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?

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?

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?

RI.
SE

Country Comparative Surveys

Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness

(Trump, Brexit...)

RI.
SE

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...)

RI.
SE

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

Alternative Methods

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

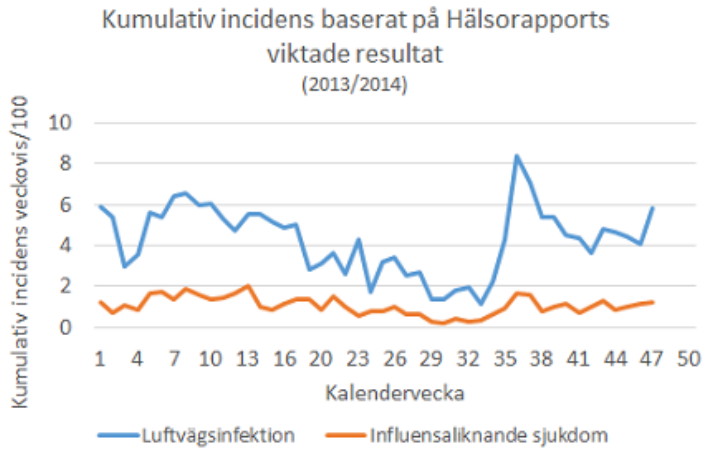
Unsolicited data from social (and other) media on the Internet

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

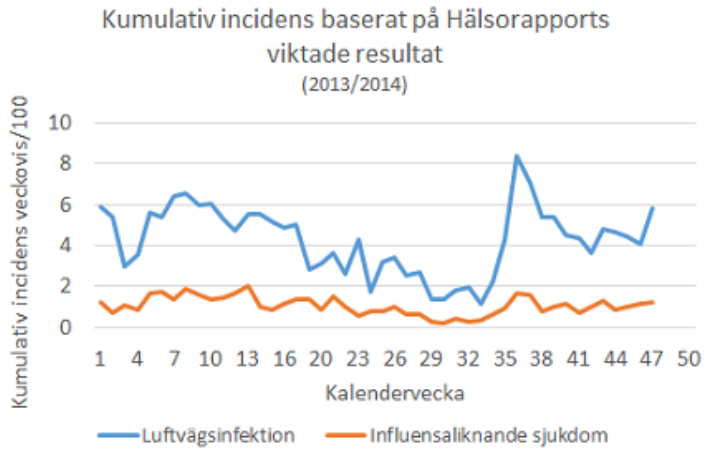
RI.
SE

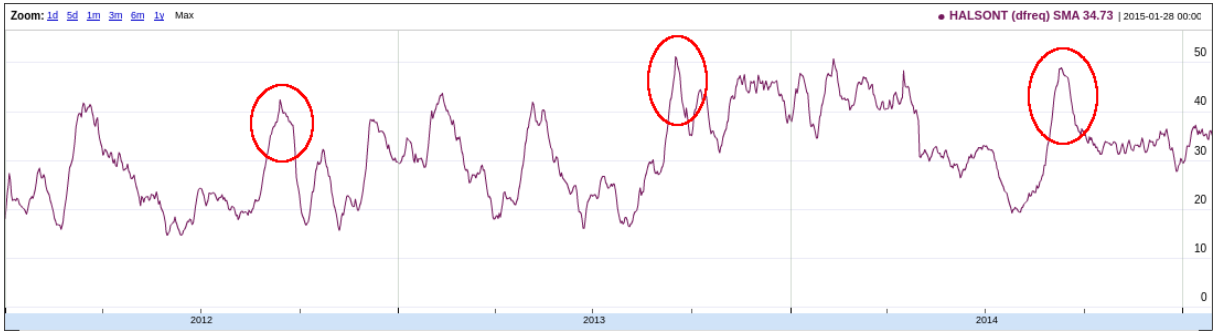
Social listening

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

RI.
SE

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

RI.
SE

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

RI.
SE

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

RI.
SE

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)

