The Use and Misuse of Semantic Space for Literary Criticism

James E. Dobson
Dartmouth College

I want to talk this afternoon about the possibilities of a computational comparative literature by thinking through some of the possibilities and problems with the concept of semantic space. Comparative literature as field has been the site of numerous important shifts and turns within the humanities and the methodologies and values found within this field have, to some degree, served as a bellwether to larger trends affecting other fields of literary study. Speaking about the possibilities of a computational comparative literature also enables me, an outsider to this symposium's discussion of non-western and pre-modern literatures, to provide some focus for my remarks.

Jonathan Culler's influential 2006 essay "Whither Comparative Literature" reconstructs the history of comparative literature and rehashes some of the major debates within the field. The field, according to Culler, began with "the study of sources and influences" and then shifted to an analysis of intertextual studies. Because it lacked the national and language focus of area studies comparative literature then shifted to problematize its own potential units of study, Culler points to genre, period, and themes as the central units. This non-bounded examination and comparison of units enabled comparative literature to become the testing ground for literary theory. In what many would characterize as our post-theory moment, comparative literature has struggled to define its place within the university. As universities have collapsed humanities departments together and dropped foreign language departments, critics have suggested that comparative literature should become the home for global anglophone literature—one of the few job market categories experiencing growth at the moment—while others suggest that comparative literature

might be a refuge for literary studies as cultural studies continues its dominance with the humanities.

The crucial disciplinary markers found within Culler's history are all present within work participating in contemporary quantitative criticism, digital humanities, distant reading, macroanalysis, computational literary studies, etc. The examination of sources and influences are found in work making using of stylistics. We see computational work addressing intertextuality appearing more frequently, especially in recent essays published by our hosts here today. The turn to literature as opposed to other cultural texts seems almost obvious as just about every computational analysis, even those emerging from that area known as cultural analytics, seems to take as its object canonical or canon-adjacent university library cataloged novels and poems. Computational criticism has mostly avoided national categories but it has, as this event acknowledges, foregrounded modern anglophone literature. British and American literature, especially those unencumbered and easily digitized and corrected nineteenth century novels, has been almost the exclusive focus in the major works of digital scholarship during the past decade.

Some of the most innovative recent findings applying computational methods to literature at scale—from Andrew Piper's account of the affective structure that he calls "conversional reading" to Ted Underwood's "life span of genre"—suggest that there might be structural features within some textual objects that are shared across language and national boundaries. The possibility that there are sets of features that form patterns of signal that enable scholars to once again map the literary field is exciting and yet potentially troubling. While some critics of quantitative formalist research call into question the robustness of such signals it is my concern today to think about the basis for the comparison of sampled data from multiple literatures and from multiple cultural moments. I invoke the phrase "cultural moment" to raise the spectre of

historicism. I have an ambivalent relationship to historicism. I find the overly rigid historicist practices, such as the cultural poetics of the new historicism reductive and simplistic. The presumption that culture operates in a synchronic fashion and that all members of "a culture," whatever that might suggest, have equal access or exposure to the discursive field has long impressed me as an unsupportable fantasy. Yet I find in much critical scholarship and especially in computational work, a need for a stronger sense of the historicity of our objects and methods. Any literary criticism belonging to the humanities rather than to the empirical laboratories of the sciences and social sciences cannot willy-nilly apply concepts and definitions from the present to the past without at least acknowledging and framing the anachronism or the critic-cum-researcher's will-to-presentism.

A key enabling technology for examining patterns of signals across cultures would be semantic space. Semantic space is one of several names for the quantified representation of relationships among sampled texts. Once defined, other representations may be fitted or aligned to this space, providing the basis for the extraction of features that enable large-scale transformations of the measured data. I'm using the "semantic space" rather than the more simplified vector space or even the generic notion of word space because I want to gesture to the way in which all such constructed spaces are meaningful—especially in the defining of the boundaries of that space. Spaces are inscribed. The hermeneutic act of understanding and meaning making, as I've argued elsewhere, are an embedded feature of all of our models. There is no model based on extracted textual features that can be said to exist *de novo.* And yet many scholars and critics still make the claim for deferred interpretative activity within computational literary studies. I could point to just about any recent writing on computational approaches to literary analysis but allow me to quote from Taylor Arnold and Lauren Tilton's essay "Distant

