Predicting Dementia Screening and Staging Scores From Semantic Verbal Fluency Performance

Nicklas Linz, Johannes Tröger and Jan Alexandersson
DFKI GmbH
Saarbrücken, Germany
<f>.<l>@dfki.de

Maria Wolters
School of Informatics - University of Edinburgh
Edinburgh, Scotland
maria.wolters@ed.ac.uk

Alexandra König and Philippe Robert
CoBTeK - IA CHU Université Côte d'Azur
Nice, France
akonig03@gmail.com

***Abstract*—The standard dementia screening tool Mini Mental State Examination (MMSE) and the standard dementia staging tool Clinical Dementia Rating Scale (CDR) are prominent methods for answering questions whether a person might have dementia and about the dementia severity respectively. These methods are time consuming and require well-educated personnel to administer. Conversely, cognitive tests, such as the Semantic Verbal Fluency (SVF), demand little time. With this as a starting point, we investigate the relation between SVF results and MMSE/CDR-SOB scores. We use regression models to predict scores based on persons' SVF performance. Over a set of 179 patients with different degree of dementia, we achieve a mean absolute error of of 2.2 for MMSE (range 0–30) and 1.7 for CDR-SOB (range 0–18). True and predicted scores agree with a Cohen's κ of 0.76 for MMSE and 0.52 for CDR-SOB. We conclude that our approach has potential to serve as a cheap dementia screening, possibly even in non-clinical settings.**

***Index Terms*—dementia, Clinical Dementia Rating Scale (CDR), Mini Mental State Examination (MMSE), machine learning, prediction of clinical scores**

I. Introduction

Alzheimer's Disease (AD) has a significant economic impact on our society: according to the World Alzheimer Report 2016, AD is about to become a *trillion dollar disease* by 2018 [1]. This is in addition to the unquantifiable mental, emotional, and physical burden that AD places on people with the illness, and their caregivers, friends, and family. AD's is a type of dementia in which the main observable symptom is characterised by a decline in cognitive functions, notably memory, as well as language and problem solving. While AD is the most common organic cause of dementia, there are many other causes, such as vascular disorders, e.g., strokes, brain tumours, traumatic brain injuries, or fronto-temporal lobe degeneration (FTLD); see also Figure 1. In order to quantify dementia's severity and prepare for its potential impact on a patient's environment, staging and screening tools have been developed. The Clinical Dementia Rating scale (CDR) [2] represents internationally the most widely applied staging tool for assessing the disease's global severity. It encompasses six domains of cognitive and functional performance: Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care [3]. The assessment is conducted in the form of semi-structured interviews with the affected person and an affiliated person/co-interviewee, e.g., a family member.

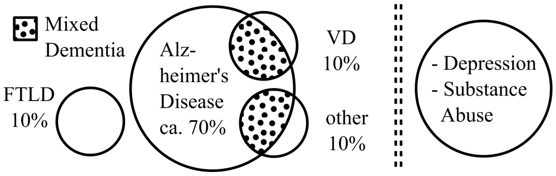

Fig. 1. The left panel shows the types of dementia, according to their cause, including Fronto-Temporal Lobar Degeneration (FTLD), and Vascular Dementia (VD); the dotted areas indicate those cases where more than one cause underlies the disorder. The right panel shows other, mostly reversible, causes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

The CDR is relatively time-consuming - interviews can take up to 90 minutes - depending on the availability of a co-interviewee and requires significant training of the raters in order to achieve good reliability [4].

The CDR is often used in combination with the Mini Mental State Examination (MMSE), a common screening tool for dementia. It takes around ten minutes and requires a trained assessor, consists of a series of tasks that cover different forms of cognitive functions, such as memory and attention, and is designed to be used as a global screening tool. However, in some applications the MMSE lacks sensitivity; especially for early stages, its items are considered to be relatively easy and are highly likely to result in ceiling effects [5]. Moreover, it has been shown, that the standard MMSE might lack sufficient intra- and interrater reliability [6].

While there are many screening tools, a reliable diagnosis of probable dementia can only be made through in-depth assessments, and a comprehensive combination of behavioural

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.

(e.g., psychometric tests) and in vivo organic assessment (e.g., functional brain imaging). Behavioural assessments typically consist of structured interviews and can also include a number of well-defined tasks to assess particular aspects of cognition, such as memory and executive function. One of these tasks is Semantic Verbal Fluency (SVF), where the participant is instructed to name as many members of a semantic category as possible in a given time period. The most common category for this task is "animals". Many neurocognitive diseases lead to a reduction in the number of items produced during an SVF task, including AD [7]–[9], Parkinson's Disease [10], schizophrenia [11], or focal brain lesions [12].

We argue that qualitative analysis of such a task allows for the deduction of corresponding dementia staging and screening scores which would allow to objectify and underpin CDR and MMSE scores, as well as to mitigate some of their aforementioned methodological caveats. In this paper, we present an analysis method that uses SVF data to predict two test scores, MMSE and CDR - Sum of Boxes (CDR-SOB), which has been used as a quantitative approximation of the CDR scale itself [2].

After reviewing related work in Section II, we outline our approach in Section III. In Section IV, we compare regression models for prediction of the MMSE and CDR-SOB. We interpret predicted scores according to common clinical thresholds and report Cohen's κ as a reliability measure. In Section V, we discuss how our algorithm can be leveraged for medical human-computer interaction applications for dementia screening, and conclude by outlining further work.

II. Related work

A. Diagnosis as a Classification Problem

The common approach for detecting signs of neurocognitive diseases from speech is to treat it as a classification problem, which is either binary or n-ary (with small n for a highly restricted number of potential diseases). The degree of manual intervention varies, from approaches that rely on manual transcriptions to a completely automated speech-based screening pipeline yielding significant discrimination results [13].