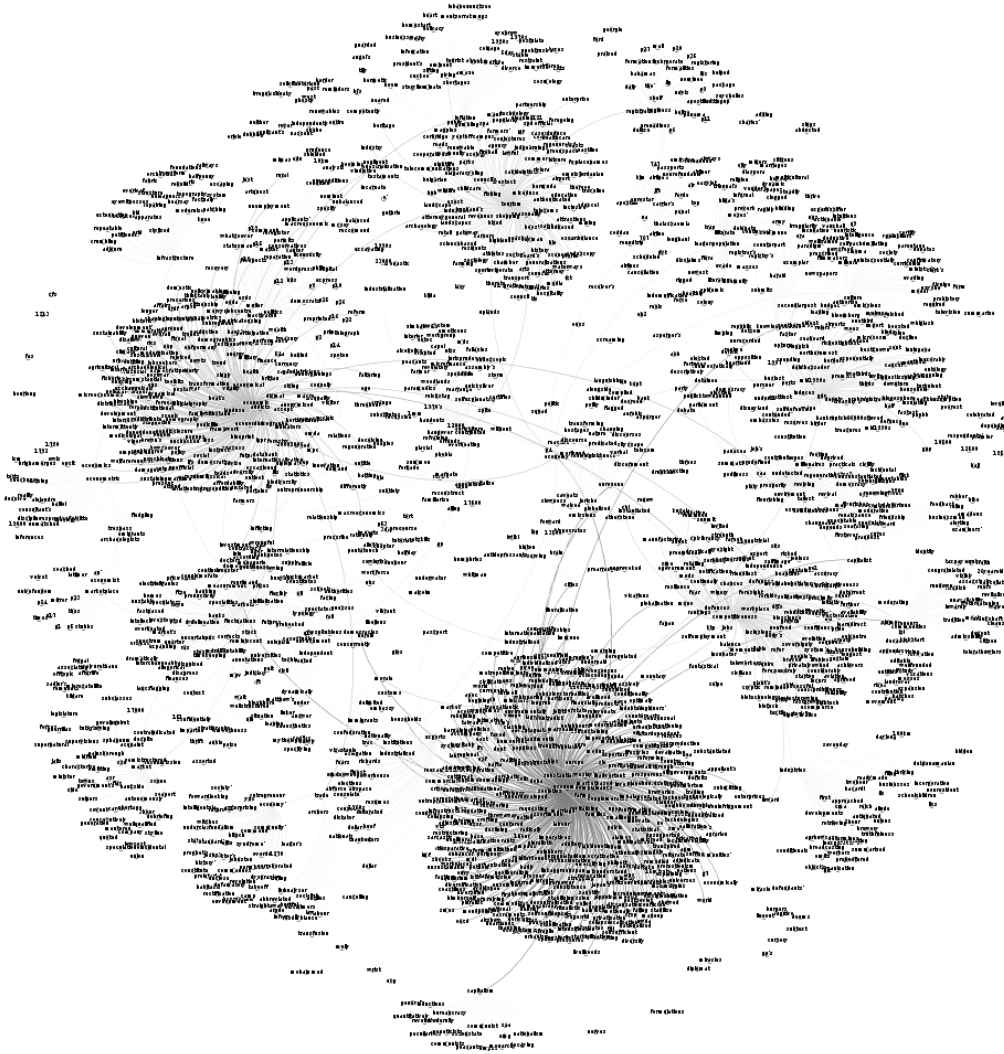
"Man is to computer programmer as woman is to homemaker"

