* Brown Bag Series hash tag: #dlbb

Modeling semantic change with word embeddings using small historical corpora

Patrícia Amaral[1], Hai Hu[2], Sandra Kübler[2]

1: Department of Spanish and Portuguese
2: Department of Linguistics

Indiana University Bloomington

March 3, 2021 @ IDAH Brown Bag

Outline

- Introduction: semantic change
- Previous work:
  - Intuition-based
  - Corpus-based
- Meaning as vectors: word embeddings
- Our situation:
  - Small historical corpus
  - No evaluation data for medieval Spanish
- Our methods and results:
  - Embedding models
  - Neighbors
  - Assessing the models
- Conclusion and future work

Semantic change

- Word meaning changes over time (*semantic change*).

- A recent example from English (Hamilton et al. 2016, a.o.):

(1) She was a fine-looking woman, cheerful and **gay** (1900, Davies 2010)

(2) “I don’t personally support **gay** marriage myself”, Edwards said. (2000, Davies 2010)

Investigating semantic change: methods

- The most prominent theories of semantic change rely on a limited number of corpus data (going through **concordances**) and on the linguist's **intuition**.
- The semanticist **detects** different meanings in different synchronies of a language and produces hypotheses for possible contexts triggering change (Traugott and Dasher 2002, Evans and Wilkins 2000, Eckardt 2006).
- Another approach: **using the word's distribution** to trace its semantic change (Sagi et al. 2012, Hamilton et al. 2016, Boleda 2020, a.o.).
- Advantages: **data-driven, replicable, measurable**.

Computational work on semantic change

Computational studies of semantic change usually rely on large corpora (Hamilton et al 2016).

What about **smaller corpora** which are more common in historical work?

Challenges: 1) small corpus, 2) multiple/ancient spellings; 3) hard to find balanced corpus.

Study the Spanish word *algo*: (Amaral 2016)

- Medieval: *goods/possessions* (noun), *something* (pron.)
- Now: *a bit* (adv.), *something* (pron.)

Comparing our results will a previous study will allow us to evaluate the methods and apply them in the future to new cases.

Our research questions

1. Replicability: How replicable are the results of word embeddings models?
2. Accuracy: How do we determine the accuracy of the different embedding models on our **medieval** data?
3. Usability of embeddings: Given the **low-resource** situation, can we find (enough) meaningful words in the embeddings to draw conclusions about semantic change?

Word embeddings

Word embeddings

- Word vectors: using a vector to represent a word
- Intuition:
- “If A and B have almost identical **environments** we say that they are synonyms.” (Harris 1954)
- “You shall know a word by the company it keeps” (Firth 1957)

What is ***tesgüino***?

A bottle of ***tesgüino*** is on the table
Everybody likes ***tesgüino***
Tesgüino makes you drunk
We make ***tesgüino*** out of corn.

Example from: http://web.stanford.edu/~jurafsky/li15/lec3.vector.pdf

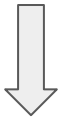
Word embeddings

- Word vectors: using a vector to represent a word

- Count the freq of *all words* in context (±2 words) of *dog*

corpus

I have a [*very cute* ***dog*** *that plays*] with me every day.
… [*, my* ***dog*** *likes stairs*] ...



Word-word freq matrix

	...	cute	play	...	stairs	vacuum	...
dog		30	45		5	0	
cat		48	50		4	0	
ascend		4	5		30	0	

Word embeddings

- Each word represented by vectors of length |V| (30000)
- dog = (... 30, 45, 5, 0 ...)
- cat = (... 48, 50, 4, 0 ...)
- ascend = (... 4, 5, 30, 0 ...)

