CHAPTER 4

CORPUS MODELING AND THE GEOMETRIES OF TEXT

Meaning Spaces as Metaphor and Method

DUSTIN S. STOLTZ, MARISSA A. COMBS, AND MARSHALL A. TAYLOR

C1S1

Introduction

C1P1 The spatial turn in computational text analysis is inciting a sprightly interdisciplinary passion that seems, at times, to entail sprinting when we ought to mosey. These advances—specifically the growing family of word embedding techniques—are based on the insight that the statistical properties of word co-occurrences form a kind of generative structure. In the middle of such exciting headway, it is wise to take a pause to think through the spatial theories and methods underlying "language modeling" techniques—namely, word embeddings (factorization, neural, and transformer). As someone probably said, "Patience is bitter, but the fruit is sweet."

C1P2 In this chapter, we stroll through the theoretical implications of *spatial metaphors* and inspect how the properties of topologies aid and inhibit our theories of textual meaning (see also Stuhler, Stoltz & Martin, in this volume).[1] In particular, we focus on the use of neural networks and deep learning to produce embedding spaces and the ways that metric spaces may distort our theoretical understanding of semantic spaces. We then saunter along the technical aspects of embedding texts into space by observing the uniting *objectives* of both classic and contemporary embedding techniques. Training objectives privilege some perspectives over others, and therefore it is more precise to describe model performance in terms of agreement as opposed to accuracy in predicting a "ground truth." In other words, ground truth for whom? This leads us to a necessary—and increasingly foregrounded—input to such techniques: the corpus. Meaning spaces are fundamentally relational such that we see "spooky semantics at a distance." The overall structure of the continuous semantic space constructed from corpora determines the "meaning" of each element in that space. While we can induce meaning from "local" structures, each act of inclusion and exclusion in corpus construction ripples throughout the totality of that space. Even with the confidence that big data lends to our analyses, and the undeniable augury of "large language models," researchers should take care to identify sources of uncertainty and carefully document each decision we make when building corpora and training models on those corpora.

C1P3 We conclude with our hope that more scholarship, institutional support, disciplinary accolades, and technical infrastructure will be directed toward carefully building, documenting, and sharing corpora, as well as machine learning (ML) models trained on those corpora. Advances in ML cannot solve omissions in our corpora. And, without sharing, these omissions may go unnoticed and undertheorized. Luckily, unlike conventional cartography, with ML techniques in computational text analysis, we can share both the maps and the territories.

C1S2 What Is the Space?

C1P4 Mining is arguably the most common metaphor for text analysis. Here, words are treated as, more or less, discrete indicators and distinguished as either signal or noise. Meaning is quarried from corpora and meaningless debris discarded. This "variables" approach to text analysis aligns with *substantialism* in social scientific inquiry. Critiques of this position (Abbott, 1988; Bourdieu, 1984, p. 103; Coleman, 1958; Emirbayer, 1997) often draw on Cassirer's distinction between "thing-concepts" and "relation-concepts" (Cassirer, 1953, p. 9). Early critics focused on conceptualizations of *individuals*, building on mid-century advances in social network analysis but soon allied with cultural sociology late in the century (Martin, 2003).

C1P5 Today, most sociological theories of culture are cozy with the notion that *meaning is relational* (Mohr et al., 2020; Spillman, 2020). Although we can consider relationalism and substantialism as two poles of a continuum—placing "different value [or emphasis] . . . upon the thing-concepts and relation-concepts" (Cassirer, 1953, p. 9)—it would be difficult to find a defender of a purely substantialist approach to meaning. Mining, then, is a poorly suited metaphor for sociological text analysis. Luckily, we can draw on (at least) three relational metaphors: systems, graphs, and spaces.[2] Here, we focus on spaces and the ML techniques used to create spaces out of texts.

C1P6 Spatial metaphors in sociology are primarily specified as *fields*, drawing on gravitational or magnetic fields, landscapes, or battlefields (Gerber, 2020; Green, 2008; Martin, 2003). Like graphs and systems, spatial metaphors of meaning proceed from the possibility that the relations between elements form a structure—that is, a topology. The shape of this topological space provides insight into meaning beyond the elements in isolation and allows for the comparison of spaces—either the same domain over time, spaces derived from distinct domains or communities (Enggaard et al., 2023; Hamilton et al., 2016; Stoltz & Taylor, 2021). Ultimately, the purpose of spatial methods for sociologists is to organize text in ways that aid interpretation and discovery. Facilitating this requires reduction: "embedding" information-rich writings in simplified spaces while remaining faithful to the patterns in those writings (Lee & Martin, 2015). What follows are some considerations when moving from a conceptual space to a metric space and, in particular, some of the ways they may clash.

C1S3
Distances and Continuity

C1P7 The most basic space is one-dimensional: a linear spectrum or continuum (Biernacki, 2012).[3] For example, the color spectrum, "thin" measures of trust or morality, and gender as a direction from "masculine" to "feminine" (Abend, 2019; Arseniev-Koehler, 2021; Taylor & Stoltz, 2021). With space (or topology) meaning is (typically) *continuous*—for example, the real number line. This is despite the elements of that space often being (roughly) discrete units, like words, documents, authors, and so on. Semantic *relations* between elements are mapped to spatial *distances* between their positions. The continuous space is infinitely divisible, so positions in such a space are also infinite. As the universe expands, energy emerges to fill that space. Similarly, the space between elements is no less meaning-*full.* Spaces thus can represent both manifest and latent meanings, with elements (e.g., words, documents) emerging like particles from quantum fields of potential significance. Following Lewin's "principle of contemporaneity" in field theory, embeddings provide insight into the "totality of the current situation" (Martin, 2003, p. 18) in which the current situation is a corpus. There is no "main effect" to be extracted from noise and randomness (Abbott, 1988). Nevertheless, we typically draw insights from inspecting local penumbras or induced subspaces rather than the topology as a whole (but see Cai et al., 2020; Ethayarajh, 2019; Rajaee & Pilehvar, 2021).