Aggregate representation of communicative behavior

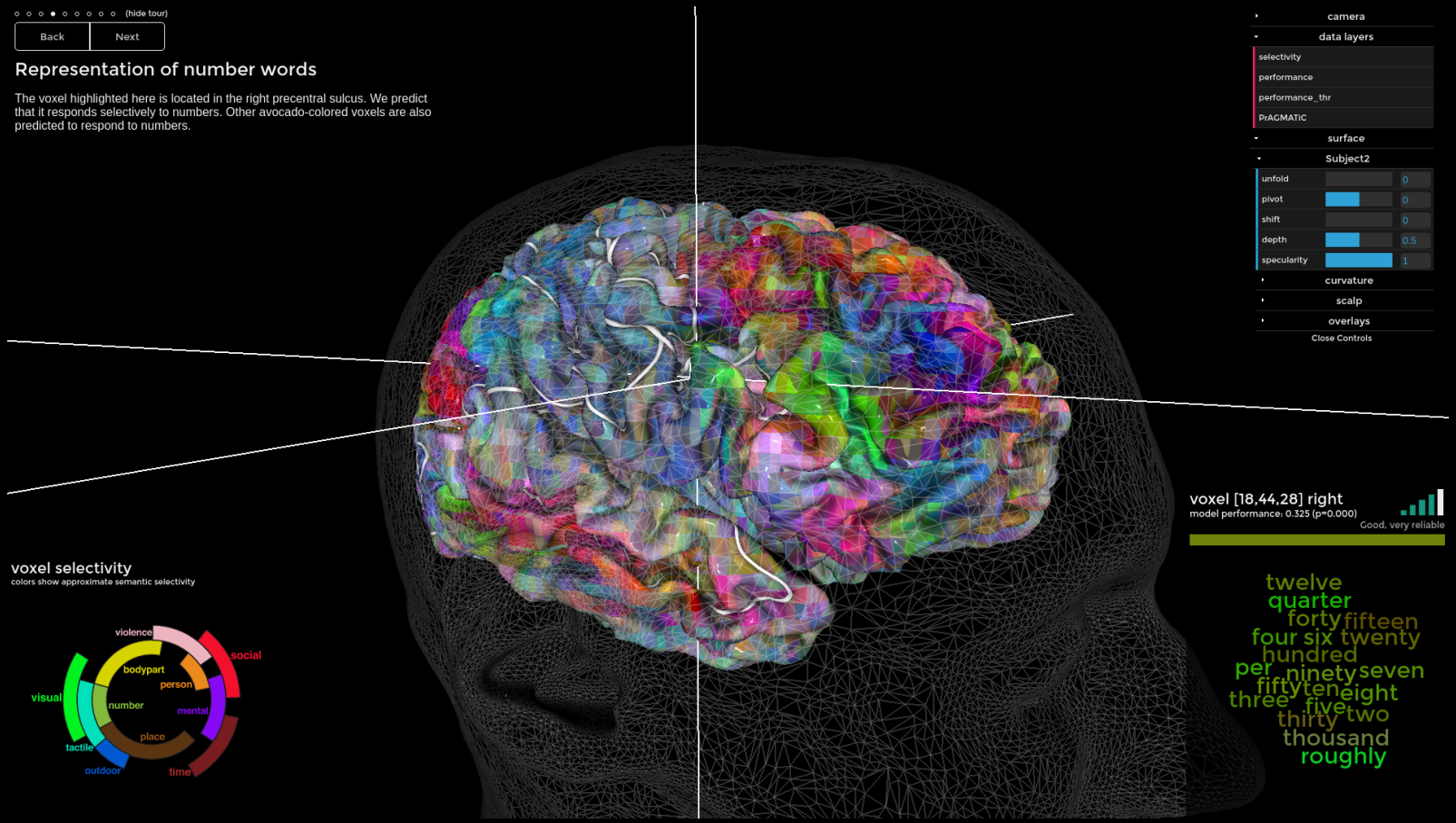
Categorize nearest neighbors

RI.
SE

Word Embeddings



Word Embeddings



Huth et al. in *Nature*, 2016

Some technical details

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

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

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

RI.
SE

Some technical details

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

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

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

RI.
SE

Some technical details

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

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 povider

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

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

RI.
SE

Example 1: Democracy

Query Term	Neighbor Term	Translated Term	Similarity	Query Term Count	Neighbor Term Count	Total Term Count
demokratie	gerechtigkeit	justice	0.7770298719406128	10400	4288	281763150
demokratie	religionsfreiheit	religious freedom	0.7728105783462524	10400	702	281763150
demokratie	freiheit	freedom	0.7708444595336914	10400	9902	281763150
demokratie	rechtsstaatlichkeit	rule of law	0.7692479491233826	10400	1015	281763150
demokratie	menschenwürde	human dignity	0.7550942301750183	10400	490	281763150
demokratie	humanität	humanity	0.7528649568557739	10400	287	281763150
demokratie	pressefreiheit	pressefreiheit	0.7386189103126526	10400	1961	281763150
demokratie	toleranz	tolerance	0.7381184697151184	10400	2354	281763150
demokratie	meinungsfreiheit	freedom of speech	0.7334141731262207	10400	1845	281763150
demokratie	gemeinschaft	community	0.73194819688797	10400	6478	281763150

