Journal of Computational Social Science (2019) 2:293–313
https://doi.org/10.1007/s42001-019-00048-6

RESEARCH ARTICLE

Concept Mover's Distance: measuring concept engagement via word embeddings in texts

Dustin S. Stoltz[1] · **Marshall A. Taylor**[2]

Received: 6 March 2019 / Accepted: 16 July 2019 / Published online: 1 August 2019

Abstract

We propose a method for measuring a text's engagement with a focal concept using distributional representations of the meaning of words. More specifically, this measure relies on word mover's distance, which uses word embeddings to determine similarities between two documents. In our approach, which we call Concept Mover's Distance, a document is measured by the minimum distance the words in the document need to travel to arrive at the position of a "pseudo document" consisting of only words denoting a focal concept. This approach captures the prototypical structure of concepts, is fairly robust to pruning sparse terms as well as variation in text lengths within a corpus, and with pre-trained embeddings, can be used even when terms denoting concepts are absent from corpora and can be applied to bag-of-words datasets. We close by outlining some limitations of the proposed method as well as opportunities for future research.

Keywords Cultural sociology · Concept Mover's Distance · Word embeddings · Natural language processing · Text analysis

Introduction

A central task in the sociological analysis of texts is tracking changes in the use of a concept overtime, or comparing engagement with a concept across texts, producers, or domains. For example, sociologists have examined the extent to which songs in different decades engage with the concept of "love," how the use of populist language varies by political position among American presidential candidates, and the effects of using "cleanliness" metaphors on user engagement among internet support

✉ Dustin S. Stoltz
dstoltz@nd.edu

[1] Department of Sociology, University of Notre Dame, Jenkins Nanovic Hall, Notre Dame, IN 46556, USA

[2] Department of Sociology, New Mexico State University, Science Hall 286, Las Cruces, NM 88003-0001, USA

groups—to name but a few [6, 19, 41]. This paper contributes to this literature by offering a straightforward method of measuring variation in engagement with a specified concept in text corpora.

We propose using a language model that represents words as positions in a continuous *n*-dimensional semantic space, referred to as word embeddings. This is a "distributional model" of word meaning, which is fundamentally relational [13, 32]. The underlying assumption follows Wittgenstein's dictum "the meaning of a word is its use;" or there is a strong connection between the semantics of terms and the linguistic contexts in which they appear [12, 14, 16, 28].

We build on a recently developed technique for measuring document similarity via word embeddings called Word Mover's Distance [25] which finds the minimum cost necessary for the embedded words of one document to "travel" to the position of all the words in another document. We introduce the notion of Concept Mover's Distance defined as the minimum cost that a document's embedded words need to travel to arrive at the position of all the words in an ideal "pseudo document" consisting of only words denoting a specified concept.

This approach (1) captures the prototypical structure of concepts; (2) is fairly robust to pruning sparse terms and variation in texts' lengths within a corpus; and when using pre-trained embeddings (3) can be used even when terms denoting concept are absent from corpus (useful for comparing texts across large time periods); (4) can be applied to bag-of-words datasets (especially useful when access to raw text is encumbered by intellectual property rights). While we rely on a high-quality source of pre-trained English word embeddings, in the conclusion we discuss how our approach can be easily adapted for corpus-trained word embeddings as well as multilingual corpora.

We use a wide variety of publicly available corpora to show the validity of our approach. First, we explore Julian Jaynes' [20] hypotheses about the presence of "consciousness" in the Iliad, Odyssey and the King James Version of the Bible. We then consider the relationship between engagement with "death" and the number of deaths in Shakespeare's plays, as compared to 200 concepts that are fundamental in 87 Indo-European languages [10, 34]. Finally, we consider George Lakoff's theory [26] regarding the competing cultural models of morality in U.S. politics by comparing engagement with the compound concepts "strict father" and "nurturing parent" in State of the Union Addresses from 1790 to 2019. It is important to note that we present these not as dispositive tests of theories, but rather as general demonstrations of the validity of, and broad uses for, the Concept Mover's Distance measure. We end by briefly clarifying possible limitations of our method, as well as opportunities for future research.

Measuring engagement with a focal concept

Background: word embeddings and Word Mover's Distances

Consider a hypothetical situation where a researcher wants to compare engagement with the concept of "authoritarianism" between George Orwell's *Animal Farm* and

1984. The researcher wants to know: Did Orwell discuss authoritarian themes more in one versus the other?

One approach to testing this would be to simply count the occurrences of "authoritarian" in the texts, divided by the total number of words in the texts. This term, however, does not occur anywhere in *Animal Farm* and only once in *1984*—despite both works clearly dealing with themes of state surveillance and abuses of power, among others [30: 101–139]. This word-counting approach also entails the assumption that all other terms do not indicate authoritarianism. We could continue to incorporate synonyms, but selection becomes increasingly arbitrary as we must consider words which are only partially related to the general concept in question, such as "control" or "power." Furthermore, this approach leaves information on the table, so to speak, because in the end terms either do or do not denote the focal concept.

A better approach, however, would account for the fuzzy, graded relationship between concepts, rather than either–or relationships [39, 46, 47]. Similarly, such an approach would allow an analyst to compare concepts through relational, rather than discrete, organization. Word embeddings provide such a foundation for this approach.

Word embeddings are models of language in which each unique word is represented as a location in a continuous, n-dimensional space [31, 36, 44]. Incorporating both the co-occurrence of words within a given context and the vectors of those words within that context, word embeddings assign a vector which corresponds to a word's meaning such that words which are closer together mean similar things. While these language models have been widely used in information retrieval research, they are just beginning to be adopted in the social sciences. For example, Garg et al. [15] use word embeddings to trace the evolution of gender and ethnic stereotypes over a century (see also [18, 24]). Given the affinities between relational and usage theories of meaning and the assumptions of word embedding models, we anticipate they will soon become a central approach to representing text in computational sociology more broadly.

Word Mover's Distance (WMD) is an adaption of Earth Mover's Distance (EMD) [40] which uses word embeddings to determine the similarity between two or more collections of words (e.g., sentences, tweets, books). To understand the underlying intuition, imagine piles of gravel which may or may not be the same volume, and which are distributed differently around a region. The EMD is the minimum cost of turning one of these piles into another, which is a function of both the weight of the load and the distance necessary to move the gravel [29]. As word embeddings prescribe locations and the words' relative frequencies in a document are weights, WMD attempts to find the "nearest neighbor" for each word such that it can minimize the "cost" of moving all the words in one collection to the positions of all the words in another collection. Thus, collections sharing many semantically similar words should have smaller distances than collections with very dissimilar words.