Work in this direction usually differs in means of the analysed corpora (free speech vs. cognitive tests vs. conversation), classification scenario (healthy vs. impaired or healthy vs. mildly impaired vs. severely impaired) and extracted features (linguistic vs. para-linguistic).

[14] worked on recordings of picture descriptions of the *Cookie Theft Picture Description Task*, extracted from the *DementiaBank* corpus [15]. They discriminate individuals with AD from healthy, age-matched, controls (HC) with an accuracy of 81% using linguistic and para-linguistic features. [16] uses language modelling techniques to calculate the perplexity of picture description tasks from *DementiaBank* to separate AD and HC individuals with an accuracy of 77.1%. [17] extracts para-linguistic features (e.g., pauses, pitch & jitter) of picture descriptions from *DementiaBank* to discriminate between AD and HC with an accuracy of 94.7%. [18] use para-linguistic markers from recordings of people performing different spoken cognitive tests (countdown, picture descriptions, sentence repetition and SVF) to classify individuals into three groups: early AD, Mild Cognitive Impairment (MCI) and HC. They train three binary classifiers with varying accuracies (HC vs. MCI: 20% ± 5; AD vs. MCI: 19% ± 5; HC vs. AD: 13% ± 3). [13] analyses spontaneous speech collected in a clinical setting through extracting temporal and para-linguistic features to separate HC from MCI patients. The resulting classifier yields an F_1 score of 86.2% and an accuracy of 82.4%. [19] extracted vocal features from a sentence reading task to discriminated between age-matched AD and HC patients with an accuracy of 84.8%. [20] uses phonetic features collected from a SVF and the East Boston memory test (EB) to discriminate between HC and MCI groups with an accuracy of 86.5%. Our own group previously extracted vocal features from cognitive tests (counting down numbers and Cookie Theft picture description) to identify patients with AD from HC with an accuracy of 89% ±3 [21].

B. Diagnosis as a Regression Problem

Neurocognitive diseases are complex and vary in their exact symptoms from person to person and from stage to stage. Therefore, it might be more useful to predict scores on screening or diagnostic tests than predicting a raw diagnosis. This makes it easier for clinical practitioners to integrate findings from an automatic analysis tool with the overall clinical picture, in particular when it comes to distinguishing between different potential causes for the same symptoms.

To our knowledge, there has been very little work on prediction of clinical scores from audio samples. [22] used semantic, acoustic and lexiosemantic features extracted from *DementiaBank* to predict MMSE scores. Using a bivariate dynamic Bayes net they achieved a mean absolute error (MAE) of 3.83, which they improved to 2.91 for patients where longitudinal data is available. The topic has received more attention in the image processing community and multiple authors have predicted clinical scores from brain imaging features, e.g., average regional grey matter density and tissue volume of MRI [23], [24]. As an example, [25] uses a Random Forest Regressor to predict clinical scores, including the MMSE and CDR-SOB, based on imaging data. This leads to a best Mean Absolute Error of 1.68 for the MMSE and 0.69 for the CDR-SOB.

C. Analysis of the SVF Task

The classical measure for SVF performance is word count per minute. In qualitative analysis of SVF performance this count can be modelled as a combination of two components: "mean cluster size" and "number of switches between clusters". *Clusters* are defined as a sequence of words that belong to the same semantic category in a person's mental lexicon. *Switches* occur at cluster boundaries, when the person switches to a new semantic category. Therefore, mean cluster size is related to the mental lexicon, whereas switches indicate executive search processes. The two measures relate to the word count as depicted below.

$$\text{Word Count} = \text{Mean Cluster Size} \times (\text{Number of Switches} + 1)$$

The semantic clustering criterion is the main determiner for both measures. An example is given below, with one switch and two clusters: pets and farm animals.

(*cat* - *dog*) - (*cow* - *horse*)
$(Cluster_1)\ Switch_1\ (Cluster_2)$

Para-linguistic features also have been shown to be of value in the analysis of SVF. [20] used the pseudo-syllable rate and average pause lengths for the analysis of SVF. [26] analysed pauses, speech rate and disfluencies in SVF. In order to differentiate between multiple pathologies, the above mentioned qualitative measures have been established which serve as additional markers next to the raw fluency word count [27], [28]. There is a broad agreement that these measures serve as indicators for underlying cognitive processes.

Pauses can occur both within clusters, as participants search their mental lexicon for more examples of a specific group, and between clusters, at the time of a switch, when a participant is searching for the next potentially productive subcategory. In the first 10-15 seconds of the task, pauses tend to be rare, and they typically become more frequent, and longer towards the end of the test.

[28] cites high reliability for their clustering annotation, and has established a list of potential semantic subcategories. However, when analysing new material, especially from different cultures [26], these subcategories need to be redefined and extended. Statistical semantic analysis can automatically and reliably provide clusters, which makes categorisation easier to replicate. Alternative approaches have been suggested based on statistical methods: Latent Semantic Analysis (LSA) [29], Explicit Semantic Analysis (ESA) [30] and Neural Word Embeddings [31].

D. MMSE and CDR as Assessment Tools

Both, MMSE and CDR-SOB, global assessment measures are the most widely used in clinical and research settings for dementia screening and staging its severity. Staging dementia is crucial for clinical trials and the development of effective pharmacological interventions. They are administered and interpreted by specially trained healthcare clinicians in order to provide appropriate patient care and to identify the effectiveness of prescribed treatment interventions in patients with dementia.

Inter-rater reliability for CDR is excellent (correlation coefficient 0.89) [32] and content validity can be assumed, as the six cognitive domains rated by the CDR are linked to validated clinical diagnostic criteria for AD [33].

