Country Comparative Surveys Using Word Embeddings

Magnus Sahlgren RISE AI magnus.sahlgren@ri.se



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?



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?



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?



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?



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?

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?

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?

Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness



Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness



Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness



Compare countries using survey questionnaires

Problem: decaying response rates

Effect: impaired representativeness



Alternative Methods

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

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



Alternative Methods

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

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



Alternative Methods

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

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



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

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



Social listening



Social listening





Social listening





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



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



How do people use terms?

Collections of term usage: word embeddings

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

Aggregate representation of communicative behavior



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



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



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



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



RI. SE



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



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



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

Split the data on source type (social vs. news), country, and language $% \left({\left[{{{\rm{social}}} \right]_{\rm{social}}} \right)$



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

Split the data on source type (social vs. news), country, and language $% \left({\left[{{{\rm{social}}} \right]_{\rm{social}}} \right)$

Example 1: Democracy

Query Term	Neighbor Term	Translated Term	ı	Similarity		Query Term Count	Neighbor Term Count	Total Term Count
demokratie	gerechtigkeit	justice		0.77702987194061	128	10400	4288	281763150
demokratie	religionsfreiheit	religious freedor	m	0.77281057834625	524	10400	702	281763150
demokratie	freiheit	freedom		0.77084445953369	914	10400	9902	281763150
demokratie	rechtsstaatlichkeit	rule of law		0.76924794912338	326	10400	1015	281763150
demokratie	menschenwürde	human dignity		0.75509423017501	83	10400	490	281763150
demokratie	humanität	humanity		0.75286495685577	739	10400	287	281763150
demokratie	pressefreiheit	pressefreiheit		0.73861891031265	526	10400	1961	281763150
demokratie	toleranz	tolerance		0.73811846971511	84	10400	2354	281763150
demokratie	meinungsfreiheit	freedom of spee	ch	0.73341417312622	207	10400	1845	281763150
demokratie	gemeinschaft	community		0.73194819688797	,	10400	6478	281763150
Query Term	Neighbor Term	Translated Term	Simi	larity	Que	ry Term Count	Neighbor Term Count	Total Term Count
Query Term демократия	Neighbor Term дипломатия	Translated Term	Simi	larity 55654191970825	Que 383	ry Term Count	Neighbor Term Count	Total Term Count 104616628
Query Term демократия демократия	Neighbor Term дипломатия держава	Translated Term diplomacy power	Simi 0.86	larity 55654191970825 95919013023376	Que 383 383	ry Term Count	Neighbor Term Count 145 313	Total Term Count 104616628 104616628
Query Term демократия демократия демократия	Neighbor Term дипломатия держава нация	Translated Term diplomacy power nation	Simi 0.86 0.85 0.84	larity 55654191970825 95919013023376 5192551612854	Que 383 383 383	ry Term Count	Neighbor Term Count 145 313 317	Total Term Count 104616628 104616628 104616628
Query Term демократия демократия демократия	Neighbor Term дипломатия держава нация русофобия	Translated Term diplomacy power nation Russophobia	Simi 0.86 0.85 0.84	Iarity Iarity 55654191970825 95919013023376 5192551612854 38273763656616	Que 383 383 383 383	ry Term Count	Neighbor Term Count 145 313 317 98	Total Term Count 104616628 104616628 104616628 104616628 104616628
Query Term демократия демократия демократия демократия	Кеідірог Тегті дипломатия держава держава нация русофобия русофобия	Translated Term diplomacy power nation Russophobia religion	Simi 0.86 0.85 0.84 0.83	larity 55654191970825 95919013023376 5192551612854 38273763656616 888894391059875	Que 383 383 383 383 383	ry Term Count	Neighbor Term Count 145 313 317 98 540	Total Term Count 104616628 104616628 104616628 104616628 104616628 104616628
Query Term Демократия демократия демократия демократия демократия	Кеlghbor Term дипломатия держава нация русофобия русофобия агрессия	Translated Term diplomacy power nation Russophobia religion aggression	Simi 0.86 0.85 0.84 0.83 0.81	Iarity 555554191970825 95919013023376 95919013023376 9 3122551612854 9 38273763656616 9 88894391059875 4 4636766910553 9	Que 383 383 383 383 383 383	ry Term Count	Neighbor Term Count 145 313 317 98 540 327	Total Term Count 104616628 104616628 104616628 104616628 104616628 104616628 104616628
