Country Comparative Surveys
Using Word Embeddings

Magnus Sahlgren
RISE AI
magnus.sahlgren@ri.se
Language Effects in Surveys

Gothenburg university, RISE, Bergen university, GESIS, Södertörn university, (University of Toronto)

Funded by the Swedish Research Council
2015–2019
2018–2021

Can we use natural language processing to investigate and control for language effects in surveys?

Can we use online data as a complement to traditional opinion polls and surveys?
Language Effects in Surveys

Gothenburg university, RISE, Bergen university, GESIS, Södertörn university, (University of Toronto)

Funded by the Swedish Research Council
2015–2019
2018–2021

Can we use natural language processing to investigate and control for language effects in surveys?

Can we use online data as a complement to traditional opinion polls and surveys?
Language Effects in Surveys

Gothenburg university, RISE, Bergen university, GESIS, Södertörn university,
(University of Toronto)

Funded by the Swedish Research Council
2015–2019
2018–2021

Can we use natural language processing to investigate and control for language effects in surveys?

Can we use online data as a complement to traditional opinion polls and surveys?
Language Effects in Surveys

Gothenburg university, RISE, Bergen university, GESIS, Södertörn university, (University of Toronto)

Funded by the Swedish Research Council
2015–2019
2018–2021

Can we use natural language processing to investigate and control for language effects in surveys?

Can we use online data as a complement to traditional opinion polls and surveys?
Language Effects in Surveys

Gothenburg university, RISE, Bergen university, GESIS, Södertörn university, (University of Toronto)

Funded by the Swedish Research Council
2015–2019
2018–2021

Can we use natural language processing to investigate and control for language effects in surveys?

Can we use online data as a complement to traditional opinion polls and surveys?
Language Effects in Surveys

Gothenburg university, RISE, Bergen university, GESIS, Södertörn university, (University of Toronto)

Funded by the Swedish Research Council
2015–2019
2018–2021

Can we use natural language processing to investigate and control for language effects in surveys?

Can we use online data as a complement to traditional opinion polls and surveys?
Language Effects in Surveys

Gothenburg university, RISE, Bergen university, GESIS, Södertörn university, (University of Toronto)

Funded by the Swedish Research Council
2015–2019
2018–2021

Can we use natural language processing to investigate and control for language effects in surveys?

Can we use online data as a complement to traditional opinion polls and surveys?
Country Comparative Surveys

Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness

(Trump, Brexit...)
Country Comparative Surveys

Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness

(Trump, Brexit...
Country Comparative Surveys

Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness

(Trump, Brexit...)
Country Comparative Surveys

Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness

(Trump, Brexit...)
Alternative Methods

Rather than asking people, what about *listening* to them?

Unsolicited data from social (and other) media on the Internet

Avantages: contemporary, freely available, in vast amounts, and in many languages
Alternative Methods

Rather than asking people, what about *listening* to them?

Unsolicited data from social (and other) media on the Internet

Avantages: contemporary, freely available, in vast amounts, and in many languages
Alternative Methods

Rather than asking people, what about *listening* to them?

Unsolicited data from social (and other) media on the Internet

Avantages: contemporary, freely available, in vast amounts, and in many languages
Rather than asking people, what about *listening* to them?

Unsolicited data from social (and other) media on the Internet

Avantages: contemporary, freely available, in vast amounts, and in many languages
Social listening
Social listening
Social listening
Word Embeddings

How do people use terms?

Collections of term usage: word embeddings
(distributional semantics: words with similar distributions have similar meanings)

“Man is to computer programmer as woman is to homemaker”

Aggregate representation of communicative behavior

Categorize nearest neighbors
Word Embeddings

How do people use terms?

Collections of term usage: word embeddings
(distributional semantics: words with similar distributions have similar meanings)

“Man is to computer programmer as woman is to homemaker”

Aggregate representation of communicative behavior

Categorize nearest neighbors
Word Embeddings

How do people use terms?

Collections of term usage: word embeddings
(distributional semantics: words with similar distributions have similar meanings)

“Man is to computer programmer as woman is to homemaker”

Aggregate representation of communicative behavior

Categorize nearest neighbors
Word Embeddings

How do people use terms?