  → using linear algebra, we know:

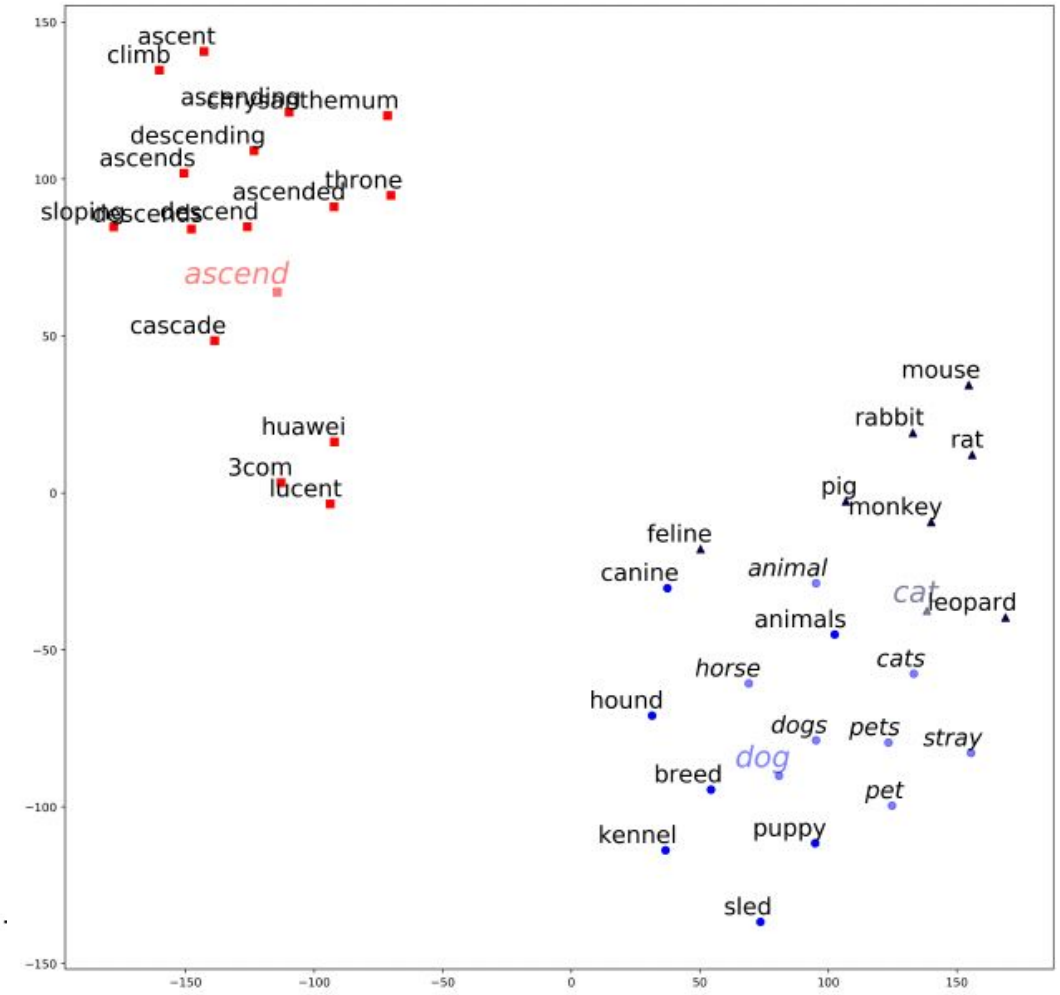
  vectors for *dog* and *cat* are more similar

- In reality:
  - People use Pointwise Mutual Information (PMI) matrix
  - Dimension reduction (PCA, SVD, etc) → 300 dimensions
  - More popular algorithms: **word2vec**