Each domain is rated on a 5-point scale of functioning as follows: 0, no impairment; 0.5, questionable impairment; 1, mild impairment; 2, moderate impairment; and 3, severe impairment (personal care is scored on a 4-point scale without a 0.5 rating available). The global CDR score is computed via an algorithm that weighs memory more heavily than the other categories[1]. The CDR-SOB score is obtained by summing each of the domain box scores, with scores ranging from 0 to 18 [34]. In general, the higher the score, the greater the severity of dementia.

The CDR-SOB score has been considered a more detailed quantitative general index than the global score and provides more subtle information than the global CDR score in patients with mild dementia, and a suitable tool for measuring the response to treatment in clinical trials of AD [35]. The advantages of the SOB method include that the CDR-SOB scores can be treated as interval data in statistical analyses, whereas global CDR scores are ordinal by the nature of the algorithm approach to condensing the data. Finally, the most significant advantage of using CDR-SOB scores for staging dementia severity is the increased precision, allowing for tracking changes over time [34].

The MMSE encompasses a variety of questions, requires minimal training and takes around 10 min. The questions are typically grouped into seven categories, representing different cognitive functions: orientation to time (5 points), orientation to place (5 points), registration of three words (3 points), attention and calculation (5 points), recall of three words (3 points), language (8 points) and visual construction (1 point) [36], [37]. Patients score between 0 and 30 points, and cutoffs of 23/24 have typically been used to show significant cognitive impairment.

Its validity has been proven and it is widely translated and used [5]. The MMSE is unfortunately sometimes misunderstood as a diagnostic tool, when it is actually a screening test with relatively modest sensitivity in detecting a mild degree of cognitive impairment. It has floor and ceiling effects and limited sensitivity to change which is becoming a particularly important issue with the recent increased focus of researchers on the milder stages of AD [38].

III. Methods

A. Data

The data used for the following experiments was collected during the *Dem@Care* [39] and ELEMENT [21] 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 [40]. Participants were asked to perform a battery of cognitive tests, including a 60 second animal SVF test. Furthermore all participants completed the MMSE and CDR. Following the clinical assessment, participants were categorised into three groups: Control participants that complained about having subjective cognitive impairment (SCI) but were diagnosed as cognitively healthy after the clinical consultation, patients with MCI and patients that were diagnosed as suffering from Alzheimer's Disease and related disorders (ADRD). AD diagnosis was determined using

[1] http://www.biostat.wustl.edu/~adrc/cdrpgm/index.html

TABLE I
DEMOGRAPHIC DATA AND CLINICAL SCORES BY DIAGNOSTIC GROUP (MEAN ± STANDARD DEVIATION). HC='HUMAN CONTROL', MCI='MILD COGNITIVE IMPAIRMENT', AD= 'ALZHEIMER'S DISEASE', MD='MIXED DEMENTIA', VD='VASCULAR DEMENTIA'.

	HC	MCI	AD	MD	VD	Other	Total
N	42	47	33	37	10	10	179
Age	72.5 ± 8.3	76.6 ± 7.7	79.2 ± 5.0	78.8 ± 7.5	78.6 ± 4.6	78.1 ± 7.2	76.8 ± 7.5
Sex	8M/34F	23M/24F	12M/21F	19M/18F	8M/2F	10M/10F	80M/109F
MMSE	28.3 ± 1.6	26.0 ± 2.5	18.9 ± 5.0	18.5 ± 4.7	20.2 ± 4.1	23.7 ± 4.8	23.2 ± 5.5
CDR-SOB	0.48 ± 0.68	1.68 ± 1.11	7.52 ± 3.95	8.05 ± 3.31	5.50 ± 4.16	3.03 ± 3.63	4.02 ± 4.16

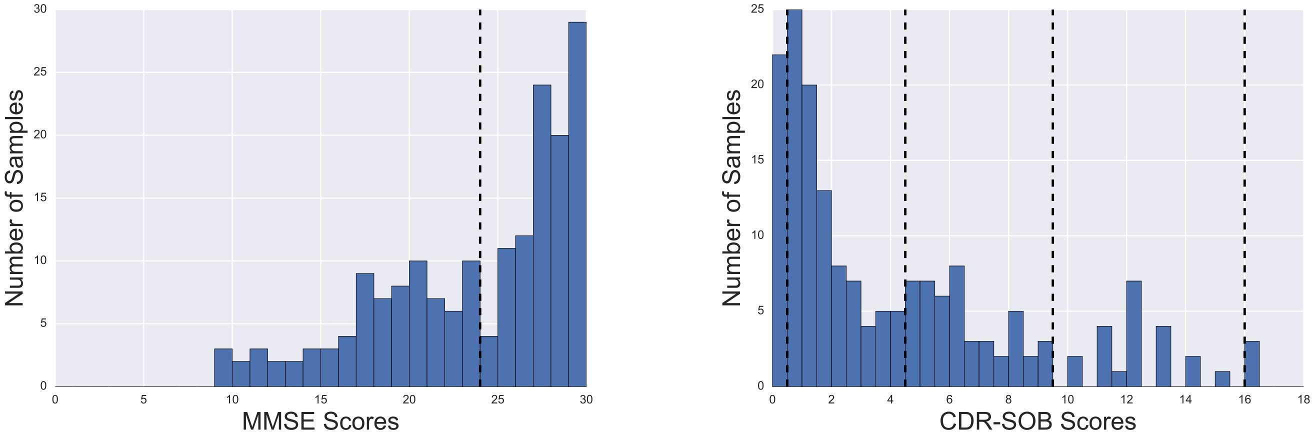

Fig. 2. Histograms of MMSE and CDR-SOB scores with cut-off values for staging are indicated by dotted lines.

the NINCDS-ADRDA criteria [41]. Mixed/Vascular dementia were diagnosed according to ICD 10 [42] criterea. For the MCI group, diagnosis was conducted according to Petersen criteria [43]. Participants were excluded if they had any major auditory or language problems, history of head trauma, loss of consciousness, psychotic or aberrant motor behaviour. Demographic data and clinical test results by diagnostic groups are reported in Table I.