Query Term демократия демократия демократия демократия демократия демократия	Neighbor Term дипломатия держава нация русофобия религия агрессия цензура	Translated Term diplomacy power ation Russophobia religion aggression censorship	Simi 0.86 0.85 0.84 0.83 0.81 0.81 0.81	Larty 556561191970825 5556511913023376 5192551612854 88894391058975 6486766910553 91827630996704 91827630996704	Que 383 383 383 383 383 383	ry Term Count	Neighbor Term Count 145 313 317 98 540 227 150	Total Term Count 104616628 104616628 104616628 104616628 104616628 104616628 104616628 104616628
Query Term демократия демократия демократия демократия демократия демократия демократия	Neighbor Term дипломатия держава нация русофобия религия агрессия цензура монархия	Translated Term diplomacy power ation Russophobia religion aggression censorship monarchy	Simi 0.86 0.85 0.84 0.83 0.81 0.81 0.81 0.81 0.80 0.80	Iarity 556541919708255 55654191970823376 5192551612854 382737363656616 88894391059875 4836766910553 91827630996704 91827630996702 81481456756592	Que 383 383 383 383 383 383 383	ry Term Count	Neighbor Term Could 145 313 317 98 540 327 150 90	Total Term Count 104616628 104616628 104616628 104616628 104616628 104616628 104616628 104616628 104616628
Query Term демократия демократия демократия демократия демократия демократия демократия демократия	Неідяbог Тегтя Дипломатия режава нация русофобия релития агрессия цензура монархия идеология	Translated Term diplomacy power atlon Russophobia religion aggression censorship censorship ideology	Simi 0.86 0.85 0.84 0.83 0.81 0.81 0.80 0.80 0.80	Larty Larty 5555-611915708255 9 95919013023376 9 5192551612854 9 88894391059875 9 963766910553 9 91827630996704 9 9141456756592 4	Que 383 383 383 383 383 383 383 383	ry Term Count	Negligibor Term Could 145 313 347 540 327 150 301 302 303	Total Term Count 104616628 104616628 104616628 104616628 104616628 104616628 104616628 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
Query Term	Neighbor Term	Translated Term	Similarity 0.7400001287460327	Query Term Count 40875	Neighbor Term Count	Total Term Count 287848026
Query Term corrupción corrupción	Neighbor Term politiquería criminalidad	Translated Term politicking criminality	Similarity 0.7400001287460327 0.7267844676971436	Query Term Count 40875 40875	Neighbor Term Count 988 2413	Total Term Count 287848026 287848026
Query Term corrupción corrupción	Neighbor Term politiquería criminalidad ilegalidad	Translated Term politicking criminality illegality	Similarity 0.7400001287460327 0 0.7267844676971436 0 0.7188566327095032 0	Query Term Count 40875 40875 40875	Neighbor Term Count 988 2413 2610	Total Term Count 287848026 287848026 287848026 287848026
Query Term corrupción corrupción corrupción	Neighbor Term politiquería criminalidad ilegalidad impunidad	Translated Term politicking criminality illegality impunity	Similarity 0.7400001287460327 0.7267844676971436 0.7188566327095032 0.7140177488327026	Query Term Count 40875 40875 40875 40875 40875	Neighbor Term Count 988 2413 2610 7451	Total Term Count 287848026 287848026 287848026 287848026 287848026
Query Term corrupción corrupción corrupción corrupción	Neighbor Term politiquería criminalidad ilegalidad impunidad corrupcion	Translated Term politicking criminality illegality impunity corruption	Similarity 0.7400001287460327 0.7267844676971436 0.7188566327095032 0.7140177488327026 0.7141505767822266	Query Term Count 40875 40875 40875 40875 40875 40875 40875	Neighbor Term Count 988 2413 2610 7451 429	Total Term Count 287848026 287848026 287848026 287848026 287848026 287848026 287848026
Query Term corrupción corrupción corrupción corrupción corrupción	Neighbor Term politiqueria criminalidad liegalidad impunidad corrupcion politización	Translated Term politicking criminality illegality impunity corruption politicization	similarity 0.7400001287460327 0.7267844676971436 0.7188566327095032 0.7140177488327026 0.7114505767822266 0.7056220769823202	Query Term Count 40875 40875 40875 40875 40875 40875 40875	Neighbor Term Count 988 2413 2610 7451 429 307	Total Term Count 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026
Query Term corrupción corrupción corrupción corrupción corrupción	Neighbor Term politiqueria criminalidad ilegalidad impunidad corrupcion politización nepotismo	Translated Term politicking cirliniality illegality impunity corruption politicization nepotism	similarity 0.7400001287460327 0.7267844676971468 0.7188566327095032 0.7140177488327026 0.714505767822266 0.7056220769882202 0.6973365545272827	Query Term Count 40875 40875 40875 40875 40875 40875 40875 40875 40875	Neighbor Term Count 988 2413 2610 7451 2629 307 289	Total Term Count 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026
