LREC 2018 Workshop

RaPID-2: Resources and ProcessIng of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments

PROCEEDINGS

Editor

Dimitrios Kokkinakis

ISBN: 979-10-95546-26-9
EAN: 9791095546269

Tuesday 8th of May 2018

Proceedings of the LREC 2018 Workshop
"Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2018)"

8th May 2018 – Miyazaki, Japan

Edited by Dimitrios Kokkinakis

http://spraakbanken.gu.se/eng/rapid-2018

ISBN: 979-10-95546-26-9
EAN: 9791095546269

Acknowledgments: This work has received support from Riksbankens Jubileumsfond - The Swedish Foundation for Humanities and Social Sciences through the grant agreement no:NHS14-1761:1; the Centre for Ageing and Health (AGECAP) and the Swedish Common Language Resources and Technology Infrastructure (Swe-Clarin).

Organising Committee

Dimitrios Kokkinakis, University of Gothenburg, Sweden

Kristina Lundholm Fors, University of Gothenburg, Sweden

Kathleen Fraser, University of Gothenburg, Sweden

Charalambos Themistocleous, University of Gothenburg, Sweden

Graeme Hirst, University of Toronto, Canada

Alexandra König, Geriatric Hospital Nice and the University of Côte d'Azur, France

Frank Rudzicz, Toronto Rehabilitation Institute and the University of Toronto, Canada

Programme Committee

Osvaldo Agamennoni, Universidad Nacional del Sur, Argentina

Jan Alexandersson, DFKI GmbH, Germany

Eiji Aramaki, Kyoto University, Japan

Heidi Christensen, University pf Sheffield, UK

Simon Dobnik, University of Gothenburg, Sweden

Marie Eckerström, University of Gothenburg, Sweden

Jens Edlund, KTH - Royal Institute of Technology, Sweden

Gerasimos Fergadiotis, Portland State University, USA

Kathleen Fraser, University of Gothenburg, Sweden

Valantis Fyndanis, University of Oslo, Norway

Peter Garrard, St George's, University of London, UK

Kallirroi Georgila, University of Southern California, USA

Lena Hartelius, the Sahlgrenska Academy, University of Gothenburg, Sweden

Graeme Hirst, University of Toronto, Canada

Kristy Hollingshead, Florida Institute for Human & Machine Cognition (IHMC), USA

Heidi Christensen, University pf Sheffield, UK

Richard Johansson, University of Gothenburg, Sweden

Masatomo Kobayashi, IBM Research, Tokyo, Japan

Dimitrios Kokkinakis, University of Gothenburg, Sweden

Alexandra König, Geriatric Hospital Nice and the University of Côte d'Azur, France

Peter Ljunglöf, Chalmers University of Technology, Sweden

Kristina Lundholm Fors, University of Gothenburg, Sweden

Marcus Nyström, University of Lund, Sweden

Paul Rayson, Lancaster University, UK

Vassiliki Rentoumi, SKEL, NCSR Demokritos, Greece

Fabien Ringeval, Université Grenoble Alpes, France

Frank Rudzicz, Toronto Rehabilitation Institute and the University of Toronto, Canada

Kairit Sirts, University of Tartu, Estonia

Hironobu Takagi, IBM Research, Tokyo, Japan

Charalambos Themistocleous, University of Gothenburg, Sweden

Athanasios Tsanas, University of Oxford, UK

Magda Tsolaki, Aristotle University of Thessaloniki, Greece

Spyridoula Varlokosta, National and Kapodistrian University of Athens, Greece

Åsa Wengelin, University of Gothenburg, Sweden

Preface

There is a growing interest among healthcare professionals and clinicians to apply non-invasive, time and cost-effective techniques as a complement to the battery of medical and clinical examinations currently undertaken for the early diagnosis or monitoring of brain and mental disorders. Previous research in this field, based on linguistic-oriented analysis of text and speech produced by such a population and compared to healthy adults, has shown promising results. Initially, work was based on written data (i.e. most commonly collected during formal assessment, and recently also datasets acquired form of blog posts, tweets, and social media in general) but there is a rapidly growing body of research based on spoken samples; and other modalities such as eye tracking; wearable and in-situ sensors data; text production measurements and digital pen strokes.

An important new area of research in natural language processing (NLP) emphasizes the processing, analysis, and interpretation of such data and current research in this field, based on linguistic-oriented analysis of text and speech produced by such a population and compared to healthy adults, has shown promising outcomes. This is manifested in early diagnosis and prediction of individuals at risk, the differentiation of individuals with various degrees of severity forms of brain and mental illness, and for the monitoring of the progression of such conditions through the diachronic analysis of language samples or other para and extra-linguistic measurements.

Nevertheless, there remains significant work to be done to arrive at more accurate estimates for prediction purposes in the future and more research is required in order to reliably complement the battery of medical and clinical examinations currently undertaken for the early diagnosis or monitoring of, e.g., neurodegenerative and other brain and mental disorders and accordingly, aid the development of new, non-invasive, time and cost-effective and objective (future) clinical tests in neurology, psychology, and psychiatry.

RaPID-2 will be an interdisciplinary forum for researchers to share information, findings, and experience on the creation and processing of data acquired or produced by people with various forms of mental, cognitive, neuropsychiatric, or neurodegenerative impairments, such as aphasia, dementia, autism, Parkinsons or schizophrenia. Particularly, the workshop focus on the creation, annotation, description, processing and analysis of linguistic, paralinguistic and extra-linguistic resources (e.g., spontaneous spoken language; audio-recorded samples and transcripts; eye tracking measurements; wearable and in-situ sensor data etc.) from individuals at various stages of these impairments and with varying degrees of severity in order to identify, extract, process, correlate, evaluate and disseminate various linguistic phenotypes and measurements and thus aid the diagnosis, monitor the progression or predict individuals at risk.

A central aim is to facilitate the study of the relationships among various levels of linguistic, paralinguistic and extra-linguistic observations (e.g., acoustic measures; phonological, syntactic and semantic features; eye tracking measurements; sensors, signs and multimodal signals). Submission of papers are invited in all of the aforementioned areas, particularly emphasizing multidisciplinary aspects of processing such data and the interplay between clinical/nursing/medical sciences, language technology, computational linguistics, natural language processing (NLP) and computer science. The workshop will act as a stimulus for the discussion of several ongoing research questions driving current and future research by bringing together researchers from various research communities.

Papers were invited in all of the areas outlined in the *topics of interest* below particularly emphasizing multidisciplinary aspects of processing such data and also on the exploitation of results and outcomes and related ethical questions. Specifically, in the call for papers we solicited papers on the following topics:

- Infrastructure for the domain: building, adapting and availability of linguistic resources, data sets and tools
- Data collection methodologies
- Acquisition and combination of novel data samples
- Guidelines, annotation schemas, annotation tools
- Addressing the challenges of representation, including dealing with data sparsity and dimensionality issues, feature combination from different sources and modalities

- Domain adaptation of NLP tools
- Acoustic/phonetic/phonologic, syntactic, semantic and pragmatic/discourse analysis of data; including modeling of perception (e.g. eye-movement measures of reading) and production processes (e.g. recording of the writing process by means of digital pens, keystroke logging etc.); use of gestures accompanying speech and non-linguistic behavior
- (Novel) Modeling and deep / machine learning approaches for early diagnostics, prediction, monitoring, classification etc. of various cognitive and psychiatric impairments
- Evaluation of the significance of features in diagnostics
- Evaluation of tools, systems, components, metrics, applications and technologies including methodologies making use of NLP; e.g. for predicting clinical scores from (linguistic) features
- Digital platforms/technologies for cognitive assessment and brain training
- Evaluation, comparison and critical assessment of resources
- Involvement of medical/clinical professionals and patients
- Ethical and legal questions in research with human data in the domain, and how they can be handled
- Deployment
- Experiences, lessons learned and the future of NLP/AI in the area

Most of these topics lie at the heart of the papers that were accepted to the workshop which features 6 oral presentations.

We would like to thank all the authors who submitted papers, as well as the members of the Program Committee for the time and effort they contributed in reviewing the papers. We are also grateful to Yasunori Yamada, PhD. Aging Research, IBM Research – Tokyo, for accepting to give an invited talk at the workshop with the title: “Behavioral features for elderly health monitoring”.

The Editor *May 2018*

Programme

Tuesday 8th of May 2018

14.00 – 16.00	**Session A**
14.00 – 14.05	Welcome and Introduction by Workshop Chair
14.05 – 14.50	**Invited keynote talk** Yasunori Yamada, Behavioral features for elderly health monitoring
15.00 – 15.15+5	*Authors:* Tolga Uslu, Lisa Miebach, Steffen Wolfsgruber, Michael Wagner, Klaus Fließbach, Rüdiger Gleim, Wahed Hemati, Alexander Henlein and Alexander Mehler *Title:* Automatic Classification in Memory Clinic Patients and in Depressive Patients
15.20 – 15.35+5	*Authors:* Christian Kohlschein, Daniel Klischies, Tobias Meisen, Björn Schuller and Cornelius Werner *Title:* Automatic Processing of Clinical Aphasia Data collected during Diagnosis Sessions: Challenges and Prospects
15.40 – 15.55+5	*Authors:* Kathleen C. Fraser, Kristina Lundholm Fors, Marie Eckerström, Charalampos Themistokleous, and Dimitrios Kokkinakis *Title:* Improving the Sensitivity and Specificity of MCI Screening with Linguistic Information
16:00 – 16:30	Coffee break
16.30 – 17.30	**Session B**
16.30 – 16.45+5	*Authors:* Nicklas Linz, Johannes Tröger, Hali Lindsay, Alexandra König, Philippe Robert, Jessica Peter and Jan Alexandersson *Title:* Language Modelling for the Clinical Semantic Verbal Fluency Task
16.50 – 17.05+5	*Authors:* Saturnino Luz, Sofia De la Fuente and Pierre Albert *Title:* A Method for Analysis of Patient Speech in Dialogue for Dementia Detection
17.10 – 17.25+5	*Authors:* Kaoru Shinkawa, Keita Shimmei and Yasunori Yamada *Title:* Detecting Dementia from Repetition in Conversational Data of Regular Monitoring Service
17.30 – 17.45	General questions and comments to the presenters and closing

Table of Contents

Automatic Classification in Memory Clinic Patients and in Depressive Patients

Tolga Uslu, Lisa Miebach, Steffen Wolfsgruber, Michael Wagner, Klaus Fließbach, Rüdiger Gleim,Wahed Hemati, Alexander Henlein, Alexander Mehler
Goethe University, Frankfurt am Main,
Department for Neurodegenerative Diseases and Geriatric Psychiatry, University of Bonn, Bonn, Germany
German Center for Neurodegenerative Diseases, Bonn, Germany
{uslu, mehler, gleim}@em.uni-frankfurt.de
{lisa.miebach, steffen.wolfsgruber, michael.wagner,klaus.fliessbach}@ukbonn.de

Abstract

In the past decade the preclinical stage of *Alzheimer's Disease* (AD) has become a major research focus. *Subjective cognitive decline* (SCD) is gaining attention as an important risk factor of AD-pathology in early stages of *mild-cognitive impairment* (MCI), preclinical AD and depression. In this context, neuropsychological assessments aim at detecting sorts of subtle cognitive decline. Automatic classification may help increasing the expressiveness of such assessments by selecting high-risk subjects in research settings. In this paper, we explore the use of neuropsychological data and interview based data designed to detect AD-related SCD in different clinical samples to classify patients through the implementation of machine learning algorithms. The aim is to explore the classificatory expressiveness of features derived from this data. To this end, we experiment with a sample of 23 memory-clinic patients, 21 depressive patients and 21 healthy-older controls. We use several classifiers, including SVMs and neural networks, to classify these patients using the above mentioned data. We reach a successful classification based on neuropsychological data as well as on cognitive complaint categories. Our analysis indicates that a combination of these data should be preferred for classification, as we achieve an F-score above 90% in this case. We show that automatic classification using machine learning is a powerful approach that can be used to improve neuropsychological assessment.

Keywords: early diagnostics, disease classification, feature selection, Alzheimers Disease, neuropsychology

1. Introduction

According to the world Alzheimer report, over 46 million people are estimated to have dementia. This number is expected to rise (Prince et al., 2015). Early detection and accurate diagnostic in preclinical stages is therefore of paramount importance. As an indicator of the earliest clinical stage of Alzheimers Disease (AD) subjective cognitive decline (SCD), defined as the individual's concerns related to cognitive functioning, is gaining interest in different settings (Jessen et al., 2014). With the growing interest in early diagnosis and early detection, SCD has been proposed as an established risk factor for AD, increased risk of future cognitive decline (Koppara et al., 2015) and abnormal AD biomarkers (Amariglio et al., 2012; Chetelat et al., 2010; Wolfsgruber et al., 2015; Buckley et al., 2017). However, in older community based samples the prevalence of memory concerns varies from 25-50% (Jonker et al., 2000) which made it difficult to distinguish AD-related cognitive complaints from those related to normal aging. Furthermore, subjective cognitive decline (SCD) is reported in the context of depression (Balash et al., 2013) and has been positively associated with SCD in different samples (Buckley et al., 2013; Benito-León et al., 2010). Some researcher therefore argued that SCD is mainly driven by depressive symptomatology than being an indicator of an underlying AD-pathology. Current investigations tries to refine the assessment of SCD with the aim to find AD-like complaints and those which may be more representative of a mood disorder or of aging in general (Molinuevo et al., 2016; Rabin et al., 2015). In line with the problematic assessment of SCD, some common-used neuropsychological screening tests such as the *Mini-Mental State Examination* (MMSE) are not sensitive enough for a reliable detection of subtle impairments presented in patients with *mild cognitive impairment* (MCI). Even when some results suggest specific types of neuropsychological deficits associated with *Major depressive Disorders* (MDD), it is still challenging for clinicians to differentiate subjective complaints as a result of a depressive symptomatology from cognitive complaints in the context of preclinical and prodromal AD (Zihl et al., 2010). In memory-clinic settings, early detection of AD is time consuming and require multiple cost intensive information (e.g. neuropsychological testing including subjective concerns and objective impairment, detailed medical history and neurological examination) as well as clinicians with a certain level of expertise. Current assessments of subjective cognitive decline are unable to capture all aspects of SCD specific for preclinical AD and could potentially confound results in the SCD field. Recently, studies started to compare specific aspects of cognitive complaints in different samples using qualitative interview based approaches (Buckley et al., 2015; Miebach et al., 2017; Miebach et al., 2018)

In conclusion, there is large room for improvement regarding the quantitative assessment of SCD and subtle cognitive decline which pose a major task for further research (Jessen et al., 2014). Automatic classification and machine learning might help detecting specific assessment strategies for preclinical AD and the refinement of neuropsychological test batteries.

We generated various neuropsychological and clinical parameters from patient conversations and examinations. To allow automatic classification using this data, we used multiple types of classifiers (SVM, neural networks) to make a diagnosis. In some cases we even managed to get a classification reaching an F-score of more than 90%.

In any event, it is very time-consuming to generate the underlying medical data. Therefore, it is of utmost importance to generate only those data that is required to pro-

duce a good classification. To find out this data, we evaluated different approaches. On the one hand, we used a genetic search over the underlying feature space to find out which subset of features leads to better results. On the other hand, we calculated distance correlation to detect dependencies between pairs of features. We discovered that in some cases, less than 50% of the features of the underlying medical study suffice to generate the best performing classification.

2. Related Work

Machine learning techniques are becoming more and more popular in clinical research and are an established technique in MRT studies (Bede, 2017). Recent studies start from optimizing neuropsychological assessment for cognitive, behavioral and functional impairment using machine learning (Battista et al., 2017). However, studies using automatic classification to distinguish AD from non-AD patients did not focus on earlier preclinical or early MCI stages (Gurevich et al., 2017). Further, modern machine learning techniques have up to now only very rarely been used for the differential diagnosis of cognitive complaints based on the results of interview data. Mehler et al. (2016), for example, automatically analyzed physician-patient talks for differentiating patients suffering from epilepsies or dissociative disorders. This was done by means of the *text2voronoi* algorithm, which is also used in this paper. Regarding the assessment of SCD, (Miebach et al., 2017) were able to confirm several qualitative complaint categories proposed by (Buckley et al., 2015) which are specific for memory-clinic and depressive patients. This suggests that the subjective experience of cognitive decline can be captured by means of a set of interview questions and categories and therefore could be useful for clinicians to detect individuals at high-risk for AD. Investigations of MCI patients self-awareness and experience of their diagnosis have revealed that qualitative approaches may well lead to a more in-depth view than quantitative measurements (Lingler et al., 2006; Roberts and Clare, 2013). However, a qualitative approach is more time consuming than a quantitative one making the diagnostic process more cost intensive. With the gaining interest in an improved detection rate of AD-pathology with less time and cost intensive screening tools, clinicians have the unique opportunity to take advantage of automated classification techniques. This exploratory example of machine learning combined neuropsychological data for the assessment of cognitive impairment and qualitative extracted interview-based features for cognitive complaints in memory-clinic-patients, depressive patients and in healthy controls.

3. Models

In the present study, we experiment with several classification models to be independent of the classifier and to assess the significance of features while being less dependent on these classifiers. As input, the classifiers are fed with neuropsychologically and clinically determined feature values. The neuropsychological part of our study includes a test battery for assessing cognitive performance and depressive symptoms. The clinically determined values are ratings based on qualitative interviews designed to capture aspects of subjective cognitive complaints in the context of preclinical dementia. In contrast to the neuropsychological data set these values are based on expert ratings instead of self-ratings or performance measures. The different group status (memory-clinic-patients, depressive patients, healthy controls) were set as output.

Since we only have a limited amount of data, we carried out a leave-one-out cross-validation for each classifier being tested. This makes sense since each patient is referred to individually for classification. With other data splitting methods, the risk of overfitting is too high (achieving good results on a given split, while performing bad on another one).

3.1. SVM

As a baseline for the experiments we trained a *Support Vector Machine* (SVM) and used it for classification. This is done by means of the SVM-light (Joachims, 1998) implementation using the radial basis function (RBF) kernel. To find optimal parameters for training, we carried out a parameter study on the gamma and the cost parameter. For the cost parameter we examined values between 0.01 and 0.000001; for the gamma parameter we considered values in the range of 1 and 1000000.

3.2. Neural Network

To carry out the same experiments using modern classification methods, neural network-based methods were indispensable. To this end, we used the framework Keras (Chollet and others, 2015). More specifically, we trained a feed-forward network to get a classifier of medical data. Here again, we conducted a parameter study to find the best performing parameters in each experiment. The following parameters were evaluated:

- optimizer: [adam, adamax, rmsprop]
- dropouts: [0.25, 0.5, 0.75]
- layersize: [50, 100, 200, 500]
- layersize2: [0, 50, 100, 200]

We achieved the best results with a dropout of 0.25, *adam* (Kingma and Ba, 2014) as optimizer and two hidden layers.

3.3. Systematic Feature Evaluation for SVM

We examine the impact of feature selection on the F-Measure. While some features may consistently contribute to good classification results, others may reduce performance. That is, we expect that using all available features will most likely not yield the best F-Measure. Since a systematic evaluation of all $2^{138} - 1$ feature combinations is impossible, we apply several approaches to determine local optimal values and to examine the overall robustness of the feature set. If not stated otherwise each evaluation of a given feature set includes a parameter study regarding the optimal *gamma* and *cost* value for the SVM. Here again, our studies are based on SVM-light (Joachims, 1998).

We start with performing a genetic search for optimal feature selection. Genetic algorithms have successfully been

used for feature selection (Li et al., 2005). In our case, a population of n variants, which have been initialized randomly, are evaluated, ranked and flipped (bitwise) over t turns. In each turn, the best ranking variants are kept and mutated to generate additional variants while worst performing instances are dropped. In this way, a hill-climbing algorithm is implemented that approaches local maxima of better performing subsets of features.

3.3.1. Top-down and bottom-up search

In order to examine the overall robustness of the feature set we gradually remove features from the entire set (top-down).

In addition we explore the effect of gradually increasing the number of features starting from an empty set (bottom-up). At each step, the feature that maximizes the performance of the remaining set is added or removed, resulting in $\frac{n^2+n}{2}$ computations. Whenever multiple variants achieve the same top value we chose one of them randomly.

Applying this methodology to feature reduction is an important step, as it not only improves the classification results but also helps reducing the computation time in further analyses.

3.4. text2voronoi

Mehler et al. (2016) have developed a new classification method which visualizes input texts and then uses the visual representation of these texts to drive the classification. The advantage of this method is that one gets a visual depiction of the underlying text that can be used by analogy to MRI scans. Instead on working on the content words of a text, *text2voronoi* is mainly working on distributions of grammatical features of words in this text. In this way, it allows for completely abstracting from text content. This is indispensable when dealing with rather short talks of doctor and patients which, though describing the same disease, may select words of a completely unrestricted semantic universe. Using grammatical information, embeddings are produced by means of word2vec (Mikolov et al., 2013). Then, a Voronoi tessellation is calculated on this data to map texts onto 2- or 3D spaces. Finally, the resulting depictions are used explored to drive the classification.

3.5. fastText

We additionally experimented with fastText (Joulin et al., 2016), an efficient text classifier, to compare it with *text2voronoi*. fastText is based on a feedforward neural network with only one hidden layer. Joulin et al. (2016) show that fastText compares with state-of-the art classifiers while being faster than its competitors.

4. Experiment

4.1. Sample description

The total sample of this study includes n=65 older subjects (mean age=70.03 years; 52.3% female). All participants were above the age of 55 and had sufficient ability to speak German. All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The study was approved by the local ethical committees of the University of Bonn, and informed written consent was obtained from all subjects.

Memory-clinic patients (MC)(n=23) were referred by their general practitioners to the Clinical Treatment and Reseach Center for Neurodegenerative Disorders (KBFZ), Department for Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn for a diagnostic work up of cognitive functioning. Diagnosis of AD Dementia or MCI was made according to the core clinical criteria of the NIA-AA (Albert et al., 2011)(McKhann et al., 1984). The diagnostic procedure included a cognitive assessment, detailed medical history, and a neurological examination. Of the total sample 15 fulfilled the core clinical criteria of mild cognitive impairment (MCI) according to the NIA-AA criteria (performance under 1.5 SD below age, gender, education adjusted norms)(Albert et al., 2011). The remaining 8 patients only had subjective concerns without objective impairment, and were classified as patients with Subjective Cognitive Decline (SCD).

Major depressive Patients (MDD) (n=21) were recruited from the Clinic of Psychiatry and Psychotherapy, University Hospital Bonn. All patients fulfilled a diagnosis of a unipolar, major depressive disorder according to ICD-10 criteria (Organization, 1993).

The Healthy control group (HC) (n=21) was recruited from a scope of a normative study of the German Center for Neurodegenerative Diseases (DZNE) Bonn that evaluated neuropsychological performance of healthy older individuals. They were excluded from the participation when they (1) were concerned about mental abilities or memory (2) had been in psychological, psychiatric or neurological treatment within the last 6 months (3) had any severe or chronic disease (e.g. diabetes or MS) (4) had experienced head injury with a loss of consciousness, (5) had a neurological disease (e.g. AD or Parkinson) (6) or had a relative with a first-degree relative with a documented diagnosis of neurodegenerative disease in their family history.

4.1.1. Clinical Rating of cognitive complaints

The Clinical rating was made based on a semi-structured interview designed to capture all complaint categories proposed by (Buckley et al., 2015). The Interview similar to a clinical routine interview, started with an open question asking whether the patient had noticed *"any changes in memory or thinking during the last years*" followed by detailed questions about the complaint itself. The interview procedure followed a semi-structured format and lasted between 8 and 31 min. Each interview had an unstructured beginning, which allowed patients to determine the initial focus of the conversation. If cognitive changes were reported, the participants were asked to give an example of their everyday life. Then the patient was asked whether he/she has noticed further cognitive problems followed by the request to give an everyday example. This process was repeated until the participant did not mention further complaints. He/she was then asked to name the most concerning symptom which was selected for further detailed questioning. If the participant reported another concerning symptom, we repeated the detailed questions

about the complaint itself. Therefore, 58% of the sample named two concerning symptoms. All Interviews were digitally recorded and later transcribed verbatim by the interviewer. Data for analyses presented in this manuscript were derived from the ratings of a single clinical psychologist (LM) who also conducted all the interviews. To capture all aspects of cognitive complaints, the clinical rating in this study was based on glossary of cognitive complaints based on a combination of the cognitive complaint categories proposed by (Buckley et al., 2015) and the complaint themes proposed by (Miebach et al., 2018). The glossary contains the following categories: *Increasing frequency, Sense of predomination and growing concern, Situational lapses, Relative absence of spatio-temporal contextualisation, burdensome coping strategies, Dismissive attitude, attentional fluctuation/vagueness, Impact on affect, Progression, an over-endorsed complaint, dependency, affective influence on memory, distractible speech, general complaints about increasing memory problems, difficulties in Action monitoring, difficulties in initiating actions, deceleration, slowing of cognitive processing speed, nonspecific overwork, forgetfulness, short-term memory problems, content memory problems, blank mind, loss-of-control experience, derealisation, formal though disorder, prospective memory, planning, learning, cognitive flexibility, increased distractibility, concentration difficulties, word finding difficulties, memory for names, dyscalculia, visual-spatial-disorientation, general decline, no changes in cognitive functions.* The categories were extracted from the interview material using inductive qualitative approaches. The complaint categories based on (Buckley et al., 2015) were related to the grounded theory (Strauss and Corbin, 1997) whereas the complaint themes extracted by (Miebach et al., 2018) were based on the interpretative phenomenological analysis (IPA) (Smith et al., 2009). Therefore the presented deductive rating of cognitive concerns is based on two different phenomenological approaches which allows to capture highly nuanced and contextualized aspects of subjective experiences (Smith et al., 2009). The Interview procedure and categorization system are described in detail in (Miebach et al., 2017; Miebach et al., 2018). For the coding process, we used a deductive category assignment approach similar to qualitative content analysis (Mayring, 2014). Participant's responses were coded using a binary coding system (i.e. 0=theme absent; 1=theme present).

