Automated Linguistic Analysis of Patients with Dementia and Mild Cognitive Impairment

Ellen Korcovelos[1], Serguei Pakhomov[2], PhD, Bridget T. McInnes[1], PhD

[1]Virginia Commonwealth University, Richmond, VA
[2]University of Minnesota, Minneapolis, MN

Background: Individuals with dementia often have difficulties with the transmission of words in written and spoken communication rather than a deficit in word knowledge and word meanings (Burke, D., & Shafto, M., 2004). Current assessment techniques to determine the stages of dementia as they relate to transmission are primarily conducted manually. The disadvantage of these techniques is that they can be subjective and may not provide the detail and precision of measurements that are required to identify the full range of syndromes that exist for dementia patients (Pakhomov, et. al 2010). The present study aims to develop automated, objective metrics to quantify the linguistic characteristics of speech. In this work, we assess 1) type to token ratios (T:T) of unique phrases, 2) presence of indefinite words using information content (IC) scores (provided by WordNet::Similarity), and 3) frequency of filled pauses (i.e. "uh" and "um", coughing and laughter).

Methods: Transcribed interviews of the Cookie Theft stimulus of University of Pittsburg's DementiaBank data (N = 520) contained participants of three categories: probable Alzheimer's disease (AD; N = 235), Mild Cognitive Impairment (N = 43), and those without dementia-like symptoms (N= 242). Each transcribed file was cleaned of extraneous data (i.e. investigator commentary, unnecessary symbols, part of speech tagging) and run through newly designed algorithms to determine T:T, IC variance, and pause word frequency.

Results: We evaluated the results of the characteristics using the unpaired T-test to identify significant differences of means between the probable AD and control group, and the MCI and control group. Significant differences were identified between the probable AD and the control group for T:T and pause words, but not IC. No significant difference was detected among the MCI and control group.

Conclusions: Automated language analysis offers great potential for identifying possible biomarkers for patients with dementia. The results show that using speech-based variables as a way to characterize cognitive function may be used to aid in detecting changes that are indicative of dementia.