Viewing: Analyzing Large Visual Corpora" that appeached this year in the journal *Digital Scholarship in the Humanities*. Arnold and Tilton gloss these approaches as such:

> The explicit code system of written language provides a powerful tool for the computational analysis of textual corpora. Methods such as topic modeling, term frequency-inverse document frequency, and sentiment analysis function directly by counting words, the smallest linguistic unit that can be meaningfully understood in isolation. The interpretive act of understanding these units may be delayed until the models are applied. It is, for example, only after applying latent Dirichlet allocation and finding a topic defined by the words 'court', 'verdict', and 'judge' that we are forced to decode the meanings of the words. (2)

For Arnold and Tilton measured semantic space resides outside of interpretive decision making. What they call the "explicit code system" of writing exists, for them, in an undifferentiated space. Tossing such varied produces such as sentiment analysis and topic modeling in the same bag, as it were, reflects an entirely undertheorized notion of text analysis. The construction of representational space in both depends upon prior decisions, selections, and shaping of the inputted objects and deployed methods and, in the case of sentiment analysis, this space depends entirely on extrinsic data that is used to defining, in terms of the model, the meaning of language. I'm critiquing this account at length because it reflects a continued blind-spot for computational critics. Models are merely encoded cultural baggage.

Michael Gavin's 2018 essay "Vector Semantics, William Empson, and the Study of Ambiguity" introduces semantic space and vector operations to skeptical humanists. Gavin's account of contextual meaning via KWIC or keyword in context raises some interesting problems. "Vector semantics," he writes, "intervenes in the critical reading process by providing

more finely grained models of individual words and phrases and by showing how meanings are transformed in context" (668). Richer context, i.e., more examples from the wild, in this notion of semantic space will provide more meaningful search results. Producing increasingly complicated vector space models from term-document matrices to windowed keywords in context from fairly well-known and defined collections of documents—he uses a collection of seventeenth century English texts found in EEBO's collection—Gavin demonstrates the power of these methods within particular contexts. Because of this contextualization, he is able to find meaningful results in what he calls composite "semantic neighborhoods" that define multiword concepts and ideas, in his case a line of Milton's *Paradise Lost* minus stopwords, by adding extracted vectors of keywords together into an "aggregate quasi word" (671) and then visualizing individual words in nearby space. Gavin's essay takes up the problem of ambiguity in language use and finds the solution to this problem in the selection of richer context.

Yet the context cannot be too rich. Adding earlier texts to Gavin's curated seventeenth century collection would most likely result in poorly defined semantic neighborhoods. If meanings of words are transformed in context, however, we humanists will surely want something better than a mapping between contexts. The notion of a single semantic space assumes what we might want to call the new historicist understanding of culture as synchronic, universal, and shared. These are the assumptions underlying much thinking about semantic space transformation including, especially, translation activities.

The most popular methods for constructing semantic space vector models at present are classified as distributional word embeddings. These embeddings were discovered among the detritus, the computational leftovers, of the Skip-gram neural network models. In 2013, Tomas Mikolov and his colleagues at Google introduced the word2vec model with some fascinating

example applications. Simple mathematical operations on vectors of words created by the model resulted in finding potentially meaningful vectors of semantically related words or n-gram phrases. Nearest neighbors in the resulting vector space provided by their pretrained model produced meaningful clusters of words and concepts but the real interest was in the vector calculations presented in their most compelling form as analogically reasoning tasks. The vector for the word Spain subtracted from that of Madrid added to vector for France resulted in a value in space closest to Paris. In a subsequent paper Mikolov et al., described the analogy King is to Man as Woman is to X, resulting in a value in space closest to Queen.

The initial critique of the Mikolov et. al. word2vec model was that it lacked enough context to disambiguate multiple meanings. The relational field established by this model and the pretrained word representations, which were produced from one million of the most common words from a six billion token Google News corpus, in particular, did not properly handle different semantic and linguistic contexts. There are two problems here. One problem is intrinsic to the archive or corpus and the other a basic assumption underlying the rather simplistic model of language used by these researchers.