4.1.2. Neuropsychological assessment

The Neuropsychological assessment included a set of different clinical measurements for global memory and cognitive performance specifically designed for early diagnosis of AD dementia. The test battery included the Free and Cued Selective Reminding Test (FCSRT) (Ivnik et al., 1997) and the German version of the neuropsychological test battery of the Consortium to Establish a Registry for Alzheimers disease (CERAD-plus;(Morris et al., 1989)) with various sub-tests (e.g. verbal fluency, Boston Naming Test, Mini Mental State Exam, Word List learning, Constructional praxis, Word List recall, word list recognition, constructional praxis recall, TMT-A, TMT-B, the symbol digit modalities test (SDMT)(Smith, 1982)). Depressive symptoms were assessed with the 15-item version of the Geriatric Depression Scale (GDS; (Yesavage et al., 1983)) and the Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2010)

4.1.3. Group characteristics and demographical differences

Analysis for the group differences were performed using IBM SPSS Version 22 (Corp, 2013). Group differences were observed for age, education, interview duration, GDS and PHQ-9 scores. Memory-clinic patients were slightly older (M=72.91 yr) compared to MDD (M=69.43 yr) and the interview duration was significantly longer (M=18.41 min) in comparison with HC (M=14.32 min) and the MDD-Group (M=12.07 min). HC were younger, performed significantly better on the MMSE (M=29) and exhibited lower levels of depressive symptomatology (GDS; M=0.62) compared to MDD and the Memory-clinic patients. The depressive group exhibited elevated levels of depressive symptomatology, significantly above the GDS cut-off for depression (M=7.00) and the PHQ-9 cut-off for moderate depression (M=10.89). Depressive patients also had significantly fewer years of education (M=12.57) compared to HC (M=15.10) and Memory-clinic patients (M=15.50).

4.2. Classification

We have used the models from section 3. to classify the patients. In doing so, we classify on the textual data and on the clinical and neuropsychologically generated data. Models 3.1. to 3.3.1. are designed for the classification with the self-generated data, while the models 3.4. and 3.5. are designed for textual classification.

4.2.1. Clinical and neuropsychological feature classification

In this experiment, we used the clinical ratings of subjective cognitive complaints based on the qualitative interview as one feature set. We also used the neuropsychological test results including data about objective cognitive performance and measurements of depression as another feature set. First of all, we used both feature sets independently for classification. We then combined both sets of features and used the combination for classification. Table 1 shows the results of the different classifiers applied to the 3 feature sets. We discovered that the combined features are always more successful than both feature sets on their own.

Model	Neuropsych.	Clinical	Combined
3.1.	0,747	0,706	0,794
3.2.	0,754	0,723	0,800
3.3.	0,870	0,821	0,881
3.3.1.	0,933	0,928	0,949

Table 1: F-scores of the classifiers with regard to the different data sets.

4.2.2. Patient talks classification

In this experiment, we analyze the texts of the cognitive complaint interviews and use them for classification (leave one out cross validation). We only used the text content of

the patients and removed the doctor's text data from the interview protocols because the doctor asks all patients similar questions, which would have a negative effect on the classification. Table 2 shows the results of the 2 methods we used for classification. It can be seen that the baseline classifier *fastText* can hardly classify the texts. However, if the texts are abstracted, as it is the case with *text2voronoi*, an improvement is achieved.

Model	F-score
3.4.	0,520
3.5.	0,340

Table 2: Results of the textual classification experiment of the 3 patient groups (MC/MDD/HC)

4.3. Feature analysis

Now that we have applied different classifiers in different experiments, we want to find out which of the used features were actually needed. To find out, we have used the following approaches.

4.3.1. Genetic feature search

As explained in chapter 3.3., we have also carried out a genetic search of the parameters to find the smallest possible subset, which provides the best results. We found out that only a fraction of the features are required to perform a good classification.

Experiment	Subset
Neuropsych.	47,30%
Clinical	40,63%
Combined	40,58%

Table 3: Subset analysis of the features using model 3.3..

In addition to the genetic search, we have also carried out two other approaches, as described in Section 3.3.. Figure 1 shows the process of this analysis. Again, it is obvious that few features are enough to get the best results. We achieve the best score with only 64 features.

4.3.2. Decision tree

A good way to analyze the features is to use Decision Trees, as it follows simple and comprehensible heuristics. The graphic representation as a tree diagram also illustrates hierarchically consecutive decisions. We have used the Python package *sklearn* (Pedregosa et al., 2011) to perform these analyses. In our best performing experiment, we have the following patient distribution:

- $[21, 23, 21]$ - (Control patients, Memory-Clinic patients, Depressive patients)

Figure 2 shows that feature 41 (SDMT - neuropsychological score developed to identify individuals with neurological impairment) is at the top of the tree. If the value of SDMT is less than or equal to 44,5, the patient distribution is divided into:

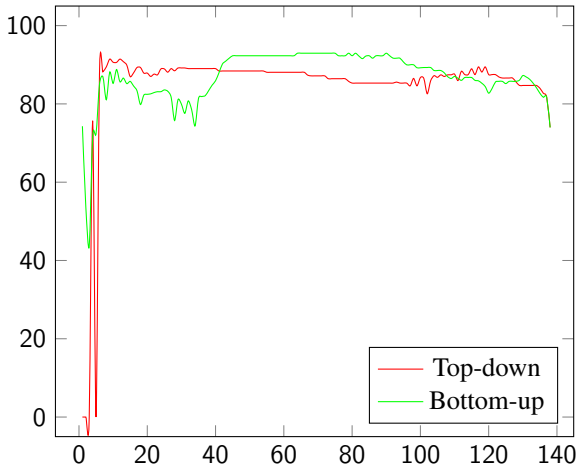


Figure 1: F-Scores based on the number of features in the example of the combined experiment (see Section 4.2.1.).

- $[0, 20, 18]$ - Only memory-clinic or depressive patients.
- $[21, 3, 3]$ - Mostly healthy controls.

Thus, we divide all control patients into a separate group. The next important parameter is feature 45 (GDS - neuropsychological measurement for depression)(Yesavage et al., 1983). This divides the group of diseases ($[0, 20, 18]$) into the following patient distribution:

- $[0, 17, 4]$ - Mostly Memory-clinic patients
- $[0, 3, 14]$ - Mostly depressive patients

This means that we could group 52 (14+17+21) patients correctly with these 2 features alone, but 13 (3+4+3+3) wrong. You can also see that these features belong to the neuropsychological features. This also makes sense, as these values also lead to better classifications (see example 4.2.1.). The further down the tree is examined at, the more precise the distributions will be. However, given the number of features (138) and the small amount of patients (65), this is rather overfitting.

4.3.3. Distance correlation

To measure the interdependence between the features as described in Section 4.1. we calculated distance correlation between pairs of features. For this we used the R package *energy* (Rizzo and Szekely, 2017) utilized by *TextImager* (Uslu et al., 2017). Interdependent features are an indicator for redundant data. These redundant features are less helpful for classification. Figure 3 shows the heatmap of the pairwise dependencies. Each cell represents the distance correlation of the features X and Y, with the green shading indicating the dependency (darker = more dependent). The diagonal is green which indicates that every feature is correlated to itself. The green squares also provide important information. The first dependency square at the top left in Figure 3, for example, contains only neuropsychological features based on the MMSE. The mini-mental state test (MMSE) is a brief screening tool for Alzheimer dementia and impairment in global cognition (Folstein et al., 1975).

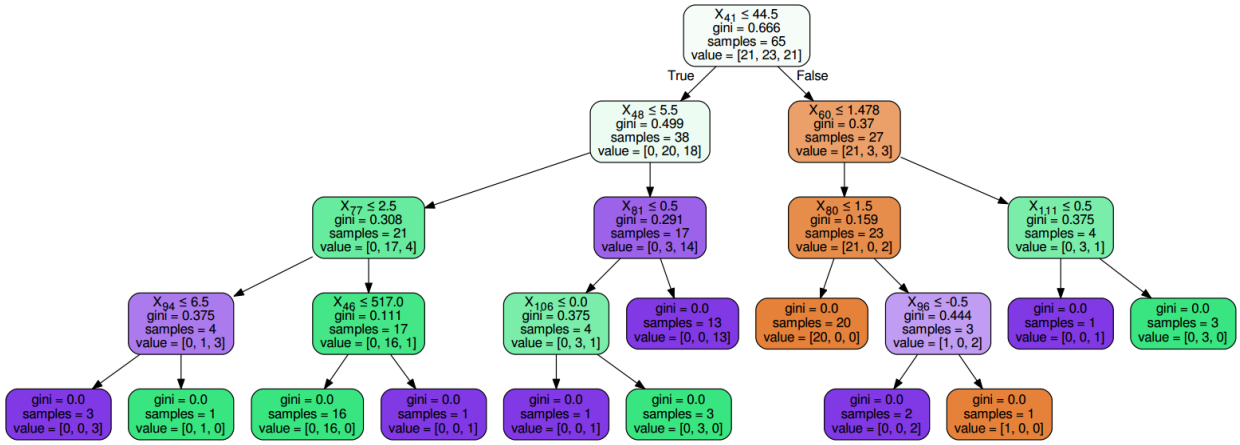


Figure 2: Result of the decision tree based on the combined example of experiment 4.2.1..

The MMSE test includes items assessing orientation, word recall and registration, attention and calculation, and language and visuospatial abilities. As a logical consequence we observed a high dependency between the different subscores of the MMSE.

As mentioned above, 58% of the sample named more than one cognitive complaint. As a result, the categories were coded for a second time. A high dependency between these features is therefore a consequence of the interview procedure. These dependencies can be seen in Figure 3 by the large green square at the bottom right.

Figure 3: Visual depiction of pairwise dependence of the features.

5. Discussion

The present study is the first to combine a qualitative text-analytical approach for cognitive complaints with an automatic classification system for three different diagnostic groups. Recently, only a few studies have explored the use of automated learning methods within the neuropsychological literature. In this proof of principle study we used a machine learning approach based on neuropsychological and interview generated cognitive complaint categories for the classification of memory-clinic patients, depressive patients and normal healthy older adults.

We aimed to replicate the diagnostic value of the recently proposed complaint categories using an automatic classification method instead of current statistical methods used in clinical research (Buckley et al., 2015; Miebach et al., 2017; Miebach et al., 2018). Cognitive complaints were elicited with a semistructured interview comparable with a typical clinical routine interview.

The current study results revealed that machine learning techniques can accurately classifying patients measured via neuropsychological test battery and via clinical rating of cognitive complaints. We found that the classification with self-generated characteristics extracted by a qualitative approach works much better than with the recorded texts in the patient conversations.

This result makes sense because patients talk about many different topics in the diagnostic interview and the content of these texts is not reliable for determining a disease. Therefore the interpretation of cognitive complaints relies on expertise of some kind which is not ideal for a wide distribution across studies.

In the case of the second experiment, the neuropsychological data outperformed the clinical ratings based on interview data. This could be explained by the heterogeneous sample including patients with mild cognitive impairment as well as patients with only subjective cognitive decline and depression. In line with current literature, the combination of neuropsychology and the clinical rating reached the best diagnostic accuracy (Molinuevo et al., 2016). A replication in a larger sample with focus on the complaint categories is needed to extract features which are truly relevant for AD-pathology.

Given the present results, we believe it is much more likely that measure incorporates both qualitative text based and quantitative neuropsychological methods will be able to

identify the preclinical AD profile. Recent studies used composite scores calculated based on z-transformed subscales of different SCD assessments to predict the taupathology in the enthorinal cortex of healthy older adults (Buckley et al., 2017).

However, in the case of textual classification, an improvement is achieved when the text is transformed into a more abstract model (*text2voronoi*). The experiments also show that the neuronal network-based approaches are usually somewhat better than the SVMs. However, the best solutions can be found with GeneticSVM and even only a subset of all features. As a result, we found out that a few features are enough to get a good classification. However, these features (SDMT; GDS) are established clinical screening tools for the measurement of memory impairment and depressive symptomatology (Yesavage et al., 1983; Smith, 1982). A feature analysis only including the cognitive complaint categories should be an important next step with a higher clinical impact in the field of AD research. We analyzed them and found out that there are some dependencies among the features. There is a need of alternative ways for the operationalization and the diagnosis of AD-relevant cognitive complaints. Using a semistructured interview based on qualitative categories seems to be promising regarding the clinical evaluation of memory complaints in non-demented elderly. Further improvement of the complaint glossary and the rating scale is needed for the detection of preclinical AD. Therefore machine learning approaches could be promising for reducing and refining neurospychological assessments. This information can save a lot of work, since the dependent features barely improve the classification.

6. Conclusion

In this paper we have used different classifiers in various patient diagnosis experiments. We have shown that a good classification can be achieved by using cognitive complaint categories based on clinical interview and neuropsychological data from standardized test batteries. We found that the combination of these data sets leads to the best results with an F-score of 80,00%. In addition, we have applied a number of different approaches to find the optimal subset of features that provide the best classification. In this case we even achieve an F-score of 94,87%. However, classification at text level is not yet particularly successful. In future work we aim at studying different abstractions of texts (as provided, for example, by *text2voronoi*) in order to detect expressive linguistic features that allow for automatically assessing the diseases under consideration.

7. Bibliographical References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., et al. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):270–279.

Amariglio, R. E., Becker, J. A., Carmasin, J., Wadsworth, L. P., Lorius, N., Sullivan, C., Maye, J. E., Gidicsin, C., Pepin, L. C., and Sperling, R. A. (2012). Subjective cognitive complaints and amyloid burden in cognitively normal older individuals. *Neuropsychologia*, 50(12):2880–2886.

Balash, Y., Mordechovich, M., Shabtai, H., Giladi, N., Gurevich, T., and Korczyn, A. D. (2013). Subjective memory complaints in elders: depression, anxiety, or cognitive decline? *Acta Neurologica Scandinavica*, 127(5):344–350.

Battista, P., Salvatore, C., and Castiglioni, I. (2017). Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: A machine learning study. *Behavioural neurology*, 2017.

Bede, P. (2017). From qualitative radiological cues to machine learning: Mri-based diagnosis in neurodegeneration.

Benito-León, J., Mitchell, A. J., Vega, S., and Bermejo-Pareja, F. (2010). A population-based study of cognitive function in older people with subjective memory complaints. *Journal of Alzheimer's Disease*, 22(1):159–170.

Buckley, R., Saling, M. M., Ames, D., Rowe, C. C., Lautenschlager, N. T., Macaulay, S. L., Martins, R. N., Masters, C. L., O'Meara, T., Savage, G., et al. (2013). Factors affecting subjective memory complaints in the aibl aging study: biomarkers, memory, affect, and age. *International Psychogeriatrics*, 25(8):1307–1315.

Buckley, R. F., Ellis, K. A., Ames, D., Rowe, C. C., Lautenschlager, N. T., Maruff, P., Villemagne, V. L., Macaulay, S. L., Szoeke, C., and Martins, R. N. (2015). Phenomenological characterization of memory complaints in preclinical and prodromal alzheimer's disease. *Neuropsychology*, 29(4):571.

Buckley, R. F., Hanseeuw, B., Schultz, A. P., Vannini, P., Aghjayan, S. L., Properzi, M. J., Jackson, J. D., Mormino, E. C., Rentz, D. M., Sperling, R. A., et al. (2017). Region-specific association of subjective cognitive decline with tauopathy independent of global β-amyloid burden. *JAMA neurology*, 74(12):1455–1463.

Chetelat, G., Villemagne, V. L., Bourgeat, P., Pike, K. E., Jones, G., Ames, D., Ellis, K. A., Szoeke, C., Martins, R. N., and O'Keefe, G. J. (2010). Relationship between atrophy and b–amyloid deposition in alzheimer disease. *Annals of neurology*, 67(3):317–324.

Chollet, F. et al. (2015). Keras. `https://github.com/keras-team/keras`.

Corp, I. (2013). Ibm spss statistics for windows, version 22.0. *Armonk, NY: IBM Corp*.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198.

Gurevich, P., Stuke, H., Kastrup, A., Stuke, H., and Hildebrandt, H. (2017). Neuropsychological testing and machine learning distinguish alzheimer's disease from other causes for cognitive impairment. *Frontiers in aging neuroscience*, 9.

Ivnik, R. J., Smith, G. E., Lucas, J. A., Tangalos, E. G.,

Kokmen, E., and Petersen, R. C. (1997). Free and cued selective reminding test: Moans norms. *Journal of Clinical and Experimental Neuropsychology*, 19(5):676–691.

Jessen, F., Amariglio, R. E., Van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois, B., Dufouil, C., Ellis, K. A., Van Der Flier, W. M., et al. (2014). A conceptual framework for research on subjective cognitive decline in preclinical alzheimer's disease. *Alzheimer's & dementia*, 10(6):844–852.

Joachims, T. (1998). Making large-scale svm learning practical. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Jonker, C., Geerlings, M. I., and Schmand, B. (2000). Are memory complaints predictive for dementia? a review of clinical and population-based studies. *International journal of geriatric psychiatry*, 15(11):983–991.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koppara, A., Wagner, M., Lange, C., Ernst, A., Wiese, B., Koenig, H.-H., Brettschneider, C., Riedel-Heller, S., Luppa, M., Weyerer, S., et al. (2015). Cognitive performance before and after the onset of subjective cognitive decline in old age. *Alzheimers & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(2):194–205.

Kroenke, K., Spitzer, R. L., Williams, J. B., and Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General hospital psychiatry*, 32(4):345–359.

Li, L., Jiang, W., Li, X., Moser, K. L., Guo, Z., Du, L., Wang, Q., Topol, E. J., Wang, Q., and Rao, S. (2005). A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, 85(1):16 – 23.

Lingler, J. H., Nightingale, M. C., Erlen, J. A., Kane, A. L., Reynolds III, C. F., Schulz, R., and DeKosky, S. T. (2006). Making sense of mild cognitive impairment: a qualitative exploration of the patient's experience. *The Gerontologist*, 46(6):791–800.

Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of alzheimer's disease report of the nincds-adrda work group* under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, 34(7):939–939.

Mehler, A., Uslu, T., and Hemati, W. (2016). Text2voronoi: An image-driven approach to differential diagnosis. In *VL@ ACL*.

Miebach, L., Wolfsgruber, S., Frommann, I., Buckley, R., and Wagner, M. (2017). Different cognitive complaint profiles in memory clinic and depressive patients. *The American Journal of Geriatric Psychiatry*.

Miebach, L., Wolfsgruber, S., Frommann, I., Fließbach, K., Jessen, F., Buckley, R., and Wagner, M. (2018). Cognitive complaints in memory clinic patients and in depressive patients: An interpretative phenomenological analysis. *The Gerontologist*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Molinuevo, J. L., Rabin, L. A., Amariglio, R., Buckley, R., Dubois, B., Ellis, K. A., Ewers, M., Hampel, H., Klöppel, S., and Rami, L. (2016). Implementation of subjective cognitive decline criteria in research studies. *Alzheimer's & Dementia*.

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., Mellits, E. D., and Clark, C. (1989). The consortium to establish a registry for alzheimer's disease (cerad): I. clinical and neuropsychological assessment of alzheimer's disease. *Neurology*.

Organization, W. H. (1993). *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*, volume 2. World Health Organization.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prince, M., Wimo, A., Guerchet, M., Ali, G., Wu, Y., and Prina, M. (2015). Alzheimer's disease international (2015). world alzheimer report 2015: The global impact of dementia: An analysis of prevalence, incidence, cost and trends. *Alzheimer's Disease International, London*.

Rabin, L. A., Smart, C. M., Crane, P. K., Amariglio, R. E., Berman, L. M., Boada, M., Buckley, R. F., Chételat, G., Dubois, B., and Ellis, K. A. (2015). Subjective cognitive decline in older adults: An overview of self-report measures used across 19 international research studies. *Journal of Alzheimer's Disease*, (Preprint):1–25.

Rizzo, M. L. and Szekely, G. J. (2017). Package 'energy'.

Roberts, J. L. and Clare, L. (2013). Meta-representational awareness in mild cognitive impairment: An interpretative phenomenological analysis. *Aging & Mental Health*, 17(3):300–309.

Smith, J., Flowers, P., and Larkin, M. (2009). Interpretative phoneomological analysis: theory, method and research.

Smith, A. (1982). Symbol digit modalities test (sdmt) manual (revised) western psychological services. *Los Angeles*.

Strauss, A. and Corbin, J. M. (1997). *Grounded theory in practice*. Sage.

Uslu, T., Hemati, W., Mehler, A., and Baumartz, D. (2017). Textimager as a generic interface to r. *EACL 2017*, page 17.

Wolfsgruber, S., Jessen, F., Koppara, A., Kleineidam, L., Schmidtke, K., FrÃ¶hlich, L., Kurz, A., Schulz, S., Hampel, H., and Heuser, I. (2015). Subjective cognitive decline is related to csf biomarkers of ad in patients with mci. *Neurology*, 84(12):1261–1268.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang,

V., Adey, M., and Leirer, V. O. (1983). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, 17(1):37–49.

Zihl, J., Reppermund, S., Thum, S., and Unger, K. (2010). Neuropsychological profiles in mci and in depression: Differential cognitive dysfunction patterns or similar final common pathway disorder? *Journal of psychiatric research*, 44(10):647–654.

Automatic Processing of Clinical Aphasia Data collected during Diagnosis Sessions: Challenges and Prospects

Christian Kohlschein*, Daniel Klischies*, Tobias Meisen*, Björn W. Schuller†, Cornelius J. Werner+
*Institute of Information Management in Mechanical Engineering (IMA), RWTH Aachen University, Germany
†GLAM – Group on Language Audio & Music, Imperial College London, United Kingdom
+Department of Neurology, Section Interdisciplinary Geriatrics, University Hospital RWTH Aachen, Germany
Corresponding authors: christian.kohlschein@ima.rwth-aachen.de and cwerner@ukaachen.de

Abstract

Aphasia is an acquired language disorder, often resulting from a stroke, affecting nearly 580,000 people Europe alone each year (Huber et al., 2013) . Depending on the type and severity, people with aphasia suffer, in varying degrees, from the impairment of one or several of the four communication modalities. To choose an appropriate therapy for a patient, the extent of the aphasia at hand has to be diagnosed. In Germany and other countries this is done using the Aachen Aphasia Test (AAT). The AAT consists of a series of tests, requiring the patient to talk, read and write over the course of up to two hours. The AAT results then have to be evaluated by a speech and language therapist, which takes around 6 hours. In order to further objectify the manual diagnosis and speed up the process, a digital support system would be highly valuable for the clinical field. To facilitate such a system, we have collected, cleaned and processed real-life clinical aphasia data, coming from AAT diagnosis sessions. Each dataset consists of speech data, a transcript and rich linguistic AAT annotations. In this paper, we report on both challenges and early results in working with the (raw) clinical aphasia data.

Keywords: Clinical Aphasia Data, Multimodal Language Data, Rich Metadata

1. Introduction

Aphasia, i. e., the full or partial loss of linguistic capabilities in adults, is usually an acquired condition, mostly due to damage inflicted to the brain by ischemic or hemorrhagic stroke, but also due to head injury, tumours or neurodegeneration. The loss of linguistic capabilities neither pertains to the motor acts of speaking or writing nor the sensory capabilities of hearing or seeing, but rather to damage to the human brain's 'supra-modal' capability of producing and comprehending language. The consequences of aphasia for the patient are immense: as language, both spoken and written, is our main tool of communication, affected persons are largely cut off from basic social interaction, leading to severe disability, social isolation, loss of health-related quality of life and depression. The socioeconomic impact also is enormous, as persons suffering from aphasia are less likely to return to their jobs (Wozniak and Kittner, 2002). Thus, every effort has to be made to keep this percentage of people dropping out of their jobs as small as possible, necessitating the need for intensive rehabilitation. However, as language is an extremely complex function of the human brain supported by a widespread network of neurons throughout the human brain (albeit with a left-hemispheric predominance), different patterns of damage to the human brain, e. g., by occlusion of different vessels or by trauma to different brain locations, will result in different aphasic syndromes (Ardila, 2010). These are marked by differential loss of putative linguistic modules (Heilman, 2006), such as syntax, semantics, phonology and finally motor speech output. Thus, it is obvious that aphasia rehabilitation is a non-trivial task, and any success in rehabilitation can only occur if and when the prominently hit modules are identified correctly, resulting in a syndromal diagnosis also encompassing the severity of the damage, as there is no general 'aphasia' rehabilitation. In order to achieve a certain level of objectivity and measurability in diagnosing and grading aphasia syndromes, clinical tests and scores are employed. In Germany and beyond, the *Aachen Aphasia Test* (AAT) (Huber et al., 2013) is regarded to be the gold standard in diagnosing and classifying aphasia. This test allows to assess different language modalities at all linguistic levels. Beyond that, it also yields information of probabilistic syndrome classification and syndrome severity. Its disadvantages are that the AAT is immensely time-consuming (up to 8 hours for one patient including data acquisition and evaluation), it does not encompass all linguistic symptoms a patient can exhibit, and it is at least in part dependent on the experience of the rater. Particularly the requirements on human resources preclude its widespread use, although it is regarded to be a prerequisite for, e. g., an intensive comprehensive aphasia program. Besides, the AAT is not very sensitive to changes over time, limiting its utility as a feedback and tracking tool.

Therefore, an automatic aphasia diagnosis system based on the AAT would be highly valuable for patients and clinicians alike. Clinicians would profit from an increased objectivity of the AAT. Having an objective system in place across different hospitals would also enable aphasia rehab units to offer individualized rehabilitation strategies to their (prospective) patients, because they could correlate their language profiles with outcomes of therapeutic success within a specific facility. Patients, e. g. mobility impaired stroke victims, would also benefit from an automatic AAT diagnosis system within their home, making it a non necessity to go the hospital every time for follow-up aphasia examinations. In order to facilitate such a system, a high-quality data and, ideally, large collection of speech and language data along with diagnosis annotations is a prerequisite. During aphasia diagnosis sessions over the course of roughly 20 years at the University Hospital Aachen, clinician-patient speech was recorded, transcribed and, along with the corresponding tests results, digitally archived. The data is in a variety of formats, not available in one homogeneous database but

rather spread over multiple systems and the speech data is a mix between clinician and patient speech. Nevertheless, to the best of our knowledge, this data is one of the richest collections of aphasia data in Germany. We therefore strive to utilize this data to built an automatic AAT system. This paper will not focus on the architecture of the system, but rather present and discuss the challenges we encountered in dealing with the clinical speech and language data itself.
The remainder of the paper is structured as follows: In Section 2. we present related work. Section 3. discusses aphasia, its diagnosis in general and introduces the Aachen Aphasia Test in its current form. Following that, Section 4. discusses our work regarding the assembly of the database and the dataset itself, including a description of its modalities. In Section 5., preliminary results will be presented and discussed. Afterwards, in Section 6., we conclude the paper. Finally, in Section 7., we outline future work.

