## Carlsen and Ralund (not yet published): Computational grounded theory revisited: from computer lead to computer assisted text analysis

C and R gives a critique of Laura Nelson’s propose for a computational grounded theory and make, and they suggest an alternative approach that they term CALM, which they claim differ by putting the researcher in charge of all operations but uses computer assisted methods to locate patterns and word similarities.

**Grounded theory**

Or GT is defined as theory derived from systematic qualitative inquiry and it differs from theorizing based on abstract and generalizing theory that are being forced on data. Thus, a focus of the article is on how to abductively derive grounded theory and not how to deductively test theories. Anthropologist often have a grounded theory approach, and there is a focus on how the proposed approach, CALM, supports more anthropological approaches to theorizing. (page. 2)

**Critique of Computational grounded theory**

Laura Nelson’s approach Computational Grounded Theory (CGT) has 3 steps pattern discovery, pattern refinement and pattern confirmation. C and R claims that the CGT relies on three assumptions: 1) unsupervised models return natural clusters of meaning 2) that the researcher can learn from minimal immersion 3) that the best way of validating a classification is through indirect measurement. The first step is pattern discovery entails that the researcher relies on a LDA Topic Model to identify patterns in their empirical material. This means that the CGT is computationally led to ensure against the biased and constrained researcher. The topic model is reproducible, and the categories that emerge from the topic model can be interpreted of the researcher. C and R criticizes the LDA topic model for the need to choose two hyperparameters: number and size of topics, and they write that it is not good at identifying small topics and because of this the small topics can be mixed with others. C and R thus states that researchers cannot rely on a LDA led approach. (page 3-4)

For the next step of CGT, pattern refinement, the researcher does what Laura Nelson terms Computationally guided deep reading. Through this step the researcher is supposed to confirm interpretation of categories without having to read the whole text corpus. In this step it is assumed that the the topics are correctly defined in the first step. C and R criticizes this step for not demanding the researcher to have any prior knowledge of the field, and to be able to do a prober analyses simply by reading little text. They say that this goes against traditions in anthropology where context is always important to consider. The most extensive critique of R and C for this step is that as the LDA cannot be trusted it does not make any sense to continue analyses solely based on this (page 5)

The last step of CGT is pattern confirmation, which means that the researcher validates their analyses. Laura Nelson recommends indirect validation strategies such as concurrent validity and/or predictive validity. (I think) Concurrent validity means judging whether top words in the topic model make sense to represent a topic. C and R argue and present evidence to show that this is not sufficient. (page 5-6)

“Summing up on our run through of computational grounded theory we believe that some aspects of the

procedure makes it less in line with concerns and ambitions within anthropology. LDA topic models can only

support weak discovery of what the researcher already” (page 6)

**Presentation of CALM – from computer led to computer assisted learning and measurement**

C and R present an alternative approach to making computationally grounded theory. They call their approach CALM and argues that it puts the human back in the loop by facilitating a computer assisted learning as opposite to the computer led learning. The methodological framework moves from 1) discovery; through 2) interpretation; and to 3) classification and measurement.

The first step, discovery, is exploratory either through qualitative methods or through unsupervised models, here they recommend HSBM as it is good at identifying clusters and smaller categories in the data. R and C clam that the main difference from the CGT is that the words clustered together by the model are only taken to be an uncertain sign of a given category. The search terms provided by the HSBM can be extended by finding similar words with Word2Vec. The end product of this first phase is a set of categories with a connected set of search terms. None of these taken to be empirically grounded yet. (page 7)

In the second step the researcher applies the search terms to identify different relevant documents to read a lot of them, make comparisons, and then to analyze each category. When that is done a coding scheme is developed. The process of developing codes and analyzing categories must be documented. (page 8)

For the third step classification and validity the coding scheme is used to classify content in order to train and test a machine learning model. At this stage, the coding process is closed. In this way the researcher’s codes are replicated, and the approach is thus assisted instead of being led.

