1994)

LECTAL VARIATION:

LECT 1

LECT 2
Structure of Lexical Variation (1994)
Structure of Lexical Variation (1994)

bandplooibroek  hippiebroek  pantalon

olifantbroek  jodhpur  harembroek
Structure of Lexical Variation (1994)

Figure 3.4(1)
The semasiological structure of *legging*

- a Reaching down to the ankles or the calves
- b Tight-fitting
- c Made of elastic material
- d Without fastening on the end of the legs
The Profile-Based Approach (1999)

FORMAL ONOMASIOLOGICAL VARIATION

LECT 1

CONCEPT

Word1

Word 2

LECT 2

CONCEPT

Word 1

Word 2
The Profile-Based Approach (1999)

AGGREGATED LEXICAL VARIATION
The Profile-Based Approach (1999)

Profile based measurement of lexical variation
Do B and NL use the same lexemes to refer to a given concept?
- define a set of synonyms = profile
- collect all instances from 2 corpora (B vs NL) + disambiguate
- Calculate relative frequency of synonyms in 2 corpora
- Overlap in relative frequency = uniformity measure

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>NL</th>
<th>overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>jeans</td>
<td>85</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>spijkerbroek</td>
<td>15</td>
<td>70</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>45</td>
</tr>
</tbody>
</table>
The Profile-Based Approach (1999)

AGGREGATED UNIFORMITY B-NL
The Profile-Based Approach (1999)

- sports and clothing terms: Geeraerts et al. 1999, 2003
- empirical, corpus-based, quantitative
- BUT time consuming, not readily scalable
  - Profiles (synonyms) manually selected
  - Occurrences of synonyms manually disambiguated

sem·metrix project (Peirsman & Heylen)

WHAT?
- semi-automatic generation of synonym sets (profiles)
- Word Sense Disambiguation for profile membership

HOW?
- Distributional semantic modelling in large corpora
Overview

1. Cognitive Sociolinguistics
   Structure of Lexical Variation
   The Profile-Based Approach

2. Onomasiological Variation
   Semantic Vector Spaces
   Different Context Models
   Lexical Sociolectometry

3. Including Semasiological Variation
   Measuring Polysemy
   Analysing Semantic Structure
   Measuring semantic change
   Token clouds

4. 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
- 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
Type-level SVS

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

<table>
<thead>
<tr>
<th></th>
<th>transport</th>
<th>train</th>
<th>commute</th>
<th>ticket</th>
<th>Scene</th>
<th>Sugar</th>
<th>Cream</th>
<th>now</th>
</tr>
</thead>
<tbody>
<tr>
<td>subway</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>underground</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>
**Type-level SVS**

weight the raw frequencies by collocational strength (pmi)

<table>
<thead>
<tr>
<th></th>
<th>transport</th>
<th>train</th>
<th>commute</th>
<th>ticket</th>
<th>scene</th>
<th>sugar</th>
<th>milk</th>
<th>now</th>
</tr>
</thead>
<tbody>
<tr>
<td>subway</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>underground</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>
Type-level SVS

calculate word by word similarity matrix

<table>
<thead>
<tr>
<th></th>
<th>subway</th>
<th>underground</th>
<th>coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>subway</td>
<td>1</td>
<td>.71</td>
<td>.08</td>
</tr>
<tr>
<td>underground</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:
- "subway" and "underground" are connected, indicating similarity.
- "coffee" is isolated, with lower similarity to the others.
- "public transport" and "drink" axes.
Different Context Models

