Referee report on Kozlowski et al. (2018)

Kozlowski et al. (2018), in their paper titled "The Geometry of Culture: Analyzing Meaning through Word Embeddings," present a computational text-analysis approach to revealing patterns in cultural meanings and relationships between cultural categories within and across cultures, synchronically and diachronically. Or, from a different perspective, they present a reading of computational text-analysis informed by sociological and related theories of meaning and culture.

While their paper is primarily a methodological, seeking to showcase the possibilities of a new (application of) a method, at the root of it is also an empirical question: Can text-analysis using word embedding models capture culturally salient valences of meaning? To address this question, they surveyed 398 respondents through Amazon Mechanical Turk about cultural associations (with respect to gender, class, race) of select tokens (words relating to different categories such as sports, music, names, food, etc.), and then compared the survey results to the performance of the word embedding model that was employed on four different datasets. Additionally, the authors work to validate their approach by showing that the model accurately sorts popular US named across 20th century by gender, or recent US politicians by political orientation. Finally, the authors use the word-embedding models to trace cultural shifts

in the US over time, as well as cultural differences between the US and the UK at the turn of the 20th century and corroborating their findings with a sociological information about the relevant contexts.

As this paper is an exposition of a novel approach, the authors' methodology is appropriate and sufficient: they use simple examples across different contexts in a way that clearly shows the potential contribution of theories of meaning to text analysis, and of this kind of text analysis to sociology and related fields. However, in its simplicity, the approach also seems a bit lacking. Most notably, potential limitations of the approach are either glossed over (e.g. the difficulty of investigating race as noted on p.34 and p.38, possibly due to changing terminologies over rime) or not discussed at all. While the paper presents the construction of dimensions of interest using 'appropriate sets of antonym pairs' (p.14) as a fairly straightforward process, it does so by using what might be the most straightforward dimensions. It is therefore unclear how many antonym pairs one has to be able to construct in order for the model to perform well (the model's sub-par performance on the n-Gram dataset with respect to race, using only four pairs might be telling here) and, by extension, it is difficult to tell how useful this approach might be with other, maybe messier, dimensions. For example, thinking about another dimension of interest to sociologists—that of religion—antonym pairs do not come to mind as easily. Additionally, being that the paper itself focuses on cultural categories and

meaning, I think it would benefit from a more detailed reflection on the choice of antonym pairs, as this choice in itself reproduces cultural categories one is interested in.

In terms of the broader context, this paper clearly positions itself with respect to previous work on text analysis, as well as relevant theories of language and culture. I am somewhat reluctant to comment more as it is unclear where the authors are seeking to publish. They mention multiple times that they present a novel method to be used in sociological analysis of culture, which would make it seem their primary audience is (computational) sociologists. If that is the case, I do find the theoretical review somewhat lacking. For example, discussing in more detail how the notion of antonym pairs draws on structuralism (and lexical semantics), while —according to the authors— 'remains free of many of the assumptions of structuralism' would help ground the dimension-construction step of the approach. Furthermore, while the authors do set up, in passing, a comparison between the word-embedding model and human learning associations through exposure to the same corpus (p.3), I do think this analogy could be productively explored further, e.g. mentioning the literatures on language acquisition and language socialization (c. Ochs, Elinor, and Bambi B. Schieffelin. 2012. The theory of language socialization. In *The handbook of language socialization*.)

I did not catch any grammatical or spelling errors, although there are some parts of the texts that could be improved. On page 5, in "and thus text has served as a key

source of data," it might be more accurate to replace text with 'language'—not only to omit reference to written text, but also to avoid unnecessarily referencing 'text' as a not-necessarily-linguistic object of inquiry in linguistic anthropology. On page 10, 'abutting' can be replaced by 'adjoining' or 'neighboring' to make it more accessible to international audiences, as 'abutting' is a less common word. Finally, the sentence starting with "Indeed, the low-dimensional projection of correspondence analysis," on page 19, the sentence starting with "We then compare focal dimensions" on page 29, and the one starting with "Having established a strong correspondence" on page 38 are all a bit difficult to parse and would benefit from being simplified.

Thinking about the extension of the paper, I would definitely like to see how the model performs (and maybe fails) with a 'messier' dimension. This could be religion (e.g. Protestant vs. Catholic, Jewish vs. Christian, Religious vs. Atheist) though I can't really think of good antonym pairs, or maybe something less salient but pervasive such as the mind-body or the reason-emotion dualism. Antonyms that would be useful for this could be along the lines of 1) 'mind-body', 'though-action', 'human-animal' and 2) 'reason-emotion', 'mind-heart', 'thinking-feeling'. These could be compared to other cultural dimensions (such as gender and class), but it would also be interesting to see where certain occupations, religions, ideologies etc, fall along these dimensions.

Furthermore, having read the paper, I've definitely had a lot of ideas about implementing this kind of an analysis in my own research on suicide. For example, I could use the word-embedding models to visualize how associated suicide is with a particular gender, race or age group, or how suicide is talked about differently for different groups (e.g. in terms of strength for men, and weakness for women). Or maybe, if I got to look at corpus of scientific articles on suicide, I would want to look at how different ideas and findings map onto the dimensions of social-individual as well as quantitative-qualitative.