Word embeddings

- Reduce to 2 dimensions for visualization
- *dog* and *cat* are more similar

Semantic Space

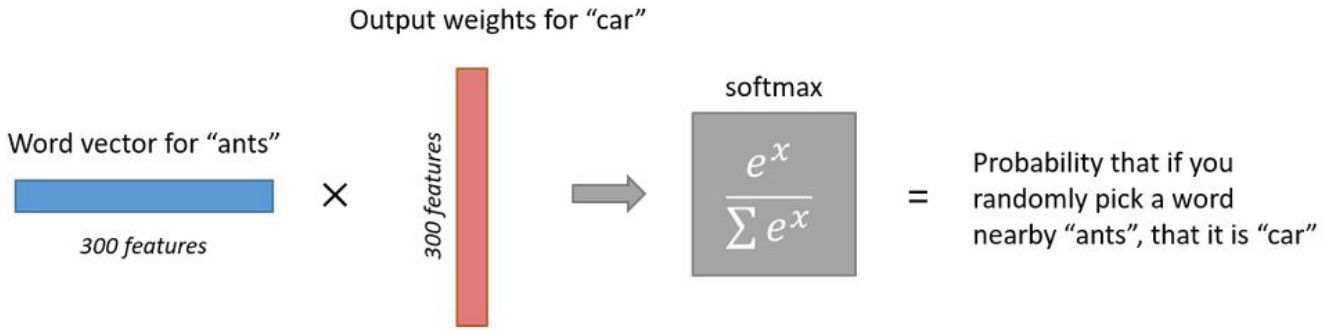


Word embeddings

- **Word2vec** (Mikolov et al. 2013a; 2013b)
- Use a ML model for placeholder task:
- For word *w*, predict whether another word is in its context

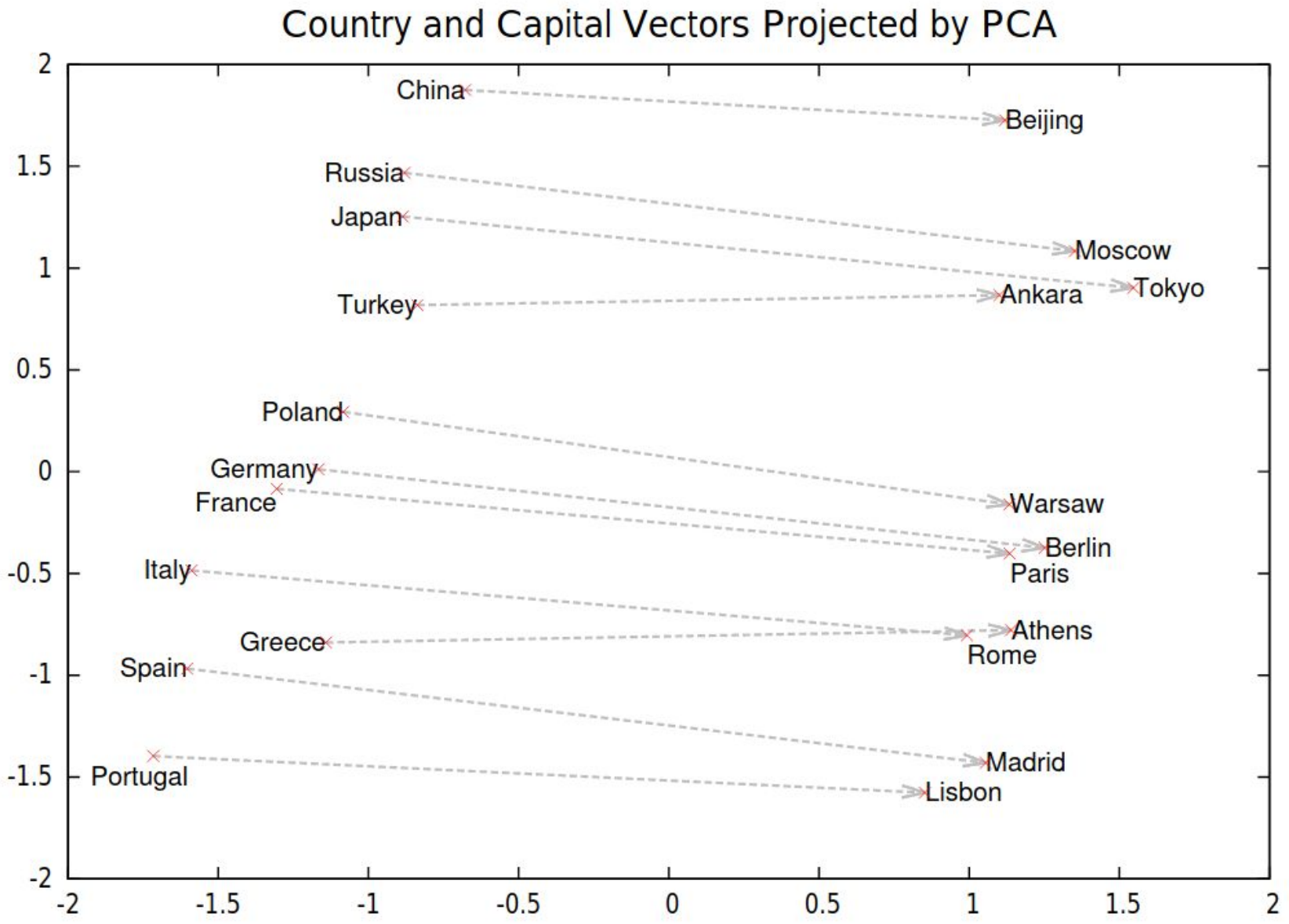
> I have a [*very cute* ***dog*** *that plays*] with me every day.
> … [*, my* ***dog*** *likes stairs*] ...