C1P8 To move this metaphor from theory to method, we can begin to turn our theoretical topological space into a *metric* space by assigning any element in this space a *vector*. When discussing spaces, a vector is a set of directions and magnitudes in each direction. To record the positions of, say, a wisdom of wombats in a pen we need the magnitudes for just two directions and two references. In our example, references may be the fence posts of the pen (absolute "Netwonian" space) or just the relative positions of the wombats themselves (relativist "Liebnizian" space). Osgood's work (e.g., Osgood et al., 1957) is a groundbreaking example of such a spatial approach to meaning. By asking people to arrange words along various "semantic dimensions," and using factor analysis, Osgood concluded that just three dimensions—good/bad, active/inactive, powerful/powerless—explained most of the variation in these judgments. Desikan et al. (2020), similarly, assigned words to three-dimensional vectors in a uniform red, green, and blue color space by summarizing the color distributions of Google Image search results associated with those words.

C1S4 Intuition and Transitivity

C1P9 In contemporary text analysis, the spaces used to represent meaning are usually "high dimensional." This entails going beyond the comfort of our three-dimensional, human-scale worlds. There is a smattering of people who can "visualize" in higher spatial dimensions. Famously, Alicia Boole learned to visualize in four dimensions (Polo-Blanco, 2008). Born with strabismus, William Thurston had to learn to "reconstruct a three-dimensional image from two-dimensional ones" and used the procedure to "see" four and five dimensions (Gabai & Kerckhoff, 2015). Many, however, have nothing in mind when thinking about theoretical constructs (in the extreme, those with aphantasia can conjure no mental imagery). Even among mathematicians, there is a decades-long discussion about the role of mental imagery in mathematical reasoning (Wheatley, 2013). So, that high-dimensional spaces tend to run against our everyday perception of spatiality is not necessarily a failure for such approaches. This is important for geometric analyses of texts, as we often deal with hundreds of dimensions.

C1P10 Adding dimensions to our conceptual "meaning space" provides a tool for thinking through the complexities of meaning. Imagine (or draw) a one-dimensional space. If wombat A moves closer to wombat B while wombat C moves closer to wombat B, then wombat A and wombat C also move closer to each other. When we imagine a higher dimensional space, it provides a way of thinking about potentially intransitive semantic relations like the example given by James (1890, p. 579): The moon and a football are similar in shape, and a moon and a gas jet are similar in luminosity, but a football and gas jet are not similar in any respect (Tversky & Gati, 1982, is foundational, but see Jones et al., 2018; Nachshon et al., 2022; Rodriguez & Merlo, 2020). Such high-dimensional spaces are perhaps uniquely suited for the complex and peculiar meanings of texts, the polysemy of words, or the multiple intersecting identities of characters and authors (e.g., Nelson, 2021). However, strict (mathematical) intransitivity is not possible with a single *metric* space.[4] To draw out these complex relations, then, we can deploy post-processing techniques to contextualize words (e.g., Arseniev-Koehler, 2021; Jones et al., 2018; Nelson, 2021; Rodriguez & Merlo, 2020; Taylor & Stoltz, 2021; Utsumi, 2020) or perhaps use several spaces or different kinds of spaces (e.g., Jäkel et al., 2008; Neelakantan et al., 2015). In other words, some meanings are latent in the space, and we must go exploring to find them.

C1S5 The Curses of Dimensionality

C1P11 As we increase dimensions, one might imagine the space will better conform to the intricacy of our texts, filling the nooks like cling wrap over a plate of mom's hyperdimensional spaghetti. Why not continue to add more dimensions? There are two considerations that give us pause; one is mathematical and the other conceptual.

C1P12 First, consider the "curse of dimensionality" (Bellman, 1961, p. 94). Bellman was referring to the ways computational limits rendered some deterministic calculations impossible, thus requiring approximate calculations in their stead. Computational resources have certainly increased since Bellman, but this remains a constraint. However, even before we hit this computational constraint, there is a point of diminishing information. Each new dimension increases the overall "volume" of the space. Increasing the volume pushes points farther apart while also restricting these points to a particular region of the space. Imagine a two-dimensional circle. If a point is selected at random, it is more likely to be near the edge than the center. If we add a dimension to create a sphere, the volume increases, so points can be more spread out. We are also more likely to randomly select a point near the equator rather than the poles. As we move to "hyperspheres" above three dimensions, both of these features compound. Points are pushed towards the "equator" and farther apart, up to the limiting case where each point is functionally the same distance from any other point, and most of the space is effectively empty. The word embedding vectors for "scotoplane" would become just as far from "fringehead" as "Gallifrey" is from "Earthseed," and thus, we learn little about the meanings of those words from their *relative positions* in the semantic space. In lieu of increasing dimensionality, we may follow similar strategies discussed above like, for example, experimenting with different kinds of spaces (e.g., Desikan et al., 2020; Linzhuo et al., 2020) or using several spaces or layers (Hamilton et al., 2016; Stoltz & Taylor, 2021).[5]

C1P13 Second, social life is indeed messy and nuanced, but many would argue that summarization is necessary for understanding, interpretation, and explanation. As it relates to sociological theory, Healy (2017) argues that the "free-floating demand that something be added" (p. 119) should be resisted: "Demands for more nuance actively inhibit the process of abstraction that good theory depends on" (p. 121). In computational sociology, we may fall prey to something akin to "nuance traps." An uncritical taste for greater computational complexity may inhibit scientific advancement by, for instance, making replication computationally unfeasible for those with limited resources, restricting the audiences conversant in such procedures, producing overly particular models, and prioritizing "shock and awe" over sound interpretation and useful theory.

C1P14 Few methods offer such a faithful realization of (linguistic) habitus (Arseniev-Koehler & Foster, 2022): "structured structures predisposed to serve as structuring structures, that is, as principles of the generation and structuring of practices and representations which can be . . . 'regular' without in any way being the product of obedience to rules" (Bourdieu, 1990, p. 53). Nevertheless, given the ways the correspondence between meaning and (metric) space frays, we should be cautious about overly realist interpretations of ML techniques and be content to use meaning spaces as a pragmatic merging of metaphor and method (see also Arseniev-Koehler, 2021). In this regard, we can look for guidance from scholars offering sobering discussions of the excesses of "AI hype" in the technology industry (e.g., Bender, 2022; Martin, 1993; Raji et al., 2021). In the final analysis, the machine does not "learn," and the AI does not "know"—we are the learners and knowers.

C1S6 What Is the Objective?