Query Term	Neighbor Term	Translated Term	Similarity	Query Term Count	Neighbor Term Count	Total Term Count
демократия	дипломатия	diplomacy	0.8655654191970825	383	145	104616628
демократия	держава	power	0.8595919013023376	383	313	104616628
демократия	нация	nation	0.845192551612854	383	317	104616628
демократия	русофобия	Russophobia	0.8338273763656616	383	98	104616628
демократия	религия	religion	0.8188894391059875	383	540	104616628
демократия	агрессия	aggression	0.814636766910553	383	327	104616628
демократия	цензура	censorship	0.8091827630996704	383	150	104616628
демократия	монархия	monarchy	0.8081481456756592	383	90	104616628
демократия	идеология	ideology	0.8045649528503418	383	367	104616628
демократия	цивилизация	civilization	0.797452986240387	383	346	104616628

Example 2: Corruption

Query Term	Neighbor Term	Translated Term	Similarity	Query Term Count	Neighbor Term Count	Total Term Count
korruption	kriminalitet	crime	0.8370269536972046	1780	2559	255900188
korruption	terrorism	terrorism	0.823310136795044	1780	2554	255900188
korruption	nepotism	nepotism	0.8047528266906738	1780	114	255900188
korruption	brottslighet	crime	0.8041348457336426	1780	2896	255900188
korruption	gängkriminalitet	gang crime	0.8011265397071838	1780	149	255900188
korruption	förföljelse	persecution	0.80067378282547	1780	732	255900188
korruption	skatteflykt	avoidance	0.788886547088623	1780	163	255900188
korruption	människohandel	trafficking	0.7856622934341431	1780	363	255900188
korruption	diskriminering	discrimination	0.781802237033844	1780	1965	255900188
korruption	antisemitism	antisemitism	0.7812042236328125	1780	641	255900188

Query Term	Neighbor Term	Translated Term	Similarity	Query Term Count	Neighbor Term Count	Total Term Count
corrupción	politiquería	politicking	0.7400001287460327	40875	988	287848026
corrupción	criminalidad	criminality	0.7267844676971436	40875	2413	287848026
corrupción	ilegalidad	illegality	0.7188566327095032	40875	2610	287848026
corrupción	impunidad	impunity	0.7140177488327026	40875	7451	287848026
corrupción	corrupcion	corruption	0.7114505767822266	40875	429	287848026
corrupción	politización	politicization	0.7056220769882202	40875	307	287848026
corrupción	nepotismo	nepotism	0.6973365545272827	40875	289	287848026
corrupción	corruptela	corruption	0.6972599029541016	40875	204	287848026
corrupción	violencia	violence	0.6748329401016235	40875	50370	287848026
corrupción	corrupcio	corruption	0.6687394380569458	40875	156	287848026

Example 3: Happiness

Query Term	Neighbor Term	Translated Term	Similarity	Query Term Count	Neighbor Term Count	Total Term Count
happy	excited	N/A	0.7659304738044739	21778	11216	188775361
happy	thrilled	N/A	0.7493855953216553	21778	2990	188775361
happy	sad	N/A	0.7033393979072571	21778	4507	188775361
happy	glad	N/A	0.6982473134994507	21778	3393	188775361
happy	blessed	N/A	0.6857538223266602	21778	1601	188775361
happy	overjoyed	N/A	0.6749811768531799	21778	165	188775361
happy	thankful	N/A	0.6693310141563416	21778	1403	188775361
happy	delighted	N/A	0.6668682098388672	21778	2048	188775361
happy	pleased	N/A	0.6647076606750488	21778	6451	188775361
happy	nice	N/A	0.6616249084472656	21778	13880	188775361
Query Term	**Neighbor Term**	**Translated Term**	**Similarity**	**Query Term Count**	**Neighbor Term Count**	**Total Term Count**
lykkelig	taknemmelig	grateful	0.7850871086120605	7168	5416	330228390
lykkelig	taknemlig	thankful	0.7487027645111084	7168	1031	330228390
lykkelig	glad	glad	0.7465753555297852	7168	85046	330228390
lykkelig	ulykkelig	unhappy	0.7455624341964722	7168	1680	330228390
lykkelig	flov	embarrassed	0.7430785894393921	7168	1210	330228390
lykkelig	henrykt	overjoyed	0.7327783107757568	7168	238	330228390
lykkelig	stolt	proud	0.7192232608795166	7168	14055	330228390
lykkelig	frustreret	frustrated	0.7172855734825134	7168	3338	330228390
lykkelig	nervøs	nervous	0.7022433280944824	7168	8347	330228390
lykkelig	utålmodig	impatient	0.7007389664649963	7168	1032	330228390

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

RI.
SE