The distribution of clinical scores in the data is shown in Figure 2. The left figure shows MMSE scores, which range from 0 (worst) to 30 (best). The most commonly used cut-off in the literature for possible dementia is 24. Other cut-offs include 17, 18, 19, 23, 25, and 26 [44]. Fewer than 10 of our participants fall below the lowest cut-off, while roughly half of them are below the traditional cut-off.

The right figure shows CDR-SOB scores, which range from 0 (normal) to 18 (worst). The stages of dementia severity corresponding to CDR-SOB scores are described in Table II (adapted from [2]). Again, most subjects are staged as normal or having possible impairment, and only few have moderate or severe dementia.

B. Features

In the following we describe which features have been computed for each sample. We compute features from three different categories: Statistical Clustering and Switching, Word Frequency Features, and Vocal Features.

Let $a_1, a_2, \ldots, a_n$ be the sequence of animals produced by patient p, with $a_i \in \mathbb{A}$ and $\mathbb{A}$ being the set of all animals.

TABLE II
CUT-OFF VALUES FOR THE CDR-SOB ACCORDING TO [2].

CDR SOB			Staging
0			normal
0.5	–	4.0	possible impairment or very mild dementia
4.5	–	9.0	mild dementia
9.5	–	15.5	moderate dementia
16.0	–	18.00	severe dementia

Word Count

$$WC = n$$

Statistical Clustering and Switching

We compute features based on using word embeddings calculated with word2vec [45] based on the french FraWac corpus [46] as described in [31]. Let $\vec{a_1}, \vec{a_2}, \ldots, \vec{a_n}$ be their representations in the vector space and let $a_1, \ldots, a_{n-1}$ form a semantic cluster. a_n is part of this cluster if

$$|\frac{\langle \vec{\mu}, \vec{a_n} \rangle}{\|\vec{\mu}\| \cdot \|\vec{a_n}\|}| > \delta_p$$

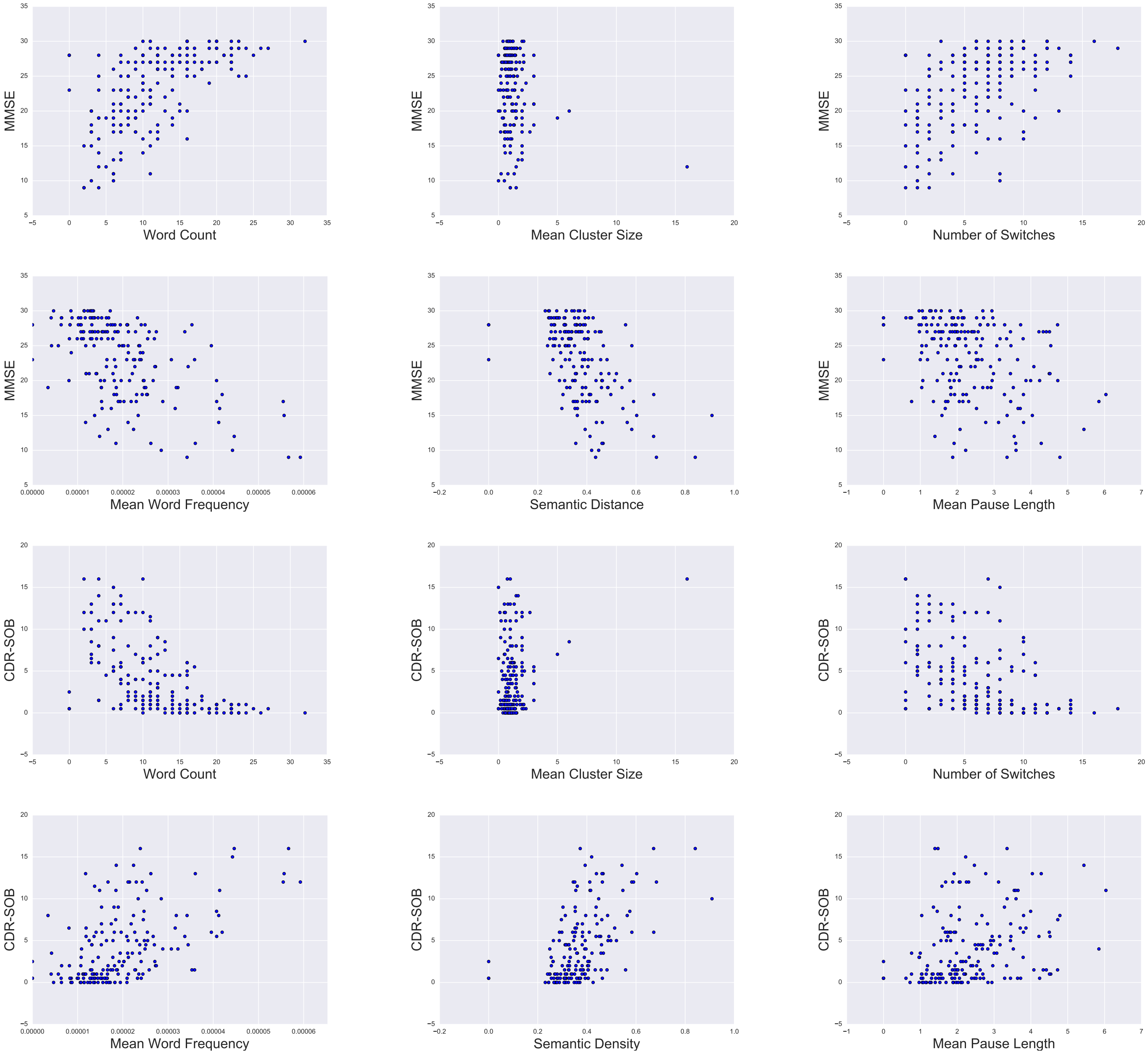

Fig. 3. Visualisation of feature distribution in relation to MMSE and CDR-SOB.