C1P15 In his influential treatise on a theory of general semantics, the linguist Alfred Korzybski (1958) wrote that "a map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness" (p. 58). Korzybski used the mapping metaphor for purposes different from our own (for him, to articulate the relationship between linguistic abstraction and real-world experience), but attention is drawn to the same underlying cartographic insight: A map is useful insofar as it is a simplification of the territory of interest.

C1P16 There are questions downstream from this insight, however. How do we want to represent the territory? After all, cartographers can map various dimensions of a territory: its geopolitical boundaries, sociodemographic distributions, or topological features, to name a few. How we represent a territory then leads to the question central to this section: What do we want to learn about the territory, given this representation? In other words, what is the objective?

C1P17 To bring this back to text, our goal here is to articulate how the form that our textual map takes is driven, in no small part, by what we hope to learn from it. If we want the texts to "tell us something about themselves," then we are typically using an algorithm that predicts something internal to the data—for example, predicting a context word given a target word, or the topic to which a word token in a document should be assigned. Alternatively, we might want to learn something about the relationship between our texts and some other (potentially non-linguistic) categories that we, as analysts, have brought to bear on the documents ourselves. In this case, we are probably using an algorithm that predicts something external to the texts. The key point is that the objective itself drives how we represent our texts. Importantly, though, the objective dictates not just the representation but also whose decisions—document producers' or analysts'—we privilege in our workflow and analysis.

C1S7 Internal and External Objectives

C1P18 At the risk of oversimplifying, ML algorithms for texts can be grouped into two categories: those that predict something *internal* to the texts and those that predict something *external* to them. Consider two popular algorithms: word2vec and logistic regression.[6] Word2vec (Mikolov et al., 2013) is a popular suite of models for training word embedding models—that is, assigning each word in a vocabulary a vector of real numbers over a set number of dimensions (often 100 to 300). Word2vec is a form of "unsupervised learning," whereby the analyst provides no preclassified data in the learning process. This can be contrasted with logistic regression, for instance, which can be thought of as "supervised learning," whereby analysts label some texts (or recruit others to do so) and use those labels to train a model that then predicts the (probably unknown) labels for a set of other texts.

C1P19 Word2vec has an "internal objective": It produces vectors through a series of word-level prediction tasks. Specifically, within a context window (i.e., a certain number of words around a target or "focal" word), the model can either predict a context word given a target word (using the "skip-gram" estimator) or predict a target word given its context words (using the continuous bag of words, or CBOW, estimator).

C1P20 Logistic regression, by contrast, has an "external objective": It attempts to learn the relationship between texts' linguistic features (typically, unique words) and analyst-provided labels by finding the weights for each feature that maximize the log-likelihood of generating our observed document labels. These feature weights form a model that can then be used to predict the unknown labels for other texts.

C1P21 The point here is not to get into the estimation details of any particular model but instead to underscore how our objective *privileges* certain types of information and therefore shapes what our textual representations can look like. With a model like word2vec—and unsupervised models for texts more generally—our analysis privileges the many linguistic decisions of the *producers* of the texts (as well as the collators of the corpus). Did the producers attempt to write with particular styles? Did they have genre conventions they were attempting to honor or subvert? Were they writing to reflect a dialect? Did they have a particular audience in mind? Did the medium (e.g., newspaper versus social media post) constrain their writing formally or practically? Are we dealing with transcriptions of verbal speech? The potential questions about the context of document production are endless. Many of these questions will remain unknown or unknowable and vacuumed into error terms. Nonetheless, unless we are willing to say that production contexts play no role in shaping the writing that ends in our corpora—something that virtually any cultural production scholar would tell us is blatantly false (Becker, 1984; Childress, 2017)—then we must come to terms with the fact that document producers' decisions and constraints can wreak havoc on the assumption that the linguistic patterns we find in our analyses are only functions of linguistic features that we can see.

C1P22 What this means practically is this: Just as cartographers must wrestle with the bounding problem (what goes inside the map and what stays outside?), so too must the text analyst think critically about which texts—and which types of texts—go into a corpus.[7] Building a scraper to collect millions of newspaper articles is easy; curating and pruning that collection is hard; knowing which articles were not digitized is even harder. "Every map is at least, whatever else it may claim to map, a map of the map-maker: her/his assumptions, skills, worldview, etc.," Korzybski (1958, p. xvii) explains.

C1P23 Methods that work with an "external objective" have to deal with these same issues surrounding the document production process, insofar as textual qualities (e.g., term frequencies, topic probabilities, part-of-speech frequencies, etc.) are used as features in the label prediction process. These methods, however, also privilege the judgments made in the hand-labeling process. Well-executed document labeling is, of course, not done willy-nilly; research teams rely on labeling protocol, regular meetings, and test labeling sessions to reach satisfactory values on various interrater agreement statistics. That said, these labels still reflect a certain definition of what constitutes a document (or whatever the lexical unit of analysis is) being inside or outside of a category. This is not meant to be read as a critique of hand-labeling text data (cf. Biernacki, 2014)—indeed, hand-labeling (or analysis in and of itself) is a vital step in many sociological text analyses (Nelson et al., 2021). Rather, we emphasize that treating hand-labeling as the "gold standard" (Nelson et al., 2021, p. 205) for text analysis in sociology might create a false sense of security or certainty. Put plainly, as hand-labeled data are often treated as the "ground truth" against which other methods are measured for "accuracy" in the social and computational sciences, this assumption of "ground-truthiness" can tempt us into glossing over an importation question: Ground truth for whom? The answer is, in the final analysis, it is the ground truth for the analysts.

C1P24 To the extent we talk of "validating" our spatial representations of corpora, *we are better off speaking of agreement rather than accuracy*. Would the producers of the text agree with a given representation? Would the audiences? Would the analyst? This is unavoidable, certainly, but it is by no means a death knell to these methods, and it is not even specific to text analysis. What it does mean is that there can be interesting, theoretically generative, and (and perhaps unknowable) interaction effects between document production contexts and analysts' decision-making on how a text is mapped and what sort of meaning can be gleaned from that map. This omission of the social positions of those producing and/or rating corpora reveals an important opportunity for future sociological analysis. It also squarely rejects the possibility that larger corpora and more complex ML algorithms will soon "solve" the problem of meaning.

C1S8

Objective or Algorithm?