Collections of term usage: word embeddings
(distributional semantics: words with similar distributions have similar meanings)

“Man is to computer programmer as woman is to homemaker”

Aggregate representation of communicative behavior

Categorize nearest neighbors
Word Embeddings

How do people use terms?

Collections of term usage: word embeddings
(distributional semantics: words with similar distributions have similar meanings)

“Man is to computer programmer as woman is to homemaker”

Aggregate representation of communicative behavior

Categorize nearest neighbors
Word Embeddings

How do people use terms?

Collections of term usage: word embeddings
(distributional semantics: words with similar distributions have similar meanings)

“Man is to computer programmer as woman is to homemaker”

Aggregate representation of communicative behavior

Categorize nearest neighbors
Word Embeddings

How do people use terms?

Collections of term usage: word embeddings
(distributional semantics: words with similar distributions have similar meanings)

“Man is to computer programmer as woman is to homemaker”

Aggregate representation of communicative behavior

Categorize nearest neighbors
Word Embeddings
Word Embeddings

Huth et al. in *Nature*, 2016
Some technical details

Web data (social and news media) from commercial data provider

Split the data on source type (social vs. news), country, and language

CBOW from word2vec with 100 dimensions and a window of 3 tokens
Some technical details

Web data (social and news media) from commercial data provider

Split the data on source type (social vs. news), country, and language

CBOW from word2vec with 100 dimensions and a window of 3 tokens
Some technical details

Web data (social and news media) from commercial data provider

Split the data on source type (social vs. news), country, and language

CBOW from word2vec with 100 dimensions and a window of 3 tokens
Some technical details

Web data (social and news media) from commercial data provider.

Split the data on source type (social vs. news), country, and language.

CBOW from word2vec with 100 dimensions and a window of 3 tokens.
### Example 1: Democracy

<table>
<thead>
<tr>
<th>Query Term</th>
<th>Neighbor Term</th>
<th>Translated Term</th>
<th>Similarity</th>
<th>Query Term Count</th>
<th>Neighbor Term Count</th>
<th>Total Term Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>demokratie</td>
<td>gerechtigkeit</td>
<td>justice</td>
<td>0.777029</td>
<td>10400</td>
<td>4288</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>religionsfreiheit</td>
<td>religious freedom</td>
<td>0.772810</td>
<td>10400</td>
<td>702</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>freiheit</td>
<td>freedom</td>
<td>0.770844</td>
<td>10400</td>
<td>9902</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>rechtsstaatlichkeit</td>
<td>rule of law</td>
<td>0.769247</td>
<td>10400</td>
<td>1015</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>menschenwürde</td>
<td>human dignity</td>
<td>0.755094</td>
<td>10400</td>
<td>490</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>humanität</td>
<td>humanity</td>
<td>0.752865</td>
<td>10400</td>
<td>287</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>pressefreiheit</td>
<td>pressefreiheit</td>
<td>0.738619</td>
<td>10400</td>
<td>1961</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>toleranz</td>
<td>tolerance</td>
<td>0.738118</td>
<td>10400</td>
<td>2354</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>meinungsfreiheit</td>
<td>freedom of speech</td>
<td>0.733414</td>
<td>10400</td>
<td>1845</td>
<td>281763150</td>
</tr>
<tr>
<td>demokratie</td>
<td>gemeinschaft</td>
<td>community</td>
<td>0.731948</td>
<td>10400</td>
<td>6478</td>
<td>281763150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query Term</th>
<th>Neighbor Term</th>
<th>Translated Term</th>
<th>Similarity</th>
<th>Query Term Count</th>
<th>Neighbor Term Count</th>
<th>Total Term Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>демократия</td>
<td>дипломатия</td>
<td>diplomacy</td>
<td>0.865565</td>
<td>383</td>
<td>145</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>держава</td>
<td>power</td>
<td>0.859591</td>
<td>383</td>
<td>313</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>нация</td>
<td>nation</td>
<td>0.845192</td>
<td>383</td>
<td>317</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>русофобия</td>
<td>Russophobia</td>
<td>0.833827</td>
<td>383</td>
<td>98</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>религия</td>
<td>religion</td>
<td>0.818889</td>
<td>383</td>
<td>540</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>агрессия</td>
<td>aggression</td>
<td>0.814636</td>
<td>383</td>
<td>327</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>цензура</td>
<td>censorship</td>
<td>0.809182</td>
<td>383</td>
<td>150</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>монархия</td>
<td>monarchy</td>
<td>0.808148</td>
<td>383</td>
<td>90</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>идеология</td>
<td>ideology</td>
<td>0.804564</td>
<td>383</td>
<td>367</td>
<td>104616628</td>
</tr>
<tr>
<td>демократия</td>
<td>цивилизация</td>
<td>civilization</td>
<td>0.797452</td>
<td>383</td>
<td>346</td>
<td>104616628</td>
</tr>
</tbody>
</table>
## Example 2: Corruption