2. Related Work

Computer programs designed to help diagnose and treat aphasia can be categorized into three different groups (Katz, 2010): Tools for 'alternative and augmentative communication (AAC)', which offer additional ways for aphasia patients to communicate, 'Computer-only treatment (COT)' such as smartphone apps designed to be used by aphasia patients to practice speaking without a therapist, and 'Computer-assisted treatment (CAT)' systems, which help therapists during the therapy. Our system is initially designed as a CAT system: While conducting a conversational speech test, the system analyses the patients speech and returns an aphasia score, as outlined in (Kohlschein et al., 2017). This contrasts many existing projects, which are designed as COT systems.
A COT system which allows patients to build sentences out of predefined clauses via a touchscreen interface, and then requests that the patient reads out the sentence was presented by (Le et al., 2016). The system aims to provide feedback to the patient, such that the patient can practice correct speech. For all predefined clauses, they recorded healthy speech during development of the application. Furthermore, this procedure provides, by design, a transcript of the sentence the patient attempted to say. Additionally, the audio file is transcribed after recording. Possession of a transcript currently leads to better detection of aphasic and especially paraphasic speech (Le et al., 2017). The transcript allows to compare healthy speech to aphasic speech on a per-word basis, and therefore to determine the fraction of correct words compared to the total number of words. Additionally, transcripts based on the recordings can be used as training data for automatic speech recognition (ASR) systems, while knowledge about which sentence the patient attempts to say constrains the search space for ASR (Le et al., 2016). Since our goal is to perform a rating on completely spontaneous clinical speech in the context of CAT systems, we do not have predefined sentences or clauses. However, we have aphasia syndrome and severeness ratings for all recordings, which were made by speech therapists or neurologists. This contrasts the ratings used by Le et al. which were made by trained students, and led to the requirement of a reduced number of severeness categories because the agreement on ratings of the same utterance between different evaluators was low. In 2013, (Fraser et al., 2013) compared different approaches to automatically identify subtypes of primary progressive aphasia. They compared two different techniques for feature detection. The first approach they tried is to perform a Welch t-test on features extracted from audio and transcript files of aphasic speech, compared to healthy speech. Then, they ranked the results based on the p-values obtained from the t-test results and selected only the most significant features. Their second approach is based on the minimum-redundancy-maximum-relevance (mRMR) technique proposed by (Peng et al., 2005). Subsequently, Fraser et al. compared a probabilistic Naive Bayes classifier to Support Vector Machines (SVMs) and Random Forests (RFs). Their results showed that, aphasia subtype detection is more accurate when combining acoustic and transcript data, compared to acoustic data alone. However, even if only acoustic data is available, classification of primary progressive aphasia patients and control group members had an average accuracy of 74.05 %, with Random Forests applied on a feature set chosen by an mRMR algorithm performed best at close to 90 % accuracy. Interestingly, the mRMR selection performed worse than the p-value feature selector when applied to a decision problem between the two aphasia subtypes.

The available aphasia speech data in the University Hospital Aachen consists of spontaneous speech interviews between a clinician and a patient. As an alternative to segmenting all the data manually, we investigated automatic systems as well, i. e., using speaker diarization. Speaker diarization can be classified into bottom-up and top-down approaches. These are based on splitting the audio sample into segments using an heuristic identifying changes in loudness, bandwidth and frequency, which implicate speaker changes. In the next step, these segments are clustered and segments in the same cluster are recombined (Tranter and Reynolds, 2006). The goal of the clustering is to form one cluster per speaker, requiring a clustering based on a method that distinguishes between speakers, but does not discriminate intra class. The top-down approach is based on starting with one cluster and iteratively differentiating it into an ideal amount of clusters, while the bottom-up approach starts with a high number of clusters and iteratively merges similar clusters (Bozonnet et al., 2010). Different approaches for clustering have been proposed. These include using Gaussian Mixture Models to model speakers (Castaldo et al., 2008) based on a sliding window and using eigenvoices as features. Eigenvoices are feature vectors in a vector space whose basis was determined using principle component analysis on the extracted features, causing a model that is based on dimensions which had a high variance in the original feature set (Kuhn et al., 2000). Another method, introduced in (Sell and Garcia-Romero, 2014), is to apply agglomerative hierarchical clustering based on scores retrieved by computing the pairwise similarity of all i-vectors using probabilistic linear discriminant analysis, merging those that are most similar. There also have been approaches based on identifying speakers by training deep neural networks to identify speakers and subsequently extracting their hidden layer feature activations, under the assumption that similar activation patterns

imply that two speakers are the same (Rouvier et al., 2015). The authors of (Isik et al., 2016) also presented an approach based on deep clustering capable of single-channel multi-speaker separation. Finally, (Zhang et al., 2017) presented a diarization approach based on paralingustic cues, e. g., age and gender.

Few collections of aphasic data are publicly available, the most prominent being the AphasiaBank (MacWhinney et al., 2011), which is mostly for the English language domain. More recently, a Greek data set (GREECAD) was made available by (Varlokosta et al., 2016). Both data sets contrast our data collection in several ways. GREECAD was assembled with scientific purposes in mind and subsequently annotated and transcribed by humans in a predefined way, thereby maximizing the agreement between evaluators to get uniform and coherent annotations. Additionally, machine readability and processability was taken into account when choosing the data format and recording the patients. In contrast, the data set of the University Hospital Aachen was solely collected for clinical diagnosis purposes during assessment sessions over a couple of years. Therefore, machine readability was not taken into account while assembling and recording the data, which in turn poses challenges for the automatic processing of it. These challenges include, but are not limited to missing or incorrect meta data, such as therapist attribution, and mono-channel recordings with low cost microphones, requiring a speaker diarization procedure capable of handling open speaker groups, with high noise tolerance and which does not rely on language models, as these do not apply to aphasic speech.

Transcripts and annotations were made by clinical speech and language therapists for the aphasia domain, whereas the Greek data set was transcribed by linguists (graduate or post graduate students). Our data currently contains transcripts roughly four times the amount of aphasic utterances in GREECAD, but does not contain a control group (due to the origin of the data). The AphasiaBank data set has similar properties as the Greek data set, albeit being larger. Additionally, the AphasiaBank contains video recordings of patients (which are not available for both GREECAD and Aachen data sets).

3. Aphasia Syndromes and Diagnosis

Due to the fact that linguistic modules usually are located in distinct neuroanatomical regions of the brain, and that the vascular supply also encompasses distinct areas, occlusion of the trunk or a particular branch of the middle cerebral artery (MCA) leads to typical combinations of linguistic symptoms, called aphasic syndromes. Testing the different linguistic domains thus allows classification of the aphasic syndrome and prediction of the location of the lesion. However, anatomical variations, incomplete or pre-existing lesions or non-vascular lesions can lead to non-standard syndromes, which are then called unclassified aphasia. Additionally, some symptoms can be mapped to anatomical areas that are not solely defined by their vascular supply (Henseler et al., 2014). Typically, however, the following syndromes will occur after an ischemic stroke: occlusion of the main trunk of the MCA (M1 segment) leads to destruction of almost all perisylvic areas concerned with speech and language and subsequent *Global aphasia*. The resulting speech is characterized by a profound loss of syntax and severe disturbances in word retrieval and semantics, sometimes leaving the patient with recurring utterances or automatisms only. Full mutism can occur and language comprehension is severely affected. Occlusion of the anterior branches usually leads to so-called *Broca's aphasia*, marked by non-fluent spontaneous speech (which is monotonous and lacking prosody) and agrammatism. Language comprehension is relatively spared. Lesions in areas supplied by posterior branches of the MCA can lead to *Wernicke's aphasia* which is characterized by fluent spontaneous speech, which however is accompanied by severe disturbances in language comprehension and the use of overshooting, long and tortuous sentences filled with neologisms and paraphasias – a symptom that is called paragrammatism. Prosody usually is preserved. *Amnestic aphasia* is caused by a prominent deficit in word-finding capabilities, while language comprehension and prosody are usually preserved.

Thus, a diagnosis of aphasia is made by testing the presence and severity of the different linguistic symptoms. For this purpose, many validated tests are available in addition to the clinician's expertise that probe variable aspects of the patient's linguistic capabilities. As outlined above, the gold standard in Germany for aphasia diagnosis is the Aachen Aphasia Test (AAT) (Huber et al., 2013). Its purpose is to assess different language modalities (i. e., understanding, writing, reading, speaking) at all linguistic levels. Beyond that, it also yields information of probabilistic aphasia syndrome classification and syndrome severity. The AAT consists of six parts in total, testing different speech and language modality impairments and differentiations. First, and most-important for our current research, an approximately 10 minutes long semi-structured interview is conducted by a clinician. The purpose of the interview is to assess the spontaneous speech capabilities of the patient. Usually, the patient gets to tell about the circumstances the aphasia syndromes first appeared (e. g., when and where a stroke happened and what they where doing), about treatment, family and job etc. The interview is followed by a series of five tests where the patients gets to read, write and has to identify certain tokens. During the AAT, the clinician records the answers on an protocol sheet and takes notes. The interview of the spontaneous speech part is recorded using a basic microphone setup and later transcribed by the clinician, typically a speech and language therapist (SLT). Both the evaluation sheet, the recording and the transcription then constitute the basis for the subsequent diagnosis, which takes up to 6 hours.

While the concrete answers of the patients for each of the five non-interview tests are not directly accessible by us, we only have their final AAT evaluation results, we have access to the raw speech recordings, transcripts and diagnosis results of the (spontaneous speech) interview section. This data forms the basis for our research and the topics discussed in this paper. Each spontaneous speech sample together with its corresponding transcript is evaluated on six different speech impairment levels and on a six point scale (with 0 being the most severe and 5 meaning no impairment) by a clinician. The levels are (Huber, 1983):

1. Communication behavior: Describes the ability of the patient to conduct a dialog, i. e., to understand questions from the clinician and respond to them, to utter speech-based information.

2. Articulation and prosody: Impairments of the speech are described in this level, in particular fluidity, vocalization, preciseness, speed, rhythm.

3. Automatic speech: Features of the speech which are produced automatically by the patient during the dialog are accounted for in this level, e. g., recurring utterances or echophrasias (e. g., repeating phrases of what the clinician said).

4. Semantic structure: This level evaluates the ability of the patient to pick words and to differentiate between their meaning. Furthermore, it evaluates if the patient picks meaningless set phrases.

5. Phonemic structure: Evaluates the order of phonemes in uttered words, e. g., if they are added, dropped, repeated or shuffled.

6. Syntactic structure: This level accounts for the completeness and complexity of sentence parts, their order and amount, and for inflections.

During diagnosis, items 1. and 2. are mostly evaluated on a qualitative level, e. g., is the patient able to communicate daily matters, while 3. – 6. are evaluated on a quantitative level, e. g., the amount of automatisms in the transcript is counted manually.

4. Clinical Aphasia Data Collection and Preprocessing

The available aphasia data in the University hospital consists of several hundred AAT sessions over the course of nearly 20 years. This data was spread over multiples systems within the aphasia ward and was not available in one homogeneous file format (i. e., a mix of txt, doc, docx and PDF documents). To make the data usable for research, we first had to consolidate this data and integrate it into one database. Furthermore, not all datasets were usable for the goal of developing an automatic AAT and had to preprocessed. Some patients had no transcripts, some had no diagnosis sheet, while others where lacking the speech recordings. After a mixture of automatic and manual consolidation, we arrived at a database of 442 complete AAT diagnosis results from 343 patients (some patients took the AAT several times, i. e., for follow-up exams). Each AAT result has a corresponding speech sample in audio format and 388 of them are transcribed. The speech sample stems from the recording of the spontaneous speech evaluation, i. e., the interview, conducted with the clinician. The following sections describes each modality in detail.

4.1. Ratings

Each patient's spontaneous speech performance is rated according to the six categories listed above (see section 3.). The corresponding rating distributions are shown in Figure 2.

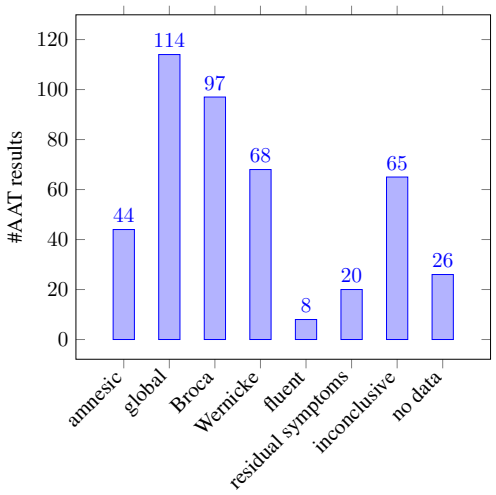


Figure 1: Distribution of aphasia syndromes in the UKA AAT database

Aphasia Type	#Patient	#Utt. Patient (Avg.)
Amnestic	40	491 (12.28)
Broca	53	1225 (23.11)
Global	61	1562 (25.61)
Wernicke	40	612 (15.30)

Table 1: Amount of transcribed utterances available for each of the four most prevalent aphasia syndromes in the UKA AAT database

Notably, there is no test result with a communication impairment rating of zero, as this would be equal to not showing any reactions at all during a conversation, including any non-verbal reactions such as gestures. Additionally, most of the of the samples contain an aphasia severeness rating and an aphasia syndrome diagnosis (e. g., a mild Broca Aphasia). The severeness is rated in five severeness levels, but apparently only mild, moderate to severe and severe are used by most therapists. The aphasia syndrome is classified in six categories, with the most prevalent syndromes being global aphasia, Broca's aphasia, and Wernicke's aphasia. There is an additional category for inconclusive syndromes, i. e., syndromes that are not clearly distinguishable between multiple categories or which do not fit into any category at all (see Figure 1). Furthermore, each AAT sheet also contains the ratings of the 5 other tests, such as the token test. About half of the available AATs also contain information on which therapist conducted the test. There are 104 different therapist names. The most involved therapist conducted 75 tests, while the overwhelming majority of therapist names occurs only once (however this information is not normalized as it was manually entered by therapists). It is entirely possible that the same therapist is referred to under different names such as initials and surname. Due to privacy concerns, information about the patients was anonymized, i. e., neither name, age or gender is given in the data.

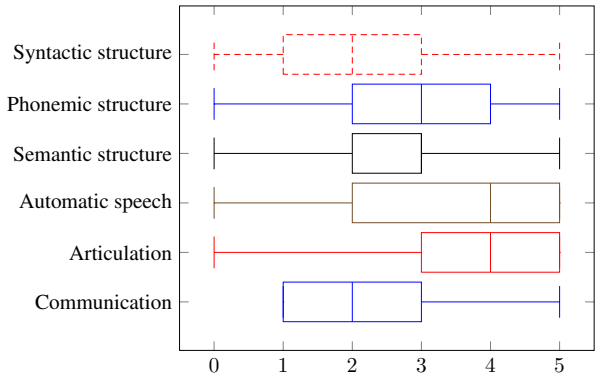


Figure 2: Distribution of AAT speech impairment ratings in the UKA AAT database

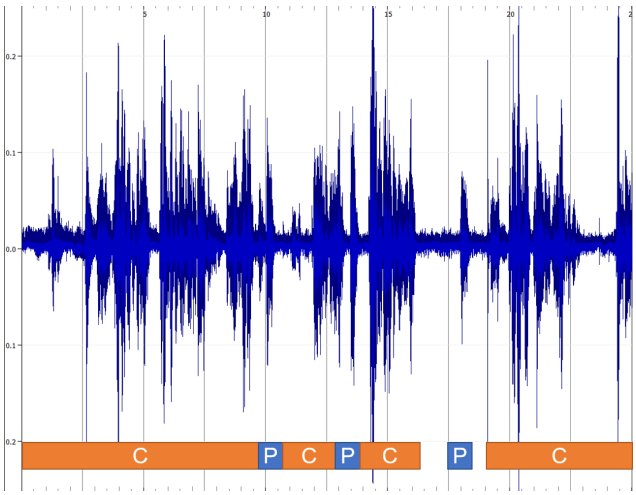


Figure 3: 25 seconds snippet of an global aphasia speech interview. To each question (e. g., "how did the disease start?") the patient (P) responds with a short "um" utterance.

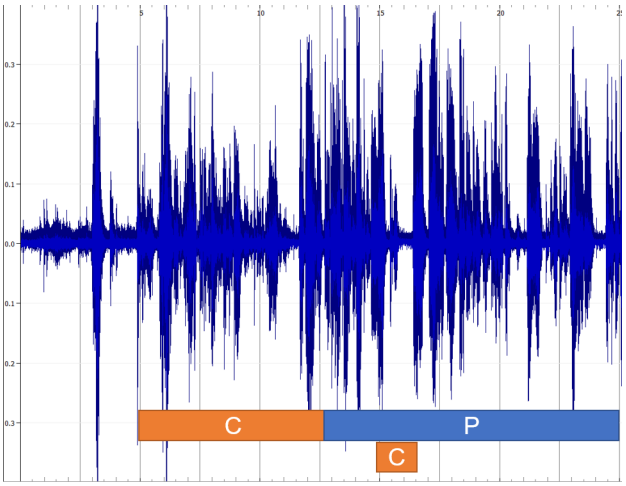


Figure 4: 25 seconds snippet of an Wernicke's aphasia speech interview. The patient (P) answers fluently, but the clinician (C) makes interjections.

4.2. Speech

Each spontaneous speech sample is available as an MP3; most of them are mono recordings. Since the data stems from the course of 20 years, we cannot state the exact type of audio setup for each recording session. As of 2017, the audio setup consists of one microphone positioned between the patient and the clinician. The recording is started manually by the clinician once the spontaneous speech test starts and stopped afterwards. The total duration of all recordings combined is around 63.7 hours. This includes both patient and clinician speech. In order to be able to evaluate aphasic speech, we needed to extract the patient portion of the interview. This can either be done manually or using an automatic speaker diarization system. A completely manual source separation is a very time consuming matter. We found that it took at least 5 – 7 minutes on average to split 1 minute of interview speech (currently, the segmentation is ongoing). Depending on the aphasia syndrome, especially in global aphasia, patients talk only briefly, sometimes uttering just an interjection, before the clinician talks again. That contributes to the necessary time invest, because one has to constantly start and pause the recordings to do the tagging. On the contrary, patients with Wernicke's aphasia tend to talk much longer, but from time to time the clinician makes a comment, leading to an overlap between patient and clinician speech. Again, this segments have to be identified by hand. For a comparison of two different aphasia speech sections see Figures 3 and 4.

As an alternative to a completely manual split of the speech data, we also tested a commercial tool and the open-source framework pyAudioAnalysis (Giannakopoulos, 2015) for speaker diarization. Neither automatic tool could provide the quality of segmentation needed for our research. We attribute this to the difficulty of speaker diarization itself and the complexity of our disease related data. Sometimes, the segmentation contained alternating patient and clinician speech, sometimes both parties were talking, sometimes a mono person segment was labeled as patient when it was in fact the clinician talking and vice versa. We experimented with counteracting the later case by building a binary classifier able to distinguish between aphasia and non-aphasia speech. For this, we extracted 45,912 utterances from the English AphasiaBank corpus ((MacWhinney et al., 2011)) and split these into a train (70 %) and test (30 %) group, based on which sub data set they belong to. Basing the split on the sub data set affiliation prevents us from training and validating based on the same therapists. This results in 25,414 utterances in the training set and 20,498 utterances in the test set (The discrepancy to our 70:30 quota is caused by different sizes of the sub data sets, and the test data set containing larger sub datasets). We subsequently extracted a feature vector for each utterance, using openSMILE (Eyben et al., 2013) with the `IS13_ComParE` feature set (Schuller et al., 2013). These feature vectors, along with the speaker labels extracted from the transcripts, have been used to train a Gradient Boosting classifier to discriminate between clinician and patient. The Gradient Boosting was implemented using scikit-learn 0.19.1 (Pedregosa et al., 2011). The resulting model was evaluated by calculating the mean accuracy of its predictions on the test set, resulting in a mean unweighted

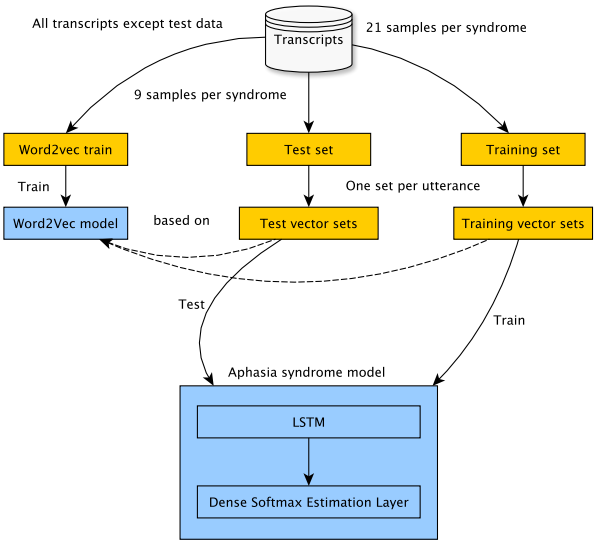


Figure 5: Transcript based aphasia syndrome classification pipeline

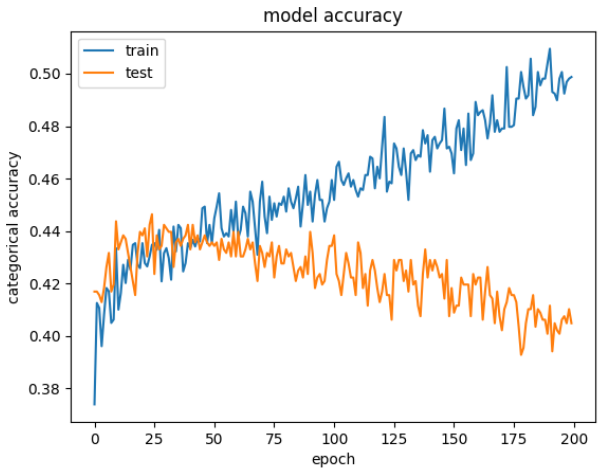


Figure 6: Categorical accuracy of an LSTM estimating the aphasia syndrome. Best performance from epoch 25 to 60, with peak accuracy of 44.3% ($\mu = 25\%$).

accuracy of 83.27 % ($\mu = 50\%$). This is too inaccurate for usage in our system. Additionally, this does not provide any segmentation, but requires a segmentation beforehand, possibly lowering its accuracy even further if the provided segmentation (using an automatic diarization system for preprocessing) is not as accurate as the segmentation of the AphasiaBank.

4.3. Transcripts

After the therapy session is completed, the clinician starts to transcribe the recording of the spontaneous speech session. The speech is transcribed as it is, including interjections like "hmm", or speech and articulation errors. Furthermore, the clinician might also include remarks like "patient is laughing" or "patient is thinking" in curly brackets within the patient portion of the transcript. The clinician also transcribes her own speech. In our data, each transcript is then an alternating list of texts, tagged with either patient or clinician. In Table 1, the amount of utterances available for each of the four standard aphasia syndromes is stated.

5. Early Results and Discussion

For an initial analysis of the data and due to the challenges with speaker diarization we described in 4.2., we started with the goal of predicting the aphasia syndrome type based on the transcripts by configuring a baseline setup. Therefore, a subset of the data has been partitioned into four groups of 30 AAT tests each, such that each group contains patients of one of the four most prevalent aphasia syndroms: global aphasia, Broca's aphasia, Wernicke's aphasia and amnesic aphasia. From each of the four groups representing syndroms, we used 70% for training and 30% for testing purposes. In order to classify the aphasia syndrome based on transcripts, we converted each patient utterance into a list of words and trained a word2vec (Mikolov et al., 2013) model. We chose a window size of three and required each word in the word2vec space to occur at least two times in our utterances. To train the word2vec model, we use our training data set described above, along with phrases from patients which we did not include in the training and test sets before, for instance because they had an inconclusive aphasia syndrome diagnosis. In order to train our aphasia syndrome classifier, we subsequently transform all training utterances into lists of 20-dimensional word vectors, padding them to a length of 30 vectors per utterance. Each of these lists has an assigned aphasia syndrome label and is used to train a pipeline of an long short-term memory (LSTM) layer, followed by a densely connected layer featuring a softmax activation function. This is implemented using Keras (Chollet and others, 2015). The LSTM has been configured to use a 30 % chance of unit dropout and 40 % chance of unit dropout in the recurrent state, while using 80 memory units. We only use a single layer LSTM configuration, as the goal is to provide a baseline for further developments. The model uses an categorical cross entropy loss function and estimates a four dimensional normalized tensor, with each dimension representing one aphasia syndrome. The result is evaluated based on categorical accuracy, which is the percentage of correctly predicted classes, with the "predicted class" being the greatest element of the softmax output tensor. The evaluation has been performed on the test set described above. Plots of accuracy and loss attributes over 200 epochs are depicted in Figures 6 and 7, while the classification pipeline is depicted in Figure 5. The increasing loss function indicates that the model overfits around 100 epochs. Further increasing loss values did not show any meaningful improvements, indicating that more training samples might be the better way to cope with this issue. In summary, the baseline setup shows both the potential and the challenges with clinical aphasia data. While it was possible to perform an initial classification, the usage in clinical scenarios depends on higher accuracies and further improvements (see Section 7.).

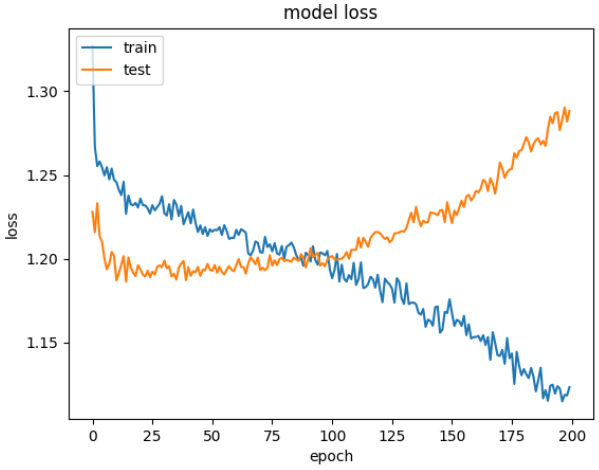


Figure 7: Loss of an LSTM estimating the aphasia syndrome.

6. Conclusion

In this paper, we presented challenges and early results in the automatic processing of real-world clinical aphasia data. We described our data collection of aphasia spanning many years of diagnosis sessions in the university hospital. Each data point in our collection consists of speech recording data, transcripts and rich meta data. The speech data consists of patient clinician interviews and has to be segmented before it can be utilized. We therefore reported on challenges with speaker diarization. The meta data was extracted from diagnosis sheets and contains aphasia syndrome and severeness classification, as well as scores and evaluations of the spontaneous speech section. The scores contain six different categories, which, among others grade the prosody, syntax and phonematic structure of the patient speech. We aim to use this data collection to build and automatic aphasia test, based on the German AAT. Such a system would both benefit clinicians and patients. E. g., patients, many of them mobility impaired stroke victims, could have a continuous spontaneous speech evaluation system at home without the need to go to the hospital every time. In our work, we started with building a baseline syndrome classifier based on an LSTM using the transcript portion of the dataset.

