automatically discovering cultural polarities in Ancient Greek word embeddings

Nick Gardner
Classics, CS

background on word embeddings

•word embeddings represent words as vectors of numbers

•think of each word as a point in 300 dimensional space; the individual points have no intrinsic meaning, but the relations between points (distance, direction) encode many aspects of the meaning-relations between the words the points represent

•overview by Jurafsky and Martin: https://web.stanford.edu/~jurafsky/slp3/6.pdf

•**key:** word embeddings capture cultural stereotypes, biases, and other associations

useful for historians, anthropologists trying to understand cultural worlds different from their own

- most work with word embeddings has focused on (“Standard”) English, but the techniques can be used to model any language/dialect, past or present, given sufficient text data for training

- since word embeddings are generated in an **unsupervised** manner from text alone, constructing them doesn’t require any pre-judgment/interpretation on the part of the researcher. The work of interpretation comes after the embeddings have been constructed by the algorithm--the researcher can then work to uncover the cultural associations that are encoded in the structure of the word embeddings

English word embeddings encode gender stereotypes

from Bolukbasi et al 2016

Figure 7: Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words *he* and *she*, and y is a direction learned in the embedding that captures gender neutrality, with gender neutral words above the line and gender specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line.

data/model used to train Ancient Greek embeddings

Data

- 25,522,507 POS tagged tokens in 1,384,550 sentences (84% of tokens lemmatized with high confidence)
- thanks to Giuseppe Celano and the Perseus Project (https://github.com/gcelano/LemmatizedAncientGreekXML)

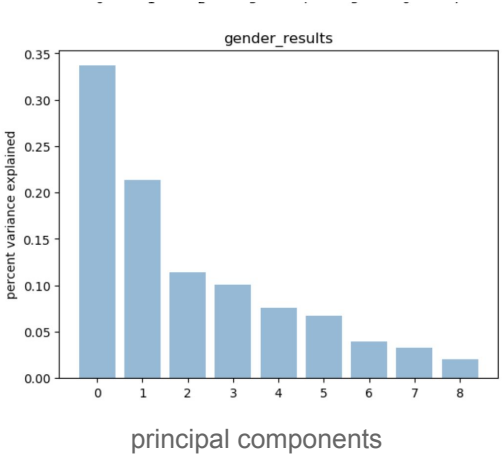
Word embedding algorithm and model parameters

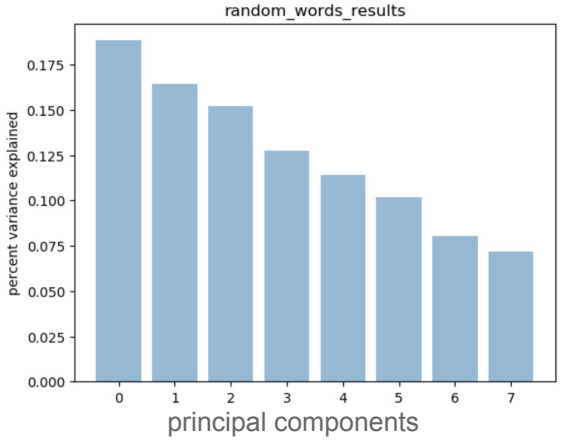
- word2vec SGNS trained on lemmatized text, 300 dimensions, context window size of 10

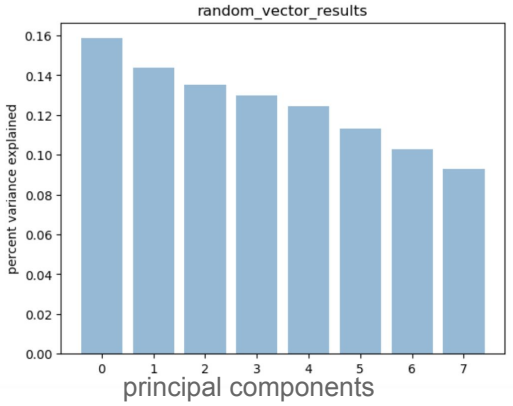
gender contrast pairs (replicating Bolukbasi et al. 2016 in Ancient Greek context)