with

$$\vec{\mu} = \frac{1}{n-1} \cdot \sum_{\vec{x}\in\{\vec{a_1},\ldots,\vec{a_{n-1}}\}} \vec{x}$$

$$\delta_p = \frac{n!}{(n-2)!} \cdot \sum_{\vec{x},\vec{y}\in\{\vec{a_1},\ldots,\vec{a_n}\}} |\frac{\langle \vec{x}, \vec{y} \rangle}{\|\vec{x}\| \cdot \|\vec{y}\|}|$$

Let c_1,c_2, $\ldots$, c_m be the sequence of clusters, determined as described above and let $|c_i|$ be their size. We compute the following metrics:

Semantic Density

$$SD = \delta_p$$

Mean Cluster Size

$$MCS = \frac{1}{m} \sum_{i=1}^{m} |c_i|$$

Number of Switches

$$NOS = m - 1$$

TABLE III

PEARSON CORRELATION OF MMSE, CDR-SOB AND COMPUTED FEATURES.

	MMSE	**CDR-SOB**	**WC**	**MCS**	**NOS**	**MWF**	**SD**	**MPL**
MMSE	1.000	-0.834^{***}	0.602^{**}	-0.176	0.486^{*}	-0.560^{**}	-0.552^{**}	-0.352^{*}
CDR-SOB		1.000	-0.569^{**}	0.226	-0.464^{*}	0.550^{**}	0.553^{**}	0.306^{*}
WC			1.000	-0.123	0.838^{***}	-0.514^{**}	-0.538^{**}	-0.398^{*}
MCS				1.000	-0.335^{*}	0.191	0.339^{*}	0.006
NOS					1.000	-0.370^{*}	-0.511^{**}	-0.376^{*}
MWF						1.000	0.642^{**}	0.311^{*}
SD							1.000	0.399^{*}
MPL								1.000

* $|\sigma| > 0.3$ ** $|\sigma| > 0.5$ *** $|\sigma| > 0.7$

Word Frequency

We approximate word frequency of animals using the Python *wordfreq* package [47], which combines resources such as Wikipedia, news and book corpora and Twitter. Let $f : \mathbb{A} \to \mathbb{R}$ be the function mapping a word to its frequency.

Mean Word Frequency

$$MWF = \frac{1}{n} \sum_{i=1}^{n} f(a_i)$$

Vocal Features

Let p_1, p_2, …,p_s be the pauses in the audio sample, determined using the Praat software [48] as intervals of absence of sound longer than 250 ms. Let $|p_i|$ be the length of a pause.

Mean Pause Length

$$MPL = \frac{1}{s} \sum_{i=1}^{s} |p_i|$$

C. Evaluation Criterion

For evaluation of the quality of prediction of regression models there are many different metrics. Popular for its mathematical sophistication and severe punishment for large errors is the *Root Mean Squared Error* (RMSE).

In our case the use of the *Mean Absolute Error* (MAE) seems more appropriate. It delivers interpretable results on the real error made by the predictive model, scaled in the same way the clinical scores are. Let y_i be the actual value of sample i, let $\hat{y_i}$ be the regression models prediction and N the number of samples. The MAE is defined as

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{y_i} - y_i|$$

In the following we will describe the results of regression models and discuss the implications of their predictions.

IV. EXPERIMENTS

In order to determine the importance of and relationship between MMSE, CDR-SOB and the computed features we examine correlations, reported in Table III, and look at scatter plots of features and MMSE/CDR-SOB in Figure 3. Correlations smaller than 0.3 are considered as weak, greater than 0.5 as moderate and greater than 0.7 as strong. Both MCS and MCL have weak correlations to MMSE and CDR-SOB. Looking at their respective scatter plot, MCS does not seem to have any predictive power for either score, whereas the MPL seems to have at least some. Therefore, we exclude MCS from our feature set for all further analysis. WC, MWF and SD have correlations greater than 0.5 with both MMSE and CDR. Inspection of their respective scatter plots shows a near linear relationship.

To predict the CDR-SOB and MMSE, we train different regression models and evaluate their performance using MAE.

A. Prediction

Regression models are trained including Support Vector Regression (SVR), Lasso (Linear Regression with L_1 regularisation), Ridge Regression (Linear Regression with L_2 regularization), Elastic Net (EN) and a Random Forest Regressor (RFR). Their implementations are provided by the *scikit-learn* python framework [49] and all are trained with the features described in Section III-B excluding MCS. Features are normalised by subtraction of their mean and division through their standard deviation. Because of the small data set size (n=179) we can not use a separate validation/test set. Instead we rely on averaging multiple shuffled k-Fold cross validations, with k set to 5. Hyper parameters are determined using a cross validation based grid search on the training folds in each iteration of the outer cross validation loop.

Results of the regression are reported in Table IV. The RFR shows the worst performance of all tested regression models. For prediction of the CDR-SOB all other models (SVR, LR-L_1, LR-L_2, EN) show similar performance with overlapping 95% confidence intervals. For the MMSE the RFR also has the worst performance and the other regressors' performance is comparable again. Especially because of the small data set we are not able to identify any clear best performing model.