Take the fictional example in Figs. 1 and 2 (adapted from [25]). Each quantity (one per arrow) indicates the contribution of that word pair to the overall WMD statistic for those two documents. These quantities are the product of two numbers: the cosine distance, $c(i, j)$, between the two words in the n-dimensional embedding space, and a weighting term indicating how much of i in one document must

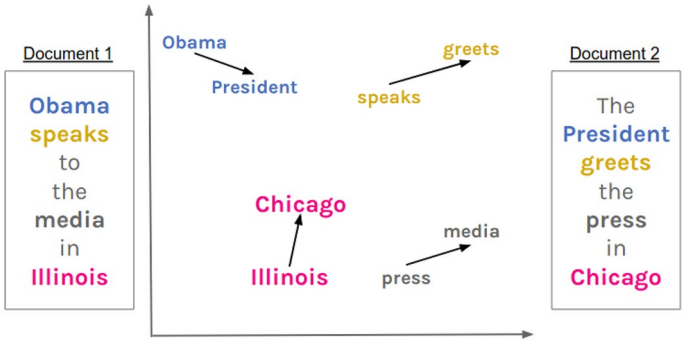

Fig. 1 An illustration of distances between word embeddings, adapted from Kusner et al. [25]. These words are represented in a two-dimensional space for illustration purposes, but words vectors are usually between 50 and 500 dimensions

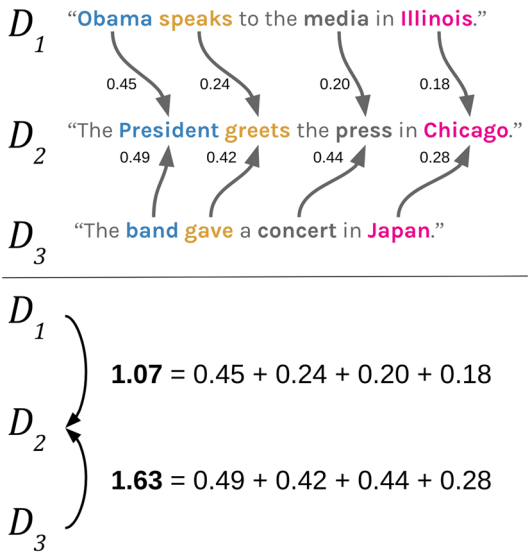

Fig. 2 An illustration of Word Mover's Distance, adapted from Kusner et al. [25]. The relative cost of moving all the words in Document 3 to the locations of the words in Document 2 is greater than moving the words in document Document 1 to Document 2

travel to word j in the other document (T_{ij}) [25: 3]. The contribution of the distance between "Chicago" and "Illinois" to the WMD between documents 1 and 2 is 0.18, whereas the contribution of the distance between "speaks" and "greets" is slightly larger at 0.24. In Fig. 2, taking the cumulative cost of moving all of the words in Document 1 to the locations of the words in Document 2 shows it to be closer and, therefore, more similar to Document 2 than is Document 3.

Formally speaking, the WMD algorithm finds the values of a matrix **T** that minimize "moving" one document, D, to the other, D' [25: 3]:

$$\text{WMD}_{ij} = \min_{\mathbf{T} \geq \mathbf{0}} \sum_{i,j=1}^{n} T_{ij} c(i,j), \tag{1}$$

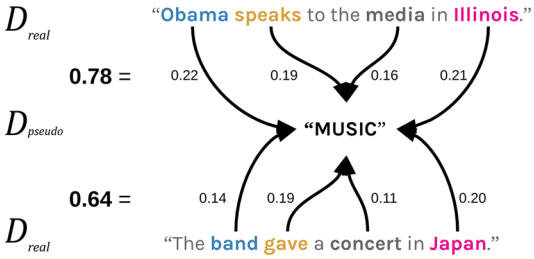

Fig. 3 An illustration of Concept Mover's Distance. The relative distance between the first sentence and the focal concept "Music" is more than for the second sentence. Therefore, the second sentence engages with the focal concept more

while under the constraints that (1) the sum of row word i in $\mathbf{T}$ is equal to the relative frequency of i in D (post any word removal), d_i, and (2) the sum of column word j in $\mathbf{T}$ is equal to the relative frequency of j in D' (again post-word removal), d'_j [25: 3]:

$$\begin{aligned} \textstyle\sum_{j=1}^{n} T_{ij} &= d_i, \ \ \forall i \in \{1, \ldots, n\} \\ \textstyle\sum_{i=1}^{n} T_{ij} &= d'_j, \ \forall j \in \{1, \ldots, n\}. \end{aligned} \tag{2}$$

If, for example, the word "text" had a relative frequency of 0.35 in a document after any word removal, then the sum of the row and column with the word "text" must each sum to 0.35. The idea, then, is to weight each ij cosine distance by "how much" of the relative frequency of i in D will move to j in D'.

Concept Mover's Distance: a text's closeness to a focal concept

We argue that a document's engagement with a specified concept can be measured as the distance between the words of that document and an ideal pseudo-document composed of only terms denoting that specified concept. Using WMD, similarities can be calculated between two collections which are of different lengths, for example, between a short query and relevant documents in a database (see [7]). In fact, and key for our approach, one collection could even be as small as a single word.[1] Consider the example in Fig. 3. Imagine measuring the two documents' closeness to the concept of "music." We would expect that a sentence about a band giving a concert would be closer than would a sentence about a politician speaking to the media, and this is exactly the results of CMD.

Simultaneously finding the minimum costs to move all words in each pairwise set of documents is fairly computationally demanding [35].[2] While there are many candidates for optimizing this calculation, Kusner et al. [25] offer a less computationally demanding procedure they call Relaxed Word Mover's Distance (RWMD), which also produced a lower error rate than eight other similarity measures in eight document classification tasks. With RWMD, the flow matrix weighting for each i, j

[1] As the document-by-term matrix used with WMD is weighted by relatively frequency this is the same as saying 100% of words are the same word.

[2] Specifically, $O(p^3 \log p)$, where p is the number of unique words in the collection.

Table 1 Pseudo-DTM with real documents (rows 1 and 2) and the pseudo-document (row 3) for Fig. 3

	Obama	Speaks	Media	Illinois	Band	Gave	Concert	Japan	Music
$\text{Doc}_r 1$	1	1	1	1	0	0	0	0	0
$\text{Doc}_r 2$	0	0	0	0	1	1	1	1	0
Doc_p	0	0	0	0	0	0	0	0	**1**

pair is solved twice: once with just the first constraint from Eq. (2) removed, and then once with just the second constraint removed. As Kusner et al. [25: 4] point out, the optimal solution when the first constraint is removed is to let T'_{ji} equal the relative frequency of j in D' if the cosine distance between j in the D' and i in D is the smallest observed. Similarly, the optimal when the second constraint is removed it to let T_{ij} equal the relative frequency of i in D if the cosine distance between i in D and j in D' is the smallest observed [25: 4]:

$$T_{ij} = \begin{cases} d_i & \text{if } \operatorname{argmin}_j c(i,j) \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

$$T'_{ji} = \begin{cases} d'_j & \text{if } \operatorname{argmin}_i c(j,i) \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The RWMD for each i, j pair is then calculated twice following Eq. (1), and the final reported RWMD score for the i, j pair is

$$\text{RWMD}_{ij} = \max\left(\min_{\mathbf{T}\geq\mathbf{0}} \sum_{i,j=1}^{n} T_{ij} c[i,j], \min_{\mathbf{T'}\geq\mathbf{0}} \sum_{j,i=1}^{n} T'_{ji} c[i,j] \right). \tag{5}$$

The document-by-document RWMD matrix is, therefore, symmetric since taking the maximum means that, in the matrix, $\text{RWMD}_{ij} \equiv \text{RWMD}_{ji}$.

Using the R statistical computing environment [8] and the text2vec R package to calculate RWMD, we wrote a function that creates a pseudo-document-by-term matrix (DTM_p) based on an input term or terms used to denote the focal concept.[3] The row for this DTM_p adds a one to the columns corresponding to specified term(s), and a zero otherwise (see Table 1). In the example from Fig. 3, the DTM_p has a 1 for the column corresponding to the focal concept "music," and a zero in all other columns corresponding to terms in the corpus. When the term denoting the focal concept is not in the corpus from which the original DTM is derived (such as the example in Fig. 3), a column corresponding to the term is added.