Vector Space Models come in many flavours

The main difference between the models lies in how they define context

Family of models

```
Word Space Models
   \-----
  |     |
  |     |
  \-----

document based        word based

              \-----
             |     |
             |     |
             \-----

bag-of-words         syntactic
```
Different Context Models

document based models

- context = stretch of text in which target word occurs
- 2 words are related when they often co-occur in text
- Landauer & Dumais 1997: Latent Semantic Analysis

word based models

- context = context words around the target word
- 2 words are related when they co-occur with the same context words, but not necessarily with each other
<table>
<thead>
<tr>
<th></th>
<th>DOC.1</th>
<th>DOC.2</th>
<th>DOC.3</th>
<th>DOC.4</th>
<th>DOC.5</th>
<th>DOC.6</th>
<th>DOC.7</th>
<th>DOC.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>ongeval</td>
<td>23</td>
<td>12</td>
<td>14</td>
<td>24</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ongeluk</td>
<td>16</td>
<td>9</td>
<td>11</td>
<td>18</td>
<td>17</td>
<td>20</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>koffie</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>12</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>auto</th>
<th>slachtoffer</th>
<th>vrachtwagen</th>
<th>file</th>
<th>gekwetst</th>
<th>suiker</th>
<th>melk</th>
<th>kop</th>
</tr>
</thead>
<tbody>
<tr>
<td>ongeval</td>
<td>120</td>
<td>424</td>
<td>388</td>
<td>82</td>
<td>270</td>
<td>11</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>ongeluk</td>
<td>154</td>
<td>401</td>
<td>376</td>
<td>99</td>
<td>305</td>
<td>20</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>koffie</td>
<td>5</td>
<td>8</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>72</td>
<td>102</td>
<td>93</td>
</tr>
</tbody>
</table>
Different Context Models

Within word based models:

**bag-of-words**

- context words in window of $n$ words left and right of target word
- a bag of unstructured context features

**syntactic features**

- context words in specific syntactic relation with target word
- takes clause structure into account
The wagging dog barked at the postman on the bike

<table>
<thead>
<tr>
<th></th>
<th>wagging</th>
<th>dog</th>
<th>bark</th>
<th>postman</th>
<th>bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>postman</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
The wagging dog barked at the postman on the bike

<table>
<thead>
<tr>
<th></th>
<th>subj.bark</th>
<th>adj.wagging</th>
<th>PC.bark.at</th>
<th>PP.on.bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>postman</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Different Context Models

Within the bag-of-words models:

1st order co-occurrences

- context = words in immediate proximity to the target
- Levy & Bullinaria 2001

2nd order co-occurrences

- context = context words of context words of target
- can generalise over semantically related context words
- Schütze 1998

NB syntactic models are also 1st order models
's Morgens veroorzaakte een ongeval een file
's Morgens veroorzaakte een ongeval een file

Hij kan 's morgens zijn bed niet uit
Neem 's morgens tijd voor een ontbijt
' s Morgens gaat de wekker altijd te vroeg af
<table>
<thead>
<tr>
<th>Overview</th>
<th>CogSoLx</th>
<th>Onomas</th>
<th>Semas</th>
<th>Conclusion</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

_Roken veroorzaakt kanker_  
_CO2 veroorzaakt klimaatopwarming_  
_De sneeuw veroorzaakt problemen_  

‘s Morgens veroorzaakte een ongeval een file

Hij kan ‘s morgens zijn bed niet uit  
Neem ‘s morgens tijd voor een ontbijt  
‘s Morgens gaat de wekker altijd te vroeg af
Roken veroorzaakt kanker
CO2 veroorzaakt klimaatopwarming
De sneeuw veroorzaakt problemen

‘s Morgens veroorzaakte een ongeval een file
Hij kan ‘s morgens zijn bed niet uit
Neem ‘s morgens tijd voor een ontbijt
‘s Morgens gaat de wekker altijd te vroeg af

Ze stond uren in de file
Op de ringweg staat een file
Met de auto sta je sowieso in de file
Roken veroorzaakt kanker
CO2 veroorzaakt klimaatopwarming
De sneeuw veroorzaakt problemen

‘s Morgens veroorzaakte een ongeval een file

Hij kan ‘s morgens zijn bed niet uit
Neem ‘s morgens tijd voor een ontbijt
‘s Morgens gaat de wekker altijd te vroeg af