C1P25 It makes sense to say that the algorithm we choose is what determines the representation that our textual map takes. After all, the algorithm takes our texts and spits out a simplified structure (space) of the texts' linguistic regularities and relationships. Some of these algorithms produce output that is literally spatial and therefore subject to the geometric qualities discussed in the previous section, while others (like logistic regression and other categorical classifiers) are less obviously so.

C1P26 But, most of the time, the choice of objective is going to matter more than the choice of algorithm. Let's focus on word embeddings as a case in point.

C1P27 There is an ever-growing number of methods that can give us word vectors.[8] It is not hard to find arguments for when to use one method over another—for example, word2vec with the skip-gram implementation is preferred over its CBOW counterpart for learning the vectors of rare words, while the fastText model (which incorporates subword information by learning subword character *n*-gram embeddings) is necessary if we want vectors for words that do not appear in our training corpus (Bojanowski et al., 2017). Each of these algorithms also has tuning parameters—such as context window size, symmetric or asymmetric windows, learning rates, and dimension length—that can affect our output.

C1P28 At the end of the day, though, these algorithms produce vectors that maximize or minimize some quantity. For GloVe (Pennington et al., 2014), another popular algorithm, that quantity is finding the word vectors in which the dot products of the vector pairs minimize the error in predicting the log of that word pair's co-occurrence frequency. For word2vec, it is finding the word vectors that maximize the average log-probability of a target and context word co-occurring. What we see here, though, is that both methods are essentially trying to maximize the same thing: *pointwise mutual information*, or the extent to which a target and context word co-occur more than we would expect if they were independent of one another in their usage. This optimization based on mutual information remains true for embedding methods based on transformer architectures (Kong et al., 2019).

C1P29 So, ultimately, the popular embedding models at our disposal (currently) are all trying to converge around the same objective function (which is why their vectors are often quite similar). Consider an analogy here to maximum likelihood models for generalized linear regression. There's a host of maximum likelihood estimators: for example, the maximum likelihood estimator, full information maximum likelihood, pseudolikelihood, and restricted maximum likelihood. But they are all trying to find an approximation of the same thing: the parameter estimates that, in the aggregate, maximize the probabilities of getting the observed data that we do. In the final analysis, the choice between them is more what the data can and cannot support (e.g., size of the sample, sparseness in select covariates, etc.), and what probably matters more is that we have the right variables given our goals and research questions.

C1P30 We can extend the analogy to the importance of variable selection over estimator selection for minimizing measurement error. Ordinary least squares, or OLS, and logit models will, in probably the majority of cases, lead us to similar conclusions, but we very well might not reach the "right" conclusions if we use the wrong variables. To bring this back to the case at hand, what good is our map if we are mapping the wrong thing? In this case, the method for making the map (algorithm) is not going to save us; *how we build and prepare our texts is far more important*.

C1S9

What Is the Corpus?

C1P31 Saul Steinberg's (1976) infamous illustration "View of the World from 9th Avenue" captures the myopic, self-absorbed mental geography of Manhattanites. From the heart of Midtown to the Pacific Ocean, the distorted world map centers the island of Manhattan, where New Jersey might as well be as far away as Kansas City. And while indeed on a torrid summer day, the walk across town might feel like a cross-continental trek, the map is but fiction. In fact, this lighthearted depiction touches on a more sinister and enduring relationship between cartography and the exercise of power. Mapping projections reflect the priorities, biases, and desires of their creators and exert a profound influence on the world polity—what political scientist Mark Neocleous (2003) calls "cartographic violence." Extending the metaphor of "mapping" to the domain of textual data analysis necessarily involves contending with these broader theoretical questions regarding the selection of data, construction of models, and downstream ethical implications.

C1S10 Implications of Different Types of Training Corpora

C1P32 All mapping methods involve a collection of texts—typically massive quantities of texts. Given that spatial methods are inherently relational, what is included and excluded in a corpus shapes the resulting space in fundamental ways. Corpora may be strictly local, using the analyst's key set of documents, or domain specific, drawing on focal documents within the same genre or field. Corpora can also attempt to be global, relying on texts that are assumed to span domains and genres so as to broadly encompass a language or country. The representativeness of corpora depends on sampling and scoping decisions made by the analysts, as well as the organizational and technical context from which documents are sampled. Articles are trashed, books burned, sentences deleted, comments blocked, pages edited, and posts censored (Ben-David, 2016; King et al., 2013; Schuster et al., 2020). Importantly, despite efforts to assure representativeness, researchers (and readers) should acknowledge the inherent incompleteness of corpora.

C1P33 Many popular pretrained embedding models rely on the largest available corpora, like the Common Crawl—a public archive of the web—based on "crawling" from link to link. Conclusions based on these models rest on unverified assumptions about representation and may reinforce hegemonic cultural understandings that are overrepresented in source documents. While best practices regarding dataset documentation and intended uses have emerged for more task-specific or local datasets, they are resource intensive to apply to massive unlabeled web-scraped text data (Dodge et al., 2021). Furthermore, these corpora can easily become so large as to make it implausible for humans to exclude "toxic" content found, for example, on Internet forums (Davidson, in this volume).

C1P34 Despite these limitations, Stoltz and Taylor (2021) argue that sociologists should default (at least at the onset) to using pretrained sources to enhance comparability across studies. Training models also require extensive computational resources and time, as well as access to quality training data to scrutinize. The aggregate computational energy utilized to run thousands of research training models on big corpora is not only an inefficient use of academic energies but may contribute to environmental degradation. Furthermore, the more researchers dissecting pretrained models, the more likely it is that their limitations will be documented. That said, if the training corpora fail to account for relevant populations or are not available for a specific community, then they may not be appropriate.

C1P35 Such trade-offs can be better understood by sociologists if sharing models becomes institutionalized in the field. More importantly, the development of corpora should also be valued as a legitimate end in itself and should be embraced with robust institutional support. This could include dedicated academic centers or interdisciplinary collaboratives tasked with constructing carefully documented corpora, suites of pretrained models, institutional review boards for dataset documentation and maintenance, and increased methodological training for graduate students and researchers.

C1S11 Documentation, Datasheets, Model Reports

C1P36 In response to the growing recognition of the implications of different types of training corpora, there have been important calls for detailed data documentation and greater caution when using massive training corpora and pretrained models relying on such corpora (Bender & Friedman, 2018; Dodge et al., 2021; Gebru et al., 2021; Hutchinson et al., 2021; Mitchell et al., 2019). This is a caution, we argue, that sociologists should adopt.