<table>
<thead>
<tr>
<th>Query Term</th>
<th>Neighbor Term</th>
<th>Translated Term</th>
<th>Similarity</th>
<th>Query Term Count</th>
<th>Neighbor Term Count</th>
<th>Total Term Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>korruption</td>
<td>kriminalitet</td>
<td>crime</td>
<td>0.8370269536972046</td>
<td>1780</td>
<td>2559</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>terrorism</td>
<td>terrorism</td>
<td>0.823310136795044</td>
<td>1780</td>
<td>2554</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>nepotism</td>
<td>nepotism</td>
<td>0.8047528266906738</td>
<td>1780</td>
<td>114</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>brottsighet</td>
<td>crime</td>
<td>0.8041348457336426</td>
<td>1780</td>
<td>2896</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>gängkriminalitet</td>
<td>gang crime</td>
<td>0.8011265397071838</td>
<td>1780</td>
<td>149</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>föröjelse</td>
<td>persecution</td>
<td>0.80067378292547</td>
<td>1780</td>
<td>732</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>skattoflykt</td>
<td>avoidance</td>
<td>0.788886547088623</td>
<td>1780</td>
<td>163</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>människohandel</td>
<td>trafficking</td>
<td>0.7856622934341431</td>
<td>1780</td>
<td>363</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>diskriminering</td>
<td>discrimination</td>
<td>0.781802237033844</td>
<td>1780</td>
<td>1965</td>
<td>255900188</td>
</tr>
<tr>
<td>korruption</td>
<td>antisemitism</td>
<td>antisemitism</td>
<td>0.7812042236328125</td>
<td>1780</td>
<td>641</td>
<td>255900188</td>
</tr>
<tr>
<td>corrupción</td>
<td>politiquería</td>
<td>politicking</td>
<td>0.7400001287460327</td>
<td>40875</td>
<td>988</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>criminalidad</td>
<td>criminality</td>
<td>0.7267844676971436</td>
<td>40875</td>
<td>2413</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>ilegalidad</td>
<td>illegality</td>
<td>0.718856327095032</td>
<td>40875</td>
<td>2610</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>impunidad</td>
<td>impunity</td>
<td>0.7140177488327026</td>
<td>40875</td>
<td>7451</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>corrupción</td>
<td>corruption</td>
<td>0.7114505767822266</td>
<td>40875</td>
<td>429</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>politización</td>
<td>politicization</td>
<td>0.7056220769882202</td>
<td>40875</td>
<td>307</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>nepotismo</td>
<td>nepotism</td>
<td>0.6973365545272827</td>
<td>40875</td>
<td>289</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>corruptela</td>
<td>corruption</td>
<td>0.6972599029541016</td>
<td>40875</td>
<td>204</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>violencia</td>
<td>violence</td>
<td>0.6748329401016235</td>
<td>40875</td>
<td>50370</td>
<td>287848026</td>
</tr>
<tr>
<td>corrupción</td>
<td>corrupcio</td>
<td>corruption</td>
<td>0.6687394380569458</td>
<td>40875</td>
<td>156</td>
<td>287848026</td>
</tr>
</tbody>
</table>
### Example 3: Happiness