7. Future Work

Our initial implementation of an automatic aphasia syndrome categorizer shows the challenge of the task of usage in a real world scenario. As higher accuracies will be needed before such systems can be used in everyday clinical settings, in the future, we aim to increase its performance in several ways, such as performing a majority vote based on the categorization of all utterances of a patient or additional layers within the classification model. These layers might use information like word histograms and utterance length distributions. Additionally, it might be possible to constrain the decision space for certain combinations of meta information. The latter could be an especially valuable approach when estimating speech impairment factors like automatic speech, as the AAT limits the possible ratings by measurable factors like misplaced words. This would help to cope with the lack of training data, since a first attempt in using an LSTM to do this expressed signs of underfitting and thus yielded a low accuracy. Regarding the segmentation of speech data, we plan to further investigate the possibility of using an automatic speaker diarization system, or at least applying a semi-automatic approach. We think that it might be helpful to include clues about one speaker having impaired speech in the process, i. e., analogous to the paralinguistic approach presented by (Zhang et al., 2017). Finally, we plan to include the speech section as well in order to build a model able to draw from both speech and transcript data. Furthermore, we plan to use the UKA AAT DB (including speech, transcript and rating data) for a challenge, e.g. ComParE at Interspeech, and release it to the research community afterwards. The DB will then include distinct portions for training, development and testing.

8. Acknowledgements

The authors would like to thank Rena Overbeck, M.Sc., for taking the time for patiently answering all questions regarding the concrete execution and evaluation of the AAT.

9. Bibliographical References

Ardila, A. (2010). A proposed reinterpretation and reclassification of aphasic syndromes. *Aphasiology*, 24(3):363–394.

Bozonnet, S., Evans, N., Fredouille, C., Wang, D., and Troncy, R. (2010). An integrated top-down/bottom-up approach to speaker diarization. In *Interspeech 2010, September 26-30, Makuhari, Japan*, pages Interspeech–2010.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., and Vair, C. (2008). Stream-based speaker segmentation using speaker factors and eigenvoices. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4133–4136. IEEE.

Chollet, F. et al. (2015). Keras. `https://github.com/keras-team/keras`.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM.

Fraser, K. C., Rudzicz, F., and Rochon, E. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *INTERSPEECH*, pages 2177–2181.

Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610.

Heilman, K. M. (2006). Aphasia and the diagram makers revisited: an update of information processing models. *Journal of Clinical Neurology*, 2(3):149–162.

Henseler, I., Regenbrecht, F., and Obrig, H. (2014). Lesion correlates of patholinguistic profiles in chronic aphasia: comparisons of syndrome-, modality-and symptom-level assessment. *Brain*, page awt374.

Huber, W., Poeck, K., and Springer, L. (2013). *Klinik und Rehabilitation der Aphasie: eine Einführung für Therapeuten, Angehörige und Betroffene*. Georg Thieme Verlag.

Huber, W. (1983). *Aachener aphasie test (AAT)*. Verlag für Psychologie Dr. CJ Hogrefe.

Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., and Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. *arXiv preprint arXiv:1607.02173*.

Katz, R. C. (2010). Computers in the treatment of chronic aphasia. In *Seminars in speech and language*, volume 31, pages 034–041. Published by Thieme Medical Publishers.

Kohlschein, C., Schmitt, M., Schuller, B., Jeschke, S., and Werner, C. J. (2017). A machine learning based system for the automatic evaluation of aphasia speech. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*.

Kuhn, R., Junqua, J.-C., Nguyen, P., and Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707.

Le, D., Licata, K., Persad, C., and Provost, E. M. (2016). Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2187–2199.

Le, D., Licata, K., and Provost, E. M. (2017). Automatic paraphasia detection from aphasic speech: A preliminary study. *Proc. Interspeech 2017*, pages 294–298.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.

Rouvier, M., Bousquet, P.-M., and Favre, B. (2015). Speaker diarization through speaker embeddings. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2082–2086. IEEE.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Sell, G. and Garcia-Romero, D. (2014). Speaker diarization with plda i-vector scoring and unsupervised calibration. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 413–417. IEEE.

Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565.

Varlokosta, S., Stamouli, S., Karasimos, A., Markopoulos, G., Kakavoulia, M., Nerantzini, M., Pantoula, A., Fyndanis, V., Economou, A., and Protopapas, A. (2016). A greek corpus of aphasic discourse: Collection, transcription, and annotation specifications. In *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016), Monday 23rd of May 2016*, number 128. Linköping University Electronic Press.

Wozniak, M. A. and Kittner, S. J. (2002). Return to work after ischemic stroke: a methodological review. *Neuroepidemiology*, 21(4):159–166.

Zhang, Y., Weninger, F., Liu, B., Schmitt, M., Eyben, F., and Schuller, B. (2017). A paralinguistic approach to speaker diarisation: Using age, gender, voice likability and personality traits. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 387–392. ACM.

Improving the Sensitivity and Specificity of MCI Screening with Linguistic Information

Kathleen C. Fraser[1], Kristina Lundholm Fors[1], Marie Eckerström[2], Charalambos Themistocleous[1], and Dimitrios Kokkinakis[1]
[1]The Swedish Language Bank, Department of Swedish, University of Gothenburg, Gothenburg, Sweden
[2]Institute of Neuroscience and Physiology, Sahlgrenska Academy, Gothenburg, Sweden
{kathleen.fraser, kristina.lundholmfors, marie.eckerstrom, charalambos.themistocleous, dimitrios.kokkinakis}@.gu.se

Abstract

The Mini-Mental State Exam (MMSE) is a screening tool for cognitive impairment. It has been extensively validated and is widely used, but has been criticized as not being effective in detecting mild cognitive impairment (MCI). In this study, we examine the utility of augmenting MMSE scores with automatically extracted linguistic information from a narrative speech task to better differentiate between individuals with MCI and healthy controls in a Swedish population. We find that with the addition of just four linguistic features, the AUC score (measuring a trade-off between sensitivity and specificity) is improved from 0.68 to 0.87 in logistic regression classification. These preliminary results suggest that the accuracy of traditional screening tools may be improved through the addition of computerized language analysis.

Keywords: language processing, machine learning, cognitive impairment, MMSE

1. Introduction

Dementia, a gradual decline in cognitive function due to neurodegeneration, is a growing concern as the global population ages. Research suggests that identifying early signs of cognitive decline may lead to better outcomes for both individuals and their caregivers (Ashford et al., 2007). Mild cognitive impairment (MCI) describes an impairment which is characterized by a clinically observable deficit in at least one area of cognition, but it not severe enough to interfere with activities of daily living (Gauthier et al., 2006; Reisberg and Gauthier, 2008). Although not everyone who is diagnosed with MCI will go on to develop dementia in their lifetimes, MCI is sometimes considered to be a prodromal stage of dementia (Ritchie and Touchon, 2000). Therefore, identifying changes associated with MCI represents a promising step towards the early detection of dementia.

Opinions differ on the value of population-wide screening for dementia; see for example Ashford et al. (2006), Solomon and Murphy (2005), and Calzà et al. (2015). However, even in the absence of large-scale screening programs, it is still of critical importance for primary care practitioners to have sensitive and accurate screening instruments to help determine when to refer an individual for more specialized assessment.

One widely-used cognitive screen is the Mini-Mental State Exam (MMSE) (Folstein et al., 1975). The MMSE contains 12 questions, covering areas such as language, recall, attention, and orientation to time and place, and takes roughly 10 minutes to administer. The test is scored out of 30, and various cut-offs have been proposed as indicating impairment. For example, Ciesielska et al. (2016) conducted a meta-analysis and found that a cut-off of 27/28 was most effective for identifying MCI, corresponding to a sensitivity of 0.66 and specificity of 0.73. Damian et al. (2011) found the optimal cut-off for detecting MCI in their dataset to also be 27/28 (sensitivity: 0.76, specificity: 0.75), while noting that these metrics can be sensitive to the proportion of patients and controls in any given data set. Other studies have considered an "abnormal" score to be anything from 25 and below (Zadikoff et al., 2008) to 28 and below (Pendlebury et al., 2012). Since performance on the MMSE is influenced by educational level and cultural background, cutoffs are not necessarily transferable from one country to another (Palmqvist et al., 2013). For Swedish, a cutoff of 24 and lower has been proposed for cognitive impairment, with a score between 25 and 27 indicating possible cognitive impairment which should be further evaluated (Palmqvist et al., 2013). While a number of researchers have argued that MMSE is not the best screening instrument for MCI, it remains the most widely used short screening tool for providing an overall measure of cognitive impairment in clinical, research and community settings (Arevalo-Rodriguez et al., 2015).

In this paper, we propose augmenting MMSE scores with additional information obtained from automated linguistic analysis, to improve the detection of MCI in a population of Swedish speakers. Our analysis currently relies on manual transcriptions, but we envision that a fully automated system incorporating speech recognition could provide a more detailed and accurate assessment of cognitive status, while requiring minimal extra effort on the part of the primary care physician.

2. Related Work

A number of studies have reported that subtle changes in speech and language may occur at the earliest stages of cognitive decline (Snowdon et al., 1996; Garrard et al., 2004; Cuetos et al., 2007; Clark et al., 2009; Le et al., 2011; Ahmed et al., 2013). According to Laske et al. (2015), language analysis is one of *the* most promising state-of-the-art diagnostic measures for MCI and Alzheimer's disease. Since manual linguistic analysis can be expensive and time-consuming, there has been interest in developing

automated methods for language analysis of clinical samples, using natural language processing and machine learning (e.g. Garrard and Forsyth (2010), Jarrold et al. (2014), Rentoumi et al. (2014), Prud'hommeaux and Roark (2015), and Kavé and Goral (2016), among others). Specifically with respect to MCI, Vincze et al. (2016) combined linguistic features extracted from patient narratives with demographic variables to achieve a classification accuracy of 0.69 using all features, and 0.75 using selected features (46 MCI, 36 controls). Asgari et al. (2017) reported an accuracy of 0.84 in distinguishing 14 MCI participants from 27 healthy controls, by extracting linguistic and psychological features from unstructured conversation.

Combining linguistic features with neuropsychological test scores has been used in the past to improve MCI classification. Roark et al. (2011) reported a study in which 37 MCI participants and 37 controls were assessed on nine neuropsychological tasks and two speech samples from a story recall task. Better classification accuracy was achieved by combining the neuropsychological and language information than by using a single modality alone. Mueller et al. (2017) correlated 280 individuals' performance on standardized neuropsychological tests with various language factors, such as grammatical complexity, fluency markers and other lexical information. Syntax was found to be weakly positively correlated with MMSE, while fluency and semantic features declined more rapidly in the MCI group than in the control group, over the course of the study period.

3. Methods

3.1. Participants and Data Acquisition

The participants were recruited from the Gothenburg MCI Study, a clinical-pathophysiologic longitudinal study investigating early and manifest phases of different dementia disorders in patients seeking medical care at a memory clinic (Wallin et al., 2016). The Gothenburg MCI Study is approved by the local ethical committee review board (reference number: L091–99, 1999; T479–11, 2011); while the currently described study is approved by the local ethical committee (decision 206–16, 2016).

A total of 31 MCI patients and 36 healthy controls were included in the present study, according to detailed inclusion and exclusion criteria (Kokkinakis et al., 2017). One control participant was excluded from the current analysis because their MMSE score was not available.

All participants were assessed by a registered nurse, who administered a number of cognitive tests, including the MMSE. If participants showed no signs of subjective or objective cognitive impairment, they were classified as healthy controls. Other participants were then categorized according to the Global Deterioration Scale following cognitive testing, and participants classified as stage 3 (MCI) were included in the current analysis. Participants classified at stage 2 (subjective cognitive impairment) were excluded. Note that the MMSE did form part of the classification procedure, which makes our analysis somewhat circular, but that other factors were also taken into account.

Participant demographics are given in Table 1. There is no significant difference between the groups on age, level of education, or proportion of males and females. There is a significant difference on MMSE score, with MCI participants scoring lower, although we observe that on average, the MCI participants score in the normal range according to the cutoff proposed by Palmqvist et al. (2013). This supports the argument that MMSE is not the most sensitive screening tool for early cognitive impairment.

	HC ($n = 35$)	MCI ($n = 31$)	Sig.
Age	68.0 (7.3)	70.1 (5.6)	n.s.
Education	13.3 (3.4)	14.1 (3.6)	n.s.
Sex (M/F)	13/22	15/16	n.s.
MMSE	29.6 (0.6)	28.2 (1.4)	$p < 0.01$

Table 1: Demographic information. Age and education are measured in years; MMSE is scored out of 30.

For the narrative speech component, participants were instructed to describe what they could see in the "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination (Goodglass et al., 1983). This image is widely used to elicit narrative speech. It shows a boy standing on a stool, trying to steal a cookie from a jar sitting on a high shelf. A girl stands nearby and a woman washes dishes, apparently unconcerned by both the children's actions and the water which overflows onto her feet.

Participants were told that they could talk for as long as they wanted and that they would not be interrupted. The narratives were audio-recorded and subsequently manually transcribed by experienced transcribers according to guidelines provided by the authors.

3.2. Features

A total of 57 linguistic features were extracted from the Cookie Theft transcripts. A complete description of these features is given in Table 2. Part-of-speech (POS) tagging, lemmatization, and dependency parsing was done using the Sparv annotation tool for Swedish (Borin et al., 2016). The constituent parse features were extracted using the CASS-Swe parser (Kokkinakis and Johansson Kokkinakis, 1999).

3.3. Classification

We take a machine learning approach to classifying the groups. We consider three classifiers in this work: support vector machines (SVM), logistic regression (LR), and random forests (RF) (Pedregosa et al., 2011).

In each classification experiment, we use a leave-one-out cross-validation framework, where a single participant narrative is held out for testing, and the classifiers are trained on the remaining 65 samples. The procedure is then repeated 66 times, and we report the average results over the folds. Within each fold, an inner 5-fold cross validation loop is used to optimize the hyperparameters of the classifiers (for SVM, we fix a linear kernel and optimize the complexity parameter C between 10^{-3} and 10^{3}; for LR we use ridge regularization and range C from 10^{-3} to 10^{3}; and for RF we fix the number of trees at 50 and optimize the maximum number of features to consider at each split to be either n or $\sqrt{n}$, where n is the number of features, and the maximum depth of the trees to be 3, 4, 5, or unlimited).

Syntactic parse features	**Phrase type proportion** and length (below) are derived from work on rating the fluency of machine translations (Chae and Nenkova, 2009). The phrase type proportion is the total number of words belonging to a given phrase type (here prepositional phrases, noun phrases, and verb groups), divided by the total number of words in the narrative. We additionally extend this feature to apply to clauses; namely main finite clauses, main infinitive clauses, and subordinate clauses, for a total of 6 distinct features. **Phrase type length** is the total number of words belonging to a given phrase or clause type, divided by the total number of occurrences of that phrase or clause type (6 features).
Dependency distance	**Dependency distance** is measured as the number of words between a given word and its dependency head, calculated for each word in the sentence. We compute average, maximum, and total dependency distance for each sentence, and then average these quantities over each sentence in the transcript (3 features).
Part-of-speech tags	**POS counts** are computed for nouns, verbs, adjectives, adverbs, prepositions, determiners, and pronouns, and then normalized by dividing by the total number of words in the narrative (7 features). **POS ratios** are also computed in some cases; namely, the ratio of nouns to verbs, the ration of pronouns to nouns, and the ratio of function words to total words (3 features).
Verb features	**Inflected verb count** includes those verb forms with morphological inflection, divided by the total number of words (1 feature). **Light verb count** includes all mentions of verb tokens from the set {*vara, ha, komma, gå, ge, ta, göra, få, flytta, lägga*}, divided by the total number of words (1 feature).
Psycholinguistic features	**Frequency** is estimated according to a word's unlemmatized frequency in the modern Swedish section of the Korp corpus (Borin et al., 2012). It is averaged over all words together, and additionally for nouns and verbs separately (3 features). **Familiarity** is estimated from a paper survey of 42 native Swedish speakers, conducted at the Gothenburg Book Fair in October, 2017. The survey contained mostly words relating to the content of the Cookie Theft image, as well as control words for which familiarity ratings already existed (Blomberg and Öberg, 2015). Correlation with the previously collected familiarity norms was $r = 0.80$, $p = 0.06$. Familiarity is averaged over all words (1 feature). **Imageability** is estimated from a paper survey, as above. Correlation with the previously collected imageability norms was $r = 0.98$, $p = 0.001$. Imageability is averaged over all words (1 feature). **Emotional valence** is estimated from a paper survey, as above. Correlation with the previously collected valence norms was $r = 0.95$, $p = 0.003$. Valence is averaged over all words (1 feature).
Vocabulary richness	**Type-token ratio (TTR)** is calculated by dividing the number of unique word types by the total number of tokens in the narrative (1 feature). **Honoré's statistic** is calculated according to $H = 100 * \log(N/(1 - V_1/V))$, where N is the total number of words used (number of tokens), V is the size of the vocabulary (number of types), and V_1 is the number of words used only once in the narrative (Honoré, 1979) (1 feature).
Information units	**Content counts** are computed for the 4 categories of information units listed in (Kavé and Levy, 2003); namely, the three *subjects*, thirteen *objects*, two *places*, and seven *actions*. These counts are extracted using a keyword-spotting method with manual post-hoc inspection. The raw features are integer-valued with no upper bound (e.g. if the speaker mentions the boy five times, then the content_count for *subjects* increases by five), and so the final features are normalized by the total number of words in the transcript. We also compute the total_content_count by summing the counts for the 4 categories (5 features). **Information counts** are computed for the 4 categories of information units listed above. These features are integer-valued, with an upper bound equal to the number of information units in each category (e.g. if the speaker mentions the boy five times, then the information_count for *subjects* still only increases by one.) We also compute the total_information_count by summing the counts for the 4 categories (to a maximum of 25) (5 features). **Content density** and **information density** are computed by dividing the total_content_count and total_information_count by the total number of words in the narrative (2 features). **Content efficiency** and **information efficiency** are computed by dividing the total_content_count and total_information_count by the total time taken to produce the narrative (2 features).
Fluency features	**Total words** is the total number of words produced (excluding filled pauses, unintelligible words, and false starts) (1 feature). **Total time** is the total time taken to produce the narrative (1 feature). **Speech rate** is measured in words per minute (total words divided by total time) (1 feature). **Hesitancy counts** are computed by counting the number of pauses, false starts, and incomplete sentences, each normalized by total number of words (3 features). **Mean length of sentence (MLS)** is the total number of words in the narrative divided by the number of sentences (1 feature). **Mean length of word (MLW)** is the average length of the words in the narrative, in letters (1 feature).

Table 2: Linguistic features extracted from the Cookie Theft transcripts.

We first train the classifiers on MMSE alone. This is equivalent to letting the classifiers learn the optimal threshold on the MMSE to separate the two groups. We then consider the effect of adding a single linguistic feature, then two linguistic features, and so on until the entire set of 57 linguistic features has been added to the classification.

The order in which features are added to the classifiers is obviously important. One possibility is to simply rank the features by computing a t-test on the training data and choosing the features which best differentiate the groups. However, initial experiments found that this could result in correlated features being selected, which had a negative effect on classifier performance. Instead, we use a wrapper method of feature selection, which selects the features based on the model itself, through recursive feature elimination (Guyon et al., 2002). In the feature selection stage, default parameter values are used, except that we again specify the linear kernel for SVM, ridge regularization for LR, and 50 trees for RF. The downside to this method is that the different models may select different features, making interpretation more difficult. The most-commonly selected features will be discussed in Section 4.2.

For evaluation, we consider accuracy, sensitivity, specificity, and the area under the curve (AUC) of the receiver

operating characteristic (ROC) curve. Accuracy, sensitivity, and specificity are computed as follows, where we consider MCI to be the positive class, and TP indicates a true positive, FP indicates a false positive, TN indicates a true negative, and FN indicates a false negative:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Sensitivity and specificity are particularly relevant in a healthcare context: a test which is highly sensitive means that not many people who actually *do* have the disease are missed, while a test which is highly specific means that not many people who *do not* have the disease are falsely indicated as having the disease. The AUC is calculated by plotting sensitivity (also known as the *true positive rate*) against $1 - \text{specificity}$ (also known as the *false positive rate*), as the decision threshold of the classifier is varied. The AUC is the area under the resulting curve. Random performance leads to an AUC of 0.5, and represents a straight line from $(0, 0)$ to $(1, 1)$.

4. Results

4.1. Classification

Figure 1 shows the sensitivity, specificity, AUC, and accuracy for each classifier as the number of linguistic features is increased from zero (MMSE score only) to 57. Looking first at Figure 1a, we see that using MMSE only, the SVM classifier has a high sensitivity of 0.81, while the LR and RF classifiers have a lower sensitivity. These results correspond to selecting a MMSE threshold of 29 (SVM) versus 28 (LR and RF). In the latter cases, the sensitivity is improved by adding language features, to maximum values of 0.77 for LR and 0.74 for RF. The SVM sensitivity is never as high as using MMSE alone, but reaches 0.77 with three linguistic features.

When we examine specificity, in Figure 1b, we see the expected trade-off between sensitivity and specificity. Using MMSE scores alone, LR and RF have specificity of 0.94 (i.e. by using a threshold of 28, very few control participants are misclassified as having MCI). The SVM classifier has a specificity of only 0.63, which can be improved to 0.80 by including only one linguistic feature.

The AUC score, shown in Figure 1c, balances the trade-off between sensitivity and specificity. For all three classifiers, performance is boosted by adding linguistic features, but all achieve maximum AUC by including only a few additional features: the SVM classifier has $\text{AUC} = 0.84$ with three language features, RF has $\text{AUC} = 0.81$ with two language features, and LR achieves the best AUC of 0.87, with four language features. Accuracy, shown in Figure 1d, presents a similar pattern.

The AUC can also be represented visually by plotting the ROC curve, as seen in Figure 2. The black points indicate the values that are achieved by simply thresholding the MMSE scores at different cutoffs, and classifying participants on that basis alone. For the purposes of illustration, we plot the ROC curves for only the best-performing configurations for each of the three classifiers. For low threshold values, the RF curve (green) lies below the black curve, indicating a higher number of false positives, but the RF classifier performs quite well at the higher threshold values. In contrast, the SVM curve (orange) lies mostly above the black curve for low threshold values, but has a lower true positive rate at high threshold values, even dipping below random performance at the far end of the range. The LR curve (blue) generally lies at or above the curve obtained using MMSE thresholds alone, indicating the improved performance at all threshold values for this classifier.

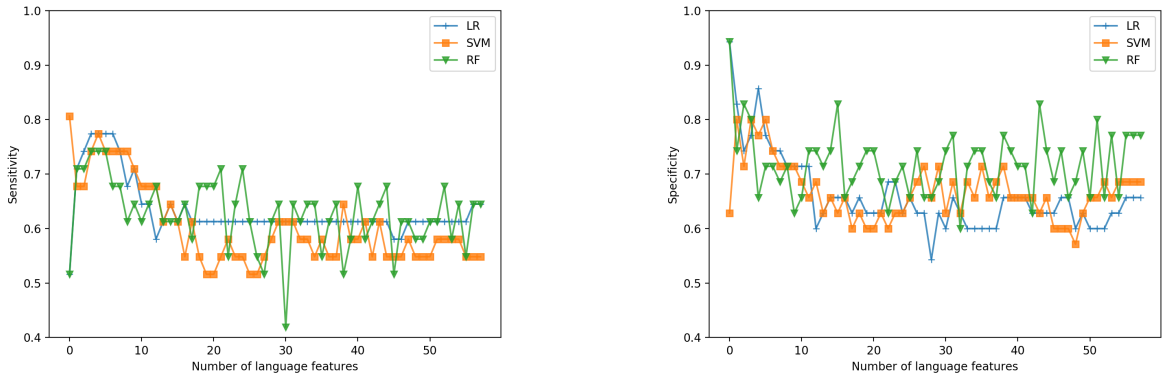
4.2. Important Features

We now consider the question of which linguistic features were the most helpful to the classifiers in improving the classification results. Rather than trying to compare classifier-specific values such as coefficients (LR or SVM) or feature importances (RF), we use the rankings produced in the feature selection stage as a measure of feature relevance. Table 3 shows the mean rank across folds for each feature, for each classifier. A higher rank generally indicates that the feature is more important in the model, while a greater standard deviation suggests a feature which may not generalize well (if it is highly ranked in some folds but not others, then it is apparently quite sensitive to the exact training set, which can lead to overfitting). Since all three classifiers reach their maximum performance with the addition of five or fewer linguistic features, we consider here only the top five ranked features.

We observe that the number of times the speaker mentions an information unit from the *place* category is ranked first for the LR classifier and second for both SVM and RF. Interestingly, this feature has a higher mean value in the MCI data than in the HC data (MCI: 0.018, HC: 0.013, uncorrected $p = 0.01$). This is in contrast to the findings of Croisile et al. (1996), who observed that healthy controls were more likely to name both of the relevant places (the kitchen and the exterior) than patients with Alzheimer's disease. Here, the effect may be driven more by the MCI participants making *repeated* references to the two locations, as on average both HC and MCI participants mention the kitchen and the exterior at least once. In the RF classifier, the number of times the speaker mentions an *action* information unit is also highly ranked, although here the difference between groups is even smaller.

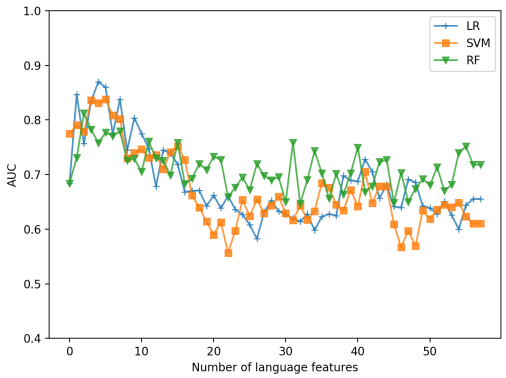
Another highly ranked set of features is the proportion of main finite clauses (lower in the MCI group), the proportion of main nonfinite clauses (higher in the MCI group), and the proportion of subordinate clauses (lower in the MCI group). Previous findings regarding changes in syntactic complexity due to mild cognitive decline are mixed; the results on our data set are discussed in more detail by Lundholm Fors et al. (2018), but require further investigation.

The remaining highly ranked features involve the count for nouns, the noun:verb ratio, word frequency, and verb frequency in particular. Our data show that the number of

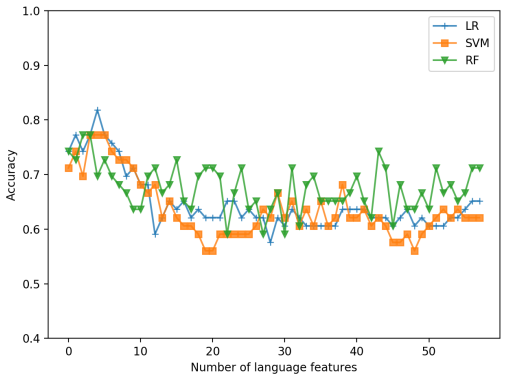


(a) Sensitivity

(b) Specificity



(c) AUC



(d) Accuracy

Figure 1: Effect on MCI-vs-HC classification results of supplementing MMSE information with linguistic features.