C1P37 At present, there are no industry or disciplinary standards for documenting ML datasets and algorithms. As a result, datasets and models are often used and shared with little insight into the processes and decisions underlying their creation. One key intervention, championed by Gebru et al. (2021), is the adoption of "datasheets," which document the contexts and contents of datasets across the data development lifecycle—from motivation to composition, implementation, maintenance, and recommended uses (Gebru et al., 2021). These datasheets should be released along with datasets. The specific components of datasheets might vary depending on the domain and organizational context, but Hutchinson et al. (2021) argue they should all include answers to the six "Ws": *who* created the dataset, *who* is potentially impacted, as well as the *what*, *where*, *when*, and *why*?

C1P38 Datasheets tend to emphasize the characteristics of data flowing into models. "Model cards" are another key intervention aimed at documenting model performance. Developed by Mitchell et al. (2019), model cards provide benchmarked evaluation across a variety of conditions, outline intended use cases, and identify potential pitfalls. Importantly, model cards can be used to document performance across cultural, demographic, or phenotypic groups relevant to the specific application domain, and they can document intersectional analyses combining multiple groups.

C1P39 Datasheets and model cards improve accountability and transparency throughout the data development lifecycle. They encourage dataset creators to be intentional throughout the development process and aid data users in the thoughtful, appropriate, and sophisticated use of datasets and models. Implementation of robust documentation practices may mitigate unwanted biases in ML models, particularly those deployed in high-stakes contexts such as criminal justice and hiring. Implementation may also reduce technical debt by creating intra-institutional memory of changes to datasets over time. This encourages a view of data as dynamic rather than static or a "versionized" artifact (Hutchinson et al., 2021). Finally, thoughtful documentation may also facilitate greater reproducibility of model results.

C1P40 The adoption of these documentation best practices among individual researchers and practitioners must be accompanied by ecological changes in the sociology of ML.

Emphasis and training on documentation, dataset creation, and maintenance must become a routine pedagogical feature of programs training dataset creators and users. Academic journals must be open and willing to accepting publications about dataset documentation, model reports, and other dataset expertise. Universities, foundations, and other prominent funding sources should fund grants for the development and maintenance of shared corpora. Reviewers and peers must demand greater methodological clarity in ML research. Sociologists should be open to interdisciplinary collaboration with ML experts.

C1S12 Ethics, Shortcomings, and Cautions

C1P41 Robust documentation practices and greater attention to the "context of discovery" (Swedberg, 2012) in ML can help buffer against some of the hidden hazards of applying ML models to textual data. However, documentation alone cannot serve as a panacea for the shortcomings of these models. Researchers should take care to identify and address, to the best of their ability, the potential computational and ethical limitations of this approach. If datasets and models are deployed uncritically, they may be inappropriately utilized in contexts that do not match their training or that reflect unwelcome social biases. At best, this uncritical deployment may result in economic or operational inefficiencies; at worst, it may have significant downstream consequences that reinforce and perpetuate historical patterns of social injustices in the criminal justice system, employment, education, and beyond.

C1P42 One issue emerging from questions of representation is the use of so-called low resource languages. These primarily oral languages are usually Indigenous or endangered, with few speakers and fewer scholars devoting due attention to them (Bird, 2022). While those using text analysis methods tend to focus on languages with larger populations of users, English in particular, recent efforts to incorporate new languages may arguably be accused of imperialism. This is especially true when engagement is solely in service of downstream tasks, like mining Indigenous languages for commercial or academic opportunity rather than producing scholarship or tools beneficial to the community or for the purposes of language preservation (Coffey, 2021). The characterization of these local languages as "under-resourced" also relies on a Eurocentric deficit model, which privileges machine readability, standardization, and scalability, and it ignores Indigenous expertise and diversity (Bird, 2022).[9] In response, Bird calls on researchers interested in nonstandardized languages to work with local language communities from the ground up to develop scholarship and to encourage the use of that ancestral language.

C1P43 Another key issue concerns the potential for the (re)production of historical patterns of social inequality if corpora used to train models contain unwanted social biases (Alegria, in this volume). While many models, and word embeddings in particular, offer a uniquely intersectional approach to meaning, they are not immune to the wider critiques stemming from postcolonial scholars about the project of cultural categorization (Borch, in this volume; Hassan, in this volume). Aspects of the word embedding approach have the potential to reinforce hegemonic cultural assumptions. Word embeddings are fundamentally built on a model of binary opposition. Feminine and masculine can be two opposites on the gender dimension, for example. As Arseniev-Koehler points out, the extent to which we believe that word embeddings model language as a relational system depends on how researchers understand meaning, interpret the relation, and operationalize specific concepts (Arseniev-Koehler, 2021, p. 14). Postcolonial scholars (e.g., Go, 2017) have highlighted the extent to which this analytic bifurcation—the West and the "Rest"—permeates social theory and remains largely unquestioned. Some embedding models can accommodate graded, hierarchical, or continuous models of opposition, but further research is needed to operationalize more flexible models of difference. Embedding models also tend to envision meaning as "unrealistically coherent" (Arseniev-Koehler, 2021, p. 21). Words have multivalent meanings that might change depending on who is interpreting the semantic space. Sociologists should be attentive to the potential polysemy of the words and concepts they wish to study.

C1P44 More generally, all embedding models reflect the biases of their training corpora and often mirror the language and ideologies of dominant social groups (Birhane & Guest, 2020). As Stoltz and Taylor (2021, p. 4) argue, while using pretrained models is productive in that it enhances comparability across studies and expands access to scholars with limited computing resources, it limits the application of these models to communities not included in the training corpora. Capturing the racial and gendered biases of these dominant groups may be an asset to sociologists as they reflect the contours of the social world but may also have negative downstream effects on marginalized communities, especially in the implementation of high-stakes decision-making (Birhane & Guest, 2020; Stoltz & Taylor, 2021, p. 3). As the field of computational text analysis continues to develop theoretically and empirically, sociologists should continue to be mindful of these limitations.

