Specifications and Methodology for Language-Related Data Acquisition and Analysis in the Domain of Dementia Diagnostics

Dimitrios Kokkinakis[a], Kristina Lundholm Fors[a], Eva Björkner[a], Arto Nordlund[b]
[a]Department of Swedish; [b]Department of Psychiatry and Neurochemistry
University of Gothenburg
firstname.lastname@gu.se

Abstract

This paper outlines the initial stages of a project that aims to build and use a corpus with data samples acquired from people diagnosed with subjective or mild cognitive impairment and healthy, age-matched controls. The data we are currently collecting consists of audio-recorded spoken language samples; transcripts of the audio recordings and eye tracking measurements. From these data we plan to extract, evaluate and model features to be used for learning classification models in order to test how well a differentiation between the aforementioned subject groups can be made. Features will also be correlated with outcomes from e.g. other language-related scores, such as word fluency, in order to investigate whether there are relationships between various variables.

1. Introduction

Efficient tools for routine dementia screening in primary health care could provide specialist centers the opportunity to engage in more demanding, advanced investigations, care and treatment. New paths of research for acquiring knowledge about Alzheimer’s disease (AD) and its subtypes using Computational Linguistic/Natural Language Processing tools based on the exploration of several complementary modalities and parameters, such as speech analysis and/or eye tracking could be integrated into established neuropsychological, memory and cognitive tests in order to explore potential (new) biomarkers for AD. This paper describes efforts to acquire data from people with subjective (SCI) and mild cognitive impairment (MCI) and healthy, age-matched controls in order to acquire, assess, analyze and evaluate potential useful linguistic and extra linguistic features and build classifiers that could differentiate between benign and malignant forms of cognitive impairment. Non-invasive and cost-effective methods that could identify individuals on an early preclinical stage remains a priority for health care providers.

2. Background

Language in all its forms plays an important role in dementia diagnosis. New findings provide important complementary information on the cognitive status of individuals and promising results have recently thrown more light on the importance of language as an essential factor that can shed light on the presence, severity and impact of the disease (Ferguson et al., 2013). The work by Snowdon et al. (2000) was one of the earliest studies which showed a strong correlation between low linguistic ability early in life and cognitive impairment in later life by analyzing autobiographies of American nuns (“The Nun Study”). Snowdon et al. could predict who could develop AD by studying the degradation of the idea density and syntactic complexity on the nuns’ autobiographical writings. Since then, research in CL/NLP in the area of processing data from subjects with mental, cognitive, neuropsychiatric, or neurodegenerative impairments has grown rapidly. Automatic spoken language analysis, including transcriptions, and eye movement measurements are two of the newer complementary diagnostic tools with great potential for dementia diagnostics (Roark et al., 2011; Laske et al., 2014). Although language is not the only diagnostic factor for cognitive impairment, several recent studies have demonstrated that automatic linguistic analysis, primarily of speech samples, produced by people with mild or moderate cognitive impairment compared to healthy individuals can identify objective evidence and measurable (progressive) language disorders (Yancheva et al., 2015; Fraser & Hirst, 2016). Analysis of eye movement is also a relevant research technology to apply, and reading texts by people with and without mild cognitive impairment may give a clear ruling on how reading strategies differ between these groups, an area that has so far not been researched to any significant extent in this domain (Fernández et al., 2013; Molitor et al., 2015).

3. Ethical Considerations

The ongoing Gothenburg mild cognitive impairment study (Wallin et al., 2016) is an attempt to conduct longitudinal in-depth phenotyping of patients with different forms and degrees of cognitive impairment using neuropsychological, neuroimaging, and neurochemical tools. The study is clinically based and aims at identifying neurodegenerative, vascular and stress related disorders prior to the development of dementia. The overall Gothenburg MCI-study and the current research are approved by the local ethical committee review board. The project aims at a homogeneous group with respect to age and education level (50 with SCI/MCI and 50 controls) and written informed consent is obtained from all participants in the study.

4. Methodology

For feature extraction the project uses various types of data, audio recordings, transcriptions and eye tracking measurements. *Audio Recordings*: for the acquisition of the audio signal we use the Cookie-theft picture from the *Boston Diagnostic Aphasia Examination* (Goodglass & Kaplan, 1983). During the presentation of the Cookie theft stimuli the subjects are asked to tell a story about the picture and describe everything that can be observed while the storytelling is recorded. The picture provides a standardized test that has been used in various studies in the past, therefore comparisons can be made based on previous results (Williams et al., 2010; Fraser & Hirst, 2016).

Moreover, in order to allow the construction of a comprehensive speech profile for each research participant, the speech task also includes reading aloud a short text from the *International Reading Speed Texts* collection presented on a computer screen (Trauzettel-Klosinski et al., 2012). Two texts from this collection are used in the eye tracking experiment (cf. section 5), but only one of those texts is read aloud and thus combined with eye-tracking recording (Meilan et al., 2014). *Verbatim Transcriptions*: the textual part consists of manually produced transcriptions of the two audio recordings previously outlined. During transcription, special attention will also be paid to non-speech acoustic events including speech dysfluencies consisting of filled pauses, false-starts, repetitions and hesitations. *Eye-tracking*: until now, eye tracking has not been used to investigate reading for individuals with SCI/MCI on a large scale, possibly due to the number of procedural difficulties related to this kind of research. On the other hand, the technology has been applied in a growing body of experiments related to other impairments such as dyslexia. For the experiments we use monocular eye tracking with head stabilization and a real-time sample access of 1000Hz. While reading, the eye movements of the participants are recorded while interest areas around each word in the text are defined by taking advantage of the fact that there are spaces between each word in the text. The eye-tracking measurements are used for the detection and calculation of fixations, saccades and backtracks. *Comparison over a two-year span*: the previously outlined experiments will be repeated 2 years after the first recordings taking place during the second half of 2016. This way we want to compare and analyze whether there are differences and at which level and magnitude between these recordings and the features acquired from them.