In contrast to normal regression, our predicted value is bound to a discrete scale, we are able to draw a confusion

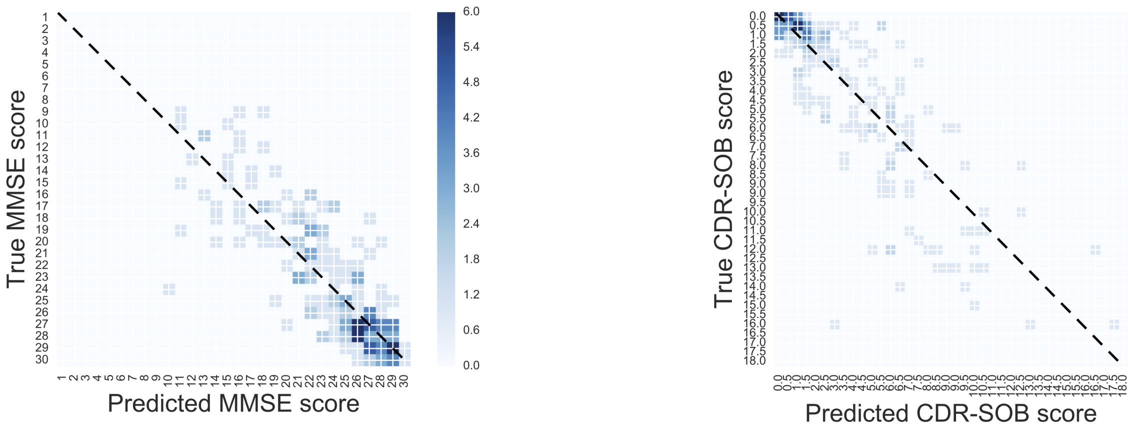

Fig. 4. Confusion matrix for MMSE and CDR-SOB predictions, as heat-map, obtained using a SVR model and rounding predictions to the nearest scale values.

TABLE IV
MEAN ABSOLUTE ERROR (MAE) AND 95% CONFIDENCE INTERVALS FOR DIFFERENT REGRESSION MODELS. BEST PERFORMANCE INDICATED IN BOLD.

	$MAE_{\textbf{MMSE}}$	$MAE_{\textbf{CDR-SOB}}$
SVR	**2.205** [1.920, 2.490]	**1.670** [1.433, 1.907]
LR - L_1	2.274 [1.988, 2.560]	1.683 [1.454, 1.912]
LR - L_2	2.289 [1.997, 2.581]	1.715 [1.485, 1.945]
EN	2.286 [1.993, 2.579]	1.688 [1.456, 1.920]
RFR	2.363 [2.073, 2.654]	1.728 [1.469, 1.986]

TABLE V
MEAN ABSOLUTE ERROR (MAE) [95% CONFIDENCE INTERVAL] AND MEAN ± STANDARD DEVIATION OF MMSE AND CDR-SOB PREDICTION FOR A SVR MODEL BY DIAGNOSIS GROUP.

	$\mu_{\textbf{MMSE}}$	$MAE_{\textbf{MMSE}}$	$\mu_{\textbf{CDR-SOB}}$	$MAE_{\textbf{CDR-SOB}}$
HC	28.244 ± 1.523	1.205 [0.905, 1.505]	0.489 ± 0.687	0.808 [0.589, 1.027]
MCI	25.679 ± 2.759	2.175 [1.678, 2.672]	1.708 ± 1.121	1.328 [1.030, 1.626]
DCI	18.914 ± 4.882	2.781 [2.311, 3.251]	7.556 ± 3.843	2.372 [1.955, 2.789]

matrix for each score by rounding predictions to the nearest value on the respective scale (1 steps for MMSe and 0.5 steps for CDR-SOB). Figure IV-A shows the confusion matrices for MMSE and SDR-SOB using predictions from the SVR model. For predictions of the MMSE score, there seems to be an underestimation for patients with an MMSE > 24 and an overestimation for patients with MMSE ≤ 24. Predictions of the CDR-SOB are overestimating for a CDR-SOB ≤ 3 and underestimating for CDR-SOB > 3.

To better understand the results we examine the MAE by diagnosis group. We define three different groups: SCI, MCI and dementia (DCI). SCI and MCI are diagnosis groups appearing in our dataset. Anyone with a confirmed diagnosis of Alzheimer's disease, Vascular Dementia or Mixed Dementia is put into the DCI group. The results are listed in Table V. At first glance it seems like the prediction error is growing with impairment of patients. But looking at the mean of each diagnosis group, one can observe that the standard deviation grows as well - meaning the values are spread further apart. This increases the complexity of the regression problem and accounts for the increased error.

B. Clinical Interpretation of Predictions

In practice, clinicians will interpret predicted scores relative to the interpretation framework they use for the actual tests. Therefore, we translated the continuous predicted test scores into categorical judgements and compared these judgments to those made on the original values using Cohen's unweighted κ [50] to measure agreement. For each case, we used the predicted value where the case was part of the test cross validation fold, not the training folds. A total of 179 cases with predicted CDR-SOB and MMSE values were available. κ was computed using the R package `psych`, Version 1.7.5.

$$\kappa = \frac{agreement_{observed} - agreement_{expected}}{1 - agreement_{expected}}$$

Since the predicted scores are continuous, we devised two strategies for mapping them onto the discrete scores required for decision making. For the MMSE, we used a strict cut-off, where all values smaller than the boundary value indicate possible dementia, and a cut-off that rounds the predicted value to the nearest integer. For CDR-SOB, we used a strict cut-off that mapped values in between two category boundaries onto the category indicating less impairment, and a cut-off where values are rounded to the nearest 0.5.

Reliability for CDR-SOB is not very high—the best agreement is 0.52, and there is a lot of overlap in the 95% confidence intervals (Table VI). As the confusion matrix shows, this is due to a tendency to slip into the next higher or next lower category. While this does not seem critical at first, in clinical practice, misdiagnosis in either direction can be highly problematic [51].