C1P45 Thus far, we have reviewed some ethical considerations regarding the application of text analysis models in the technology industry. Just as important, however, are ethical considerations internal to research projects within sociology. Scholars should be mindful of the usual issues regarding identifiability, confidentiality, sensitivity, and consent for data collection. Consent for web-scraped data is of particular concern, as the use of those data might not reflect the original intention of the poster. Scholars and universities should seek to develop data infrastructure that can accommodate changing perspectives or desires of individuals whose data are used in their models—particularly for data gathered on overstudied populations. Although private data management companies may be better suited to this task than individual scholars are, there is the potential risk of monetizing sensitive data and acting as gatekeepers to researchers or other stakeholders.

C1P46 A final caution on research design. Even highly skilled researchers motivated to do work ethically and with integrity can fall short of being perfect. A growing body of literature on the inherent uncertainties in research design and practice has found that, even in the absence of structural incentives and individual biases, idiosyncrasies in the research process may lead to wide variability in our findings (Breznau et al., 2022; Hariri et al., 2019). Given all the decision points available from the creation of corpora to model specification and the interpretation of results, the problem of uncertainty is even more complex for researchers who are leveraging ML. Del Giudice and Gangestad (2021) argue that, like the curse of dimensionality, small decisions inaccurately deemed arbitrary can lead to a combinatorial explosion of misleading or meaningless findings. But do not let thorny, uncertain brambles leave you tangled in the "garden of forking paths" (Gelman & Loken, 2014). Documentation practices like datasheets and model cards offer concrete ways to account for uncertainty and to document decision nodes encountered throughout the research process. We echo Breznau et al.'s (2022) calls for "epistemic humility": Even with the confidence that big data lends to our analyses, researchers should take care to identify sources of uncertainty and to interrogate decisions that they deem arbitrary. As the field continues to develop, communities of practice will be responsible for making sense of these challenges at a tacit level.

C1S13

Computational Methods as Community Endeavors

C1P47 We end with one final thought. Methods in computational text analysis are often lauded for their ability to scale to large datasets. Indeed, reducing the petabytes of text contained in the Common Crawl to a semantic space requiring mere gigabytes is no trivial feat. These methods are also uniquely suited for deeply collaborative research. A key feature of maps is that they are "smaller than the objects" they represent (Lee & Martin, 2015, p. 12).

C1P48

> The map, then, is an instrument that lets us project complex questions into a reduced space to allow simple cognitive processes to reach adequate understandings that are, within an allowable margin, equivalent to those that could be reached by more cumbersome, and error-prone, processes. . . . We impoverish to make the artifact "small enough" so that . . . the writer *and* reader can (metaphorically speaking) both lean over and examine it at the same time, pointing to key features. . . . We do this not with a "full-size" *part*, but with a "reduced-size" *whole*.
>
> (Lee & Martin, 2015, p. 12)

C1P49 It is our hope that tools designed for software production pipelines can be repurposed for hosting, sharing, and maintaining assemblages of training corpora, models, and documentation in ways that enhance transparency, foster inclusion, and allow scholars to "lean over" at the same time. We argue that, given the myriad decisions analysts must make when preparing texts, it is imperative that sharing models and corpora becomes the disciplinary standard—in particular, because of the potential environmental impact of training new models when pretrained models will do (Scoville et al., in this volume). Sharing also affords unaffiliated teams the capacity to scrutinize models' shortcomings and to incorporate this into the living documentation of the models.

C1P50 Our field should consider the benefits of a future in which well-documented corpora, and large models trained on those corpora, are not siloed in private enterprises setting limited access to pretrained models as a paid service. We argue that more scholarship, institutional support, disciplinary accolades, and technical infrastructure should be directed toward carefully building, documenting, and sharing corpora and models.

Notes

1. Malinowski (2015) once wrote, "Social science is still burdened with the superstition that words contain their meanings" (p. 86), and the way we write about meaning here might imply that we subscribe to such superstition. We do not. We presume that textual meaning involves producers who (roughly) use general lexical conventions to (roughly) encode an intended meaning and readers who can (roughly) decode this intended meaning provided that they (roughly) know these conventions. The patterns we find in texts reflect, in part, these conventions, intentions, producers, and audiences.
2. System metaphors typically pilfer Saussure's theory of phonology (Stoltz, 2021). Per this theory, in its strong form, discovering the meaning of one element of the system necessarily requires grasping the entire system; that is, there is no "local structure," rendering infinite regress and boundary-specification critical problems (Harnad, 1990). Furthermore, taken to the extreme, the very idea of a position in this system is *entirely constituted* by relations of differences (e.g., Alexander & Smith, 1993), leading to the fundamental impossibility of knowing anything (Smith, 2011). Exemplars of this position include Jameson's "prison house of language" and Derrida's formulation of knowledge as "traces" in an otherwise undifferentiated substance (Lizardo, 2013). In sociology, graph metaphors grow out of "organicist" notions of social structure (Lizardo, 2013), formalized primarily in social network analysis. In language, graph metaphors are used to model dependencies between aspects of language, such as Barthes and Duisit's (1975) "levels of meaning," Levi-Strauss's organization of a myth by relations between mythemes, or the operation of *fabula* and *syuzhet* in Russian Formalism. Graphs thrive on absences. Advances in social network analysis are built on the notion of structural *holes*: not just who or what is connected but who and what are not connected. The structure formed by connections between elements is not informative if all elements are equally connected.
3. Biernacki rightly critiques sociologists for relying on spatial metaphors and yet leaving implicit these metaphors' contributions to theoretical reasoning. Biernacki, however, elides all possible spatial notions to this one-dimensional, linear spectrum and fails to appreciate the spookiness of more complex spaces.
4. By definition, a true metric space must satisfy the triangle inequality that disallows intransitivity (Griffiths et al., 2007). Some would argue, however, that intransitive meanings of words—such as those found by Tversky and Gati in free association tasks—are an artifact of *implicit contextualizing* on the part of respondents (Nachshon et al., 2022). That is,

respondents never really think of a word completely isolated from a context, but they might not report this assumed context. So a justification for post-processing techniques is to explicitly contextualize words when making comparisons of similarity. Nevertheless, this may turn out to be another instance in which there is tension between conceptualization and operationalization.