('μήτηρ', 'πατήρ')	mother, father
('ἀδελφή', 'ἀδελφός')	brother, sister
('παρθένος', 'νεανίσκος')	“maiden”, young man
('παρθένος', 'μειράκιον')	“maiden”, “lad”
('νύμφη', 'ἀνήρ')	young wife, man/husband
('θεράπαινα', 'θεράπων')	female servant, male servant
('γυνή', 'ἀνήρ')	woman, man/husband
('θυγάτηρ', 'υἱός')	daughter, son
('κόρη', 'νεανίσκος')	young woman, young man

gender contrast pairs



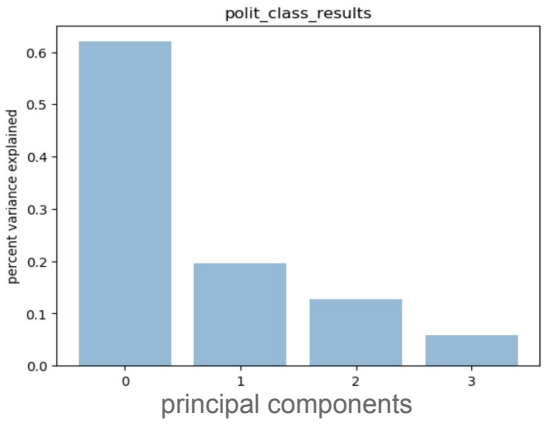


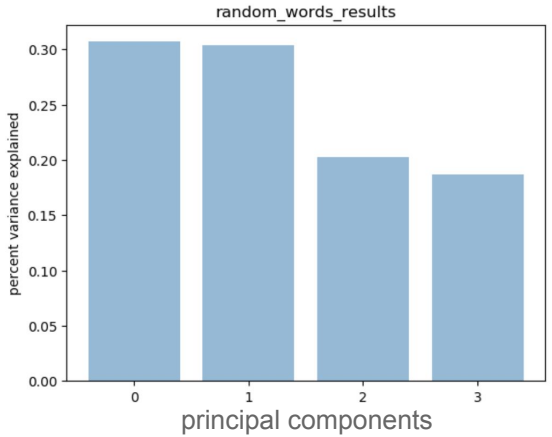


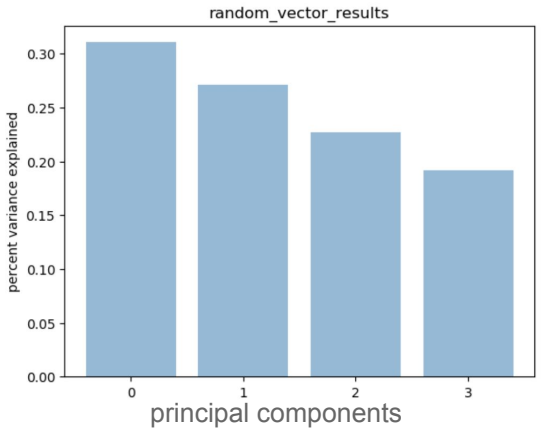
political contrast pairs

('δημοκρατία', 'ὀλιγαρχία')	democracy, oligarchy
('δῆμος', 'ὀλίγος')	the people, the few/elite
('δῆμος', 'γνώριμος')	the people, the eminent
('δημοτικός', 'ὀλιγαρχικός')	popular, oligarchic