Ze stond uren in de file
Op de ringweg staat een file
Met de auto sta je sowieso in de file
Roken veroorzaakt kanker
CO2 veroorzaakt klimaatopwarming
De sneeuw veroorzaakt problemen

‘s Ochtends veroorzaakte een ongeval een file
Hij kan ‘s ochtends zijn bed niet uit
Neem ‘s ochtends tijd voor een ontbijt
‘s ochtends gaat de wekker altijd te vroeg af

Ze stond uren in de file
Op de ringweg staat een file
Met de auto sta je sowieso in de file
Different Context Models

Comparing models

- syntactic model
- first-order bag-of-word models with context sizes 1, 3 and 5
- second-order bag-of-word models with context sizes 1, 3 and 5

Parameters

- data: Twente Nieuws Corpus (400M words)
- 2000 dimensions: the most frequent features in the corpus
- weighting scheme: point-wise mutual information
- similarity calculated with cosine
Different Context Models

1 neighbour

- syn
- c1
- c3
- c5
- cc1
- cc3
- cc5

<table>
<thead>
<tr>
<th>Word Space Model</th>
<th>number of nearest neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>syn</td>
<td>2500</td>
</tr>
<tr>
<td>c1</td>
<td>1500</td>
</tr>
<tr>
<td>c3</td>
<td>1000</td>
</tr>
<tr>
<td>c5</td>
<td>500</td>
</tr>
<tr>
<td>cc1</td>
<td>250</td>
</tr>
<tr>
<td>cc3</td>
<td>150</td>
</tr>
<tr>
<td>cc5</td>
<td>100</td>
</tr>
</tbody>
</table>

- synonym
- hyponym
- hyperonym
- cohyponym
Different Context Models

- Syntactic model > first-order bag-of-words > second-order bag-of-words
- The “stricter” the definition of context, the more semantically similar words.
- demo: https://perswww.kuleuven.be/~u0042527/demo.html

AUTOMATIC GENERATION OF SYNONYM PROFILES
Extension: Finding Equivalents

Bilectal Word Space Models

- Extend Word Space Models from one corpus to two corpora
- Word co-occurrences are relatively constant across varieties
- Identification of synonyms in different varieties of the same language
  - BE dessert – NL toetje, UK nappy – US diaper

<table>
<thead>
<tr>
<th></th>
<th>road</th>
<th>drive</th>
<th>engine</th>
<th>traffic</th>
<th>baby</th>
<th>clean</th>
<th>toilet</th>
<th>legs</th>
</tr>
</thead>
<tbody>
<tr>
<td>car (UK)</td>
<td>120</td>
<td>424</td>
<td>221</td>
<td>189</td>
<td>6</td>
<td>11</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>car (US)</td>
<td>152</td>
<td>389</td>
<td>178</td>
<td>167</td>
<td>15</td>
<td>9</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>nappy (UK)</td>
<td>3</td>
<td>12</td>
<td>16</td>
<td>9</td>
<td>270</td>
<td>189</td>
<td>318</td>
<td>167</td>
</tr>
<tr>
<td>diaper (US)</td>
<td>8</td>
<td>11</td>
<td>13</td>
<td>7</td>
<td>301</td>
<td>205</td>
<td>324</td>
<td>179</td>
</tr>
</tbody>
</table>
Two varieties
Lexical Sociolectometry (Ruette)

Large-scale Profile-based approach

- Measure distances between language varieties based on onomasiological variation.
- Aggregated over automatically selected profiles of near-synonyms

Corpora and varieties

<table>
<thead>
<tr>
<th></th>
<th>Usenet</th>
<th>Popular news</th>
<th>Quality news</th>
<th>Legalese</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE</td>
<td>22 million</td>
<td>905 million</td>
<td>373 million</td>
<td>70 million</td>
<td>1.4 billion</td>
</tr>
<tr>
<td>NL</td>
<td>26 million</td>
<td>126 million</td>
<td>161 million</td>
<td>115 million</td>
<td>428 million</td>
</tr>
<tr>
<td>Total</td>
<td>48 million</td>
<td>1 billion</td>
<td>499 million</td>
<td>185 million</td>
<td>1.8 billion</td>
</tr>
</tbody>
</table>
Lexical Sociolectometry (Ruette)

Clustering by Committee

1. All pairwise similarities between 10K nouns in the corpus are calculated.
2. A cluster algorithm (Pantel et al., 2002) finds sets of similar nouns (profile).
3. A trained linguist goes through these sets and filters the sets that contain synonyms.