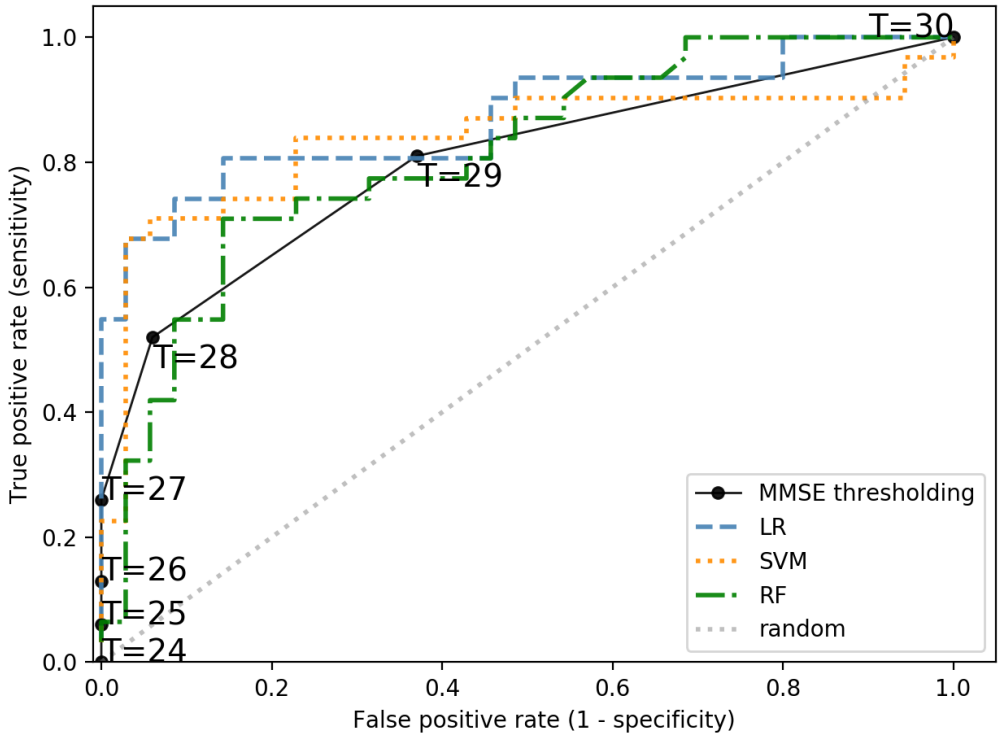


Figure 2: ROC curves. The values corresponding to thresholding the MMSE scores manually are shown in black (e.g. $T = 28$ indicates a split between 28/29) . The coloured curves represent the performance obtained by varying the decision threshold from 0.0 to 1.0, for the best configuration for each of the three classifiers.

LR		RF		SVM	
Feature	Rank	Feature	Rank	Feature	Rank
content count: *places*	1.3 (0.3)	noun count	1.9 (2.1)	MAIN-FIN proportion	1.5 (0.8)
MAIN-FIN proportion	2.2 (0.5)	content count: *places*	2.2 (0.9)	content count: *places*	2.4 (1.0)
SUB proportion	3.2 (0.5)	MAIN-INF proportion	4.1 (1.3)	SUB proportion	2.7 (1.1)
verb frequency	5.0 (4.0)	noun:verb ratio	5.7 (4.9)	noun:verb ratio	10.0 (7.5)
noun count	7.0 (6.9)	frequency	9.1 (3.6)	content count: *actions*	10.2 (7.7)

Table 3: The average ranking of the top five features across folds, for each classifier. Standard deviations are given in parentheses.

nouns is reduced in the MCI group, while there is a corresponding (but very small) increase in the number of verbs. As a result, the noun:verb ratio is slightly higher in the HC group than in the MCI group (HC: 1.07, MCI: 0.95, uncorrected $p = 0.08$). This pattern is consistent with the neurophysiology of Alzheimer's disease, in that areas connected with noun processing tend to be affected in the earliest stages of the disease (Vigliocco et al., 2011). Participants with MCI also tended to use higher-frequency verbs, and higher-frequency words in general.

However, we note that of these features, only the proportion of main clauses with nonfinite verbs, the proportion of main clauses with finite verbs, and the content count for *places* varied significantly between the groups before correcting for multiple comparisons, and no differences were significant after Bonferroni correction.

5. Conclusions and Future Work

In this study, we examined the utility of adding automated language analysis to improve MCI classification, relative to using MMSE scores alone. The results were positive, showing that all three classifiers could improve AUC by including a few language features. The best result was achieved using logistic regression, which improved from $\mathrm{AUC} = 0.68$ using MMSE alone to 0.87 by allowing the classifier to include four language features in addition to the MMSE score.

However, none of the features showed a significant difference between groups, and many features reported to be relevant by previous studies were not found to be so here. We attribute this mainly to the high level of cognitive function in our MCI group, and the small sample size, which together mean we lack the statistical power needed to uncover very small differences between the groups.

We also consider the possibility that the Cookie Theft task is not particularly difficult for highly-educated, professional individuals at a very early stage of cognitive decline. In our next round of data collection, we plan to include language tasks which also incorporate aspects of memory and attention, and which elicit dialogue as opposed to just monologue. We expect that these additional tasks may offer a broader assessment of the speaker's cognitive status.

Nonetheless, we consider this a promising result that can offer additional diagnostic value, and a step towards improving the accuracy of screening tools by augmenting traditional methods with computer technology.

6. Acknowledgements

This work has received support from Riksbankens Jubileumsfond – The Swedish Foundation for Humanities & Social Sciences, through the grant agreement no: NHS 14-1761:1.

7. Bibliographical References

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737.

Arevalo-Rodriguez, I., Smailagic, N., i Figuls, M. R., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., Pedraza, O. L., Cosp, X. B., and Cullum, S. (2015). Mini-mental state examination (MMSE) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *BJPsych Advances*, 21(6):362–362.

Asgari, M., Kaye, J., and Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228.

Ashford, J. W., Borson, S., O'Hara, R., Dash, P., Frank, L., Robert, P., Shankle, W. R., Tierney, M. C., Brodaty, H., Schmitt, F. A., Kraemer, H. C., and Buschke, H. (2006). Should older adults be screened for dementia? *Alzheimer's & Dementia*, 2(2):76–85.

Ashford, J. W., Borson, S., O'Hara, R., Dash, P., Frank, L., Robert, P., Shankle, W. R., Tierney, M. C., Brodaty, H., Schmitt, F. A., et al. (2007). Should older adults be screened for dementia? It is important to screen for evidence of dementia! *Alzheimer's & Dementia: the Journal of the Alzheimer's Association*, 3(2):75–80.

Blomberg, F. and Öberg, C. (2015). Swedish and English word ratings of imageability, familiarity and age of acquisition are highly correlated. *Nordic Journal of Linguistics*, 38(3):351–364.

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp - the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 474–478.

Borin, L., Forsberg, M., Hammarstedt, M., Rosen, D., Schäfer, R., and Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University, 17-18 November*.

Calzà, L., Beltrami, D., Gagliardi, G., Ghidoni, E., Marcello, N., Rossini-Favretti, R., and Tamburini, F. (2015).

Should we screen for cognitive decline and dementia? *Maturitas*, 82(1):28–35.

Chae, J. and Nenkova, A. (2009). Predicting the fluency of text with shallow structural features: Case studies of machine translation and human-written text. In *12th EACL*, pages 139–147.

Ciesielska, N., Sokolowski, R., Mazur, E., Podhorecka, M., Polak-Szabela, A., and Kedziora-Kornatowska, K. (2016). Is the Montreal Cognitive Assessment (MoCA) test better suited than the Mini-Mental State Examination (MMSE) in mild cognitive impairment (MCI) detection among people aged over 60? *Psychiatria Polska*, 50(5):1039–1052.

Clark, L. J., Gatz, M., Zheng, L., Chen, Y.-L., McCleary, C., and Mack, W. J. (2009). Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimer's disease. *American Journal of Alzheimer's Disease & Other Dementias*, 24(6):461–468.

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., and Trillet, M. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language*, 53(1):1–19.

Cuetos, F., Arango-Lasprilla, J. C., Uribe, C., Valencia, C., and Lopera, F. (2007). Linguistic changes in verbal expression: A preclinical marker of Alzheimer's disease. *Journal of the International Neuropsychological Society*, 13(3):433–439.

Damian, A. M., Jacobson, S. A., Hentz, J. G., Belden, C. M., Shill, H. A., Sabbagh, M. N., Caviness, J. N., and Adler, C. H. (2011). The Montreal Cognitive Assessment and the Mini-Mental State Examination as screening instruments for cognitive impairment: item analyses and threshold scores. *Dementia and Geriatric Cognitive Disorders*, 31(2):126–131.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

Garrard, P. and Forsyth, R. (2010). Abnormal discourse in semantic dementia: A data-driven approach. *Neurocase*, 16(6):520–528.

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2004). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128(2):250–260.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., et al. (2006). Mild cognitive impairment. *The Lancet*, 367(9518):1262–1270.

Goodglass, P., Barresi, B., and Kaplan, E. (1983). Boston Diagnostic Aphasia Examination. Philadelphia: Lippincott Williams and Willkins. A Wolters Kluwer Company.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pages 27–36.

Kavé, G. and Goral, M. (2016). Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, 38(9):958–966.

Kavé, G. and Levy, Y. (2003). Morphology in picture descriptions provided by persons with Alzheimer's disease. *Speech, Language, and Hearing Research*, 46(2):341–352.

Kokkinakis, D. and Johansson Kokkinakis, S. (1999). A cascaded finite-state parser for syntactic analysis of swedish. In *Proceedings of the 9th EACL*, pages 245–248, Oslo, Norway.

Kokkinakis, D., Lundholm Fors, K., Björkner, E., and Nordlund, A. (2017). Data collection from persons with mild forms of cognitive impairment and healthy controls—infrastructure for classification and prediction of dementia. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, number 131, pages 172–182. Linköping University Electronic Press.

Laske, C., Sohrabi, H. R., Frost, S. M., López-de Ipiña, K., Garrard, P., Buscema, M., Dauwels, J., Soekadar, S. R., Mueller, S., Linnemann, C., et al. (2015). Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's & Dementia*, 11(5):561–578.

Le, X., Lancashire, I., Hirst, G., and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*, 26(4):435–461.

Lundholm Fors, K., Fraser, K. C., and Kokkinakis, D. (2018). Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment. In *Proceedings of the Medical Informatics Europe (MIE) Conference*.

Mueller, K., Koscik, R., Hermann, B., Johnson, S., and Turkstra, L. (2017). Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Frontiers in Aging Neuroscience*, 9(437):1–14.

Palmqvist, S., Terzis, B., Strobel, C., and Wallin, A. (2013). MMSE-SR: Mini Mental State Examination - Svensk Revidering.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pendlebury, S. T., Mariz, J., Bull, L., Mehta, Z., and Rothwell, P. M. (2012). MoCA, ACE-R, and MMSE versus the National Institute of Neurological Disorders

and Stroke–Canadian Stroke Network vascular cognitive impairment harmonization standards neuropsychological battery after TIA and stroke. *Stroke*, 43(2):464–469.

Prud'hommeaux, E. and Roark, B. (2015). Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578.

Reisberg, B. and Gauthier, S. (2008). Current evidence for subjective cognitive impairment (SCI) as the pre-mild cognitive impairment (MCI) stage of subsequently manifest Alzheimer's disease. *International Psychogeriatrics*, 20(1):1–16.

Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C. A., and Garrard, P. (2014). Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease*, 42(S3):S3–S17.

Ritchie, K. and Touchon, J. (2000). Mild cognitive impairment: conceptual basis and current nosological status. *The Lancet*, 355(9199):225–228.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Journal of the American Medical Association*, 275(7):528–532.

Solomon, P. R. and Murphy, C. A. (2005). Should we screen for Alzheimer's disease? *Geriatrics*, 60(11):26–31.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., and Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and Biobehavioral Reviews*, 35(3):407–426.

Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M., and Kálmán, J. (2016). Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 181–187.

Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., Stålhammar, J., Eckerström, M., Kern, S., Börjesson-Hanson, A., Carlsson, M., Olsson, E., Zetterberg, H., Blennow, K., Svensson, J., Öhrfelt, A., Bjerke, M., Rolstad, S., and Eckerström, C. (2016). The Gothenburg MCI study: Design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow and Metabolism : Official journal of the International Society of Cerebral Blood Flow and Metabolism*, 36(1):114–31.

Zadikoff, C., Fox, S. H., Tang-Wai, D. F., Thomsen, T., de Bie, R., Wadia, P., Miyasaki, J., Duff-Canning, S., Lang, A. E., and Marras, C. (2008). A comparison of the Mini Mental State Exam to the Montreal Cognitive Assessment in identifying cognitive deficits in Parkinson's disease. *Movement disorders*, 23(2):297–299.

Language Modelling for the Clinical Semantic Verbal Fluency Task

Nicklas Linz[1], Johannes Tröger[1], Hali Lindsay[1], Alexandra König[2], Philippe Robert[2], Jessica Peter[3] and Jan Alexandersson[1]
[1]German Research Center for Artificial Intelligence (DFKI), Germany
[2]Memory Clinic, Association IA, CoBTek Lab - CHU Université Côte d'Azur, France
[3]University Hospital of Old Age Psychiatry and Psychotherapy, University of Bern, Switzerland
nicklas.linz@dfki.de

Abstract

Semantic Verbal Fluency (SVF) tests are common neuropsychological tasks, in which patients are asked to name as many words belonging to a semantic category as they can in 60 seconds. These tests are sensitive to even early forms of dementia caused by e.g. Alzheimer's disease. Performance is usually measured as the total number of correct responses. Clinical research has shown that not only the raw count, but also production strategy is a relevant clinical marker. We employed language modelling (LM) as a natural technique to model production in this task. Comparing different LMs, we show that perplexity of a persons SVF production predicts dementia well ($F_1 = 0.83$). Demented patients show significantly lower perplexity, thus are more predictable. Persons in advanced stages of dementia differ in predictability of word choice and production strategy - people in early stages only in predictability of production strategy.

Keywords: Dementia, Alzheimer's Disease, Semantic Verbal Fluency, Language Modelling, Machine Learning

1. Introduction

Verbal fluency is among one of the most widely used neuropsychological standard tests. Category fluency, or semantic verbal fluency (SVF), requires a participant to produce as many different items from a given category, e.g. animals, as is possible, in a given time frame. Over the past years, a growing body of research substantiates the discriminative power of semantic fluency for multiple different pathologies: neurodegenerative diseases such as Alzheimer's disease (Pakhomov et al., 2016; Raoux et al., 2008; Auriacombe et al., 2006; Gomez and White, 2006; Henry et al., 2004), Parkinson's disease (Henry and Crawford, 2004), psychiatric disorders such as schizophrenia (Robert et al., 1998), Primary Progressive Aphasia (PPA) and its subforms (Bonner et al., 2010; Marczinski and Kertesz, 2006), as well as focal lesions (Troyer et al., 1998). Traditionally, SVF is one of the most broadly used test to diagnose dementia and it's multiple subforms (see Figure 1).

As is standard clinical procedure, performance in this test is evaluated as the raw word count (count of correct responses). In order to differentiate between multiple pathologies, qualitative measures have been established which serve as additional indicators in tandem with the raw word count (Gruenewald and Lockhead, 1980; Troyer et al., 1997). There is broad evidence that those qualitative SVF measures serve as indicators for underlying cognitive processes; this has been investigated to the extent that verbal fluency can be considered as a multifactorial task, comprising both executive control and memory retrieval processes (Henry et al., 2005; Robert et al., 1998; Troyer et al., 1997).

Considering the involvement of the two distinct cognitive processes, Troyer et al. (1997) first introduced a systematic framework to calculate measure for both processes from the response behaviour of a subject. In general, production of words is organised in *spurts*—temporal clusters—followed by pauses, implying the lexical search for semantic fields or subcategories between clusters, and retrieval/production of words within clusters (Gruenewald and Lockhead, 1980; Troyer et al., 1997). This means, that between temporal clusters, executive search processes—switching—and within temporal clusters, semantic memory retrieval processes—clustering—are engaged. The underlying notion is that temporal clusters correspond to semantic clusters; in other words, "words that comprise these temporal clusters tend to be semantically related" (Troyer et al., 1997, p. 139).

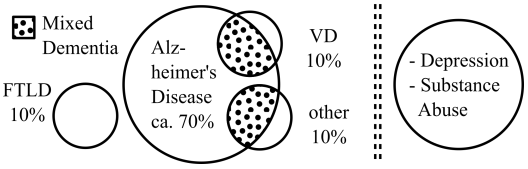


Figure 1: The left panel shows different dementia types and their underlying causes, including Fronto-Temporal Lobar Degeneration (FTLD), and Vascular Dementia (VD); the dotted areas indicate those cases where more than one cause underlies the disorder. The right panel shows other, mostly reversible, causes for dementia-like symptoms.

In this paper we use statistical language models (LMs) as a tool for modelling production of SVF responses of healthy patients, those with a diagnosis of mild cognitive impairment (MCI[1]) and Alzheimer's disease or related dementia (ADRD). LMs intuitively model production of words in SVF, as production of the next word depends on the previously produced words. Given a corpus of SVF performances, we use LMs to learn these probabilities from data, and then test the model, by estimating the likelihood of a patient's SVF performance. We use the LM's perplexity of a given SVF performance — a score for how well the

[1]MCI is associated with an increased risk to develop manifest dementia

model is able to predict a given sequence — as a feature for classification of a person's cognitive health.

This paper is structured as follows: Section 2. discusses prior work on clinical applications of language models and perplexity scores. Section 3. introduces language models. Section 4. describes the data for further experiments, how the language models were trained and evaluated in a classification experiment. Section 5. presents results of the conducted experiments. Lastly, Section 6. discusses implications and concludes the paper.

2. Related Work

There is a growing body of research using language modelling and perplexity scores for classification of neurocognitive disorders including Alzheimer's disease (AD), varying types of dementia, and frontotemporal lobar degeneration (FTLD).

In previous work, perplexity scores have been used to automatically classify between AD patients' and healthy controls' speech (Wankerl et al., 2017). Language models were built on transcripts from spontaneous speech of subjects describing the Cookie Theft Picture from the Boston Diagnostic Aphasia Examination battery. The resulting language models based on AD speech and control subjects' speech were then used to compute different perplexity scores per patient including perplexity of an AD language model given an AD speech sample and perplexity of an AD language model given a control speech sample. The authors conclude that perplexity in such a free speech task is higher for AD samples than healthy controls, which could be interpreted as evidence for the deterioration of expressive language capabilities over the course of AD.

Using free speech from autobiographic interviews — a more liberal scenario for natural language — Weiner et al. used perplexity scores to automatically discriminate between general dementia patients and healthy controls (Weiner et al., 2017). Multiple-hour interviews (98 subjects, 230 hours) were cleaned of experimenter speech interventions and transcribed both manually and by an automatic speech recognition (ASR) system. Based on the raw audio signal and transcripts, the authors compared classification results using both automatically and manually generated feature sets divided into acoustic features, linguistic features and ASR features. Perplexity scores were reported as ASR features, differentiated into within and between subject perplexity. The authors concluded that automatic classification is feasible and report within/between speaker perplexity as two of their best performing features.

Similarly to Wankerl et al.(2017), other researchers used manual transcripts from speech of the Cookie Theft Picture description task and language models built on healthy controls' speech to differentiate between different forms of FTLD (Pakhomov et al., 2010). Results show that perplexity scores discriminate well between different subforms of FTD: behavioural variant of the FTLD and semantic dementia. This is in line with the notion that the behavioural FTD variant manifests not primarily in corrupted language but semantic dementia does. The authors also correlated perplexity scores with results from common neuropsychological tests, such as SVF: the free speech task perplexity scores negatively correlate with the SVF task. This is perfectly in line with the semantic retrieval problems in semantic dementia, manifesting in a very low SVF word count (i.e., high perplexity due to corrupted free speech and low SVF score).

The underlying latent objective of free speech tasks is, by nature, to produce syntactically correct speech. Using a language model trained on healthy controls, perplexity measures how people are not able to produce such an output following the given objective. In the semantic verbal fluency task however, the inherent objective is to produce as many items as possible which necessarily requires to exploit deeper semantic stock. As the objective is also to not produce repetitions, to be successful one has to produce sequences of increasingly rare items to maintain a high production rate towards the end of the task; this follows as the common easy-to-access semantic items are typically produced at the beginning of the timed task.

There is broad evidence, proving that demented persons have significant difficulties in the SVF task which manifests not only in a lower SVF raw count, but also in inefficient semantic stock exploitation strategies. In other words, demented patients are, especially towards the second half of such a task, not able to produce rare/repetition-free sequences of correct item responses. This lack of strategic semantic memory exploitation can be observed through multiple computational approaches (Woods et al., 2016), allowing to automatically compute semantic exploitation measures which compare the patient's sequence of words to a global semantic representation inferred from large text corpora leveraging either graph theory (Clark et al., 2016) or neural word embeddings (Linz et al., 2017a).

Recent work on the qualitative computational analysis of the SVF in demented patients shows that features based on neural word embeddings discriminate well between healthy controls and dementia types. Especially semantic density—the lexical coverage of a patients semantic exploitation—and word frequency—the lexical rareness of a patients produced items—have been shown to be very predictive and highly significant features in this task (Linz et al., 2017b). In general, demented persons are less successful in the SVF task as they are less able to systematically exploit a large distributed semantic stock and produce sequences of relatively rare items.

Therefore the aim of this study was to explore the possibility of a SVF language model to detect inefficient SVF production strategies, thus dementia. This represents a novel approach, as to the authors' knowledge, perplexity has so far only been used to detect language corruption.

3. Background

Statistical Language Models are a common tool for representing the probability distribution of language data, in either written or spoken form. After computing these models, they can be used to determine the probability of a given sequence of words.

To train a model, a corpus is split into a list of n-grams, a sequence of words of length n, $N = (w_1 \dots w_n)$. The probability of the ngram, N, is determined using maximum

likelihood estimation (MLE):

$$P(N) = P(w_n|w_1...w_{n-1}) = \frac{P(w_1...w_n)}{P(w_1...w_{n-1})} \quad (1)$$

The model stores the counts of all the n-grams in the corpus, thus 'training' it. To evaluate the probability of getting a certain sequence of words of length m, $S = (w_1...w_m)$, from our model, based on the Markov assumption, we can multiply the probability of each ngram in the sequence.

$$P(S) = \prod_{i=1}^{m} P(w_i|w_1...w_{i-1}) \quad (2)$$

Unigram models are simple models where the probability of every type, or unique word, is equivalent to the relative frequency of the word in the training set. Because unigrams assume that every word does not depend on any of the previous words, they does not capture the relationships between words. This is why we continue with the bigram and trigram models, where conditional probabilities are used in training.

One challenge of language modelling is data sparsity as we will never encounter every possible combination of n-gram that can be generated during training. Data sparsity makes it likely that our model will encounter unseen n-grams during testing and assign them a probability of zero, causing $P(S) = 0$. To counter this, language models employ a technique known as smoothing, in which some of the probability mass of seen n-grams is shifted to unseen n-grams.

Lidstone smoothing (Lidstone, 1920) is an additive smoothing technique in which an 'unknown' token is added, as a placeholder, to our training set. Then, a predetermined α is added to every n-gram count. Any n-grams that appear in testing, and that were not seen in training, will be accounted for by the 'unknown' token. The counts of the n-grams are then normalized by adding the count of the n-gram's history, $C(w_1...w_{n-1})$, to the size of the vocabulary of the n-gram's history, V, multiplied by α. After smoothing, the probability of an n-gram is represented by:

$$P(w_n|w_1...w_{n-1}) = \frac{C(w_1...w_n) + \alpha}{C(w_1...w_{n-1}) + V\alpha} \quad (3)$$

After calculating the smoothed probability distribution of a training set, language models can be evaluated on a test sample using a measure called perplexity. Perplexity is a score that shows how well a trained model predicts a test sample by taking the probability of the test sample and normalizing it by the number of words in the test sample. Perplexity is computed by the following equation:

$$PPL(S) = \frac{1}{\sqrt[m]{\prod_{n=1}^{m} P(w_n|w_1...w_{n-1})}} \quad (4)$$

Perplexity and probability are inversely related, so when perplexity is minimized, probability is maximized. This means a low perplexity indicates that the model fits the test sample well.

	HC	MCI	ADRD
N	40	47	79
Age	72.65 (8.3)	76.59* (7.6)	79.0* (6.1)
Sex	8M/32F	23M/24F	39M/40F
Education	11.35 (3.7)	10.81 (3.6)	9.47* (4.5)
MMSE	28.27 (1.6)	26.02* (2.5)	18.81* (4.8)
CDR-SOB	0.47 (0.7)	1.68* (1.11)	7.5* (3.7)

Table 1: Demographic data and clinical scores by diagnostic group; mean (standard deviation); Significant difference ($p < 0.05$) from the control population in a Wilcoxon-Mann-Whitney test are marked with *; HC='Healthy control', MCI='Mild cognitive impairment', ADRD= 'Alzheimer's disease and related disorders'; MMSE='Mini-Mental-State-Examination'; CDR-SOB='Clinical Dementia Rating Scale - Sum of boxes'.

4. Methods

4.1. Data

The data used for the following experiments was collected during the *Dem@Care* (Karakostas et al., 2014) and ELEMENT (Tröger et al., 2017) projects. All participants were aged 65 or older and were recruited through the Memory Clinic located at the Institute Claude Pompidou in the Nice University Hospital. Speech recordings of elderly people were collected using an automated recording app on a tablet computer and were subsequently transcribed following the CHAT protocol (MacWhinney, 1991). Participants completed a battery of cognitive tests, including a 60 second animal SVF test. Furthermore, all participants completed the MMSE (Folstein et al., 1975) and CDR (Morris, 1997). Following the clinical assessment, participants were categorised into three groups: Control participants (HC) diagnosed healthy after assessment, patients with MCI and patients that were diagnosed as having Alzheimer's Disease or related disorders (ADRD). AD diagnosis was determined using the NINCDS-ADRDA criteria (McKhann et al., 2011). Mixed/Vascular dementia was diagnosed according to ICD 10 (World Health Organization, 1992) criteria. For the MCI group, diagnosis was conducted according to Petersen criteria (Petersen et al., 1999). Participants were excluded if they had any major auditory or language problems, history of head trauma, loss of consciousness, or psychotic or aberrant motor behaviour. Demographic data and clinical test results by diagnostic groups are reported in Table 1.

4.2. Language Modelling

Based on our three patient populations (HC, MCI, ADRD), we construct three LMs: (1) trained only on the healthy population, (2) trained only on the impaired population (MCI + ADRD) and (3) trained on all patient data, regardless of diagnosis.