- Positive: (dog, very), (dog, cute), (dog, that), (dog, plays)
- Negative: (dog, green)



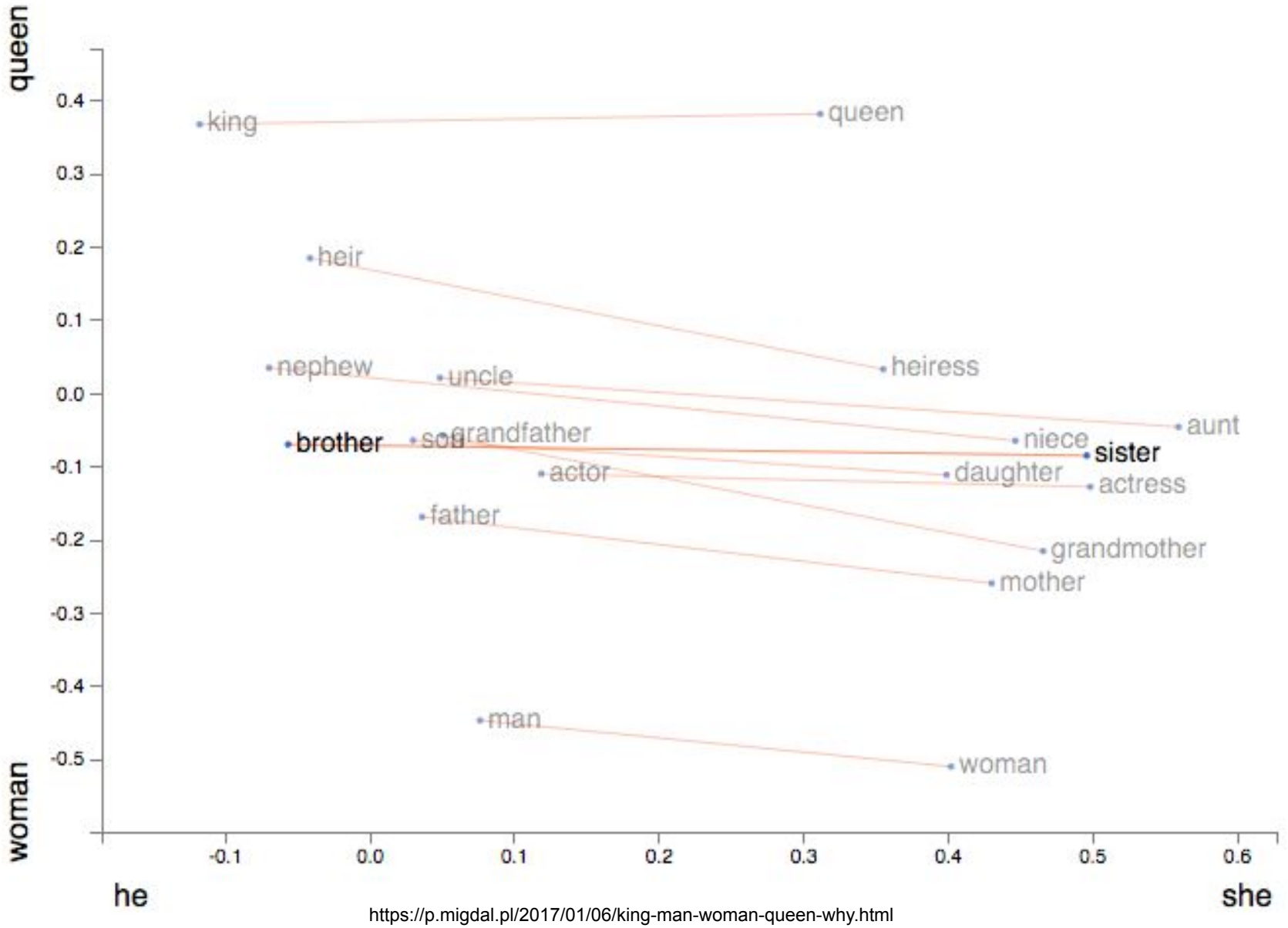
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word embeddings capture analogy relations!