5. Researchers have also focused on the extent to which bidirectional encoder representations from transformers, or BERT, and similar transformer-based models may be overparameterized (Ethayarajh, 2019) and may overdifferentiate "types" such that two tokens of the same type are erroneously classified by the model as different types (Pavlick, 2022).
6. Using terms as predictors, for instance, would result in a model in which each term is assigned a weight in predicting a binary outcome—that is, a one-dimensional space. Here we discuss logistic regression using the sigmoid function, which is appropriate for binary classification tasks. There are other options for multiclass classification, including the multinomial generalization of the logistic regression model using the softmax function.
7. There can be consequences for not understanding the production contexts behind a corpus. See, for example, a 2021 study using in the Google Books *N*-Gram Corpus (Bollen et al., 2021), in which the authors claimed to have evidence that there has been a rise in "cognitive distortions" in English, Spanish, and German over the past two decades. The issue here is that the 2020 version of the corpus included a larger collection of fictional works because of changing partnerships between Google and other entities in the book publishing industry. There was a larger collection of fiction post-2000, meaning also a potential rise in embellished, colorful, and dramatic language, which could confound the researchers' thesis (Schmidt et al, 2021).
8. For instance, singular value decomposition on a term-co-occurence matrix (TCM); global log-bilinear regression on a distance-weighted TCM; neural network models of word prediction tasks; and variants of these methods that account for token positions, subword information, bidirectional representations.
9. Similarly, most text analysis tools available require a "segmentalism" be imposed on (written) language such that texts can be divided into discrete tokens—a Westernized and overly "crisp set" conceptualization of language use in general (Aronoff, 1992; Lowe, 2020; Port, 2007).

C1S14 References

C1P51 Abbott, A. (1988). Transcending general linear reality. *Sociological Theory*, *6*(2), 169–186.

C1P52 Abend, G. (2019). Thick concepts and sociological research. *Sociological Theory*, *37*(3), 209–233.

C1P53 Alexander, J. C., & Smith. P. (1993). The discourse of American civil society: A new proposal for cultural studies. *Theory and Society*, *22*(2), 151–207.

C1P54 Aronoff, M. (1992). Segmentalism in linguistics. *Linguistics of Literacy*, *21*, 71.

C1P55 Arseniev-Koehler, A. (2021). Theoretical foundations and limits of word embeddings: What types of meaning can they capture? *arXiv*. https://doi.org/10.48550/arXiv.2107.10413

C1P56 Arseniev-Koehler, A., & Foster, J. G. (2022). Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. *Sociological Methods & Research*, *51*(4), 1484–1539.

C1P57 Barthes, R., & Duisit, L. (1975). An introduction to the structural analysis of narrative. *New Literary History*, *6*(2), 237–272.

C1P58 Becker, H. S. (1984). *Art worlds*. University of California Press.

C1P59 Bellman, R. E. (1961). *Adaptive control processes*. Princeton University Press.

C1P60 Ben-David, Anat. (2016). What does the web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society*, *18*(7), 1103–1119.

C1P61 Bender, E. M. (2022, May 11). Look behind the curtain: Don't be dazzled by claims of "artificial intelligence." *The Seattle Times*.

C1P62 Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604.

C1P63 Biernacki, R. (2012). Rationalization processes inside cultural sociology. In J. C. Alexander, R. N. Jacobs, & P. Smith (Eds.), *The Oxford handbook of cultural sociology*. *Essay*. Oxford University Press.

C1P64 Biernacki, R. (2014). Humanist interpretation versus coding text samples. *Qualitative Sociology*, *37*(2), 173–188.

C1P65 Bird, S. (2022). Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Vol.* 1. *Long papers* (pp. 7817–7829). Association for Computational Linguistics.

C1P66 Birhane, A., & Guest. O. (2020). Towards decolonising computational sciences. arXiv. https://doi.org/10.48550/arXiv.2009.14258

C1P67 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.

C1P68 Bollen, J., Thij, M. T., Breithaupt, F., Barron, A. T. J., Rutter, L. A., Lorenzo-Luaces, L., & Scheffer, M. (2021). Historical language records reveal a surge of cognitive distortions in recent decades. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(30), Article e2102061118. http://doi.org/10.1073/pnas.2102061118

C1P69 Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Routledge & Kegan Paul.

C1P70 Bourdieu, P. (1990). *The logic of practice*. Stanford University Press.

C1P71 Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C. . . . Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *PNAS*, *119*(44), Article e2203150119.

C1P72 Cai, X., Huang, J., Bian, Y., & Church, K. (2020). Isotropy in the contextual embedding space: Clusters and manifolds. *International Conference on Learning Representations*.

C1P73 Cassirer, E. (1953). *Substance and function and Einstein's theory of relativity*. Dover.

C1P74 Childress, C. (2017). Under the cover: The Creation, production, and reception of a novel. Princeton University Press. https://doi.org/10.2307/j.ctt1vxm7zv

C1P75 Coffey, D. (2021, April 28). Maori are trying to save their language from Big Tech. WIRED UK. https://www.wired.co.uk/article/maori-language-tech#:~:text=Thanks%20to%20a%20%2413%20million,integrity%20of%20the%20original%20sound

C1P76 Coleman, J. (1958). Relational analysis: The study of social organizations with survey methods. *Human Organization*, *17*(4), 28–36.

C1P77 Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–15.

C1P78 Desikan, B. S., Hull, T., Nadler, E. O., Guilbeault, D., Abubaker Kar, A., Chu, M., & Ruggiero Lo Sardo, D. (2020). Comp-syn: Perceptually grounded word embeddings with color. arXiv. https://doi.org/10.48550/arXiv.2010.04292

C1P79 Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv. https://doi.org/10.48550/arXiv.2104.08758

C1P80 Emirbayer, M. (1997). Manifesto for a relational sociology 1. *American Journal of Sociology*, *103*(2), 281–317.

C1P81 Enggaard, T., Lohse, A., Pedersen, M. A., & Lehmann, S. (2023). Dialectograms: Machine learning differences between discursive communities. arXiv. https://arxiv.org/abs/2302.05657. https://doi.org/10.48550/arXiv.2302.05657

C1P82 Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv. https://doi.org/10.48550/arXiv.1909.00512

C1P83 Gabai, D., & Kerckhoff, S. (2015). William P. Thurston, 1946–2012. *Notices of the American Mathematical Society*. *American Mathematical Society*, *62*(10), 1318–1332.

C1P84 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92.

C1P85 Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

C1P86 Gerber, A. (2020). Black hole suns: Binarism and gravity in cultural fields. *Culture Unbound*, *12*(2), 393–411.

C1P87 Go, J. (2017). Decolonizing sociology: Epistemic inequality and sociological thought. *Social Problems*, *64*(2), 194–199.