5. Analysis

The envisaged analysis and exploration intends to extract, evaluate and combine a number of features from the three modalities selected to be investigated. These are speech-related features, text/transcription-related features and eye tracking-related features. *Speech-related features* (pause frequency, filled pauses, total pause duration and idea density, formants, pitch, volume and spectral noise measures); cf. Roark et al., 2011; Yancheva et al., 2015. *Text/transcription-related features* (linguistic variables such as Frazier and Yngve scores syntactic complexity; dependency distance; word types and lexical distribution measures) cf. Pakhomov et al. (2010). *Eye Tracking-related features* (*fixations*: i.e. the state the eye remains still over a period of time; *saccades*: i.e. the rapid motion of the eye from one fixation to another and *backtracks:* i.e. the relationship between two subsequent saccades where the second goes in the opposite direction than the first) cf. Fernández et al., 2013; Molitor et al. (2015). *Correlation analysis*: we intend to further perform correlation analysis with the features previously outlined and various measures from language-related tests performed in the Gothenburg MCI-study, and often used when assessing possible dementia. Such tests include the token test, subtest V, is a test of syntax comprehension; the Boston naming test and the word fluency FAS test (the number of words initiated by the letters F, A, S). This investigation intends to identify whether there are features (highly) correlated with i.e. the MCI class, yet uncorrelated with each other i.e. the healthy controls or SCI.

6. Conclusions and Future Work

We have outlined work in progress towards the design and development of data resources and a set of features to be used for experimentation in both machine learning tasks (differentiation between SCI/MCI and healthy adults); as benchmark data for future research in the area and we also intend to repeat the experiments two years after in order to assess possible changes at each level of analysis.

Acknowledgements

This work has received support from Riksbankens Jubileumsfond - The Swedish Foundation for HSS through the grant agreement no: NHS14-1761:1

References

A. Ferguson, E. Spencer, H. Craig and K. Colyvas. 2014. Propositional Idea Density in women’s written language over the lifespan: Comp. analysis. *Cortex* 55. 107-121.

G. Fernández et al. 2013. Eye Movement Alterations during Reading in Patients with early Alzheimer Disease. *Inv Ophth&Vis Sc*. Vol. 54, 8345-52

K. Fraser and G. Hirst. 2016. *Detecting semantic changes in Alzheimer’s disease with vector space models*. Workshop: RaPID-2016. Pp. 1-8. Portorož Slovenia.

H. Goodglass and E. Kaplan. 1983. *The Assessment of Aphasia and Related Disorders*. Lea&Febiger, PA, USA.

C. Laske et al. 2014. Innovative diagnostic tools early detection of Alzheimer’s disease. *Alz & Dementia.* 1-18.

J. JG. Meilán et al. 2014. Speech in Alzheimer’s Disease: Can Temporal and Acoustic Parameters Discriminate Dementia? *Dement Geriatr Cogn Disord*;37:327–334.

RJ. Molitor, PC. Ko and BA. Ally. 2015. Eye Movements in Alzheimer’s Disease. *J of Alz Dis* 44, 1–12. IOS.

S. VS Pakhomov et al. 2010. A computerized technique to assess language use patterns in patients with frontotemporal dementia. *J Neuroling*. 23(2): 127–144.

B. Roark, M. Mitchell, J-P. Hosom, K. Hollingshead and J. Kaye. 2011. *Spoken Language Derived Measures for Detecting Mild Cognitive Impairment*. IEEE Trans Audio Speech Lang Processing. 19(7): 2081–2090.

DA. Snowdon, L. Greiner and WR. Markesbery. 2000. Linguistic ability in early life and the neuropathology of Alz. disease and cerebrovascular disease. Findings from the Nun Study. *Annals of the NY Ac of Sc*. 903:34-8.

S. Trauzettel-Klosinski, K. Dietz and the IReST Study Group. 2012. Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST. *Investigative Ophth&Visual Sc*. Vol. 53:9.

A. Wallin et al. 2016. The Gothenburg MCI study. *J Cereb Blood Flow Metab*. 36(1):114-31.

C. Williams et al. 2010. *The Cambridge Cookie-Theft Corpus: A Corpus of Directed and Spontaneous Speech of Brain-Damaged Patients and Healthy Individuals*. 7th LREC. Pp. 2824-2839. Malta.

M. Yancheva, K. Fraser and F. Rudzicz. 2015. *Using linguistic features longitudinally to predict clinical scores for Alzheimer’s disease and related dementias*. 6th SLPAT Workshop. pp. 134–139, Dresden, Germany.