https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

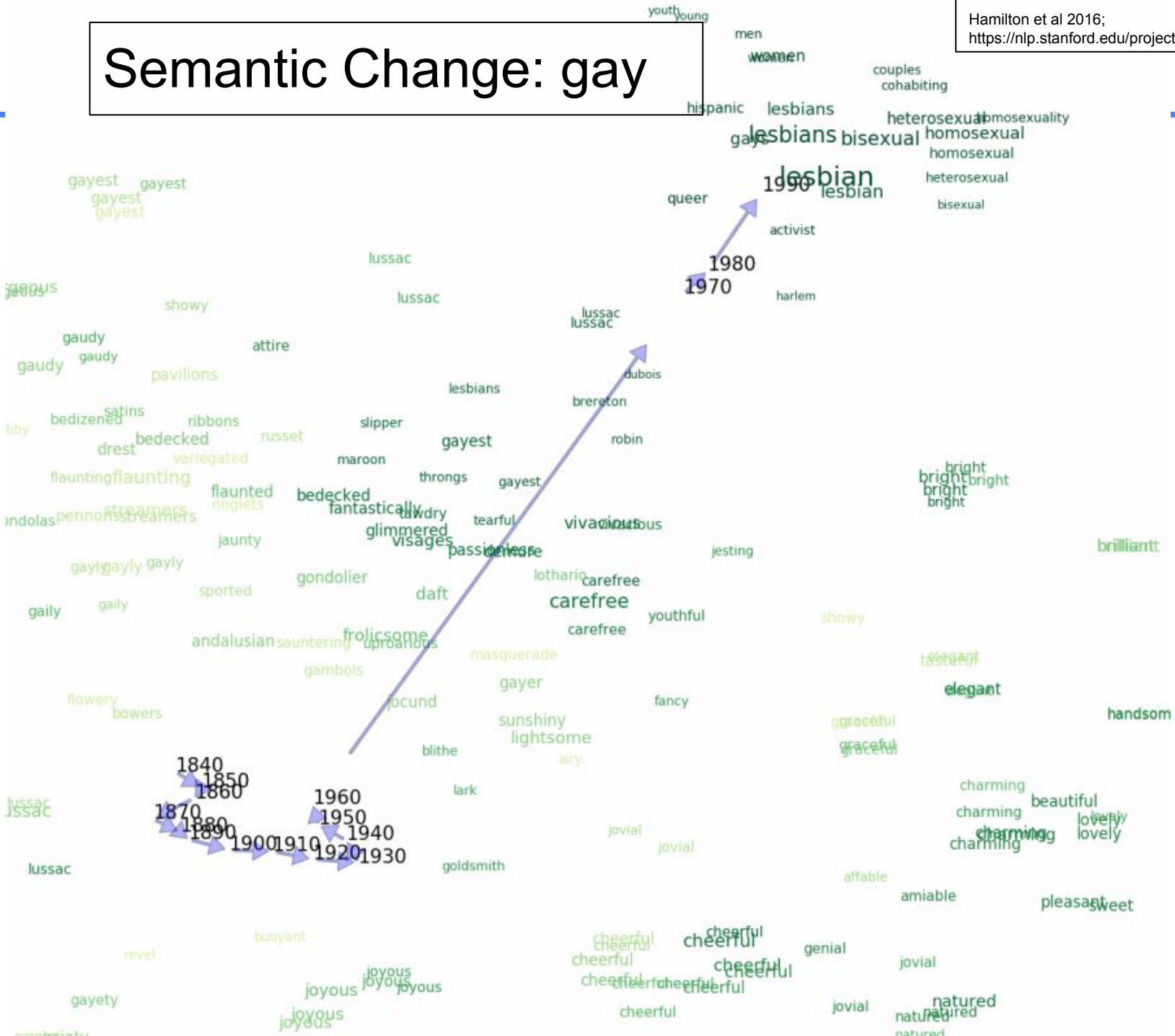
Word embeddings capture analogy relations!