Incorporating the normal procedure for calculating RWMD in text2vec, we measure the distance between a real DTM_r (derived from a natural language corpus) and the DTM_p. Since the most basic implementation of CMD is used to find a document's distance from a pseudo-document consisting of only one term denoting a focal concept, the RWMD optimization process behind finding the flow matrix **T** is

[3] Replication materials are available at https://github.com/dustinstoltz/concept_movers_distance_jcss.

simplified. Namely, since the cosine distance of i in D to j in D' will always satisfy $\text{argmin}_j c(i,j)$ when D' only consists of j, the T_{ij} weighting in the simplest implementation of CMD (i.e., CMD with no compound concepts, which we discuss later) will always be the relative frequency of i in D and the T'_{ji} weighting will always be 1. Therefore, in the example in Fig. 3, the contribution of the distance between "Obama" and "music" to the CMD for the top document, 0.22, is the result of multiplying the cosine distance between these two words in the embedding space (0.87) by the relative frequency of "Obama" in the (preprocessed) real document, 0.25 (since there are four words).

The output of CMD is a list of distances. We invert these scores to get "closeness" for interpretability—meaning that larger CMD values indicate more concept engagement. Letting the RWMD of real document D from the pseudo-document be RWMD_D, we standardize the distances and multiply the values by -1 to arrive at the CMD for document D from a focal concept represented by the pseudo-document:

$$\text{CMD}_D = \left(\frac{\text{RWMD}_D - \overline{\text{RWMD}}}{\sqrt{\frac{\sum_{D=1}^{n} \text{RWMD} - \overline{\text{RWMD}}}{n-1}}} \right) \times -1. \tag{6}$$

One important difference to note between the paper that introduced RWMD [25] and the RWMD function as implemented in text2vec is that the former used Euclidean distance, but the default for the latter uses cosine distance (compare [25: 3] to [44: 26]). As cosine is a more widely used method for comparing the distance between two vectors in computational text analysis, we used this distance measure; however, future work could examine the robustness of results to the use of different distances metrics when calculating CMD.

In what follows, we rely on a high-quality source of pre-trained English word embeddings: fast text embeddings, created by Facebook's AI Research team [5, 22] and trained on Wikipedia 2017, UMBC webbase corpus, and statmt.org news datasets, containing a total of 1 million word vectors. We rely on the fasttextM package[4] to download and prepare only word vectors corresponding to terms in the DTM_r and the DTM_p. In the conclusion, we discuss using other sources of pre-trained word embeddings, as well as corpus-trained embeddings. We use standard pre-processing techniques to clean the texts, which included removing stopwords,[5] punctuation, capitalization, numbers, and non-ASCII characters. However, we do not stem or lemmatize the words as the fastText embeddings are neither stemmed nor

[4] https://github.com/statsmaths/fasttextM.

[5] We compared the difference between including and removing stopwords on a variety of terms and corpora. Overall the results were highly correlated, but the larger the initial corpus size, the higher the correlation. However, including stopwords tended to make the distances much more stark, i.e., documents which were close became much closer and documents which were far became much further. Therefore, we chose to remove stopwords throughout. This is certainly an area for further research.

Table 2 Descriptive statistics for illustrative corpora

Corpus	N_D	N_{TW}	N_{UW}	$\overline{TW}$	SD_{TW}
Iliad and Odyssey (0.99)	48	114,771	11,215	2391.06	1126.83
The KJV Bible (0.999)	66	273,308	12,109	4141.03	4407.09
Shakespeare's plays (0.99)	37	363,402	22,782	9821.67	1782.94
SOTUs (0.65)	239	423,275	954	1771.03	1375.00
SOTUs (0.75)	239	512,722	1589	2145.28	1671.35
SOTUs (0.85)	239	602,228	2756	2519.78	1953.67
SOTUs (0.95)	239	701,701	6258	2935.99	2275.75
SOTUs (0.99)	239	748,722	13,264	3132.73	2427.46

Numbers in parentheses are sparsity factors

Iliad, Odyssey, KJV Bible, and Shakespeare plays were scraped from Project Gutenberg [37]. State of the Union addresses are from The American Presidency Project [48] and obtained from the quanteda.corpora R package [2]

N_D number of documents, N_{TW} number of total words, N_{UW} number of unique words, $\overline{TW}$ mean number of total words per document, and SD_{TW} standard deviation of the number of total words per document

lemmatized and doing so needlessly reduces semantic information (and without a noticeable boost in performance). Descriptive statistics for the datasets used are provided in Table 2.

Demonstrations of Concept Mover's Distance

'Thought' and 'Action' in the *Iliad* and the *Odyssey*

The proposed method was inspired by recent attempts [10, 38] to test claims made by the psychologist Julian Jaynes [20]. We use this debate as a touchstone for demonstrating the usefulness of our approach.

The question Jaynes poses is this: If consciousness—i.e., awareness of our awareness—is unique to modern humans, when did it emerge and what preceeded it? His answer: humans experienced an intermediary stage between being unconscious and conscious creatures which he refers to as the "bicameral" mind. He argues further that evidence of bicamerality can be seen in the earliest human writings beginning around 2000 BCE, and began to break down slowly in the Bronze Age, during major transformation among human civilizations often referred to as the "Axial Age" (cf. [33]). This bicameral mentality gave way fully to conscious humans by around 600–300 BCE.

To support his theory, Jaynes engages in a close reading of several iconic works of the literature, specifically the books of the Christian Bible and classic Greek and Roman texts. He argues that the earlier Iliad lacked engagement with "consciousness" as compared to the later Odyssey, and a similar pattern emerges when comparing the earliest books of the Christian Bible to later books such as those in the New Testament. The implications of his theory are far reaching—touching nearly every

discipline in the human sciences and the humanities—however, we will focus on his specific claims regarding early human texts.

A cornerstone of the evidence Jaynes marshals to support his theory is a comparison of the Homeric epics the Iliad and the Odyssey. Both were likely composed toward the very beginning of the Bronze Age in the eighth century BCE, but may have as much as a century between them. Jaynes contends that the two epics straddle the fading of the proto-conscious "bicameral" human mind and the earliest glimmer of fully conscious human minds (see also [11, 43]). While the more recent Odyssey is filled with references to characters' self-reflection, "There is in general no consciousness in the Iliad" [20: 69]. The Iliad primarily describes action which, according to Jaynes, is instigated not by an agentic decision of the individual's will, but rather by automatic habit or the occasional auditory hallucination of "gods that pushed men around like robots" [21: 73]. Therefore, we should see a stark contrast with how the two texts engage the concepts of thought and action. Specifically, the Iliad should be low on the first, but high on the second. Using CMD, we consider this hypothesis by breaking down each poem into its separate 24 books[6] and calculating its closeness to "action" and "thought" (see Fig. 4).

It is clear from the top panel in Fig. 4 that the Odyssey engages with thought much more than the Iliad, and holds the reverse relationship regarding action (bottom panel). Notice that Book 22 is closer to 'thought' than any other book in the Iliad. This is a book which Jaynes considers a later edition to the epic. Citing Leaf's *A Companion to the Iliad*, Jaynes agrees with Leaf [27: 82] who argues that "many reasons combine to show that these [sections in Book 22] belong to a late epoch" [27: 356].