political contrast pairs



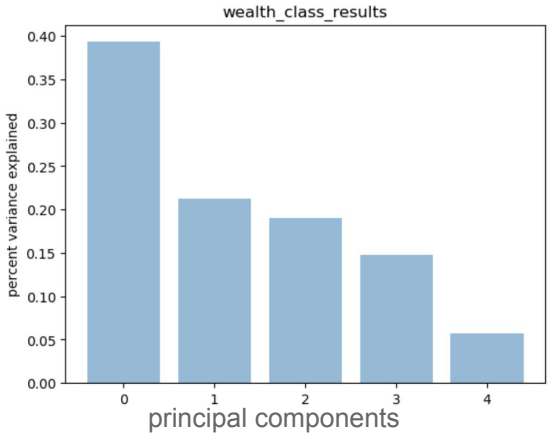


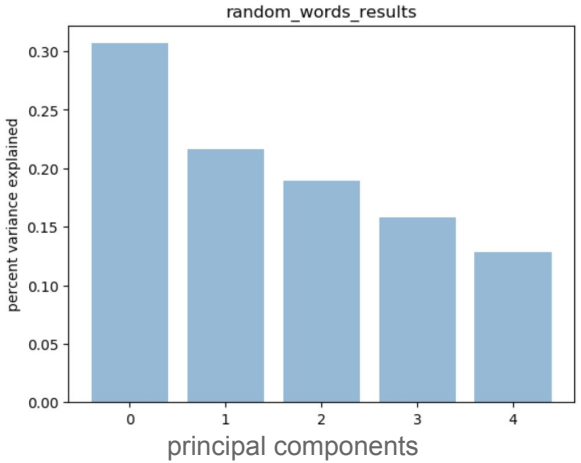


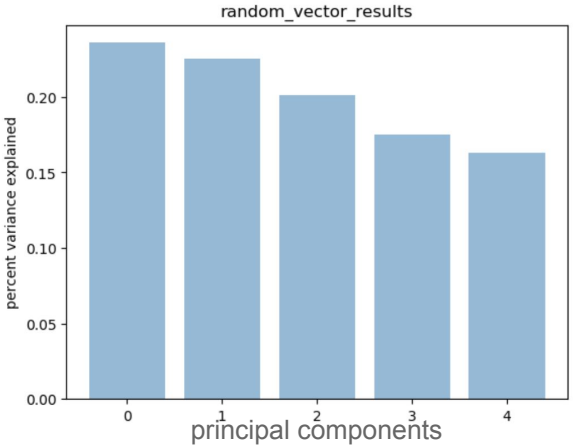
wealth contrast pairs

('πλούσιος', 'πένης')	wealthy, poor
('πλουτέω','πένομαι')	be wealthy, be poor/toil
('πλουτίζω','ἄνολβος')	make wealthy, poor
('πλοῦτος', 'πενία')	wealth, poverty
('πλούσιος', 'πτωχός')	wealthy, beggar

wealth contrast pairs



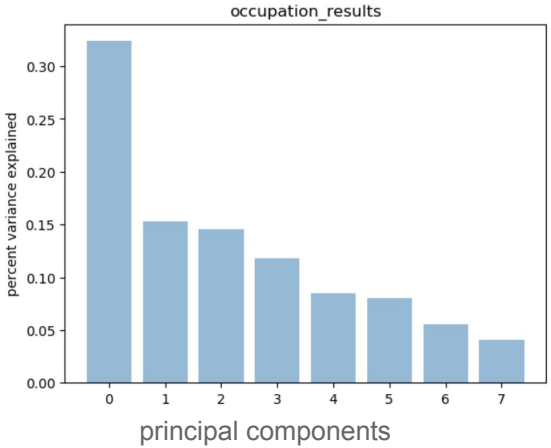


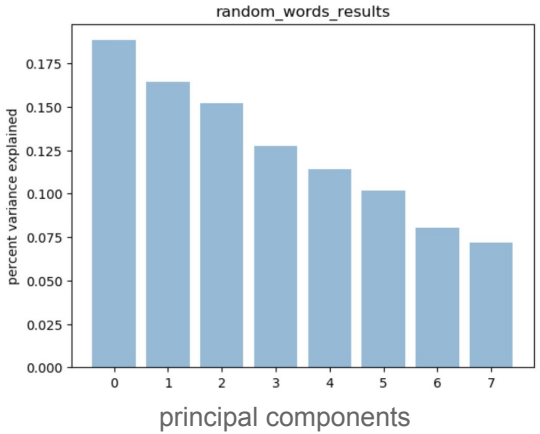


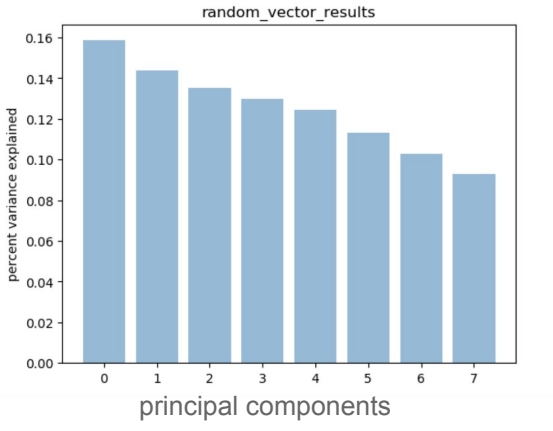
political activity vs. work contrast pairs