For each training set we build unigram, bigram and trigram models. We stop at trigrams, since given our vocabulary (n=238) the possible number of trigrams is 13,481,272 and our corpus only contains 2,203 trigram tokens, leading to extreme sparsity. We apply Lidstone smoothing to the

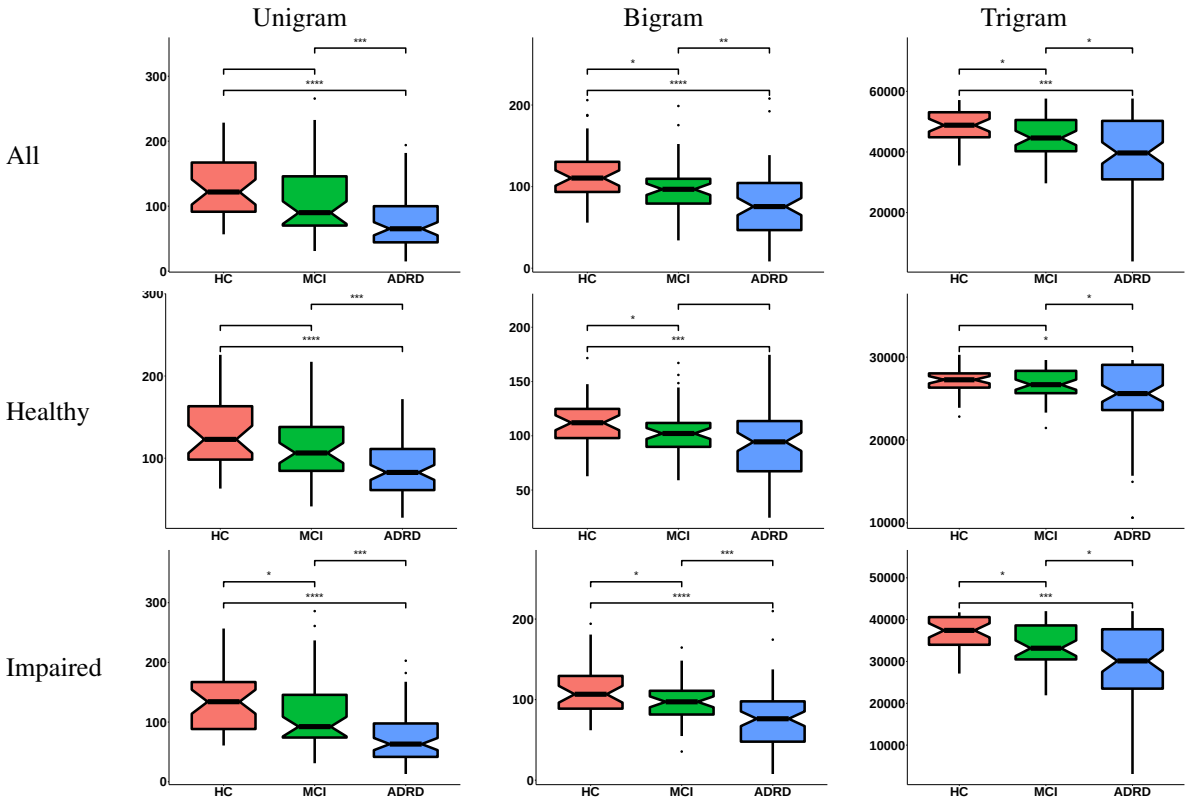


Figure 2: Boxplots of perplexity in relation to diagnostic criteria for all three sets of language models. The HC group is depicted in red, the MCI group in green and the ADRD group in blue. Horizontal brackets indicate group comparisons by a Wilcoxon-Mann-Whitney test ($^{*}: p \leq 0.05$, $^{**}: p \leq 0.01$, $^{***}: p \leq 0.001$, $^{****}: p \leq 0.0001$).

model with $\alpha = 1$.

Due to the nature of our training samples, lists of animals, and leave one out method of cross validation, we have a small vocabulary and do not expect a high amount of unseen tokens in the testing sequence, compared to natural language, making this a justifiable method of smoothing on this data set.

Perplexity is calculated as described in Equation 4. For models (1) and (2) we discriminate between the training population and the rest. Let $A_t = a_1, ..., a_m$ be the training population and $A_r = a_{m+1}, ..., a_n$ the rest of the samples. Then we perform leave-one-out cross validation on A_t, generating one perplexity value for the held-out sample a_i and each sample in A_r, per iteration. In the end, every sample in A_t has one perplexity value and every sample in A_r has m perplexity values. Averaging the m values per sample, leaves us with one perplexity value per sample. For (3) we perform a simple leave-one-out cross validation on the complete set $a_1, ..., a_n$, yielding one perplexity value per patient.

4.3. Classification

To confirm the diagnostic power of perplexity, we perform a simple classification experiment. Each person in the database was assigned a label relating to their diagnosis (HC, MCI and ADRD). Perplexity values from different models were used as input to classification models. All features were normalised using z-standardisation.

In all scenarios we use Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) implemented in the scikit-learn framework (Pedregosa et al., 2011). We use a radial bases kernel, since there is only one feature (Hsu et al., 2010) and 10-fold cross validation was used for testing. To find a well-performing set of hyperparameters, parameter selection using cross-validation on the training set of the inner loop of each cross validation iteration was performed. Performing cross validation on small data sets only once leads to performance fluctuations between different iterations. To work around this problem, cross validation was performed multiple times and then the mean of all performance metrics was calculated.

5. Results

Figure 2 displays boxplots of perplexity values by diagnostic groups. Each column corresponds to either uni-, bi- or trigram models. Rows indicate the training scenario. In general the perplexity decreases with disease progression - from HC, to MCI, to ADRD.

People with ADRD have significantly smaller perplexity values compared to the HC population, regardless of the context history length considered and training material. The same is true for people with ADRD in comparison to the MCI population. A significant difference between the HC and the MCI population for unigrams is only visible in the 'Impaired' model, (3). Bigram models all show significant differences between both populations. Trigrams only show this effect for models trained on the whole population or the impaired part. Overall, trigrams show less differences between populations and high perplexity values,

Scenario	Model	F_1
HC vs. MCI	U_{all}	0.62
	B_{all}	**0.71**
	T_{all}	0.67
HC vs. ADRD	U_{all}	**0.83**
	B_{all}	0.81
	T_{all}	0.72
MCI vs. ADRD	U_{all}	0.75
	B_{all}	**0.76**
	T_{all}	0.69

Table 2: Classification results for different scenarios and models as F_1 scores. U_{all} = Unigram model trained on all samples; B_{all} = Bigram model trained on all samples; T_{all} = Trigram model trained on all samples.

which can be attributed to the extreme sparseness of these models given our small data set.

Table 2 shows classification results for different models and scenarios. Following inspection of Figure 2, only models trained on all samples in the population were used in classification experiments, as the inter-group effects seem consistent between different training material. Between the HC and the ADRD group, as well as the MCI and ADRD populations, the unigram and bigram model show comparable performance. For classification of the HC and the MCI population the bigram model clearly shows the best performance.

6. Discussion and Conclusion

A general result of this study is that people with MCI or dementia show significantly lower perplexity values in SVF compared to a healthy population, meaning the n-gram LMs, regardless of training corpus, are more suited to model a demented person's speech versus that of a healthy person. Thus people with dementia are more predictable in their production of words in the SVF task.

This differs from findings about perplexity of demented patients in free speech tasks, where perplexity values of demented speech have been shown to be higher than that of healthy controls (Wankerl et al., 2017). This can be explained by the different scenarios where language modelling is applied: on natural language, a LM and its resulting perplexity can be interpreted as a measure for syntactic normality/correctness. When training on and predicting SVF performances, in which production of word sequences is motivated semantically, the perplexity can be viewed as a measure for effective semantic retrieval strategy.

Furthermore, we found word production in SVF differed in advancing stages of dementia syndromes. Unigram perplexity approximated on the SVF task, can be seen as a measure of predictability of word choice. Perplexity values of unigram models were found to be good indicators to separate the ADRD group from the HC group, but not the MCI population from the HC. Thus, word choice in SVF is more predictable in late stage dementia and not in early stage. Perplexity of bigram models trained on SVF productions—and for that matter any ngram where $n \geq 2$—can be seen as a measure for predictability of production strategy in the task. Both ADRD and MCI groups show significant differences in perplexity of bigram models to the HC group. Consequently, both populations show more predictable production strategies.

When modelling with trigrams, we would expect to see effects of context length—such as people with dementia using less contextual information. Unfortunately, this study is limited in the conclusions that can be drawn about the trigram models as it lack sufficient amounts of SVF data and therefore those models are severly undertrained.

In future experiments, we would like to gather more data to generate well-trained trigram models and possibly draw a more definitive conclusion on the effects of context length in SVF. We would also like to try different smoothing techniques, possibly interpolated methods such as Witten-Bell, that are not as coarse as the Lidstone technique.

Based on the trends shown in the unigram and bigram models, demented patients show significantly lower perplexity values, regardless of training data, and are therefore more predictable. Furthermore, persons in advanced stages of dementia differ in predictability of word choice — as shown by the unigram models — and production strategy — as shown by the bigram models — where as people with mild cognitive impairment only show significant predictability in their production strategy.

Perplexities from both the unigram and bigram models also function as adequate diagnositic features in classification tasks where the unigram model differentiates the best between HC and ADRD and the bigram model differentiates best between the more fine-grained distinctions of MCI versus the healthy controls or more severly demented patients.

7. Acknowledgements

This work was partially funded by the EIT Digital Wellbeing Activity 17074, *ELEMENT*. The data was collected during the EU FP7 *Dem@Care* project, grant agreement 288199.

8. Bibliographical References

Auriacombe, S., Lechevallier, N., Amieva, H., Harston, S., Raoux, N., and Dartigues, J.-F. (2006). A Longitudinal Study of Quantitative and Qualitative Features of Category Verbal Fluency in Incident Alzheimer's Disease Subjects: Results from the PAQUID Study. *Dementia and geriatric cognitive disorders*, 21(4):260–266.

Bonner, M. F., Ash, S., and Grossman, M. (2010). The New Classification of Primary Progressive Aphasia into Semantic, Logopenic, or Nonfluent/Agrammatic Variants. *Current Neurology and Neuroscience Reports*, 10(6):484–490.

Clark, D. G., McLaughlin, P. M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., Eastman, J., Dukes, R. M., Kapur, P., DeRamus, T. P., and Apostolova, L. G. (2016). Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimers Dement (Amst)*, 2:113–122.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-Mental State". A Practical Method for Grading the

Cognitive State of Patients for the Clinician. *J Psychiatr Res*, 12(3):189–198.

Gomez, R. G. and White, D. A. (2006). Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology*, 21(8):771 – 775.

Gruenewald, P. J. and Lockhead, G. R. (1980). The Free Recall of Category Examples. *Journal of Experimental Psychology: Human Learning and Memory*, 6:225–240.

Henry, J. D. and Crawford, J. R. (2004). Verbal fluency deficits in parkinson's disease: A meta-analysis. *Journal of the International Neuropsychological Society*, 10(4):608–622.

Henry, J. D., Crawford, J. R., and Phillips, L. H. (2004). Verbal fluency performance in dementia of the alzheimer's type: a meta-analysis. *Neuropsychologia*, 42(9):1212–1222.

Henry, J. D., Crawford, J. R., and Phillips, L. H. (2005). A meta-analytic review of verbal fluency deficits in huntington's disease. *Neuropsychology*, 19(2):243–252.

Hsu, C.-W., Chang, C.-C., and jen Lin, C. (2010). A Practical Guide to Support Vector Classification.

Karakostas, A., Briassouli, A., Avgerinakis, K., Kompatsiaris, I., and Tsolaki, M. (2014). The Dem@Care Experiments and Datasets: a Technical Report. Technical report, Centre for Research and Technology Hellas (CERTH).

Lidstone, G. J. (1920). Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192.

Linz, N., Tröger, J., Alexandersson, J., and König, A. (2017a). Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

Linz, N., Tröger, J., Alexandersson, J., Wolters, M., König, A., and Robert, P. (2017b). Predicting Dementia Screening and Staging Scores From Semantic Verbal Fluency Performance. In *IEEE International Conference on Data Mining (ICDM)-Workshop on Data Mining for Aging, Rehabilitation and Independent Assisted Living*, pages 719–728.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum Associates, Inc.

Marczinski, C. A. and Kertesz, A. (2006). Category and letter fluency in semantic dementia, primary progressive aphasia, and alzheimer's disease. *Brain and Language*, 97(3):258 – 265.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., et al. (2011). The diagnosis of dementia due to Alzheimer'??s disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, 7(3):263–269.

Morris, J. C. (1997). Clinical Dementia Rating: A Reliable and Valid Diagnostic and Staging Measure for Dementia of the Alzheimer Type. *International Psychogeriatrics*, 9(S1):173–176.

Pakhomov, S. V., Smith, G. E., Marino, S., Birnbaum, A., Graff-Radford, N., Caselli, R., Boeve, B., and Knopman, D. S. (2010). A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of Neurolinguistics*, 23(2):127–144.

Pakhomov, S. V., Eberly, L., and Knopman, D. (2016). Characterizing Cognitive Performance in a Large Longitudinal study of Aging with Computerized Semantic Indices of Verbal Fluency. *Neuropsychologia*, 89:42–56.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild Cognitive Impairment: Clinical Characterization and Outcome. *Arch. Neurol.*, 56(3):303–308.

Raoux, N., Amieva, H., Goff, M. L., Auriacombe, S., Carcaillon, L., Letenneur, L., and Dartigues, J.-F. (2008). Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study. *Cortex*, 44(9):1188–1196.

Robert, P. H., Lafont, V., Medecin, I., Berthet, L., Thauby, S., Baudu, C., and Darcourt, G. (1998). Clustering and switching strategies in verbal fluency tasks: Comparison between schizophrenics and healthy adults. *Journal of the International Neuropsychological Society*, 4(6):539–546.

Tröger, J., Linz, N., Alexandersson, J., König, A., and Robert, P. (2017). Automated Speech-based Screening for Alzheimer's Disease in a Care Service Scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*.

Troyer, A. K., Moscovitch, M., and Winocur, G. (1997). Clustering and Switching as Two Components of Verbal Fluency: Evidence From Younger and Older Healthy Adults. *Neuropsychology*, 11(1):138–146.

Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M. P., and Stuss, D. (1998). Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia*, 36(6):499 – 504.

Wankerl, S., Nöth, E., and Evart, S. (2017). An N-Gram Based Approach to the Automatic Diagnosis of Alzheimer's Disease from Spoken Language. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. in press.

Weiner, J., Engelbart, M., and Schultz, T. (2017). Manual and Automatic Transcriptions in Dementia Detection from Speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 3117–3121.

Woods, D. L., Wyma, J. M., Herron, T. J., and Yund, E. W. (2016). Computerized Analysis of Verbal Fluency: Nor-

mative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PLOS ONE*, 11(12):1–37.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization.

A Method for Analysis of Patient Speech in Dialogue for Dementia Detection

Saturnino Luz, Sofia de la Fuente, Pierre Albert
Usher Institute of Population Health Sciences & Informatics
Edinburgh Medical School
The University of Edinburgh, Scotland, UK
{s.luz,sofia.delafuente,pierre.albert}@ed.ac.uk

Abstract

We present an approach to automatic detection of Alzheimer's type dementia based on characteristics of spontaneous spoken language dialogue consisting of interviews recorded in natural settings. The proposed method employs additive logistic regression (a machine learning boosting method) on content-free features extracted from dialogical interaction to build a predictive model. The model training data consisted of 21 dialogues between patients with Alzheimer's and interviewers, and 17 dialogues between patients with other health conditions and interviewers. Features analysed included speech rate, turn-taking patterns and other speech parameters. Despite relying solely on content-free features, our method obtains overall accuracy of 86.5%, a result comparable to those of state-of-the-art methods that employ more complex lexical, syntactic and semantic features. While further investigation is needed, the fact that we were able to obtain promising results using only features that can be easily extracted from spontaneous dialogues suggests the possibility of designing non-invasive and low-cost mental health monitoring tools for use at scale.

Keywords: Dementia diagnosis and prediction, Alzheimer's disease, dialogue analysis, speech features, vocalisation graphs, content-free analysis.

1. Introduction

Research into early detection of Alzheimer's disease (AD) has intensified in the last few years, driven by the realisation that in order to implement effective measures for secondary prevention of Alzheimer's type dementia (ATD) it may be necessary to detect AD pathology decades before a clinical diagnosis of dementia is made (Ritchie et al., 2017). While imaging (PET, MRI scans) and cerebrospinal fluid analysis provides accurate diagnostic methods, there is an acknowledged need for alternative, less invasive and more cost-effective tools for AD screening and diagnostics (Laske et al., 2015). A number of neuropsychological tests have been developed which can identify signs of AD with varying levels of accuracy (Mortamais et al., 2017; Ritchie et al., 2017). However, the proliferation of technologies that enable personal health monitoring in daily life points towards the possibility of developing tools to predict AD based on processing of behavioural signals.

Speech is relatively easy to elicit and has proven to be a valuable source of clinical information. It is closely related to cognitive status, having been used as the primary input in a number of applications to mental health assessment. It is also ubiquitous and can be seamlessly acquired. In recent years, combinations of signal processing, machine learning, and natural language processing have been proposed for the diagnosis of AD based on the patient's speech and language (Fraser et al., 2016). Models built on phonetic, lexical and syntactic features have borne out the observation that these linguistic processes are increasingly affected as the disease progresses (Kirshner, 2012). However, most machine learning research in this area has employed either recorded narrative speech (Lopez-De-Ipiña et al., 2012), or recorded scene descriptions (Luz, 2017; Fraser et al., 2016) collected as part of a neuropsychological assessment test, such as the Boston "cookie theft" picture description task (Becker et al., 1994).

In contrast to those methods, our approach employs spontaneous conversational data, exploring patterns of dialogue as basic input features. Content-free interaction patterns of this kind were first used in the characterisation of psychopathology by Jaffe and Feldstein (1970), who represented therapist-patient dialogues as Markov chains. Here, we build on these ideas to analyse patient data from the Carolina Conversations Collections (CCC) (Pope and Davis, 2011). We trained machine learning models on these data to differentiate AD and non-AD speech. This work is, to the best of our knowledge, the first to employ low-level dialogue interaction data (as opposed to lexical features, or data from narrations other forms of monologue) as a basis for AD detection on spontaneous speech.

2. Background

One of the greatest challenges facing developed countries, and increasingly the developing world, is the challenge of improving the quality of life of older people. In 2015, the First Ministerial Conference of the WHO on Global Action Against Dementia estimated that there are 47.5 million cases of this condition in the world. Cohort studies show between 10 and 15 new cases per each thousand people every year for dementia, and between 5 and 8 for Alzheimer's Disease. Prognosis is usually poor, with an average life expectancy of 7 years from diagnosis. Less than 3% diagnosed live longer than 14 years. Current statistics predict that the population aged over 65 is expected to triple between years 2000 and 2050 (World Health Organization and others, 2015). This will lead to structural and societal changes, accentuating what is already becoming a highly demanding issue for health care systems.

Dementia is therefore set to become a very common cause of disability which places a heavy burden on carers and patients alike. While there are currently neither a cure nor a way to entirely prevent the progress of the disease, it is

hoped that a better understanding of language and communication patterns will contribute to secondary prevention. A characterisation of communication patterns and their relation to cognitive functioning and decline could be useful in the design of assistive technologies such as adaptive interfaces and social robotics (Wada et al., 2008). These technologies might help provide respite to carers, and stimulate cognitive, physical and social activity, which can slow disease progression and improve the patient's quality of life (Middleton and Yaffe, 2009). Collecting relevant real life observational data and assembly of prior and current knowledge (Wada et al., 2008) could lead to new effective and personalised interventions.

Assessing people's behaviour in natural settings might also contribute to earlier detection (Parsey and Schmitter-Edgecombe, 2013; Mortamais et al., 2017). Language impairment is a common feature of dementia, implying signs such as word-finding and understanding difficulties, blurred speech or disrupted coherence (American Psychiatric Association, 2000). Although language is a good source of clinical information regarding cognitive status, manual analysis of language by mental health professionals for diagnostic purposes is challenging and time-consuming. Advances in speech and language technology could help by providing tools for detecting reliable differences between patients with dementia and controls (Bucks et al., 2000), distinguishing among dementia stages (Thomas et al., 2005) and differentiating various types of dementia (Fraser et al., 2016).

Features such as grammatical constituents, vocabulary richness, syntactic complexity, psycholinguistics, information content, repetitiveness, acoustics, speech coherence and prosody, have been explored in conjunction with machine learning methods to identify Alzheimer's and other types of dementia through the patient's speech. This is not only because language is impaired in these patients, but also because language relies on other cognitive functions, such as executive functions, which allow us to interact in a sound and meaningful way. These functions are responsible for decision making, strategy planning, foreseeing consequences and problem solving, which are essential to successful communication, but are impaired by AD (Fraser et al., 2016; Marklund et al., 2009; Satt et al., 2013). Although hardly perceptible to the speakers themselves, patterns of impairment are thought to occur even in informal and spontaneous conversations (Bucks et al., 2000; Cohen and Elvevåg, 2014).

Our hypothesis in this paper is that people with an AD diagnosis will show identifiable patterns during dialogue interactions. These patterns include disrupted turn taking and differences in speech rate. These indices relate to the fact that, in general, patients with AD show poorer conversation abilities and their normal turn-taking is repeatedly interrupted. Therefore, we expect less conversational fluidity overall in the AD group dialogues, as compared to non-AD group. Our approach, which does not rely on transcription but only on speech-silence patterns and basic prosodic information, obtains levels of accuracy comparable to state-of-the-art systems that rely on more complex feature sets.

3. Related work

Potential applications of the kind of speech technology described in this paper include the development of interactive assistive technologies, and monitoring of users for signs of cognitive decline with a view to mitigating further decline. From the perspective of potential applications of automatic speech analysis to technology-assisted care, there is evidence (Rudzicz et al., 2014b) that it is psychologically more acceptable for a user to be aided by another person or a robot than from ambient sensors and devices which are unable to offer meaningful interaction. Therefore, the development of such assistive applications involves research on speech processing for natural conversations rather than scripted speech or monologues (Conway and O'Connor, 2016).

From the perspective of monitoring for early detection, it is known that AD leads to disruption of one's ability to follow dialogues, even in simple, routine interactions. At later stages of the disease, failure to perform meaningful interactions appears (Watson, 1999). This has a negative impact on tasks such as following instructions regarding household activities and medication, as well as preventing rewarding social interactions. Here, once again the focus should be on natural interaction data, as scripted talk cannot be compared to spontaneous conversation in terms of information richness and external validity of results (Kato et al., 2013). Over the last decades, different approaches have targeted early detection of AD on spontaneously generated data through automatic and non-invasive intelligent methods. Some of these approaches have focused on speech parameters analysis: automatic spontaneous speech analysis (ASSA), emotional temperature (ET), (Lopez-De-Ipiña et al., 2012), voiceless segments, and phonological fluency have been shown to explain significant variance in neuropsychological test results (García Meilán et al., 2012). These methods are not only non-invasive and free from side-effects, but also relatively cheap in time and in terms of resources. Another approach that rely on easily extracted acoustic features, such as the ones we propose in this paper, though not in dialogical or spontaneous speech settings is presented by Satt et Al. (2013). This approach extracts a number of voice features (voiced segments, average utterance duration, etc.) from recordings of picture description, sentence repetition, and repeated pronunciation of three syllables used in diadochokinetic tests in succession. The method achieves accuracy levels of over 80% in detection of AD and mild cognitive impairment (MCI).

Other approaches have used time-aligned transcripts and syntactic parsing, extracting speech features and using them for classifying healthy elderly subjects from subjects suffering AD or MCI, as well as other tasks. This classification has been done either by comparing impaired to healthy speech performance (speech quality in terms of lexicon, coherence, etc.), or by comparing classifier performance when only neuropsychological tests are included against performance when such tests are used together with speech features, generally with statistically significant improvements (Roark et al., 2011; Fraser et al., 2016).

Analysis performed on similar corpora provide good insight of the performances achieved using different features.

A first analysis (Fraser et al., 2016), based on a monologue corpus (DementiaBank), identified four different linguistic factors as main descriptors: syntactic, semantic, and information impairments, and acoustic abnormality. They achieved accuracy of up to 92.05% using full scale analysis of 25 features, selected amongst an original feature set of 370 features after extensive experimentation.

An analysis of the CCC corpus by Guinn et al (Guinn and Habash, 2012) used similar linguistic features. Unlike the work presented in this paper, Guinn's analysis was focused on the differences between interviewers and subjects in the subset of patients with AD. They achieved a combined accuracy of 75-79.5 % using decision trees, with a large discrepancy between AD (38-42 %) and non-AD (74-100 %) recognition accuracy.

Works on dialogue so far have identified features such as conversational confusion (AD increases confusion rates, and this relates to slower and shorter speech; (Rudzicz et al., 2014a), prosodic measures (Gonzalez-Moreira et al., 2015), and emotion (Devillers et al., 2005). These studies used machine learning methods (neural networks, Naïve Bayes, and random forests, respectively), reporting accuracy in the 70-90 % range. Although these results are promising, they are difficult to generalise. This is because they are primarily content dependent. That is, they employ lexical, and sometimes syntactic information, which present a number of potential disadvantages. The content of a conversation is likely to change greatly depending on whether a participant belongs to the control group or to the group with Alzheimer's Disease, especially if the conversational partner is their doctor. In addition, such content is difficult to acquire in spontaneous speech settings. Despite the advances in automatic speech recognition, recognition (word) error rates in unconstrained settings are still over 11%, even for fairly clear, telephone dialogues (Xiong et al., 2016). Another difficulty with these approaches is the fact that they are language-dependent, and therefore require building different models for different languages, which in the context of global mental health could be a major shortcoming. Therefore, these models should aim to be as content-independent as possible to be generalisable (Satt et al., 2013). In contrast to content-based approaches, our method focuses on the interaction patterns themselves, rather than on characteristics of the speech and language content as such.

4. Methods

4.1. Dataset

We have conducted our analysis using the Carolina Conversations Collection (Pope and Davis, 2011). The dataset is a digital collection of recordings of conversations about health, including both audio and video data, with corresponding transcriptions. The corpus consists of natural conversations involving an older person (over the age of 65) with a medical condition. Several demographic and clinical variables are also available, including: age range, gender, occupation prior to retirement, disease diagnosed, and level of education (in years). The interviewers were gerontology and linguistic students or researchers to whom the patients spoke at least twice a year. A unique alias was assigned to each patient to protect their identity.