https://p.migdal.pl/2017/01/06/king-man-woman-queen-why.html

Semantic Change: gay

Hamilton et al 2016;
https://nlp.stanford.edu/projects/histwords/



Methods

Data

Chronicles corpus	Spanish Billion Words corpus
1300s-early 1500s	Contemporary
7 million tokens	1.5 billion tokens
Narrative texts	Various genres
Annotated medieval manuscripts	Wikipedia, Europarl, etc.
Digital Library of Old Spanish Texts http://hispanicseminary.org/	https://crscardellino.github.io/SBWCE/

Methods

- SGNS:
  - Skip-Grams with Negative Sampling (Mikolov et al 2013); gensim lib by Řehůřek
- GloVe:
  - original implementation (Pennington et al 2014)
- $SVD_{PPMI:}$
  - word-word PPMI table → SVD; hyperword implementation (Levy et al 2015)

Methods

- Replicability:
  - Run SGNS and GloVe three times, with different seeds.
  - Obtain the neighbors of *algo*, ranked by cosine similarity
  - Compute the mean rank for three runs, and for three algorithms
- Accuracy:
  - Create an analogy test tailored for medieval Spanish
  - Make sure the words show up in medieval corpus
- Usability:
  - Analyze the neighbors
  - Compare with previous work on *algo*

Results

Results: accuracy → need for tailored analogy tests

MTS: the Mikolov et al (2013) test Translated to Spanish

→ only 331 out of the 14k examples valid for medieval corpus

Medieval:

Analogy test	# total	# valid	# correct			% correct		
			SVD_{PPMI}	SGNS	GloVe	SVD_{PPMI}	SGNS	GloVe
MTS	14 764	331	47	47	31	14.2	14.2	9.4
MTS+ours	9 220	4 950	1 178	1 351	985	23.8	27.3	19.9

Contemporary:

Analogy test	# total	# valid	# correct			% correct		
			SVD_{PPMI}	SGNS	GloVe	SVD_{PPMI}	SGNS	GloVe
MTS	14 764	14 672	4725	7 424	7 805	32.2	50.6	53.2
MTS+ours	9 220	8 146	2038	3 720	3 486	25.0	45.7	42.8

Results: replicability--contemporary corpus

- Top 10 neighbors of *algo* in contemporary corpus
- Stable across different seeds, and algorithms
- 1.5 billion tokens

Algorithm seed	Analogy accuracy	Most similar words
		SBW corpus
SGNS 1	45.7	nada, realmente, eso, mucho, bastante, cosa, alguien, porque, aspecto, demasiado
SGNS 2	44.6	nada, realmente, mucho, eso, bastante, cosa, alguien, porque, aspecto, tan
SGNS 3	45.6	nada, realmente, eso, mucho, bastante, cosa, alguien, porque, tan, demasiado
GloVe 1	42.8	nada, eso, parece, cosa, mucho, poco, parecido, bastante, cierto, porque
GloVe 2	43.1	nada, eco, parece, bastante, poco, mucho, porque, cosa, realmente, cierto
GloVe 3	43.5	nada, eso, parece, mucho, alguien, cosa, proque, bastante, poco, realmente
SVD_{PPMI}	25.0	nada, eso, mucho, pensar, realmente, imaginarlo, bueno, quizá, parece, quizás

Results: replicability--contemporary corpus

SGNS		GloVe		SVD_{PPMI}		Three algorithms		
word	*Rank*	word	*Rank*	word	*Rank*	word		*Rank*
nada	1	nada	1	nada	1	nada	'nothing'	1
realmente	2	eso	2	eso	2	eso	'that'	2.57
eso	3.33	parece	3	mucho	3	mucho	'much'	4.14
mucho	3.67	mucho	5	pensar	4	realmente	'really'	6.57
bastante	5	cosa	6	realmente	5	bastante	'enough'	8.14
cosa	6	poco	6.67	imaginarlo	6	parece	'it seems'	10.14
alguien	7	bastante	6.67	bueno	7	porque	'because'	11.43
porque	8	porque	8	quizá	8	alguien	'someone'	13.43
aspecto	10	alguien	10	parece	9	quizás	'maybe'	13.71
tan	10	cierto	11	quizás	10	tan	'so much'	13.86

(a) *Rank* of words trained on the **SBW** corpus.