For now, semi-automatic variable set generation

- The SVS approach ensures automatic, bottom-up generation of candidate variables, with high recall, but low precision.
- The manual correction of the trained linguist ensures the quality of the variable set.
## Lexical Sociolectometry

<table>
<thead>
<tr>
<th><strong>CONCEPT</strong></th>
<th><strong>Items</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>MANNER</td>
<td>wijze, manier</td>
</tr>
<tr>
<td>GENOCIDE</td>
<td>volk_moord, genocide</td>
</tr>
<tr>
<td>POLL</td>
<td>peiling, opiniepeiling</td>
</tr>
<tr>
<td>MARIHUANA</td>
<td>cannabis, marihuana</td>
</tr>
<tr>
<td>PUTSCH</td>
<td>staatsgreep, coup</td>
</tr>
<tr>
<td>MENINGITIS</td>
<td>hersenvliesontsteking, meningitis</td>
</tr>
<tr>
<td>DEMONSTRATOR</td>
<td>demonstrant, betoger</td>
</tr>
<tr>
<td>AIRPORT</td>
<td>vliegveld, luchthaven</td>
</tr>
<tr>
<td>COLDNESS</td>
<td>koude, kou</td>
</tr>
<tr>
<td>TORTURE</td>
<td>marteling, foltering</td>
</tr>
<tr>
<td>VICTORY</td>
<td>zege, overwinning</td>
</tr>
<tr>
<td>HOMOSEXUAL</td>
<td>homo, homoseksueel</td>
</tr>
<tr>
<td>SAXOPHONE</td>
<td>sax, saxofoon</td>
</tr>
<tr>
<td>INTERNETPROVIDER</td>
<td>provider, internetprovider, internetaanbieder</td>
</tr>
<tr>
<td>AIRCONDITIONING</td>
<td>airconditioning, airco</td>
</tr>
<tr>
<td>RELIGION</td>
<td>religie, godsdienst</td>
</tr>
<tr>
<td>THE OTHER SIDE</td>
<td>overkant, overzijde</td>
</tr>
<tr>
<td>EXPLOSION</td>
<td>explosie, ontploffing</td>
</tr>
<tr>
<td>RESTROOM</td>
<td>toilet, wc</td>
</tr>
</tbody>
</table>
Lexical Sociolectometry (Ruette)

Distance metric

- “different lexemes referring to the same concept”
- concept naming: if two subcorpora use the same word to refer to a certain concept, they are considered to be similar
- similarity between two subcorpora is measured by summing the City-block distance for every concept

Distance metric

\[
D_{CB,L}(V_1, V_2) = \frac{1}{2} \sum_{i=1}^{n} |R_{V_1,L}(x_i) - R_{V_2,L}(x_i)|
\]

(1)

\[
D_{CB}(V_1, V_2) = \sum_{i=1}^{m} (D_{Li}(V_1, V_2)W(L_i))
\]

(2)
Lexical Sociolectometry (Ruette)

Non-parametric MDS

be.pop.news  be.qua.news  be.usenet  be.legal  nl.legal

Dimension 1

Dimension 2

Dimension 3
Lexical Sociolectometry (Ruette)

Individual difference scaling
Overview

1. Cognitive Sociolinguistics
   Structure of Lexical Variation
   The Profile-Based Approach

2. Onomasiological Variation
   Semantic Vector Spaces
   Different Context Models
   Lexical Sociolectometry

3. Including Semasiological Variation
   Measuring Polysemy
   Analysing Semantic Structure
   Measuring semantic change
   Token clouds