'Introspection' in the books of the King James Bible

The term "introspection" rather succinctly refers to what Jaynes considers the essential property of human consciousness [9]—not just awareness but awareness of awareness, not just cognition but meta-cognition—however, this term is not very common in general, let alone in historical texts. A major strength of our measure of conceptual engagement is that concepts can be denoted with terms that do not actually occur in the corpus (provided we are using pre-trained, rather than corpus-trained, word embeddings). This is because the distance between two word vectors is not just a function of co-occurrence, but also shared context. Words need not co-occur for their vectors to be positioned close together. For example, one can search for iPod in newspapers from the 1980s and identify articles discussing the Walkman, or the Discman if searching in the 1990s.

To show this, we measure engagement with "introspection" in the 66 books of the King James Version of the Bible. Jaynes similarly devotes much attention to the philology of the Biblical texts, arguing that earlier books lacked evidence of

[6] CMD works with any size document; therefore, we could have compare the two works as a whole (or even sentence by sentence), rather than by individual chapters. Our choice is entirely for illustrative purposes, more specifically to show variation across more observations.

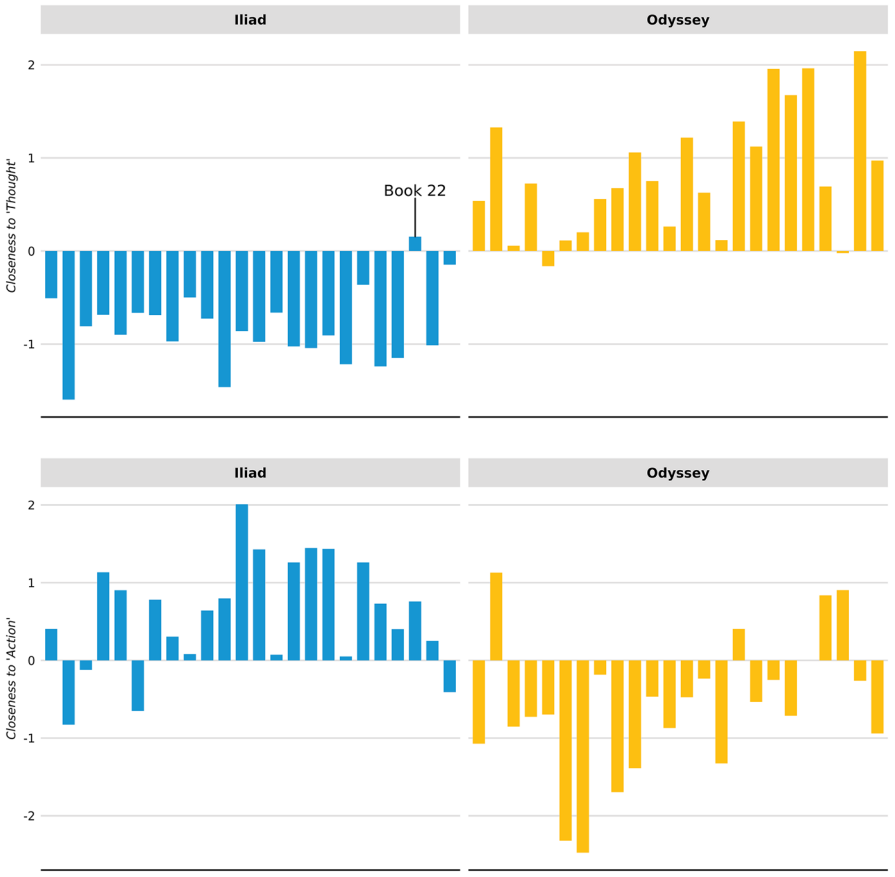

Fig. 4 Closeness to 'thought' and 'action' in the Iliad and Odyssey. Note: the Iliad (left, blue) and the Odyssey (right, yellow) are divided into their 24 books. Each bar is the CMD for one book. All data plots made with some combination of ggplot2 [51] and ggally [42]. The theme for all plots come from the Urban Institute [49]

consciousness, while privileging action compelled by god(s). Figure 5 shows each book arranged from the oldest (bottom) to most recent (top). Two books from the Old Testament are highlighted blue, the Book of Amos and the Book of Ecclesiastes. Jaynes singles these two out for comparison [20: 295-7] as Amos was likely composed around the same time as the Iliad in the early eighth century BCE, while Ecclesiastes is one of the newer texts of the Old Testament, composed circa 300 BCE. Jaynes argues [20: 296]:

> In Amos there are no words for mind or think or feel or understand or anything similar whatever; Amos never ponders anything in his heart... Ecclesiastes is the opposite on all these points. He ponders things as deep in the paraphrands of his hypostatic heart as is possible. And who but a very subjective man could