For the MMSE, however, agreement is much better. Depending on the threshold and the cut-off mechanism used, κ

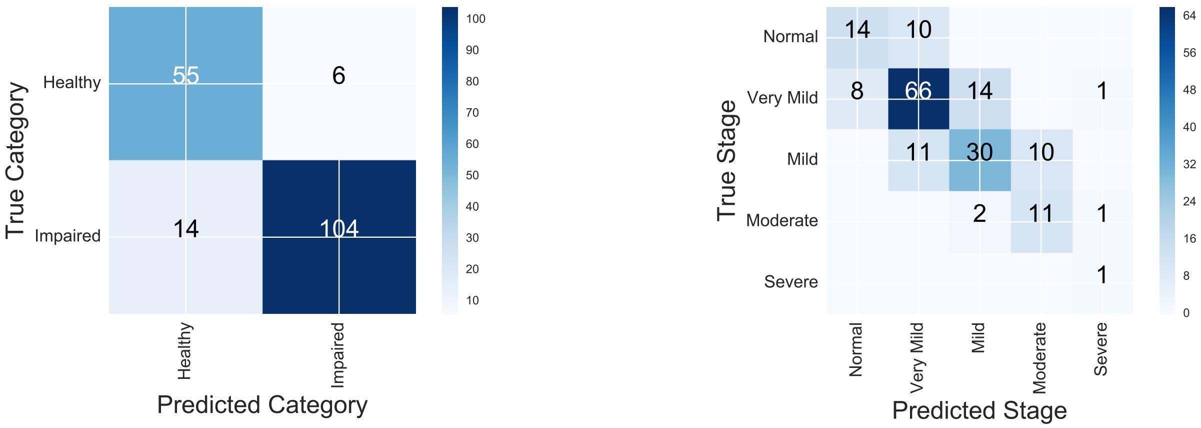

Fig. 5. Confusion Matrix for MMSE categories and CDR-SOB stages, as heat-map.

TABLE VI
κ FOR CDR-SOB STAGING, DIFFERENT CUT-OFF STRATEGIES. ESTIMATED VALUE WITH 95% CONFIDENCE INTERVAL. BEST PERFORMANCE INDICATED IN BOLD.

Cut-Off	κ
Strict	0.47 [0.36, 0.57]
Rounded	**0.52** [0.41, 0.62]

TABLE VII
κ FOR MMSE STAGING, ROUNDING TO NEAREST INTEGER. ESTIMATED VALUE WITH 95% CONFIDENCE INTERVAL. BEST PERFORMANCE INDICATED IN BOLD.

Threshold	κ
17	0.59 [0.40, 0.78]
23	**0.76** [0.66, 0.86]
24	0.74 [0.64, 0.84]
26	0.64 [0.53, 0.75]

varies between 0.59 (95% CI [0.4, 0.77]) for a threshold of 17 and 0.76 (95% CI [0.66, 0.86]) for a threshold of 23. Table VII shows agreement values for four thresholds, 17 (lowest), 23 (best), 24 (traditional), and 26 (highest), using the rounding strategy to match thresholds. As we can see from the confusion matrix, decisions based on the MMSE scores estimated by our approach would lead to slightly more people being screened.

V. Discussion

A. *Translation Into Clinical Practice*

In principle, it is desirable to detect dementia at an early stage, so that the person with the disease and their family can take steps to maximise their quality of life. However, coming to terms with a diagnosis of dementia can be very difficult [52], [53]. Even if a person is referred for additional screening on the basis of a test such as the MMSE, and is found to be healthy, there can be negative consequences, such as people taking screening results less seriously, or becoming more anxious to bother their doctor for nothing [51], [54]. Therefore, once we have established that a machine learning approach has promise, we need to consider how it is best integrated into practice to avoid unnecessary harm.

While SVF clearly contains some information that can be useful when establishing the stage of a person's dementia, the most promising results are those for predicting MMSE scores. This makes sense clinically, as SVF does not reflect all of the dimensions on which people with dementia can be impaired, and the trajectory of decline can be very different depending on the person and the subtype of dementia they have.

At the moment, for the MMSE, we achieve good agreements with traditional judgements using manual features. Problems might arise when automating the scenario. [55] saw the performance of their classifiers deteriorate when using Automatic Speech Recognition (ASR) but this is likely to improve as ASR modules are specially developed for clinical data.

Since administering SVF requires minimal training, this makes the test ideal for deployment in a telehealth scenario. Recordings of patients can be obtained by carers, case workers, social workers, and nurses, and they can take place in a quiet room in the patient's home or a convenient clinic room. After automatic analysis, the results can be sent automatically to the patient's General Practitioner and their specialist geriatrician or old age psychiatrist.

It is even possible to fully automate the SVF test as part of an in-home kiosk or tablet app. However, for this use case, algorithms would need to be calibrated with additional training data, as people with moderate to severe dementia may find it difficult to follow the instructions of an automated app.

VI. Conclusion and Future Work

In this paper, we explored the possibility to predict MMSE and CDR-SOB scores based on linguistic and vocal features extracted from a SVF task. We were able to train a regression model with a MAE of 2.2 for the MMSE and 1.7 for the CDR-SOB. We discussed how these predictions could be used in clinical practice and that the agreement of MMSE predictions and real scores were high enough for a potential use as a screening tool. For predictions of the CDR-SOB the SVF task does not seem to capture all dimensions of impairment found in dementia.

These promising results are first steps in the direction of formulating diagnosis and cognitive assessment as a regression problem. To additionally reliably predict severity of dementia progression, in-depth analysis of more than one cognitive test might be needed.

References

[1] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World Alzheimer Report 2016 Improving Healthcare for People living with Dementia. Coverage, Quality and Costs now and in the Future," Tech. Rep., 2016.