('σχολή', 'ἀσχολία'),	leisure, busy-ness/work
('ἐλεύθερος', 'ἀνελεύθερος'),	free, unfree
('πολίτης', 'τεχνίτης'),	citizen, technician
('πόλις', 'ἐργαστήριον'),	city-state, workshop
('βουλή', 'ἐργαστήριον'),	council, workshop
('ἀγορεύω', 'ἐργάτης'),	speak (in public), laborer
('βουλευτής', 'σκυτοτόμος'),	councillor, shoemaker
('πολιτεύω','μισθόω')	participate in politics, earn a wage

political activity vs. work contrast pairs







automatically discovering cultural polarities in word embedding spaces

•polar concepts within a given culture (like a binary gender spectrum, poor vs. rich, hot vs. cold) are encoded in pairs of words that are similar to each other in every dimension of meaning except the dimension representing the polarity in which they differ (Standard English “father” and “mother” differ in meaning primarily in a binary gender dimension)

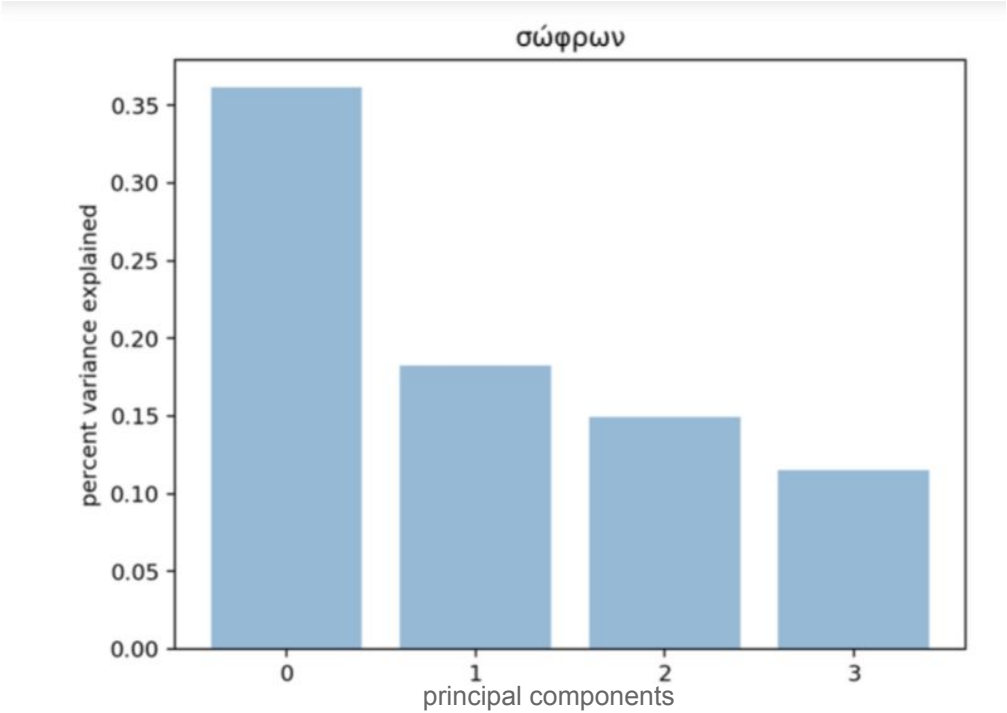
•since such word pairs have similar meanings in all but one dimension, they will **end up near one another in the word embedding space**