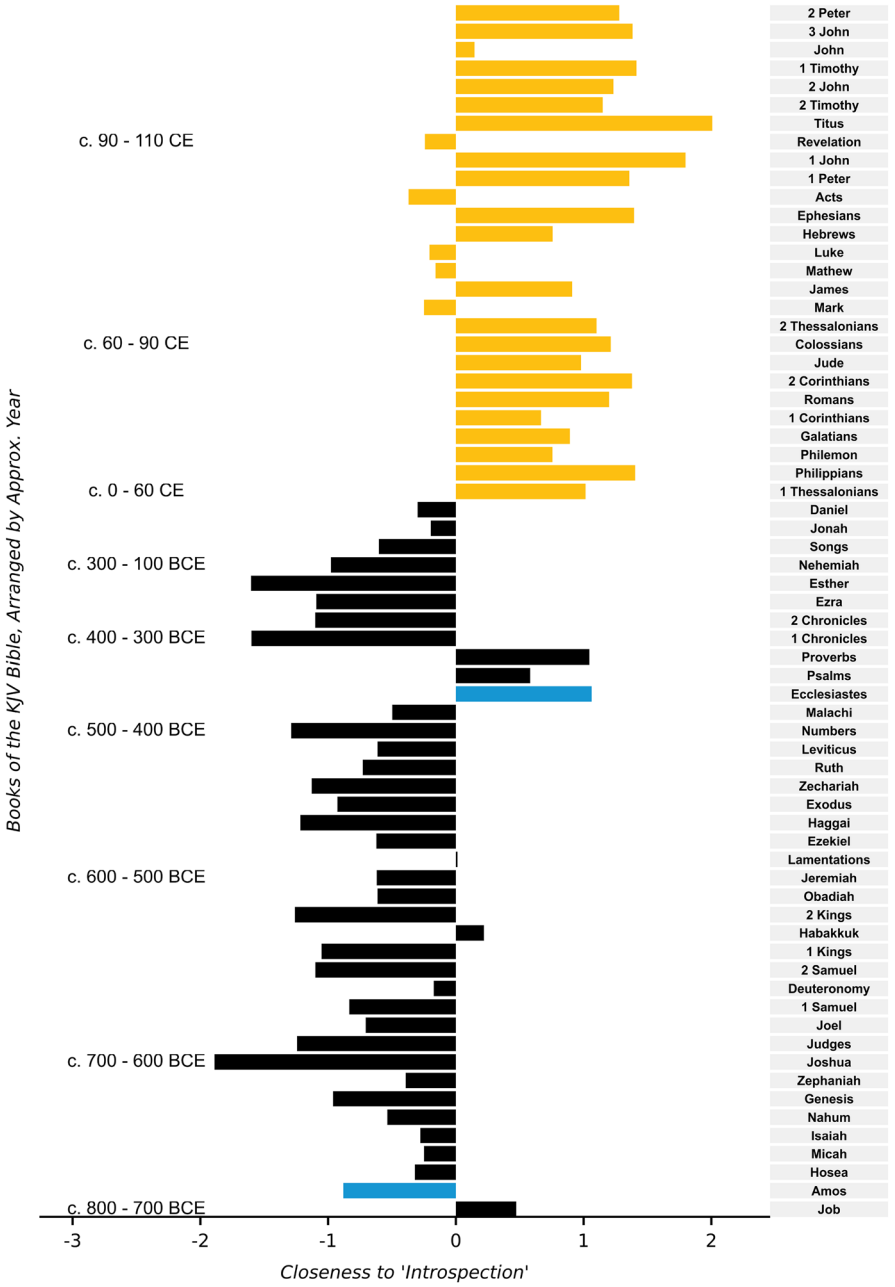

Fig. 5 Closeness to 'Introspection' in the Books of the Bible. Note: the books of the King James Version of the Bible, arranged by the approximate year of composition, with the bottom being the oldest books and the top being the most recent. Each bar is the CMD for one book

> say, 'Vanity of vanities, all is vanity,' ...These then are the extremes in the Old Testament.

Our results seem to uphold Jaynes' close reading of the texts. The most damning results for Jaynes' theory is the Book of Job. It is considered one of the oldest—if not the oldest—book of the Christian Bible. In fact, Job is remarkably similar to the Babylonian story *Ludlul-Bel-Nimeqi*, dated to around 1700 BCE. And yet, it shows relatively high engagement with introspection. In many ways, Job could be considered an exemplar of the loss of bicamerality as Job spends the length of the text reflecting on why god has forsaken him.[7] Jaynes has little to say about Job, however, other than the god of the book is pompous [20: 85] and scholars mistranslated a word in it as "bottle" when it should have been "familiar spirit" [20: 310].

'Death' and dead bodies in Shakespeare's plays

As another demonstration of CMD, we can consider how well engagement with a concept tracks identifiable events within a text. As an example, consider how closely Shakespeare's plays' closeness to the concept of "death" correlates with the number of actual deaths in the plays, as well as the genre. We downloaded 37 plays from Project Gutenberg, and compare the closeness of each play to "death." The genre of each play refers to how it was categorized in the First Folio: as a history, tragedy, or comedy. This is no small point as many books have been written on how precisely to categorize the plays. For example, the literary critic F. S. Boas, in *Shakespeare and His Predecessors* [4], focused on the unique qualities of All's Well that Ends Well, Measure for Measure, and Troilus and Cressida, arguing that they "cannot be strictly called comedies or tragedies" [4: 345], and referred to them as "problem plays."

The number of deaths are best guesses based on the authors' prior knowledge, supplemented by skimming each play, and reading summaries. However, there are many faked, implied, or presumed deaths in Shakespeare's work, which lead us to hypothesize only a moderately strong relationship. In our scatterplot (Fig. 6) showing the Loess-smoothed fit line between the number of deaths in a play and its engagement with death as a concept, we represent the three problem plays referenced above as triangles. There is a positive relationship between the two axes, with comedies clustering near the bottom, left corner. The three problem plays are outliers as they relate to their genres: the two comedies engaging death more than the other comedies (save Much Ado About Nothing), and the tragedy engaging death far less than the other tragedies.

It may be, however, that our measure is picking up on some general features of the texts themselves, and not the relationship between the events (deaths) and the concept (death) in the plays. Therefore, following Diuk et al. [10], we compare the relationship between "death" and deaths against a list of 200 concepts which where were found to be "fundamental" in 87 Indo-European languages [34]. Within

[7] It is outside the scope of this paper to unpack this further, however it is worth noting that Jaynes saw a direct connection between "gods forsaking" people and the breakdown of bicamerality (see [21]).

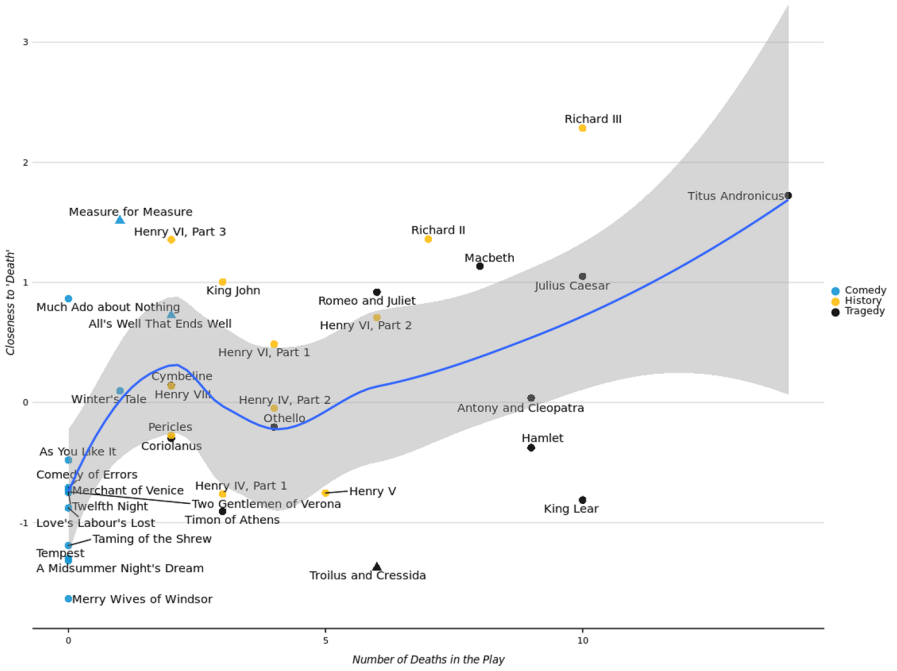

Fig. 6 Closeness to 'death' and dead bodies in Shakespeare. Note: plays are colored by how they are categorized in the First Folio. The three triangles correspond to plays which F.S. Boas considered "problem plays." Lines are smoothed with LOESS. Gray band is the 95% confidence interval

these lists, there are a few semantically similar to death: die, kill, stab, and blood. Therefore, in Fig. 7, we highlight death with blue, and die, kill, stab, and blood with magenta, while all other concepts are gray. The main takeaway is that CMD is not measuring a general lexical trend across the texts, but rather extracting specific conceptual engagement, and that this pattern holds across different yet semantically similar terms.