<table>
<thead>
<tr>
<th>Query Term</th>
<th>Neighbor Term</th>
<th>Translated Term</th>
<th>Similarity</th>
<th>Query Term Count</th>
<th>Neighbor Term Count</th>
<th>Total Term Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>happy</td>
<td>excited</td>
<td>N/A</td>
<td>0.7659304738044739</td>
<td>21778</td>
<td>11216</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>thrilled</td>
<td>N/A</td>
<td>0.7493855953216553</td>
<td>21778</td>
<td>2990</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>sad</td>
<td>N/A</td>
<td>0.7033393979072571</td>
<td>21778</td>
<td>4507</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>glad</td>
<td>N/A</td>
<td>0.6982473134994507</td>
<td>21778</td>
<td>3393</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>blessed</td>
<td>N/A</td>
<td>0.6857538223266602</td>
<td>21778</td>
<td>1601</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>overjoyed</td>
<td>N/A</td>
<td>0.6749811768531799</td>
<td>21778</td>
<td>165</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>thankful</td>
<td>N/A</td>
<td>0.6693310141563416</td>
<td>21778</td>
<td>1403</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>delighted</td>
<td>N/A</td>
<td>0.666868020838672</td>
<td>21778</td>
<td>2048</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>pleased</td>
<td>N/A</td>
<td>0.6647076606750488</td>
<td>21778</td>
<td>6451</td>
<td>188775361</td>
</tr>
<tr>
<td>happy</td>
<td>nice</td>
<td>N/A</td>
<td>0.6616249084472656</td>
<td>21778</td>
<td>13880</td>
<td>188775361</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query Term</th>
<th>Neighbor Term</th>
<th>Translated Term</th>
<th>Similarity</th>
<th>Query Term Count</th>
<th>Neighbor Term Count</th>
<th>Total Term Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>lykkeilig</td>
<td>taknemmelig</td>
<td>grateful</td>
<td>0.7850871086120605</td>
<td>7168</td>
<td>5416</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>taknemlig</td>
<td>thankful</td>
<td>0.7487027645111084</td>
<td>7168</td>
<td>1031</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>glad</td>
<td>glad</td>
<td>0.7465753555297852</td>
<td>7168</td>
<td>85046</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>ulykkeilig</td>
<td>unhappy</td>
<td>0.7455624341964722</td>
<td>7168</td>
<td>1680</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>flov</td>
<td>embarrassed</td>
<td>0.743078594393921</td>
<td>7168</td>
<td>1210</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>henrykt</td>
<td>overjoyed</td>
<td>0.7327783107757568</td>
<td>7168</td>
<td>238</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>stott</td>
<td>proud</td>
<td>0.71922232608795166</td>
<td>7168</td>
<td>14055</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>frustreret</td>
<td>frustrated</td>
<td>0.7172855734825134</td>
<td>7168</td>
<td>3338</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>nervos</td>
<td>nervous</td>
<td>0.7022432380944824</td>
<td>7168</td>
<td>8347</td>
<td>330228390</td>
</tr>
<tr>
<td>lykkeilig</td>
<td>utålmødig</td>
<td>impatient</td>
<td>0.7007369664499653</td>
<td>7168</td>
<td>1032</td>
<td>330228390</td>
</tr>
</tbody>
</table>
Challenges

Is the data:
- sufficient?
- comparable?
- representative?
Challenges

Is the data:

- sufficient?
- comparable?
- representative?
Challenges

Is the data:

- sufficient?

- comparable?

- representative?
Challenges

Is the data:

- sufficient?
- comparable?
- representative?
Challenges

Is the data:
- sufficient?
- comparable?
- representative?
Summary

Word embeddings as a complement to traditional survey methods (in cases that concern how citizens conceive certain concepts)

Current research focus: validity, reliability and representativity of online data
Summary

Word embeddings as a complement to traditional survey methods (in cases that concern how citizens conceive certain concepts)

Current research focus: validity, reliability and representativity of online data
Summary

Word embeddings as a complement to traditional survey methods (in cases that concern how citizens conceive certain concepts)

Current research focus: validity, reliability and representativity of online data