- *algo* = "a bit" / "something" (contemporary Spanish)
- No neighbors meaning "possessions"
- But we have more adverbs ("a bit" is an adverb)

Results: replicability--medieval corpus

Top 10 neighbors for *algo* in Chronicles:

Algorithm seed	Analogy accuracy	Most similar words
		Chronicles corpus
SGNS 1	27.3	aueres, pro, heredades, demas, mayordomos, ualiesse, dones, soldadas, prometer, heredamjentos
SGNS 2	28.1	aueres, abenagit, dones, criar, heredades, demas, pro, mayordomos, soldadas, conducho
SGNS 3	25.5	aueres, demas, pro, soldadas, ualiesse, criar, dones, abondados, enbiolos, en-bargo
GloVe 1	19.9	demas, auer, nada, quanto, farie, dones, dar, daua, pan, comer
GloVe 2	19.8	dones, sabor, demas, quanto, ganar, auer, aueres, comer, pan, nada
GloVe 3	20.2	auer, ganar, quanto, dones, nada, demas, dar, darie, farie, sabor
SVD_{PPMI}	23.8	nada, mester, quanto, dar, ello, demas, gelos, gelo, recabdo, rentas

- Not very stable for SGNS (5 overlap) and GloVe (5 overlap)
- SVD_{PPMI} is quite different from SGNS and GloVe
- There is a need for an averaging method

Results: replicability--medieval corpus

SGNS		GloVe		SVD_{PPMI}		Three algorithms		
word	*Rank*	word	*Rank*	word	*Rank*	word		*Rank*
aueres	1	auer	3	nada	1	demas	'more'	4
pro	4	demas	3.33	mester	2	dones	'goods'	6.14
demas	4	quanto	3.67	quanto	3	nada	'nothing'	11
dones	5.67	dones	3.67	dar	4	aueres	'possessions'	12.57
heredades	6.67	nada	6	ello	5	auer	'possessions'	15.71
soldadas	7	ganar	6	demas	6	sabor	'wish'	29.43
criar	7	dar	8.67	gelos	7	dioles	'gave them'	32.57
ualiesse	8.67	sabor	8.67	gelo	8	precio	'price'	44.86
abenagit	9.33	farie	9	recabdo	9	dineros	'monies'	46.14
mayordomos	10.67	delo	13.33	rentas	10	pro	'gain'	46.57

(b) *Rank* of words trained on the **Chronicles** corpus.

fijos de algo
fijos de Abenagit

- We also rank across the algorithms (SVD_{PPMI} is deterministic)
- Mean rank is high in last column b/c too much variance among 3 algorithms

Results: Usability

- Medieval: 6 out of 10 neighbors: words related to property and value, consistent with previous work showing *algo* = 'possessions'

- Contemporary: no 'possessions' (noun), but adverbs

- Both medieval and contemporary: *nada*: indefinite pronoun

→ Corroborates previous work (Amaral, 2016).

Conclusion and future work

- Embeddings work for small corpus
  - But larger variation in the results
  - May need methods for stabilization/removing noise
  - E.g., multiple runs, taking average
- Tailored evaluation set needed for historical data
- For *algo*, our results corroborate w/ previous work.

Future work:

- Normalize spelling
- Obtain a more balanced corpus

Thank you!
Questions and comments are welcome!