•so it should be possible to **automate discovery of polarities** by systematically comparing the nearest neighbors of words (at least words that are likely to participate in some cultural/semantic contrast relationship with other words)

automatically discovering cultural polarities in word embedding spaces

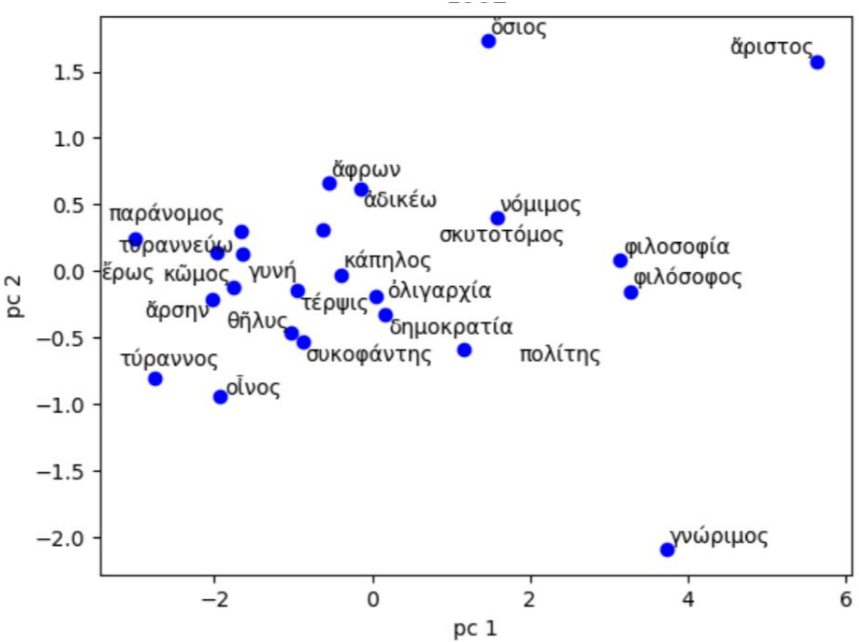
- one way to operationalize this idea:

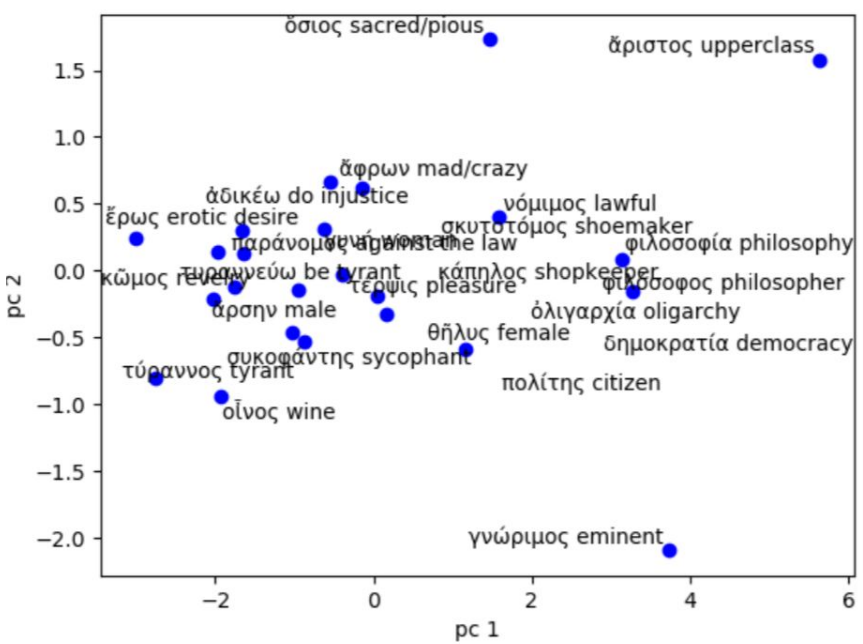
> Given a seed word *w* likely to participate in a cultural polarity, form a set *S* of the *n* nearest neighbors of *w* in the embeddings space (including *w* itself, and with *n* even), and then compute the PCA of all possible complete pairings of the words in *S* (i.e. each complete pairing consists of *n/2* pairs). Return a list of these PCA results and associated word pairings, ranked by the strength (percent variance explained) of the first principal component.



Shows percentage of variance explained by the highest ranking top 4 principal components returned by PCA on 7 difference vectors formed from pairings of the 14 words nearest to *sofron* 'temperate, self-controlled' (including *sofron* itself among them). Here "highest ranking" means that this pairing of the 14 words and their associated difference vectors resulted in the highest percentage of variance explained by the first principal component. The last three principal components are excluded for space.

projecting words onto the discovered polarity space





Thank you!