Query Term corrupción corrupción corrupción corrupción corrupción corrupción corrupción corrupción corrupción	Neighbor Term politiqueria criminalidad ilegalidad limpunidad corrupcion politización nepolismo corruptela	Translated Term politicking ciriminality illegality impunity corruption politicization nepotism corruption	Similariy 0.740001287460327 0.7287844676971408 0.718856632709032 0.7140177483270204 0.71450576722260 0.705923760982202 0.6973365545278227 0.897299029541010	Query Term Count 40875 40875 40875 40875 40875 40875 40875 40875 40875 40875 40875	Neighbor Term Could 988 2413 7451 7452 747 747 748 749 749 749 749 749 749 749 749 749 749 749 749 749 749 749 749 749	Total Term Count 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026
Query Tem corrupción corrupción	Neighbor Term politiqueria criminalidad ilegalidad impunidad corrupcion politización nepotismo corrupteia violencia	Translated Term politicking ciriminality illegality impunity corruption politicization nepolism corruption violence	Similarity 0.7400001287460324 0.72678446789714804 0.718856327095032 0.714057748232704 0.71450577622264 0.714505747224204 0.713455452728227 0.8733655452728227 0.8733655452728227 0.8733655452728227 0.87325942054517 0.8743659427842784 0.87436594278427	Query Term Count 40875 40875 40875 40875 40875 40875 40875 40875 40875 40875 40875 40875 40875 40875	Neighbor Term Could 988 2413 2410 2410 2410 2421 2421 2421 2421 2432 2432 2434	Total Term Count 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 287848026 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
Query Term	Neighbor Term taknemmelig	Translated Term	Similarity 0.7850871086120605	Query Term Count 7168	Neighbor Term Count 5416	Total Term Count 330228390
Query Term lykkelig lykkelig	Neighbor Term taknemmelig taknemlig	Translated Term grateful thankful	Similarity 0.7850871086120605 0.7487027645111084	Query Term Count 7168 7168	Nelghbor Term Count 5416 1031	Total Term Count 330228390 330228390
Query Term lykkelig lykkelig lykkelig	Neighbor Term taknemmelig taknemlig glad	Translated Term grateful thankful glad	Similarity 0.7850871086120605 0.7467027645111084 0.7465753555297852	Query Term Count 7168 7168 7168	Neighbor Term Count 5416 1031 85046	Total Term Count 330228390 330228390 330228390 330228390
Query Term lykkelig lykkelig lykkelig	Neighbor Term taknemmelig taknemlig glad ulykkelig	Translated Term grateful thankful glad unhappy	Similarity 0.7850871086120605 0.7487027645111084 0.7465753555297852 0.7455624341964722	Query Term Count 7168 7168 7168 7168	Neighbor Term Count 5416 1031 85046 1680	Total Term Count 330228390 330228390 330228390 330228390 330228390
Query Term lykkelig lykkelig lykkelig lykkelig lykkelig	Neighbor Term I taknemmelig taknemlig glad ulykkelig flov	Translated Term grateful thankful glad unhappy embarrassed	Similarity 0.7850871086120605 0.7487027645111084 0.7465753555297852 0.7455624341964722 0.7430785894393921	Query Term Count 7168 7168 7168 7168 7168 7168 7168	Neighbor Term Count 5416 1031 85046 1680 1210	Total Term Count 330228390 330228390 330228390 330228390 330228390 330228390 330228390
Query Term lykkelig lykkelig lykkelig lykkelig lykkelig	Neighbor Term taknemmelig taknemlig glad ulykkelig flov henrykt	Translated Term grateful thankful glad unhappy embarrassed overjoyed	Similarity 0.7850871086120605 0.7487027645111084 0.7465753555297852 0.7455624341964722 0.7430785894393921 0.7327783107757568	Query Term Count 7168 7168 7168 7168 7168 7168 7168 7168	Nelghbor Term Count 5416 1031 85046 1680 1210 238	Total Term Count 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390
Query Term lykkelig lykkelig lykkelig lykkelig lykkelig lykkelig	Neighbor Term taknemmelig taknemlig glad ulykkelig flov henrykt stolt	Translated Term grateful thankful glad unhappy embarrassed overjoyed proud	Similarity 0.7850871086120605 0.748702745111084 0.746578355297852 0.7430785943939921 0.7327783107757568 0.7192232608795166	Outery Term Countil 7168 7168 7168 7168 7168 7168 7168	Heighbor Term Count 5416 1031 85046 1210 238 14055	Total Term Count 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390
Query Term lykkelig lykkelig lykkelig lykkelig lykkelig lykkelig lykkelig	Heighbor Termi taknenmleija taknenmlig glad ulykkelig flov henrykt stolt fustreret	Translated Term grateful thankful gled unhappy ombarrassed overjoyed proud fustrated	Similarity 0.7850871086120050 0.7487027645111084 0.7487027645111084 0.748702764106472 0.749702764010757508 0.74970278040795164 0.7182527480275442514	Auery Term Count 7168 7168 7168 7168 7168 7168 7168	Negheor Term Court 5416 1031 85046 1060 210 203 10455	Total Term Count 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390
Query Term) hykkelig hykkelig hykkelig hykkelig hykkelig hykkelig hykkelig	Heighbor Termi taknemmelig taknemlig glad ulykkelig flov henrykt stolt fusteret nervøs	Translated Term grateful thankful glad unhappy embarrassed overjoyed proud frustrated nervous	Simularity 0.785087108612005 0.74870274451104 0.74575355537582 0.7455624341964722 0.7437085494399021 0.7437783107757568 0.71723573402131 0.71723573402134	buey Term Count 7168 7168 7168 7168 7168 7168 7168 7168 7168	Neighbor Term Coult 5416 1031 85046 1040 1210 238 14055 3334 8547	Total Term Count 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390 330228390

RI SE

- sufficient?
- comparable?
- representative?



- sufficient?
- comparable?
- representative?



- sufficient?
- comparable?
- representative?



- sufficient?
- comparable?
- representative?



- 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