{huhai,pamaral,skuebler}@indiana.edu

Backup slides

Pre-processing Chronicles

Before:

[fol. 1r]
{CB1.
{IN5.} Estas canonicas fizo escribir el Reuerent
en lh<es>u xp<ist>o padre don fray garcia de Eugui ob<is>po
de Bayona delos fechos que fuero<n> fechos anti-
gament en espan~a segunt se trueba por sc<r><<i>>pto
en diuersos libros antigos & por que mellor se
p<ar>ta deuedes saber que los sabios antigos p<ar>tiero<n> todos los t<iem>pos
pasados despues que dios formo ad adam en vj hedades et
por esto aqui digamos que cosa es hedat. Et Responden los
sabios antigos que antigam<en>t qu<an>do porel mundo achaesc'ia
algun grant. fecho estrayn~o que nu<n>qua oviesse achaec'ido
fazien enel dep<ar>timj<en><<to>>. del t<iem>po hedat & clamaua<n> hedat al t<iem>po
pasado & exo mesmo clamaua<n>. hedat al t<iem>po por venir et
agora digamos dela p<r><<i>>mera hedat & qu<an>tos an~os turo.
{RUB. La p<r><<i>>mera hedat}
{IN3.} Deuedes saber. que la p<r><<i>>mera hedat enpesco
qu<an>do n<uest>ro sen~or dios creo el mundo et formo. a adam
et turo esta p<r><<i>>mera hedat fasta el diluuio que noe
/ por mandamj<en><<to>>. de n<uest>ro s<en><<or>>. dios se puso. con. sus. iij. fillos et
con sus mulleres en larq<u><<a>>. et fuero<n>. por todos. viij<<o>>. p<er>sonas et
ouo en esta p<r><<i>>m<<e>><r>a hedat. segunt la biblia que oy es et segu<n>t
el conto que fazen los judios mil dc.l.vj. an~os. mas.
segunt. los. lxx. jnterpretadores. dela ley. obo. ij. mil.cc.
l<<o>>. an~os destos. lxx. jnterpretadores dela ley dezir sea aua<n>t

Pre-processing Chronicles

After:

amen estas canonicas fizo escribir el reuerent en ihesu xpisto padre don fray garcia de eugui obispo de bayona delos fechos que fueron fechos antigament en españa segunt se trueba por scripto en diuersos libros antigos y por que mellor se parta deuedes saber que los sabios antigos partieron todos los tiempos pasados despues que dios formo ad adam en vj hedades et por esto aqui digamos que cosa es hedat et responden los sabios antigos que antigament quando porel mundo achaesc'ia algun grant fecho estrayño que nunqua oviesse achaec'ido fazien enel departimjento del tiempo hedat y clamauan hedat al tiempo pasado y exo mesmo clamauan hedat al tiempo por venir et agora digamos dela primera hedat y quantos años turo deuedes saber que la primera hedat enpesco quando nuestro señor dios creo el mundo et formo a adam et turo esta primera hedat fasta el diluuio que noe por mandamjento de nuestro senor dios se puso con sus iij fillos et con sus mulleres en larqua et fueron por todos viijo personas et ouo en esta primera hedat segunt la biblia que oy es et segunt el conto que fazen los judios mil dc.l.vj años mas segunt los lxx jnterpretadores dela ley obo ij mil.cc lo años destos lxx jnterpretadores dela ley dezir sea auant

Removed paleographical annotations.
Lower-cased all words.
Removed punctuations.

Analogy tests for medieval Spanish

- MTS+ours, tailored for medieval Spanish:
  - Selected the MTS + those created by us
  - Total: 9220 questions

Source	Category	Example	#Questions
MTS	Morphology nouns: kinship terms	padre madre : hijo hija	506
	Morphology verbs: third person singular	comer come : ir va	650
	Morphology verbs: infinitive to participle	saber sabido : tomar tomado	1190
	Morphology verbs: gerund to participle	sabiendo sabido : tomando tomado	1190
ours	Morphology adj.: singular to plural	negra negras : rica ricas	992
	Morphology adj.: singular to plural	negro negros : rico ricos	992
	Morphology adj.: masc to fem	negro negra : negros negras	992
	Morphology adj.: masc to fem	negros negras : ricos ricas	992
	Morphology nouns : singular to plural	casa casas: capilla capillas	1332
	Morphology/Semantics: antonyms	feliz infeliz : posible imposible	42
	Semantics: antonyms	cerca lejos : bien mal	342
Total			9220