'Nurturing parents' and 'strict fathers' in the U.S. state of the union address

The previous demonstration relied on single terms to denote a focal concept. However, it may be useful to use more than one term to reference more complex or specified concepts. Importantly, this is not searching for distances to bigrams, but rather how each document engages with each of two or more terms. The mathematical basis for CMD with a compound concept is the same as before, but the optimization process for finding T_{ij} and T'_{ji} more closely approximates Eqs. (3) and (4) since there is now more than one term per pseudo-document. Recall that, with RWMD, T_{ij} equals the relative frequency for word i in D if $c(i, j)$ is the smallest observed cosine distance for word i in D from any word j in document D' (and a 0 otherwise). As such, the T_{ij} for CMD with a compound concept is simply the relative frequency for i

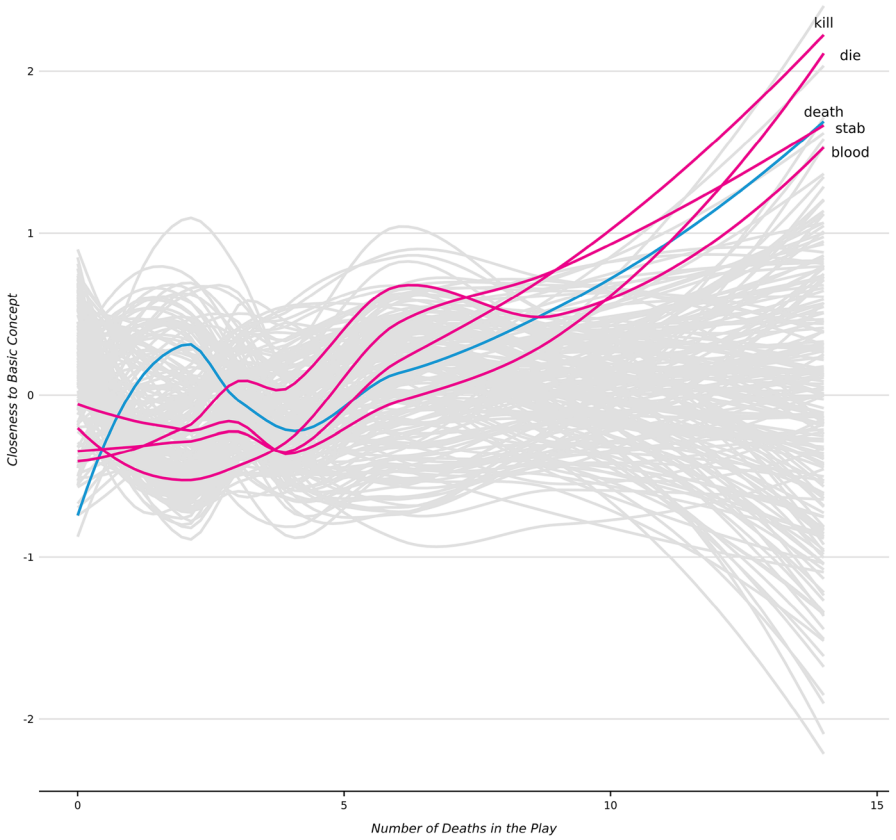

Fig. 7 Closeness to 'Death' and 200 basic concepts by dead bodies in Shakespeare. Note: 'Death' is colored blue, while 'kill,' 'die,' 'blood,' and 'stab' are colored magenta. The remaining 197 concepts are colored gray. Lines are smoothed with LOESS

in D for the smallest cosine distance between word i in real document D and the component words denoting the compound concept in the pseudo-document D' (and a 0 otherwise). Similarly, the T'_{ji} for CMD with a compound concept is equal to $1 / k$ for the smallest $c(i,j)$ (and a 0 otherwise), where k is the number of words in the pseudo-document.

As an example, consider George Lakoff's theory about the morality of politics in the United States (2010). Building on his decades of work in cognitive linguistics, Lakoff studied the underlying conceptual metaphors characterizing both liberal and conservative political discourse. The "family" is a foundational conceptual metaphor for the entire political spectrum; however, he contends differences emerge with how this metaphor is elaborated—i.e., how people believe the family is best organized. In the ideal-typical conservative world view, this family is organized around a "strict father," while in the ideal-typical liberal world view, this family is organized around a "nurturing parent."

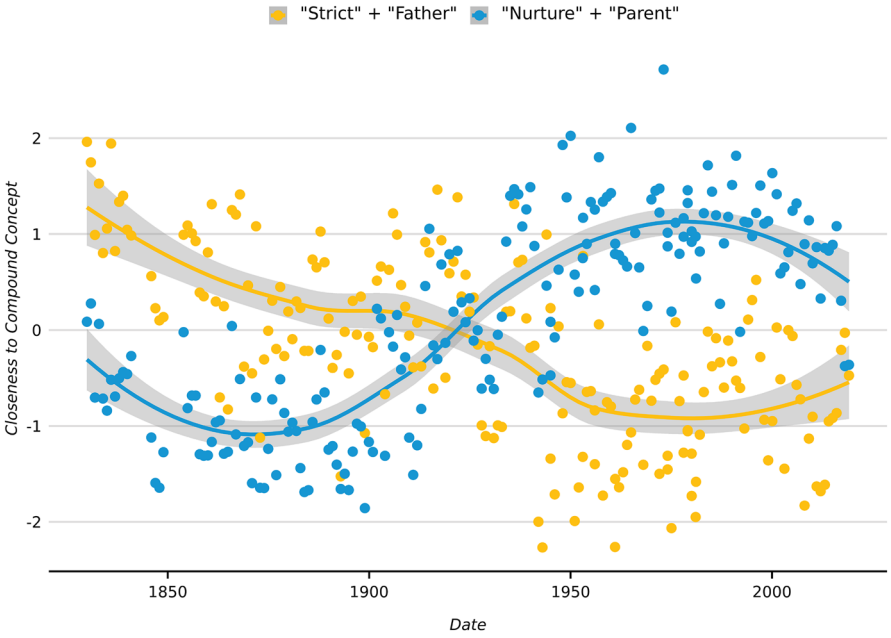

Fig. 8 Closeness to 'strict father' and 'nurturing parent' in 239 State of the Union Addresses. Note: all State of the Union Addresses from 1790 to 2018 are included. Every speech has a closeness to each compound concept; as such, every vertical line has both blue and yellow markers. Lines are smoothed with LOESS. Gray bands are 95% confidence intervals

While we leave aside the specifics of the theory, we do consider the extent to which engagement with these two compound concepts are opposed (although see Lakoff [26]: 313–314). We test the opposition between the strict father and nurturing parent within the U.S. State of the Union address from 1790 to 2018. Figure 8 reveals the hypothesized opposition such that if a speech is close to one it tends to be far from the other. Furthermore, the figure reveals an historical trend in which the strict father model was dominant throughout the nineteenth century and early twentieth century. The nurturing parent model gains dominance in the 1930s, but is currently in a state of decline.

Finally, we consider how changing the threshold of rare terms—how sparse we allow our document-by-term matrix to be—may impact the measure. Removing terms at a certain level of sparsity refers to a threshold of relative document frequency for each term. For example, sparsity of 99% means removing only terms that are more sparse than 0.99, i.e., they only appear in 0.01 of the documents in the corpus or less. Removing sparse words decreases the time to complete computations while also retaining the terms that are most representative of the corpus as a whole. As such, we test the extent to which the general relationship produced by CMD is robust to pruning higher levels of sparse terms. Figure 9 shows the correlations between CMD calculated for 'strict father' on document term matrices with terms removed at 99–65% sparsity. The results demonstrate that, at least for the present corpus and focal concept, the results are fairly robust to reducing sparse terms.