C1P88 Green, A. I. (2008). The social organization of desire: The sexual fields approach. *Sociological Theory*, *26*(1), 25–50.

C1P89 Griffiths, T., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

C1P90 Hamilton, W., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Vol. 1. Long papers* (pp. 1489–1501). Association for Computational Linguistics.

C1P91 Hariri, R., Fredericks, E., & Bowers, K. (2019). Uncertainty in big data analytics: Survey, opportunities, and challenges. *Journal of Big Data*, *6*, 44.

C1P92 Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346.

C1P93 Healy, K. (2017). Fuck nuance. *Sociological Theory*, *35*(2), 118–127.

C1P94 Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560–575). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445918

C1P95 Jäkel, Frank, Bernhard Schölkopf, & Felix A. Wichmann. (2008). Similarity, Kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*(5), 297–303.

C1P96 James, W. (1890). *The principles of psychology* (Vol. 1). Henry Holt.

C1P97 Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2018). In defense of spatial models of semantic representation. *New Ideas in Psychology*, *50*, 54–60.

C1P98 King, G., Pan, J., & Roberts, M. E. (2013). How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326–343.

C1P99 Kong, L., de Masson d'Autume, C., Ling, W., Yu, L., Dai, Z., & Yogatama, D. (2019). A mutual information maximization perspective of language representation learning. arXiv. https://doi.org/10.48550/arXiv.1910.08350

C1P100 Korzybski, A. (1958). *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. Institute of General Semantics.

C1P101 Lee, M., & Martin, J. L. (2015). Coding, counting and cultural cartography. *American Journal of Cultural Sociology*, *3*(1), 1–33.

C1P102 Linzhuo, L., Lingfei, W., & Evans, J. (2020). Social centralization and semantic collapse: Hyperbolic embeddings of networks and text. *Poetics*, Article 101428.

C1P103 Lizardo, O. (2013). Re-conceptualizing abstract conceptualization in social theory: The case of the "structure" concept. *Journal for the Theory of Social Behaviour*, *43*(2), 155–180.

C1P104 Lowe, J. J. (2020). Segmental phonology in ancient India? *Language*, *96*(2), e97–e113.

C1P105 Malinowski, B. (2015). *Freedom and civilization*. Routledge. (Original work published 1947).

C1P106 Martin, C. D. (1993). The myth of the awesome thinking machine. *Communications of the ACM*, *36*(4), 120–133.

C1P107 Martin, J. L. (2003). What is field theory? *American Journal of Sociology*, *109*(1), 1–49.

C1P108 Mikolov, T., Yih, W.-T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 746–751). Association for Computational Linguistics.

C1P109 Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery.

C1P110 Mohr, J. W., Bail, C. A., Frye, M., Lena, J. C., Lizardo, O., McDonnell, T. E., Mische, A., Tavory, I., & Wherry, F. F. (2020). *Measuring culture*. Columbia University Press.

C1P111 Nachshon, Y., Cohen, H., & Maril, A. (2022). Empirical evidence for a semantic distance in a patch: Investigating symmetry and the triangle inequality violations. PsyArXiv. https://doi.org/10.31234/osf.io/fhkq8

C1P112 Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. (2015). Efficient non-parametric estimation of multiple embeddings per word in vector space. arXiv. https://doi.org/10.48550/arXiv.1504.06654

C1P113 Nelson, L. K. (2021). Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century U.S. south. *Poetics*, *88*, Article 101539.

C1P114 Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, *50*(1), 202–237.

C1P115 Neocleous, M. (2003). Off the map: On violence and cartography. *European Journal of Social Theory*, *6*(4), 409–425.

C1P116 Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

C1P117 Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, *8*(1), 447–471.

C1P118 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). Association for Computational Linguistics.

C1P119 Polo-Blanco, I. (2008). Alicia Boole Stott, a geometer in higher dimension. *Historia Mathematica*, *35*(2), 123–139.

C1P120 Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, *25*(2), 143–170.

C1P121 Rajaee, S., & Pilehvar, M. T. (2021). How does fine-tuning affect the geometry of embedding space: A case study on isotropy. arXiv. https://doi.org/10.48550/arXiv.2109.04740

C1P122 Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. arXiv. https://doi.org/10.48550/arXiv.2111.15366

C1P123 Rodriguez, M., & Merlo, P. (2020). Word associations and the distance properties of context-aware word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 376–385). Association for Computational Linguistics.

C1P124 Schmidt, B., Piantadosi, S. T., & Mahowald, K. (2021). Uncontrolled corpus composition drives an apparent surge in cognitive distortions. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45), 1–21.

C1P125 Schuster, R., Schuster, T., Meri, Y., & Shmatikov, V. (2020). Humpty Dumpty: Controlling word meanings via corpus poisoning. arXiv. https://doi.org/10.48550/arXiv.2001.04935

C1P126 Smith, C. (2011). *What is a person? Rethinking humanity, social life, and the moral good from the person up*. University of Chicago Press.

C1P127 Spillman, L. (2020). *What is cultural sociology?* John Wiley.

C1P128 Steinberg, S. (1976). *View of the world from 9th avenue* [Cover illustration]. *The New Yorker*, March 29.

C1P129 Stoltz, D. S. (2021). Becoming a dominant misinterpreted source: The case of Ferdinand de Saussure in cultural sociology. *Journal of Classical Sociology*, *21*(1), 92–113.

C1P130 Stoltz, D. S., & Taylor, M. A. (2021). Cultural cartography with word embeddings. *Poetics*, *88*, Article 101567.

C1P131 Swedberg, R. (2012). Theorizing in sociology and social science: Turning to the context of discovery. *Theory and Society*, *41*(1), 1–40.

C1P132 Taylor, M. A., & Stoltz, D. S. (2021). Integrating semantic directions with concept mover's distance to measure binary concept engagement. *Journal of Computational Social Science*, *4*(1), 231–242.

C1P133 Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*(2), 123–154.

C1P134 Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, *44*(6), Article e12844.

C1P135 Wheatley, G. H. (2013). Reasoning with images in mathematical activity. In L. D. English (Ed.), *Mathematical reasoning: Analogoies, metaphors, and images* (pp. 281–298). https://doi.org/10.4324/9780203053485-18/reasoning-images-mathematical-activity-grayson-wheatley