4. Conclusion
Semasiological Variation

Profile-Based approach

LECT 1

CONCEPT

Word1

Word 2

LECT 2

CONCEPT

Word 1

Word 2
Semasiological Variation

Polysemy:
Measuring Polysemy (Bertels)

- Univocity hypothesis in Wuesterian terminology: Technical terms are monosemous
- Bertels: correlation between keyness and monosemy
- Monosemy-measure: overlap of second-order co-occurrences

![Figure 1: Schéma: mot-clé + cooccurrences de premier et de deuxième ordre.](image)
Measuring Polysemy (Bertels)

Inverse relation between keyness and monoosemy!
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 $\approx$ 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

[Diagram showing a semantic structure with nodes such as 'crime', '2002 riots', 'education', 'probleem', 'criminaliteit', 'pak_aan', 'concentratie', 'buurt', 'overval', 'islamitisch', 'Marokkaans', 'Turks', 'AEL', 'Jahjah', 'rel DIM', 'onderwijs', 'school', 'groep', 'handicap', 'doel_groep', 'jongere', 'kansarm']
Analysing Semantic Structure
Analysing Semantic Structure

Diagram showing semantic structure relationships involving terms such as 'integration', 'policies', and various nouns and adjectives related to integration and policies.
Analysing Semantic Structure

Contextually defined "semantic features" that constitute the prototypical structure of the concept
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></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>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</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
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

QUALITY

POPULAR

RACISM
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

NEW-COMERS

QUALITY

POPULAR
Measuring semantic change in registers

MIGRANT SPECIALIZES RELATIVE TO ALLOCHTOON

QUALITY

POPULAR

VOTING RIGHTS
Measuring semantic change in registers

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT

POLICY

QUALITY

POPULAR
Measuring semantic change in registers

ALLOCHTOON SPECIALIZES RELATIVE TO MIGRANT
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.
Token clouds
How are the individual tokens/exemplars structured?
≈ Word Sense Disambiguation
Token-level SVS

Make a vector for each occurrence of the variants

the teacher saw the dog chasing the cat
Token-level SVS

Make a vector for each occurrence of the variants

3.2 4.3 0.8 7.1
5.1 2.2 3.7 0.1
0.2 3.5 2.3 0.3
3.1 1.9 2.9 4.1
4.7 0.2 1.3 3.1
2.2 3.1 4.1 3.8

the teacher saw the dog chasing the cat
Token-level SVS

Make a vector for each occurrence of the variants

<table>
<thead>
<tr>
<th>teacher</th>
<th>saw</th>
<th>dog</th>
<th>chasing</th>
<th>cat</th>
<th>AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>4.3</td>
<td>0.8</td>
<td>7.1</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>2.2</td>
<td>3.7</td>
<td>0.2</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>3.5</td>
<td>2.3</td>
<td>0.3</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>1.9</td>
<td>2.9</td>
<td>4.1</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>0.2</td>
<td>1.4</td>
<td>3.1</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>3.1</td>
<td>4.1</td>
<td>3.8</td>
<td>3.3</td>
<td></td>
</tr>
</tbody>
</table>
## Token-level SVS

### Weighting

<table>
<thead>
<tr>
<th>teacher</th>
<th>saw</th>
<th>dog</th>
<th>chasing</th>
<th>cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMI weights</td>
<td>0.4</td>
<td>0.8</td>
<td>2.1</td>
<td>1.5</td>
</tr>
</tbody>
</table>

```
3.2  4.3  0.8  7.1  
5.1  2.2  3.7  0.1  
0.2  3.5  2.3  0.3  
3.1  1.9  2.9  4.1  
4.7  0.2  1.3  3.1  
2.2  3.1  4.1  3.8  
```

Context words are not equally informative for the meaning of dog.
## Token-level SVS