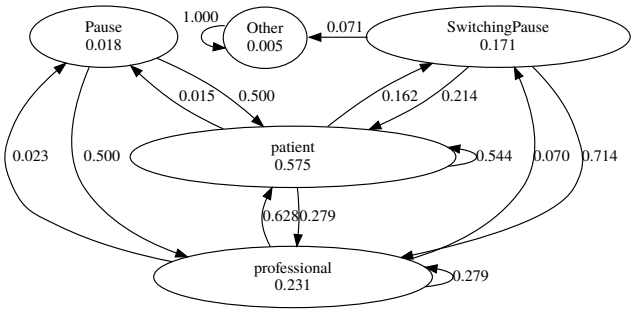


Figure 1: Vocalisation diagram for a patient dialogue.

Access to the data was provided after complying with the ethical requirements of the University of Edinburgh and the Medical University of South Carolina. In order to ensure that the results described here are reproducible we will provide, on request, the identifiers for the dialogues used in our experiments so that interested researchers can recreate our dataset upon being granted access to the CCC. The source code used for processing the data is available at a University of Edinburgh gitlab server[1].

For the research described here we selected a total of 38 patient dialogues: 21 patients had a diagnosis of Alzheimer's disease (15 females, 6 males), and 17 patients (12 females, 5 males) had other diseases (diabetes, cardiac issues, etc., excluding neuropsychological conditions), but not AD. These groups were selected for matching age ranges and gender frequencies so as to avoid statistical bias. The dataset also included time-aligned transcripts, which we did not use except for the computation of an alternative speech rate feature as described below.

4.2. Data Preparation

The speech data selected as previously described were preprocessed in order to generate *vocalisation graphs* — that is, Markov diagrams encoding the first-order conditional transition probabilities between vocalisation events and steady-state probabilities (Luz, 2013).Vocalisation events are classified as speech by either the patient or the interviewer/others, joint talk (overlapping speech), or silence events (also known as 'floor' events, which are further in the diagrams as pauses and switching pauses, according to whether the floor is taken by the same speaker or another speaker, respectively). An example of vocalisation graph is shown in Figure 1.

Vocalisation and pause patterns have been successfully employed in the analysis of dialogues in a mental-health context (Jaffe and Feldstein, 1970), segmentation (Luz and Su, 2010) and classification of dialogues, and more recently on characterisation of participant role and performance in collaborative tasks (Luz, 2013). Models that employ basic turn-taking statistics have also been proposed for dementia diagnosis (Mirheidari et al., 2016), though not in a systematic content-fee framework as in our proposed method.

The distributions of event counts according to vocalisation

[1] https://cybermat.tardis.ed.ac.uk/pial/CCCdataset

events is shown in Figure 2. It can be observed that patients with AD tend to produce more vocalisation events than their interviewers (and, consequently, produce more silence events). This is consistent with findings in the literature on language changes in AD (American Psychiatric Association, 2000).

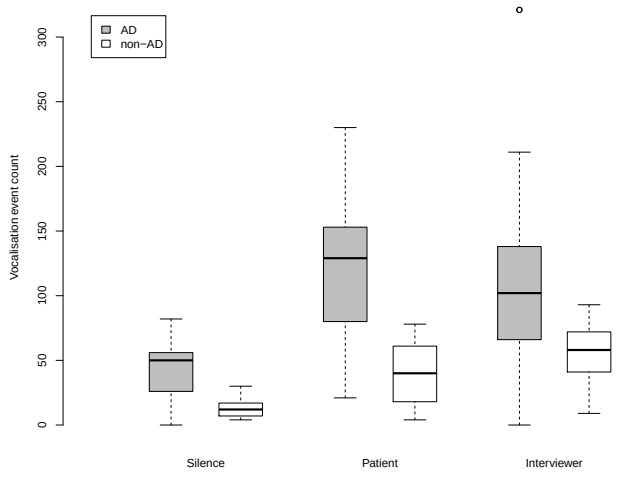


Figure 2: Distribution of vocalisation event counts for patients with and without AD in CCC dialogues.

Speech rate was estimated using De Jong's syllable nuclei detection algorithm (Jong and Wempe, 2009), which is an unsupervised method – that is, it can be applied directly to the acoustic signal, with no need of human annotation. However, as the audio quality of the CCC recordings is uneven, and as the dataset provides no gold standard against which one could assess syllable count, we decided to validate the use of De Jong's method against the time-stamped transcripts provided. Using these transcripts one could, in principle, estimate average words per minute (WPM) for individual utterances, as is sometimes done (Hayakawa et al., 2017). However, this method of measuring WPM based on transcription has a number of limitations. Words have variable length, and their articulation can vary greatly due to a number of speech-related phenomena, such as phonological stress, frequency, contextual predictability, and repetition (Bell et al., 2009). In order to mitigate these problems, we instead produced *speech rate ratio* estimates normalised through a speech synthesizer, employing the methods proposed by Hayakawa et al. Hayakawa et al. (2017). These estimates represent deviations from a "normalised" pace of 160 words per minute (WPM) synthesised using the MaryTTS system (Schröder and Trouvain, 2003). We therefore computed the ratio of the synthesised speech to the actual duration of the patient's speech. The speech rate ratio correlated well with the syllable per minute rate extracted using only the recorded audio ($\rho = 0.502$, $t(30) = 3.19, p = 0.003$) indicating that speech rate can be estimated with an acceptable level of reliability through the unsupervised method, even in fairly noisy settings.

A Python script was employed to extract basic speaker turn time stamps, speaker role information, and transcriptions from the original XML-encoded CCC data. The resulting data were then processed using the R language in order to detect silence intervals, and categorise turn transitions and pause events.

Table 1: Descriptive statistics on dialogue turn-taking (duration given in seconds).

Feature	non-AD	AD
Dialogue duration	4107.3	7628.4
Dialogue duration TTS	7618.8	7618.8
Avg turn duration	97.3	255.8
Total turn duration	1654.3	4348.3
Norm. total turn duration	3.0	4.1
Avg turn duration TTS	107.6	238.0
Total turn duration TTS	1829.7	4046.1
Norm. total turn duration TTS	3.0	4.2
Avg number of words	314.6	742.5
Total number of words	5348.0	12622.0
Avg words per minute	155.9	166.5

Some descriptive statistics on the dialogues can be seen in Table 1. These statistics include: average turn duration (how many seconds a participant speaks on average), total turn duration (how many seconds did the participant's turns lasted in total), normalised turn duration (the ratio of a participant's turn duration to the total duration of AD or non-AD dialogues, according the participant's class), number of words generated (total per class and on average per class' participant), and number of words per minute (average per class participant).

Contrary to our expectations, we did not observe a statistically significant difference between the speech rate in syllables per minute between patients with and without AD (Welch two sample t-test $t(30.5) = 1.15, p = 0.28$), even though the mean for non-AD ($M = 180.8$ syllables/min, $sd = 28.4$) was higher than that for patients with AD ($M = 168$ syllables/min, $sd = 35.6$).

Two alternative data representations were generated. The first (henceforth referred to as VGO) was based on the vocalisation graphs only. That is, VGO encodes the probabilities of each possible pair of transitions, including self-transitions, which tend to dominate Markov chains sampled, and the steady-state probabilities for each vocalisation event. The second form of representation (VGS) simply consists of the VGO with information about the participant's speech rate (mean and variance) added to the vocalisation statistics. With the exception of speech rate ratio, which necessitates transcription, all the information needed to build VGO and VGS instances can be extracted through straightforward signal processing methods.

4.3. Machine learning

The data instances in the two alternative representation schemes were annotated for presence or absence of Alzheimer's Disease (AD). A supervised learning procedure was employed in order to train classifiers to predict such annotations on unseen data.

We trained a boosting model (Schapire and Freund, 2014) using decision stumps (i.e. decision trees with a single split node) as weak learners. The training process consisted

of 10 iterations whereby, for each training instance (x_i), a weak classifier $\hat{f}_m$ was fitted using weights on the data which were iteratively computed so that the instances misclassified in the preceding step had their weights increased by a factor proportional to the weighted training error. In this case class probability estimates $P(ad = 1|data)$ were used to compute these weights and to weigh the final classification decision (additive logistic regression) following the Real Adaboost algorithm (Friedman et al., 2000):

$$\hat{F}(x) = sign\left[\sum_{m=1}^{M} \hat{f}_m(x)\right] \quad (1)$$

Classification performance was assessed through a 10-fold cross validation procedure. As the dataset is reasonably balanced, results were assessed in terms of accuracy, precision (the ratio of the number of true positives to the number of instances classified as AD), recall (or sensitivity, the ratio of true positives to the number of AD cases) and F_1 score (the harmonic mean of precision and recall). Micro (μ) and macro (M) averages for these scores are given by taking means over the entire set of classification decisions and over individual classifiers respectively, across the 10 folds. As the data set is fairly small, we also ran a leave-one-out cross validation (LOOCV) procedure to obtain better estimates of generalisation accuracy. This consisted of selecting one instance for testing, and building a classification model on the remaining instances, and iterating this procedure until all instances have been selected as testing instances. Macro averages are uninformative in LOOCV, so we only report overall accuracy figures for this procedure.

ROC curves showing the relationship between true positive and false positive rates as the classification threshold is varied were also plotted. Simulation was employed in order to smooth these ROC curves by running 10 rounds of 10-fold cross validation tests with a randomised selection of instances making up the hold-out sets.

5. Results

Our first approach, based on the VGO data representation scheme, produced promising results. Accuracy levels were well above the baseline, with overall accuracy reaching 81.1%, showing that turn taking patterns can provide useful cues to the detection of AD in dialogues. The results for the VGO-based classification are shown in Table 2. The corresponding ROC curve is shown in Figure 3.

Adding speech rate information (VGS representation) contributed to further enhancing AD detection, bringing the overall accuracy score to about 86.5%. Detailed evaluation metrics are shown in Table 3. The ROC curve for the VGS-based classification approach is shown in Figure 4. It can be seen that the addition of features for mean and variance of speech rate ratio over dialogues had the effect of improving classification trade-offs, particularly reducing the false positives while increasing the true positives at low threshold cut-offs.

For comparison we ran the same testing procedure using some of the other classifiers employed in the literature reviewed in section 3., namely, logistic regression,

Table 2: AD detection results for the VGO data representation scheme.

AD		non-AD	
Accuracy_μ	0.812	Accuracy_μ	0.714
Precision_μ	0.765	Precision_μ	0.769
Recall_μ	0.812	Recall_μ	0.714
$F_{1,\mu}$	0.788	$F_{1,\mu}$	0.741
Precision_M	0.667	Precision_M	0.792
Recall_M	0.722	Recall_M	0.729
$F_{1,M}$	0.685	$F_{1,M}$	0.721
Overall accuracy (LOOCV): 0.811			

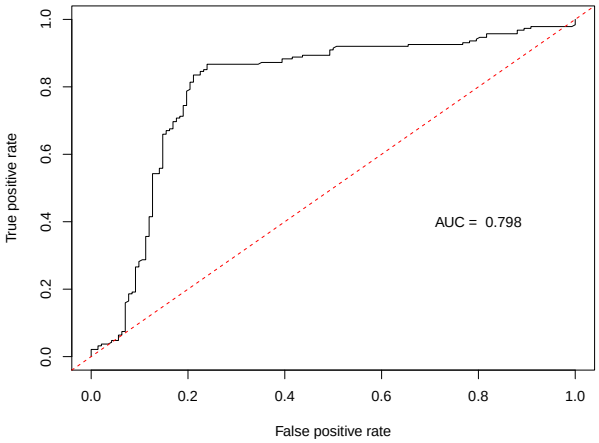


Figure 3: ROC curve for VGO-based classifiers.

naïve Bayes (Gaussian kernel), decision trees (C4.5 algorithm), SVM trained using sequential minimal optimisation, with a polynomial kernel (Platt, 1998), and random forests (Breiman, 2001), Weka implementation (Hall et al., 2009). The overall (LOOCV) accuracy figures are shown in Table 4. There is little difference in performance between our chosen method (Real Adaboost) and other methods used in the literature, except for logistic regression, which underperforms the machine learning methods. Real Adaboost slightly outperforms SVM and random forests classifiers, and matches C4.5 decision trees, with a slight advantage over the latter on the target AD class ($F_m = 0.878$ vs. $F_m = 0.872$).

Although there is considerable room for improvement upon this level of classification performance, the levels obtained with these simple models are comparable to the accuracy

Table 3: Results for the VGS data representation scheme.

AD		non-AD	
Accuracy_μ	0.882	Accuracy_μ	0.769
Precision_μ	0.833	Precision_μ	0.833
Recall_μ	0.882	Recall_μ	0.769
$F_{1,\mu}$	0.857	$F_{1,\mu}$	0.800
Precision_M	0.796	Precision_M	0.708
Recall_M	0.833	Recall_M	0.708
$F_{1,M}$	0.811	$F_{1,M}$	0.700
Overall accuracy (LOOCV): 0.865			

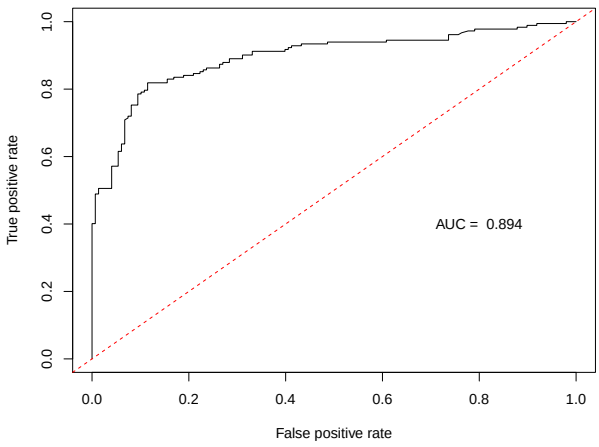


Figure 4: ROC curve for VGS-based classifiers.

Table 4: Compared accuracy results obtained with different classification algorithms, on VGS-based datasets.

Classification method	Accuracy (LOOCV)
Logistic regression	75.7%
Real Adaboost	86.5%
Decision trees	86.5%
SVM	83.7%
Random forests	81.1%

of approaches that employ more detailed linguistic information, which are presumably harder to acquire in everyday conversational situations, as they would involve a level of speech recognition accuracy which is beyond the capabilities of current systems for spontaneous speech in noisy environments.

6. Conclusion and Further Work

Dementia prevention and life quality in elderly care are important societal challenges. Automatic detection of signs of AD in speech can provide useful tools for the design of technologies for care-giving and cognitive health monitoring to help address these challenges.

This paper presented initial results of a new method to automatically recognise the first signs of disrupted communication using dialogue features. This method obtained an overall accuracy of 0.83, with a micro F-measure of 0.83 and a macro F-measure of 0.76 on the classification of patients as "AD" and "non-AD". Although it is difficult to compare these results directly to related works (Fraser et al., 2016; Guinn and Habash, 2012), our accuracy figures are situated within a similar range, 0.70-0.80, with a smaller discrepancy between the classification of the two groups, while relying on features that can be more robustly extracted from spontaneous speech.

Thanks to the increasingly important role of social technology, longitudinal studies may become richer in terms of the amount of variables measured, frequency of measurements and places where measures are taken (living settings), allowing for larger datasets. As more data are gathered in natural settings, we expect to obtain more reliable and generalisable results.

There are several linguistic parameters that are promising for the assessment of cognitive functioning. In current approaches, these features have been typically extracted from data collected through structured interviews, storytelling or picture descriptions. The work presented here contributes a new perspective to feature extraction by focusing on spontaneous dialogues. Dialogue processing provides a convenient framework for the analysis of natural conversations, in which readily available predictors, such as turn taking behaviour, have already yielded satisfactory results. We plan to further analyse verbal and non-verbal parameters to obtain a better characterisations of AD in order to infer neurosychological assessment results through speech and language processing, and subsequently to combine such features with actual neuropsychological evaluations and other relevant variables, building accurate models to achieve detection of AD at the time of onset.

The data set used in the present study has some limitations. Due to its constraints, the study was performed on a restricted subset of 21+17 sessions. In addition, the interview setting includes a degree of bias, as the interviewer's objective is to get the patient to perform a certain task (e.g. description of a picture, driving the discussion) therefore influencing the patient's speech. In order to mitigate these limitations, we plan to collect further data in more spontaneous dialogue in the near future.

7. Acknowledgements

Sofia De la Fuente and Pierre Albert are supported by the Medical Research Council (MRC). The authors would like to acknowledge Charlene Pope and Boyd H. Davis, from the Medical University of South Carolina, who host the Carolinas Conversation Collection, for providing access to the dataset and help in completing the required procedures.

8. Bibliographical References

American Psychiatric Association. (2000). Delirium, dementia, and amnestic and other cognitive disorders. In American Psychiatric Association, editor, *Diagnostic and Statistical Manual of Mental Disorders, Text Revision (DSM-IV-TR)*, chapter 2. Arlington, VA, 4^{th} edition.

Becker, J., Boiler, F., Lopez, O., Saxton, J., and McGonigle, K. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. 60(1):92–111.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bucks, R., Singh, S., Cuerden, J., and Wilcock, G. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Cohen, A. S. and Elvevåg, B. (2014). Automated Computerized Analysis of Speech in Psychiatric Disorders. *Current opinion in psychiatry*, 27(3):203–209.

Conway, M. and O'Connor, D. (2016). Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*, 9:77–82.

Devillers, L., Vidrascu, L., and Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, 49(2):407–422, October.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, April.

García Meilán, J. J., Martínez-Sánchez, F., Carro, J., Sánchez, J. a., and Pérez, E. (2012). Acoustic Markers Associated with Impairment in Language Processing in Alzheimer's Disease. *The Spanish Journal of Psychology*, 15(2):487–494.

Gonzalez-Moreira, E., Torres-Boza, D., Kairuz, H., Ferrer, C., Garcia-Zamora, M., Espinoza-Cuadros, F., and Hernandez-Gómez, L. (2015). Automatic prosodic analysis to identify mild dementia. *BioMed Research International*.

Guinn, C. I. and Habash, A. (2012). Language analysis of speakers with dementia of the alzheimer's type. In *AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*, pages 8–13.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hayakawa, A., Vogel, C., Luz, S., and Campbell, N. (2017). Speech rate comparison when talking to a system and talking to a human: A study from a speech-to-speech, machine translation mediated map task. In *Proc. Interspeech 2017*, pages 3286–3290.

Jaffe, J. and Feldstein, S. (1970). *Rhythms of dialogue*. Personality and Psychopathology. Academic Press, New York.

Jong, N. H. d. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, May.

Kato, S., Endo, H., Homma, A., Sakuma, T., and Watanabe, K. (2013). Early detection of cognitive impairment in the elderly based on Bayesian mining using speech prosody and cerebral blood flow activation. *Proceedings of the 37^{th} Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2013:5813–6.

Kirshner, H. S. (2012). Primary Progressive Aphasia and Alzheimer's Disease: Brief History, Recent Evidence. *Current Neurology and Neuroscience Reports*, 12(6):709–714.

Laske, C., Sohrabi, H. R., Frost, S. M., de Ipiña, K. L., Garrard, P., Buscema, M., Dauwels, J., Soekadar, S. R., Mueller, S., Linnemann, C., Bridenbaugh, S. A., Kanagasingam, Y., Martins, R. N., and O'Bryant, S. E. (2015). Innovative diagnostic tools for early detection of alzheimer's disease. *Alzheimer's & Dementia*, 11(5):561–578.

Lopez-De-Ipiña, K., Alonso, J., Solé-Casals, J., Barroso, N., Faundez, M., Ecay, M., Travieso, C., Ezeiza, A., and Estanga, A. (2012). Alzheimer disease diagnosis based on automatic spontaneous speech analysis. In *Proceedings of the 4th International Joint Conference on Computational Intelligence*, pages 698–705.

Luz, S. and Su, J. (2010). The relevance of timing, pauses and overlaps in dialogues: Detecting topic changes in scenario based meetings. In *Proceedings of INTERSPEECH 2010*, pages 1369–1372, Chiba, Japan. ISCA.

Luz, S. (2013). Automatic Identification of Experts and Performance Prediction in the Multimodal Math Data Corpus through Analysis of Speech Interaction. *Proceedings of the 15th ACM on International conference on multimodal interaction, ICMI'13*, pages 575–582.

Luz, S. (2017). Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data. In *Computer Based Medical Systems*, pages 45–46. IEEE Press.

Marklund, P., Sikström, S., Bååth, R., and Nilsson, L. G. (2009). Age effects on semantic coherence: Latent Semantic Analysis applied to letter fluency data. *3rd International Conference on Advances in Semantic Processing - SEMAPRO 2009*, pages 73–76.

Middleton, L. E. and Yaffe, K. (2009). Promising strategies for the prevention of dementia. *Arch Neurol*, 66(10):1210–1215.

Mirheidari, B., Blackburn, D., Reuber, M., Walker, T., and Christensen, H. (2016). Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of Interspeech 2016*, pages 1220–1224. ISCA.

Mortamais, M., Ash, J. A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., Pose, C., Albala, B., Ropacki, M., Ritchie, C. W., and Ritchie, K. (2017). Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimer's & Dementia*, 13(4):468–492.

Parsey, C. M. and Schmitter-Edgecombe, M. (2013). Applications of technology in neuropsychological assessment. *The Clinical neuropsychologist*, 27(8):1328–1361.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, et al., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Pope, C. and Davis, B. H. (2011). Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

Ritchie, K., Carrière, I., Su, L., O'Brien, J. T., Lovestone, S., Wells, K., and Ritchie, C. W. (2017). The midlife cognitive profiles of adults at high risk of late-onset alzheimer's disease: The PREVENT study. *Alzheimer's & Dementia*, 13(10):1089–1097.

Roark, B., Mithcell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *The New England journal of medicine*, 19(7):2081–2090.

Rudzicz, F., Chan Currie, L., Danks, A., Mehta, T., and Zhao, S. (2014a). Automatically Identifying Trouble-indicating Speech Behaviors in Alzheimer's Disease. In *16th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 241–242.

Rudzicz, F., Wang, R., Begum, M., and Mihailidis, A. (2014b). Speech recognition in Alzheimer's disease with personal assistive robots. *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies*, pages 20–28.

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., and Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August):1692–1696.

Schapire, R. E. and Freund, Y. (2014). *Boosting: Foundations and Algorithms*. The MIT Press, January.

Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *IEEE International Conference Mechatronics and Automation, 2005*, 3(February):1569–1574.

Wada, K., Shibata, T., Musha, T., and Kimura, S. (2008). Robot Therapy for Elders affected by Dementia. (August).

Watson, C. M. (1999). An analysis of trouble and repair in the natural conversations of people with dementia of the alzheimer's type. *Aphasiology*, 13(3):195–218.

World Health Organization et al. (2015). First who ministerial conference on global action against dementia: meeting report, who headquarters, geneva, switzerland, 16-17 march 2015.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). Achieving human parity in conversational speech recognition. *CoRR*, abs/1610.05256.

Detecting Dementia from Repetition in Conversational Data of Regular Monitoring Service

Kaoru Shinkawa[1], **Keita Shimmei**[1,2], **Yasunori Yamada**[1]
[1]IBM Research - Tokyo, Tokyo, Japan
19-21 Nihonbashi Hakozaki-cho Chuo-ku, Tokyo, Japan
[2]Department of Preventive Medicine and Public Health, Graduate School of Medicine, Keio University
35 Shinanomachi, Shinjuku-ku, Tokyo, Japan
kaoruma@jp.ibm.com, shinmeikeita@keio.jp, ysnr@jp.ibm.com

Abstract

Language dysfunctions are recognized as prominent signs of dementia, and previous computational studies have shown that measuring such dysfunctions can serve as a sensitive index of cognitive decline. These features of measuring language dysfunctions have been investigated in conversational data collected during neuropsychological tests but not in data collected during daily conversations. In this study, we used data obtained from a daily monitoring service for eight elderly people, including two who had been reported as having dementia, and investigated the features that characterize repetition in conversations on different days as well as single conversations on the same day. Through the analyses, we found that features for measuring repetition significantly increase for dementia patients in terms of topic and words. The results suggest that using the repetition features over the regular conversational data is a promising approach for detecting dementia sufferers.

Keywords: Monitoring Service, Linguistic Dysfunctions, Daily Conversation, Topic Similarity, Vocabulary Richness

1. Introduction

As the world's elderly population increases, the number of people living with dementia is rising rapidly, making dementia an increasingly serious health and social problem (Prince et al., 2013). Globally, around 47 million people were living with dementia in 2015, corresponding to about 7.6% of the world's over-65-year-olds (Prince et al., 2013). Although dementia is the fifth-biggest cause of death in high-income countries, it incurs the highest annual global cost to manage (estimated to be as high as US$818 billion) because patients require constant and costly care for years (Dolgin, 2016; Prince et al., 2015). Japan is one high-income country facing a severe aging problem. The prevalence of dementia for persons 65 years or older is estimated at around 15%, and the annual cost spent on care for dementia patients was around US$120 billion (14.5 trillion JPY) in 2014 (Shikimoto et al., 2016).

Technological innovations in monitoring services for older adults as well as dementia care are expected to help people with dementia and their carers. These include diagnostic, monitoring, assistive, therapeutic, and carer supporting technologies (Livingston et al., 2017). In particular, interest is growing in technologies for early diagnosis as a possible way of improving dementia care because of recent failures in both clinical trials and laboratory work (Sperling et al., 2011). However, many people with dementia remain undiagnosed, and diagnostic coverage worldwide remains low (Prince et al., 2016). Even in high-income countries, only 40-50% of dementia sufferers have received a diagnosis (Prince et al., 2016). For example, only 45% of dementia sufferers in the United States have received clinical cognitive evaluations (Kotagal et al., 2015). Timely diagnosis is a prerequisite for good dementia care and helps people benefit from interventions, social support, and treatments. From this perspective, monitoring technology able to detect early signs of dementia in everyday situations might have great potential for supporting earlier diagnosis. Available published work has shown the usefulness of monitoring technologies for inferring an individual's state, such as stress (Lu et al., 2012), and mental fatigue (Yamada and Kobayashi, 2017a; Yamada and Kobayashi, 2017b); assessing behavioral characteristics, such as sleep quality (Rahman et al., 2015) and activities (Cook, 2010); and screening for diseases, such as bipolar disorder (Faurholt-Jepsen et al., 2016) and Parkinson's diseases (Tsanas et al., 2010). However, dementia remains difficult to detect from data collected on a daily basis for various reasons, such as people misconstruing the symptoms as a normal part of ageing.

One of the most promising ways to assess the health of dementia patients in everyday situations is identifying the evolution of language change over the course of dementia's progression. Although the most typical symptom of dementia is memory impairment due to the medial temporal lobe shrinking (Kirshner, 2012; MacKay et al., 2008), dementia is characterized by a decline from a previously attained level of performance in one or more cognitive domains such as memory, learning, executive function, and language (American Psychiatric Association, 2013). As for language function, both retrospective analyses and prospective cohort studies have shown that language problems are prevalent dating from presymptomatic periods (Van Velzen and Garrard, 2008; Oulhaj et al., 2009). In addition, studies on pathologically proven Alzheimer's disease (AD) patients have shown that they exhibited significant language changes such as syntactic simplification and impairments in lexico-semantic processing at the time of diagnosis (Ahmed et al., 2012; Ahmed et al., 2013).