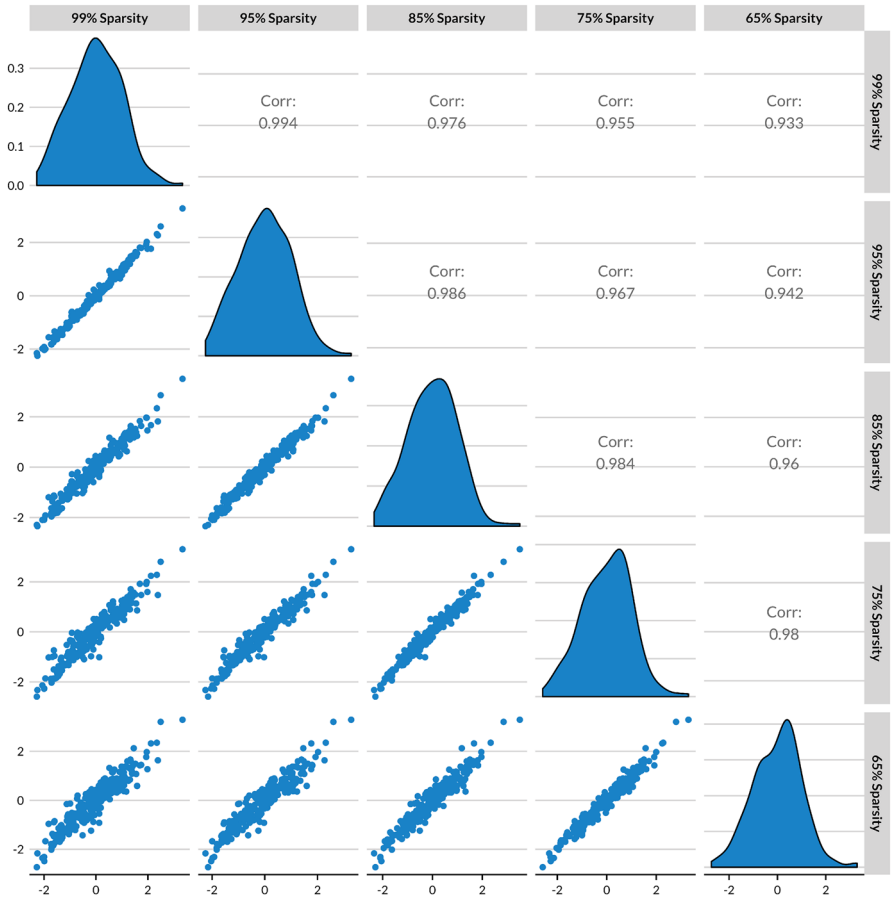

Fig. 9 Correlations between CMD for 'strict father' at higher levels of pruning. Note: using pairwise Pearson's correlations for CMD at different levels of pruning.

Conclusion

In this paper, we offered a relatively straightforward method of measuring variation in engagement with a specified concept in a corpus, which we call Concept Mover's Distance (CMD). This method uses a distributional model of language which represents words as positions in a semantic space, wherein words which appear in similar contexts, and thus mean similar things, are placed closer together. Next, we built upon a technique for measuring document similarity called Word Mover's Distance [25], which finds the minimum cost necessary for the embedded words of one document to "travel" to the position of all the words in another document. CMD is then the minimum cost that a document's embedded words need to travel to arrive at the position of all the words in an ideal "pseudo document" consisting of only words

denoting a specified concept. Below, we briefly clarify possible limitations as well as opportunities for future research presented by CMD.

Generic and specified concepts

An important consideration is where CMD fits into the broader text analysis toolkit. We see CMD as complementary to more inductive approaches, like latent Dirichlet allocation (LDA; [3]), which allow patterns to emerge from text without concepts selected prior to the analysis. CMD, by contrast, is useful when theory or prior literature suggests a particular concept is important at the onset. Therefore, CMD is an alternative to dictionary approaches, specifically those that use relative proportions of a pre-specified list of terms. When one is looking for very specific concepts, however, word frequencies based on a dictionary may be more appropriate.

As we discussed previously, a strength of our approach is that the terms used to denote a concept do not need to occur in a corpus. This is because the word embeddings associated with terms which are contextually similar will still be very close to concept terms. What this means methodologically is that CMD is best suited for measuring engagement with more generic concepts. For example, if one were to consider how much a set of texts engages "airplane," one should interpret this as engagement with the more generic concept(s) to which airplane is associated, such as flight and transportation. As we demonstrated with "strict father" and "nurturing parent," an analyst can add additional terms which further specify the concept. Nevertheless, these are still fairly generic concepts, and thus the analyst should take this into account when considering whether CMD would be the best tool for the job.

Closeness to opposing concepts

Due to the nature of word embeddings' underlying affinity to relational theories of meaning, words denoting saliently opposed concepts [17] will be located closer than if they are completely unrelated. What this means for CMD is that texts which are close to one concept will also be close to these opposing words. For example, in Shakespeare, we find that closeness to life and closeness to death are reasonably correlated (0.42), while closeness to love and closeness to hate are even more correlated (0.78). One strategy to distinguish whether a text is engaging one side of a binary is to specify the concept further, e.g., "romantic love" rather than just "love." Another strategy is to simply subtract the closeness to one side of the binary from closeness to the other side. For example, while closeness to "love" has a negative correlation of -0.21 to dead bodies in Shakespeare's plays, subtracting closeness to "hate" from closeness to "love" increases this negative correlation to -0.40.

Corpus-trained word embeddings

Our technique is dependent on word embeddings, but it is not dependent upon the specific embedding model we use (i.e., fastText). Future research should, nevertheless, examine the sensitivity of the measure to the use of different word embedding models,

that is, although fastText currently provides the most accurate pre-trained word embeddings available, there are reasons an analyst may choose to use corpus-trained word embeddings in their analysis. For example, measuring closeness to a concept using corpus-trained embeddings will likely be more sensitive to how that concept is used within that corpus, as compared to how that concept is understood more generally. This is perhaps especially important for examinations, like ours, of texts from 800 BCE or even the seventeenth century.

Aside from which model to use, when preparing corpus-trained embeddings, there are several choices one makes along the way—such as how many dimensions to make the vectors or how large to make the context window (for skip-gram models). How varying these parameters will impact word representations is currently under-studied, especially as it relates to questions relevant for social scientists. Nevertheless, while the substantive interpretations of results would vary, there is no reason CMD would be less effective on corpus-trained embeddings or embeddings using alternative models and parameters.

Multilingual Concept Mover's Distance

One exciting opportunity for future research will be measuring closeness to concepts in a corpus with more than one language. What this requires in practice is aligned multilingual word embeddings. This process entails training on languages separately, and then using a procedure to align the separate vector spaces into a common space. While several procedures have been suggested for optimizing alignment [1, 23, 45, 50], Facebook's AI Research team recently released aligned word vectors for 44 languages that out performed several state-of-the-art word translation procedures on standard benchmarks. Using aligned multilingual vectors, one could replicate, for example, our analyses of the Bible, but instead use versions of the books in their original translations, as well as translated into languages other than English, to determine whether the relationship between year and concept remains.