### Weighted vectors

<table>
<thead>
<tr>
<th>teacher</th>
<th>saw</th>
<th>dog</th>
<th>chasing</th>
<th>cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2x0.4</td>
<td>4.3x0.8</td>
<td>0.8x2.1</td>
<td>7.1x1.5</td>
<td>4.3</td>
</tr>
<tr>
<td>5.1x0.4</td>
<td>2.2x0.8</td>
<td>3.7x2.1</td>
<td>0.2x1.5</td>
<td>3.0</td>
</tr>
<tr>
<td>0.2x0.4</td>
<td>3.5x0.8</td>
<td>2.3x2.1</td>
<td>0.3x1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>3.1x0.4</td>
<td>1.9x0.8</td>
<td>2.9x2.1</td>
<td>4.1x1.5</td>
<td>3.8</td>
</tr>
<tr>
<td>4.7x0.4</td>
<td>0.2x0.8</td>
<td>1.4x2.1</td>
<td>3.1x1.5</td>
<td>2.4</td>
</tr>
<tr>
<td>2.2x0.4</td>
<td>3.1x0.8</td>
<td>4.1x2.1</td>
<td>3.8x1.5</td>
<td>4.4</td>
</tr>
</tbody>
</table>
Token clouds

Calculate similarity between all tokens
use MDS and googlevis to plot in 2D
Demo: https://perswww.kuleuven.be/~u0038536/googleVis
Calibration

Semantic Vector Spaces, and especially token-level SVSs are parameter-rich.

Examples of parameters

- Bag-of-Words $\leftrightarrow$ Dependency Models
- Size of the context window for co-occurrences
- Size of the context window for weights
- Weighting scheme:
  - Pointwise Mutual Information $\leftrightarrow$ Log-Likelihood Ratio
- Include $\leftrightarrow$ exclude highly-frequent (function words) words
Token Clouds

- Calibration could benefit from visual analytics of the different solutions.
- Using manually disambiguated data facilitates the visual evaluation as we can color-code the tokens for their different meanings.
- Misclassified tokens are quickly identified.
- We built our own, customisable tool to explore these token clouds.
Token Clouds

Dutch noun monitor

- Manually disambiguated data for the concept of BEELDSCHERM (display)
- Semasiological study of a polysemous word with lectal variation (geographical and register)
4. Visual Analytics
4. Visual Analytics
4. Visual Analytics

**7-7 weights context window**

- computer screen
- youth leader

**monitor**

isoMDS, k=2
stress=22.80%
4. Visual Analytics

Cut-off to 0.5 of weights below 1

monitor
isoMDS, k=2
stress=21.10%

computer screen
youth leader
4. Visual Analytics

Log-Likelihood Ratio with cut-off

monitor
isoMDS, k=2
stress=24.84%

computer screen
youth leader
4. Visual Analytics

monitor
isoMDS, k=2
stress=22.33%

dim 1
dim 2

References

- Dat monitoren system is geïntroduceerd door dokter Alfred Steinschneider.
Overview

1. Cognitive Sociolinguistics
   Structure of Lexical Variation
   The Profile-Based Approach

2. Onomasiological Variation
   Semantic Vector Spaces
   Different Context Models
   Lexical Sociolectometry

3. Including Semasiological Variation
   Measuring Polysemy
   Analysing Semantic Structure
   Measuring semantic change
   Token clouds

4. Conclusion
Conclusion

Lexicology in Cognitive Sociolinguistics

- Variant level: Structure of Lexical Variation
- Variety level: Lexical Sociolectometry

Distributional Semantic Modelling

- to find synonym profiles
- measure polysemy
- structure collocations
- measure semantic change
- examplar level modelling
For more information:
http://wwwling.arts.kuleuven.be/qlvl
kris.heylen@arts.kuleuven.be
dirk.geeraerts@arts.kuleuven.be
dirk.speelman@arts.kuleuven.be


11th International Conference on Statistical Analysis of Textual Data (JADT 2012).


Wielfaert, Thomas, Heylen, Kris, & Speelman, Dirk. 2013. Visualisations interactives des espaces vectoriels sémantiques