Previous computational studies attempted to characterize such language dysfunctions by using acoustic, prosodic, and linguistic features extracted from data gathered while participants performed neuropsy-

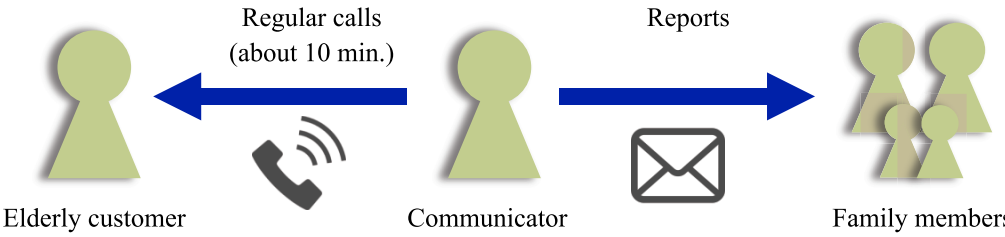


Figure 1: Overflow of regular monitoring service. A communicator calls an elderly customer once or twice a week, transcribes the conversations, and e-mails the transcripts to family members. We analyzed the conversation transcripts this service provided.

			Data duration				
Status	Gender	Age	Start	End	No. of calls	Ave. call time Mean (SD) [min.]	Ave. no. of words per conversation
Control	F	75-77	2015 Mar	2017 Apr	75	11.21 (8.85)	395.13 (124.18)
	F	80-83	2014 Jul	2017 Apr	109	16.63 (4.47)	734.34 (195.10)
	F	87-89	2016 Jan	2017 May	104	11.15 (4.46)	418.86 (235.12)
	M	66-70	2014 Jul	2017 Apr	133	10.62 (2.32)	482.89 (118.95)
	M	78-81	2014 Dec	2016 Mar	72	12.06 (2.83)	554.69 (119.03)
	M	82-85	2014 Nov	2017 Apr	226	17.75 (6.29)	572.12 (235.49)
Dementia	F	85-86	2014 Jul	2015 Nov	40	9.29 (2.15)	462.28 (204.12)
	F	88-88	2014 Jul	2014 Nov	13	7.77 (1.72)	277.94 (151.47)

Table 1: Specifications of conversational data of participants provided by the regular monitoring service.

chological tests by professionals such as medical doctors (Bucks et al., 2000; Hoffmann et al., 2010; Guinn and Habash, 2012; Fraser et al., 2016). For example, the short-term memory loss associated with dementia often brings about word-finding and word-retrieval difficulties (Henry et al., 2004; Kavé and Goral, 2018). These difficulties have typically been characterized by measuring fillers including non-words and short phrases (e.g., "umm" or "uh") (Guinn and Habash, 2012; König et al., 2015). Dementia patients tend to reduce the tempo of and articulation rates in their speech (Hoffmann et al., 2010). These reductions have been measured by phonemes per second in patients' speech. Dementia patients also tend to reduce the expressiveness of their speech. This reduction has been characterized by using linguistic features such as the decrease in adjective proportion and indices related to vocabulary richness (Bucks et al., 2000; Chinaei et al., 2017). These features have been extracted from spontaneous speech data during neuropsychological tests such as image descriptions, which might be useful for characterizing everyday conversations (Tomoeda et al., 1996; Fraser et al., 2016; Chinaei et al., 2017). Studies have recently started investigating whether these language dysfunctions observed in neurodegenerative diseases including dementia can be extracted in conditions close to those of everyday life and have garnered increasing attention (Masrani et al., 2017; Shinkawa and Yamada, 2018b; Shinkawa and Yamada, 2018a). However, these studies remain limited, and further research is required to detect dementia from language data gathered from everyday situations such as social media posts, family conversations, and conversations during monitoring services for older adults.

In this study, we analyzed conversational data of older adults with and without dementia obtained from a regular monitoring service for elderly people in Japan. We focused on repetition in conversations on different days in addition to single conversations on the same day on the basis of previous observational and descriptive studies that reported atypical repetitions as one of the prominent characteristics observed in dementia patients in everyday conversations (Cook et al., 2009). The analyses revealed that features for measuring repetition of both words and topics on different days increase for people with dementia compared with controls.

2. Materials & Methods

To gain insight into how dementia affects language features extracted from daily conversations, we analyzed conversational data collected in a regular monitoring service for elderly people. For the language features, we focused on topic and word repetition in separate conversations. In this section, we first describe the conversational data we used for analysis. We next explain how to calculate features to capture topic and word repetition on different days.

2.1. Conversational data from a regular monitoring service

We used conversational data obtained from the regular monitoring service for elderly people provided by Cocolomi Co., Ltd. (http://cocolomi.net/). Their service is intended to help families living separately to catch up on the lives of their older members. A communicator calls elderly people once or twice a week and encourages them to talk

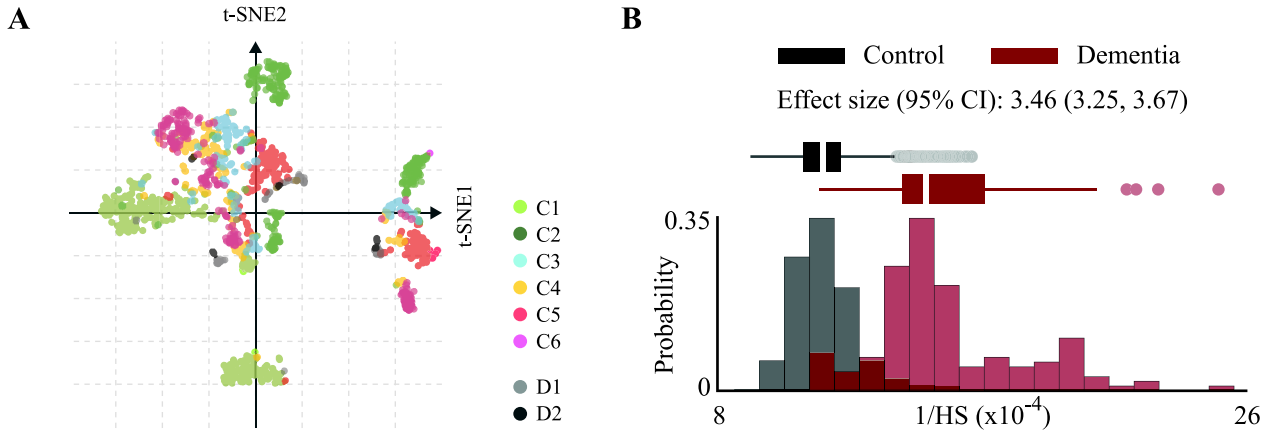


Figure 2: Representations of topic repetition. (A) Two-dimensional visualization of the topics using t-SNE. Circles are positioned on the basis of their position in the t-SNE1 and t-SNE2 dimensions. Circles represent each topic. Grey and black circles are topics extracted from the dementia participants' conversations, and the others are from controls' conversations. (B) Histogram and boxplot of the feature related to topic repetition on different days. Boxes denote the 25th (Q1) and 75th (Q3) percentiles. The line within the box denotes the 50th percentile, while whiskers denote the upper and lower adjacent values that are the most extreme values within Q3+1.5(Q3-Q1) and Q1-1.5(Q3-Q1), respectively. Filled circles show outliers.

about their daily life. The conversation is transcribed by the communicator and e-mailed to the family members (Figure 1). The communicator typically transcribes the conversation in a spoken word format omitting incomplete words and fillers.

We used the transcribed texts collected from eight Japanese people (five females and three males aged 66-89 years, i.e., 82.37 $\pm$ 5.91 years old). Two of them were reported by their families as suffering from dementia. Table 1 shows the duration the service was used the number of the reported calls, the average duration of each call, and the average word length of each report. In total, 458,738 words in 772 documents were used for the analysis. All reports were written in Japanese.

For preprocessing, we performed word segmentation, part-of-speech tagging, and word lemmatization on the transcribed texts. Required part-of-speech words were extracted from the transcribed texts, and predefined stop words were eliminated. We used the Japanese morphological analyzer MeCab (Kudo, 2005).

2.2. Feature related to topic repetition on different days

To obtain a feature related to topic repetition, we first extracted N topics from conversational data of three successive phone calls that were arranged in time sequence of data collection date during the monitoring service. We then calculated topic repetition by using topic similarities between the two sets of conversational data.

To extract topics, we used latent Dirichlet allocation (LDA), an unsupervised Bayesian probabilistic model commonly used in text analysis (Blei et al., 2003). It assumes that all documents are probabilistically generated from a set of N topics, where each topic is a multinomial distribution over the words (β) and the documents are a mixture of these topics (θ). LDA assumes every document in the corpus is generated using the following generative process:

1. A document specific topic distribution $\theta_c \sim Dir(\alpha)$ is drawn,

2. and for the ith word in the document; a topic assignment $z_i \sim \theta_c$ is drawn, and a word $w_i \sim \beta_{z_i}$ is drawn and observed.

For any given data, LDA automatically infers the latent document distribution θ_c for each document $c \in D$ and the topic distribution β_k for each of the $k = \{1, ..., N\}$ topics. The probability of the ith word in a document c is:

$$p(w_i, \theta_c) = \sum_k p(w_i|\beta_k)p(z_i = (k|\theta_c)).$$

Two sets of conversational data for calculating topic similarities were picked as two to six conversational data separated. Topic similarity was measured as cosine similarity of the word probability vectors for each topic, and the word probabilities less than 0.001 were padded with zero. Topic similarities were calculated for all possible combinations of topics extracted from different set of conversational data, and their maximum values were used as measures of topic repetition. The N was set to 5 in this study.

2.3. Features related to word repetition on same day and different days

To quantify the word repetition behavior, previous computational studies focused on sentence similarities such as calculating the cosine distance between each pair of sentences in a conversation (Fraser et al., 2016). In this study, we focused on the feature of vocabulary richness as word repetition could result in a small number of distinct words being used in a conversation (Manschreck et al., 1981). Vocabulary richness was calculated by three typical measures (Bucks et al., 2000; Honoré, 1979): type-token ratio (TTR), Brunet's index (BI), and Honoré's statistic (HS). TTR is computed by dividing the total number of words

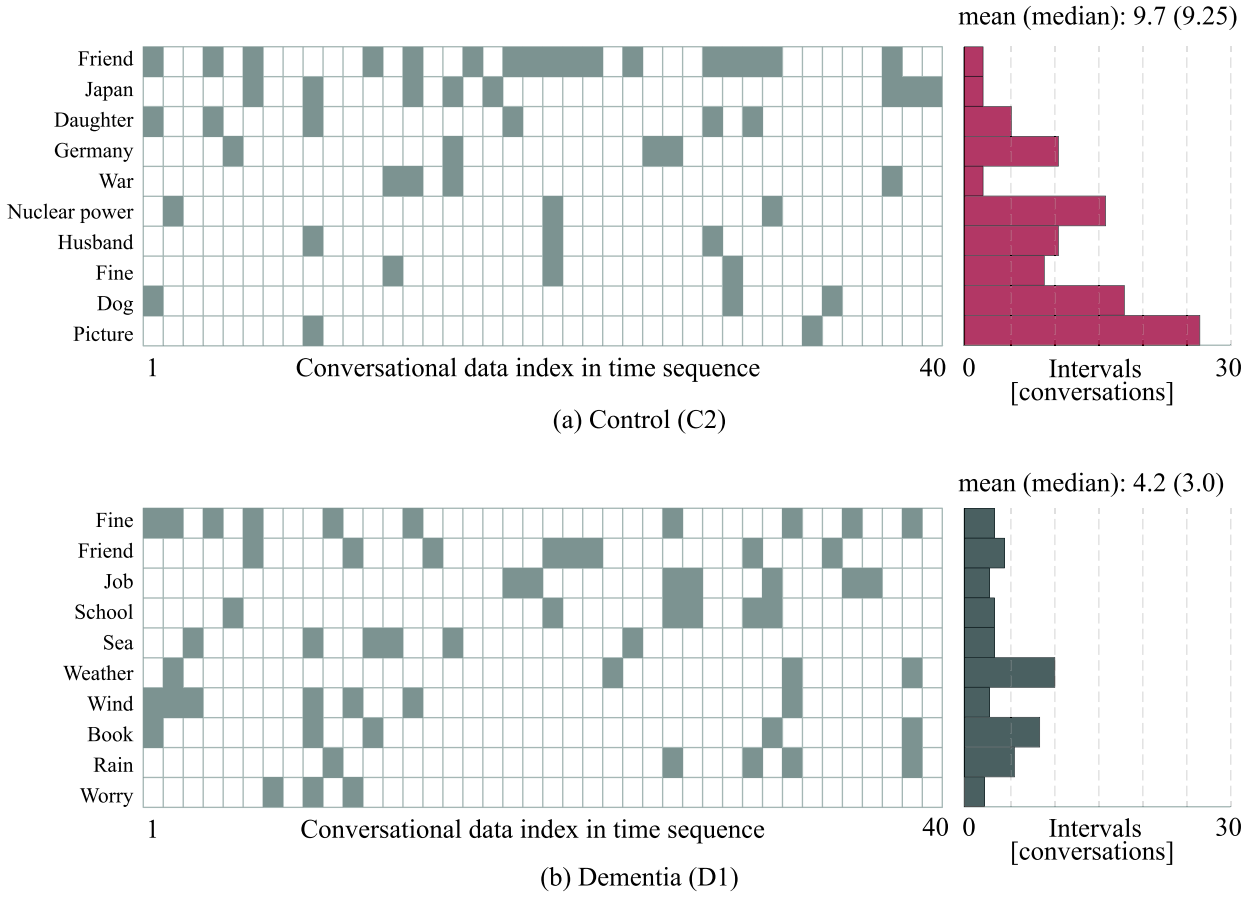


Figure 3: Appearance and interval of featured words in conversation. Featured words were translated from Japanese into English. Matrix with grey cells shows the word appearance in each conversation. The conversational data indices are ordered in time sequence from left to right. Red bars show the intervals between words.

(N) into the number of different word types (U). By using the same U and N, BI is also defined as BI=$N^{U^{-0.165}}$. Unlike with other measures related to vocabulary richness, the vocabulary richness becomes greater as BI becomes smaller. HS gives particular importance to unique vocabulary items used only once (N_{uni}). HS is defined as HS=$100 \log N/(1 - N_{\mathrm{uni}}/U)$. Previous studies reported that HS relatively showed a higher discrimination power than other measures for detecting dementia (Fraser et al., 2016). Therefore, we focused on HS and used its inversed number as a measure of repetition.

We first obtained pairs of conversational data D_i and D_j separated by t days ($T - M \leq t \leq T + M$). D_i and D_j contain all the words except numerals and symbols. HS_i and HS_j were extracted from D_i and D_j by calculating HS. Next, we defined D_{ij} as a combined document of D_i and D_j and extracted HS_{ij} as a feature of repetitiveness in conversations on different days. As a feature of repetitiveness in conversations on the same day, HS_i^{-1} (HS_j^{-1}) was used.

3. Results

We investigated how topics in conversations in regular monitoring services differed between older adults with and without dementia. We applied LDA to conversational data of each participant and obtained pairs of topics and word probability vectors. We obtained 1,240 word probability vectors represented in 795 dimensions. We investigated the similarities of the word probability vectors extracted from the conversations of each individual throughout the period of the monitoring service. To this end, we used the method of t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008), which is widely used for visualizing high-dimensional datasets. Specifically, it models each high-dimensional object by two- or three-dimensional points in such way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points. Figure 2A shows the two-dimensional representation of the word probability vectors. The topics extracted from the conversation of participants with dementia are localized near each other in each participant. This could be considered to be because similar topics appear more frequently in conversations of individuals with dementia than in conversations of those without.

We next quantified such topic similarities by using the topic-repetition feature between conversations on different days. We investigated if the feature shows the difference between individuals with and without dementia. The feature was measured by using an effect size (Cohen's d) (Nakagawa and Cuthill, 2007). For Cohen's d, a 0.8 effect size is large, 0.5 medium, and 0.2 small. We observed that the feature extracted from the conversations of individuals with dementia was significantly higher than that extracted

from the conversations of individuals without (effect size of 3.46, 95% Confidence interval (CI): 3.25-3.67; Figure 2B). Topic repetition as the behavior of dementia sufferers might be captured in conversations of the daily monitoring service.

For a prior investigation of the word-repetition feature, we focused on the repetition interval of the featured words in conversations of two participants with and without dementia. We selected the top ten words on the basis of the word probabilities of topics extracted by LDA. We investigated their intervals between repetitions by calculating the mean duration of each word occurrence. From the results, the mean durations were 9.7 for the control and 4.2 for the participant with dementia (Figure 3). This result indicates that word-repetition intervals in conversations of individuals with dementia might be shorter than those in conversations of controls.

To objectively measure the repetition of words, we investigated word-repetition features in both single conversations on the same day and paired conversations on different days. We observed higher repetition in conversations of individuals with dementia than in conversations of those without in both single and paired conversational data (effect size of 1.58, 95% CI 1.28-1.88 for single conversation and effect size of 2.67, 95% CI 2.10-3.24 for paired conversation; Figure 4). The difference was larger in paired conversational data, which suggests that monitoring word repetition in conversations on different days may help to detect signs of dementia in daily life.

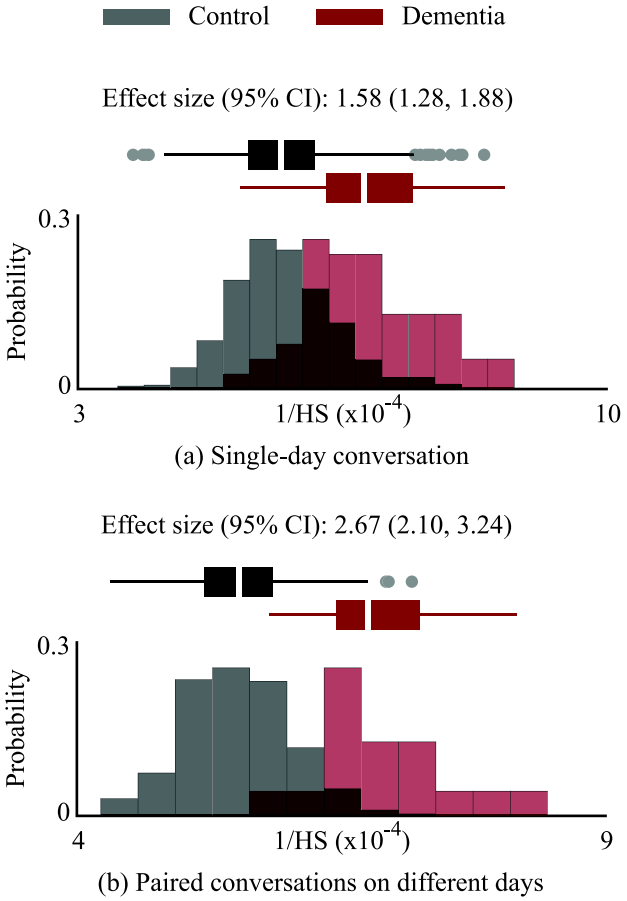


Figure 4: Comparison of histogram and boxplot between the word-repetition feature, extracted from single and paired conversations. Boxes denote the 25th (Q1) and 75th (Q3) percentiles. The line within the box denotes the 50th percentile, while whiskers denote the upper and lower adjacent values that are the most extreme values within Q3+1.5(Q3-Q1) and Q1-1.5(Q3-Q1), respectively. Filled circles show outliers.

4. Conclusion

We investigated word and topic repetition in daily conversations as a sensitive index of cognitive decline in dementia. We focused on repetition in conversations on different days in addition to single conversations on the same day on the basis of previous observational and descriptive studies that reported atypical repetitions as one of the prominent characteristics observed in dementia patients in everyday conversations.

We investigated topic- and word-repetition features by using conversational data obtained from a regular monitoring service. First, we visualized the topic probabilities using t-distributed stochastic neighbor embedding (t-SNE) to obtain an overview of topic distribution of each participant. Next, we investigated the feature of topic repetition in separate conversations on different days. We observed higher repetitiveness for participants with dementia than those without. This result indicates that the topic-repetition feature was able to capture atypical repetition in daily conversation of participants with dementia. For the word-repetition feature, we first investigated the intervals between the featured words in conversations. The intervals between repeated words were shorter in conversations of participants with dementia than in conversations of those without. Next, we investigated the word-repetition feature in single-day and different-day conversations. We observed the increase in features for measuring repetition in patients with dementia compared to those without in both topic and words. The feature related to repetition in conversations in regular monitoring services might be useful for discriminating individuals with dementia from controls.

One of the limitations in this study was its small number of participants. In future work, we will need to confirm our results on a larger number participants. Another limitation was the participants' labels for dementia sufferers and healthy controls. As mentioned in the Materials & Methods section, in this study, the participants' labels were based on not clinical assessments including dementia types and severities but reports from the participants' families.

5. Acknowledgements

We sincerely thank A. Kamiyama, Y. Masuda, J. Hayakawa, and K. Cho at Cocolomi Co., Ltd. for providing all of the data used in this study, and appreciate valuable comments and suggestions.

6. Bibliographical References

Ahmed, S., de Jager, C. A., Haigh, A.-M. F., and Garrard, P. (2012). Logopenic aphasia in Alzheimer's disease: clinical variant or clinical feature? *J Neurol Neurosurg Psychiatry*, 83(11):1056–1062.

Ahmed, S., de Jager, C. A., Haigh, A.-M., and Garrard, P. (2013). Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology*, 27(1):79–85.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Ed. (DSM-5)*. American Psychiatric Pub.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J Mach Learn Res*, 3(Jan):993–1022.

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Chinaei, H., Currie, L. C., Danks, A., Lin, H., et al. (2017). Identifying and avoiding confusion in dialogue with people with Alzheimer's disease. *Computational Linguistics*, pages 377–406.

Cook, C., Fay, S., and Rockwood, K. (2009). Verbal repetition in people with mild-to-moderate Alzheimer disease: a descriptive analysis from the VISTA clinical trial. *Alzheimer Dis Assoc Disord*, 23(2):146–151.

Cook, D. J. (2010). Learning setting-generalized activity models for smart spaces. *IEEE intelligent systems*, 2010(99):1.

Dolgin, E. (2016). How to defeat dementia. *Nature*, 539(7628):156–158.

Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., et al. (2016). Voice analysis as an objective state marker in bipolar disorder. *Transl Psychiatry*, 6(7):e856.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis*, 49(2):407–422.

Guinn, C. I. and Habash, A. (2012). Language analysis of speakers with dementia of the Alzheimer's type. In *AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*, pages 8–13.

Henry, J. D., Crawford, J. R., and Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia*, 42(9):1212–1222.

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., et al. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *Int J Speech Lang Pathol*, 12(1):29–34.

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin*, 7(2):172–177.

Kavé, G. and Goral, M. (2018). Word retrieval in connected speech in Alzheimer's disease: a review with meta-analyses. *Aphasiology*, 32(1):4–26.

Kirshner, H. S. (2012). Primary progressive aphasia and Alzheimer's disease: brief history, recent evidence. *Curr Neurol Neurosci Rep*, 12(6):709–714.

König, A., Satt, A., Sorin, A., Hoory, R., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst).*, 1(1):112–124.

Kotagal, V., Langa, K. M., Plassman, B. L., Fisher, G. G., et al. (2015). Factors associated with cognitive evaluations in the United States. *Neurology*, 84(1):64–71.

Kudo, T. (2005). Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*.

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., et al. (2017). Dementia prevention, intervention, and care. *Lancet*, 390(10113):2673–2734.

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., et al. (2012). Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proc ACM Int Conf Ubiquitous Comput*, pages 351–360. ACM.

MacKay, D. G., James, L. E., and Hadley, C. B. (2008). Amnesic HM's performance on the language competence test: Parallel deficits in memory and sentence production. *J Clin Exp Neuropsychol*, 30(3):280–300.

Manschreck, T. C., Maher, B. A., and Ader, D. N. (1981). Formal thought disorder, the type-token ratio and disturbed voluntary motor movement in schizophrenia. *Br J Psychiatry*, 139(1):7–15.

Masrani, V., Murray, G., Field, T., and Carenini, G. (2017). Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. *BioNLP 2017*, pages 232–237.

Nakagawa, S. and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc*, 82(4):591–605.

Oulhaj, A., Wilcock, G. K., Smith, A. D., and de Jager, C. A. (2009). Predicting the time of conversion to MCI in the elderly role of verbal expression and learning. *Neurology*, 73(18):1436–1442.

Prince, M., Guerchet, M., and Prina, M. (2013). *The global impact of dementia 2013-2050*. Alzheimer's Disease International.

Prince, M., Wimo, A., Guerchet, M., Ali, G.-C., et al. (2015). *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International.

Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M., and Karagiannidou, M. (2016). World Alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future.

Rahman, T., Adams, A. T., Ravichandran, R. V., Zhang, M., et al. (2015). DoppleSleep: A contactless unobtrusive sleep sensing system using short-range doppler radar. In *Proc ACM Int Conf Ubiquitous Comput*, pages 39–50.

Shikimoto, R., Sado, M., and Mimura, M. (2016). The social costs of dementia in Japan: Focusing on the informal care cost. *Brain and nerve*, 68(8):939–944.

Shinkawa, K. and Yamada, Y. (2018a). Topic repetition in conversations on different days as a sign of dementia. In *Medical Informatics Europe*.

Shinkawa, K. and Yamada, Y. (2018b). Word repetition in separate conversations for detecting dementia: A preliminary evaluation on regular monitoring service. In *Proceedings of AMIA Informatics Summit*.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A.,

et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.*, 7(3):280–292.

Tomoeda, C. K., Bayles, K. A., Trosset, M. W., Azuma, T., et al. (1996). Cross-sectional analysis of Alzheimer disease effects on oral discourse in a picture description task. *Alzheimer Disease & Associated Disorders*, 10(4):204–215.

Tsanas, A., Little, M. A., McSharry, P. E., and Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Trans Biomed Eng*, 57(4):884–893.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *J Mach Learn Res*, 9:2579–2605.

Van Velzen, M. and Garrard, P. (2008). From hindsight to insight–retrospective analysis of language written by a renowned Alzheimer's patient. *Interdiscip Sci Rev*, 33(4):278–286.

Yamada, Y. and Kobayashi, M. (2017a). Detecting mental fatigue from eye-tracking data gathered while watching video. In *Proceedings of Conference on Artificial Intelligence in Medicine in Europe*, pages 295–304. Springer.

Yamada, Y. and Kobayashi, M. (2017b). Fatigue detection model for older adults using eye-tracking data gathered while watching video: Evaluation against diverse fatiguing tasks. In *IEEE International Conference on Healthcare Informatics*, pages 275–284. IEEE.