Solutions to or methods capable of ameliorating some aspects of the second problem are being developed. Mikolov's sophomore effort, a method he calls fastText, uses what are called sub-word units or wordpieces, these are partial word objects, to fit either unknown words, words that are not in the model's vocabulary, or misspelled words—Mikolov has left Google for Facebook and apparently there are more misspelled words encountered on this social media platform—to the larger semantic space. FastText might be an improvement on word2vec because of this ability to point roughly to the space where a word should fit but this sub-word model introduces new assumptions—it, like word2vec, was designed for and trained on English

language resources—and potentially many errors when comparing semantic space models trained on different input data resources. The word pieces, too, may risk amplifying culturally problematic semantic information contained in roots.

One may, of course, as a solution to the first problem, trivially train their own model on their own collection of texts leading, potentially, to a better model of sample space or the risk of not having enough data to produce a complex and large enough space to generate useful measurements and semantic retrieval. A recent approach called "a la carte embedding" published by Mikhail Khodak and Nikunj Saunshi et. al., uses the same linear transformation used by others to fit independently produced models derived from historically partitioned input data to "learn feature representations from the average word embeddings in the feature's available contexts." Such methods assume a relative compatibility or rather comparability between semantic spaces. Semi-supervised neural network models including ELMo, which models each token in relation to its embedded input sentence (as opposed to the stripped continuous bag of words approach used in the original word2vec) also propose combining modeled spaces. These and other recent efforts suggest that comparative work of modeled space across semantic and cultural contexts to identify similar or different literary units is going to become increasingly easier.

Now obviously demonstrating the cross cultural or transhistorical validity of Mikolov et. al's gender binary finding is interesting but what does that really tell us? And is that the criteria by which we want to make claims about cultural artifacts? Is that the hill on which anyone would choose to die?

The challenge for a computational comparative literature is located precisely within the problematic notion of semantic space. Given a large enough historicizing, periodizing archive, it

can be useful for literary studies. It seems likely that there will be an increasing number of critical experiments devoted to showing the existence of table patterns of signals that can be used to produce comparisons across literatures. Linear regression or other methods can align and warp the difference between sampled literatures and will no doubt improve the strength of predictive models and comparative studies. But how might such semantic spaces be vulnerable to some of the existing theoretical and critical interventions into comparative literature as a subject? I mean to say that in addition to the obvious archival gaps and silences within the digitized written historical record, the concept of comparability itself might be rather suspect. Jonathan Culler, to return by way of conclusion to my beginning cautions comparativists that frequently the basis for comparison turns to be an understanding of the world that is deeply Eurocentric. Culler argues that "[the] intertextual nature of meaning – the fact that meaning lies in the differences between one text or one discourse and another – makes literary study essentially, fundamentally comparative, but it also produces a situation in which comparability depends upon a cultural system, a general field that underwrites comparison." The coding and the desire to code, to use the language of Arnold and Tilton, of the semantic space needs to be decoded prior to the statistical transformation of this space and subsequent decodings. Let me end with a historicizing provocation: why are we comparing and what are the conditions of possibility of comparison as such in this particular moment?

Query triplet (A - B + C)? berlin germany france
paris 0.884514
dubourg 0.776413
louveciennes 0.770261
bourges 0.767376
charlesbourg 0.764689
dessaulles 0.753519
pompignan 0.752995
valenciennes 0.751361
faubourg 0.750989
maubourg 0.749332

Query triplet (A - B + C)? man king queen
woman 0.765593
blonde 0.725058
girl 0.706779
mannequin 0.696525
girly 0.679173
sexy 0.673635
showgirl 0.667835
sleuthing 0.666855
caress 0.665466
irishwoman 0.659996

./fasttext nn result/fil9.bin [woman]
madwoman 0.809427
everywoman 0.805153
girl 0.793177
womans 0.787899
womanhood 0.780942
englishwoman 0.778515
irishwoman 0.777088
countrywoman 0.772686
henchwoman 0.757244
washerwoman 0.755074