Applications for sociological inquiry

Lastly, our proposed method can be used on any sociological data that take the form of text: not just works of literature, religious texts, or speeches, but interviews, news articles, meeting transcripts, online forums, diaries, judicial rulings/dissents, or open-ended survey responses. Furthermore, in addition to measuring conceptual engagement across broad domains or collective actors, CMD can be used to measure variation in individual engagement. CMD can be used to consider such questions as, for example, whether political parties are more or less likely to support poverty initiatives based on their engagement with certain moral foundations such as fairness or authority, or whether extremist groups that engage closely with violent concepts are more likely to enact violence, or whether hearing sermons which are close to the concept of community are associated with volunteering or voting. Finally, as CMD can be used to measure variation within groups as well, an important application would be measuring cultural or ideological diversity within a population, as well as cultural convergence or divergence over time.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2289–2294). Austin: Association for Computational Linguistics.
2. Benoit, K., & Watanabe, K. (2019). quanteda.corpora: A Collection of Corpora for quanteda. R package version 0.86. https://github.com/quanteda/quanteda.corpora. Accessed 18 Feb 2019.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.
4. Boas, F. S. (1896). *Shakespeare and his predecessors*. London: John Murray.
5. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.
6. Bonikowski, B., & Gidron, N. (2016). The populist style in American politics: Presidential campaign discourse, 1952–1996. *Social Forces*, *94*, 1593–1621.
7. Brokos, G. -I., Malakasiotis, P, & Androutsopoulos, I. (2016). Using centroids of word embeddings and Word Mover's Distance for biomedical document retrieval in question answering. arXiV preprint arXiv:1608.03905.
8. Core R Team. (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
9. Dennett, D. C. (1991). *Consciousness explained*. Boston: Back Bay Books.
10. Diuk, C. G., Fernandez Slezak, D., Raskovsky, I., Sigman, M., & Cecchi, G. A. (2012). A quantitative philology of introspection. *Frontiers in Integrative Neuroscience*, *6*, 1–12.
11. Dodds, E. R. (1951). *The Greeks and the irrational*. Berkeley: The University of California Press.
12. Ellis, N. C. (2019). Essentials of a theory of language cognition. *The Modern Language Journal*, *103*, 39–60.
13. Emirbayer, M. (1997). Manifesto for relational sociology. *American Journal of Sociology*, *103*, 281–317.
14. Firth, J. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis* (pp. 168–205). Oxford: Blackwell.
15. Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*, E3635–E3644.
16. Garvin, P. L. (1962). Computer participation in linguistic research. *Language*, *38*(4), 385–389.
17. Greimas, A. (1983). *Structural semantics: An attempt at a method*. Lincoln: University of Nebraska Press.
18. Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In: *Proceedings of the 54thAnnual Meeting of the Association for Computational Linguistics* (pp. 1489–1501). Berlin: Association for Computational Linguistics.
19. Ignatow, G. (2009). Culture and embodied cognition: Moral discourses in internet support groups for overeaters. *Social Forces*, *88*, 643–670.
20. Jaynes, J. (1976). *The origins of consciousness in the breakdown of the bicameral mind*. Boston: Houghton Mifflin.
21. Jaynes, J. (1986). Consciousness and the voices of the mind. Lecture given at the Canadian Psychological Association Symposium on Consciousness. Halifax: Canadian Psychological Association.
22. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.
23. Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In: *Proceedings of COLING 2012: Technical Papers* (pp. 1459–1474). Mumbai: Association for Computational Linguistics.
24. Kozlowski, A. C., Taddy, M., & Evans, J. A. (2018). The geometry of culture: Analyzing meaning through word embeddings. arXiv preprint arXiv:1803.09288.

25. Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From word embeddings to document distances. In: *Proceedings of the 32nd International Conference on Machine Learning*. Lille: International Machine Learning Society.
26. Lakoff, G. (2002). *Moral politics: How liberals and conservatives think*. Chicago: The University of Chicago Press.
27. Leaf, W. (1892). *A companion to the iliad, for English readers*. London: MacMillan and Co.
28. Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171.
29. Levina, E., & Peter, B. (2001). The Earth Mover's Distance is the mallows distance: Some insights from statistics. In: *IEEE Proceedings of the Eighth IEEE International Conference on Computer Vision*. Vancouver: Institute of Electrical and Electronics Engineers.
30. Meyers, V. (1991). *George Orwell*. London: MacMillan.
31. Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In: *Proceedings of NAACL-HLT 2013* (pp. 746–751). Atlanta: Association for Computational Linguistics.
32. Mohr, John W. (1998). Measuring meaning structures. *Annual Review of Sociology*, *24*, 345–370.
33. Mullins, Daniel Austin, Hoyer, Daniel, Collins, Christina, Currie, Thomas, Freeney, Kevin, François, Pieter, et al. (2018). A systematic assessment of 'Axial Age' proposals using global comparative historical evidence. *American Sociological Review*, *83*, 596–626.
34. Pagel, Mark, Atkinson, Quentin D., & Meade, Andrew. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *49*, 717–721.
35. Pele, O., & Werman, M. (2009). Fast and Robust Earth Mover's Distances. In: *2009 IEEE 12th International Conference on Computer Vision* (pp. 460–467). Kyoto: Institute of Electrical and Electronics Engineers.
36. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). Doha: Association for Computational Linguistics.
37. Project Gutenberg. (2019). *Project Gutenberg*. https://www.gutenberg.org/wiki/Main_Page. Accessed 18 Feb 2019.
38. Raskovsky, I., Fernández Slezak, D., Diuk, C. G., & Cecchi, G. A. (2010). The emergence of the modern concept of introspection: A quantitative linguistic analysis. In: *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas* (pp. 68–75). Los Angeles: Association for Computational Linguistics.
39. Rosch, Eleanor, & Mervis, Carolyn B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.
40. Rubner, Y., Tomasi, C., & Guibas, L. J. (1998). A metric for distributions with applications to image databases. In: *Proceedings of the 1998 IEEE International Conference on Computer Vision*. Bombay: Institute of Electrical and Electronics Engineers.
41. Scheff, Thomas J. (2011). *What's love got to do with it? Emotions and relationships in pop songs*. New York: Routledge.
42. Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Larmarange, J. (2018). "GGally: Extension to 'ggplot2.'" R package version 1.4.0. https://cran.r-project.org/web/packages/GGally/GGally.pdf. Accessed 18 Feb 2019.
43. Snell, B. (2013). *The Discovery of the Mind: The Greek Origins of European Thought*. Translated by T. G. Rosenmeyer. Tacoma: Angelico Press **(1953)** .
44. Selivanov, D., & Wang, Q. (2018). text2vec: Modern text mining framework for R." R package 0.5.1 documentation. https://cran.r-project.org/web/packages/text2vec/text2vec.pdf. Accessed 16 Feb 2019.
45. Smith, S., Turban, D., Hamblin, S., & Hammerla, N. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859.
46. Taylor, John R. (2003). *Linguistic categorization*. New York: Oxford University Press.
47. Taylor, Marshall A., Stoltz, Dustin S., & McDonnell, Terence E. (2019). Binding signicance to form: Cultural objects, neural binding, and cultural change. *Poetics*, *73*, 1–16.
48. The American Presidency Project. (2018). *Annual Messages to Congress on the State of the Union (Washington 1790—Trump 2018)*. https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/annual-messages-congress-the-state-the-union. Accessed 3 Feb 2019.
49. Urban Institute Research. (2019). *urbnthemes: Urban Institute's ggplot2 Theme and Tools*. https://github.com/UI-Research/urbnthemes. Accessed 18 Feb 2019.
50. Xing, C., Wang, D., Liu, C., & Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In: *Proceedings of the 2015 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies (pp. 1006–1011). Denver: Association for Computational Linguistics.
51. Wickham, Hadley. (2016). *ggplot2: Elegant graphics for data science*. New York: Springer.