[2] S. E. O'Bryant, L. H. Lacritz, J. Hall, S. C. Waring, W. Chan, Z. G. Khodr, P. J. Massman, V. Hobson, and C. M. Cullum, "Validation of the New Interpretive Guidelines for the Clinical Dementia Rating Scale Sum of Boxes Score in the National Alzheimer's Coordinating Center Database," *Archives of Neurology*, vol. 67, no. 6, pp. 746–749, 2010.

[3] C. P. Hughes, L. Berg, W. L. Danziger, L. A. Coben, and R. L. Martin, "A New Clinical Scale for the Staging of Dementia," *The British Journal of Psychiatry*, vol. 140, no. 6, pp. 566–572, 1982.

[4] J. C. Morris, "Clinical Dementia Rating: A Reliable and Valid Diagnostic and Staging Measure for Dementia of the Alzheimer Type," *International Psychogeriatrics*, vol. 9, no. S1, pp. 173–176, 1997.

[5] T. N. Tombaugh and N. J. McIntyre, "The Mini-Mental State Examination: A Comprehensive Review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[6] D. W. Molloy, E. Alemayehu, and R. Roberts, "Reliability of a Standardized Mini-Mental State Examination compared with the traditional Mini-Mental State Examination," *American Journal of Psychiatry*, vol. 148, no. 1, pp. 102–105, 1991.

[7] S. V. Pakhomov, L. Eberly, and D. Knopman, "Characterizing Cognitive Performance in a Large Longitudinal study of Aging with Computerized Semantic Indices of Verbal Fluency," *Neuropsychologia*, vol. 89, pp. 42–56, 2016.

[8] N. Raoux, H. Amieva, M. L. Goff, S. Auriacombe, L. Carcaillon, L. Letenneur, and J.-F. Dartigues, "Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study," *Cortex*, vol. 44, no. 9, pp. 1188–1196, 2008.

[9] S. Auriacombe, N. Lechevallier, H. Amieva, S. Harston, N. Raoux, and J.-F. Dartigues, "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*, vol. 21, no. 4, pp. 260–266, 2006.

[10] J. D. Henry and J. R. Crawford, "Verbal fluency deficits in parkinson's disease: A meta-analysis," *Journal of the International Neuropsychological Society*, vol. 10, no. 4, pp. 608–622, 2004.

[11] P. H. Robert, V. Lafont, I. Medecin, L. Berthet, S. Thauby, C. Baudu, and G. Darcourt, "Clustering and switching strategies in verbal fluency tasks: Comparison between schizophrenics and healthy adults," *Journal of the International Neuropsychological Society*, vol. 4, no. 6, pp. 539–546, 1998.

[12] J. D. Henry and J. R. Crawford, "A Meta-Analytic Review of Verbal Fluency Performance Following Focal Cortical Lesions." *Neuropsychology*, vol. 18, no. 2, pp. 284–295, 2004.

[13] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán, "Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech using ASR," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 1–5.

[14] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic Features Identify Alzheimer's Disease in Narrative Speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[15] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for Studying Discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[16] S. Wankerl, E. Nöth, and S. Evart, "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*, 2017, in press.

[17] S. Al-hameed, M. Benaissa, and H. Christensen, "Simple and robust audio - based detection of biomarkers for Alzheimer' s disease," in *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016, pp. 32–36.

[18] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[19] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's Disease: Can Temporal and Acoustic Parameters Discriminate Dementia?" *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5–6, pp. 327–334, 2014.

[20] B. Yu, T. F. Quatieri, J. R. Williamson, and J. C. Mundt, "Cognitive impairment prediction in the elderly based on vocal biomarkers," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 3734–3738.

[21] J. Tröger, N. Linz, J. Alexandersson, A. König, and P. Robert, "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*, 2017.

[22] M. Yancheva, K. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias," in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015.

[23] K. Thung, C.-Y. Wee, and P.-T. Yap, "Identification of Alzheimer's Disease Using Incomplete Multimodal Dataset via Matrix Shrinkage and Completion Chapter," in *4th International Workshop on Machine Learning in Medical Imaging, MLMI*, 2013, pp. 163–170.

[24] X. Zhu, H.-I. Suk, and D. Shen, "A Novel Matrix-Similarity Based Loss Function for Joint Regression and Classification in AD Diagnosis," *NeuroImage*, vol. 100, pp. 91–105, 2014.

[25] L. Huang, Y. Jin, Y. Gao, K. H. Thung, and D. Shen, "Longitudinal Clinical Score Prediction in Alzheimer's Disease with Soft-Split Sparse Regression based Random Forest," *Neurobiol. Aging*, vol. 46, pp. 180–191, 2016.

[26] M. K. Wolters, N. Kim, J. H. Kim, S. E. MacPherson, and J. C. Park, "Prosodic and Linguistic Analysis of Semantic Fluency data: A Window into Speech Production and Cognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016, pp. 2085–2089.

[27] P. J. Gruenewald and G. R. Lockhead, "The Free Recall of Category Examples," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 6, pp. 225–240, 1980.

[28] A. K. Troyer, M. Moscovitch, and G. Winocur, "Clustering and Switching as Two Components of Verbal Fluency: Evidence From Younger and Older Healthy Adults," *Neuropsychology*, vol. 11, no. 1, pp. 138–146, 1997.

[29] K. Ledoux, T. D. Vannorsdall, E. J. Pickett, L. V. Bosley, B. Gordon, and D. J. Schretlen, "Capturing additional information about the organization of entries in the lexicon from verbal fluency productions," *Journal of Clinical and Experimental Neuropsychology*, vol. 36, no. 2, pp. 205–220, 2014.

[30] D. L. Woods, J. M. Wyma, T. J. Herron, and E. W. Yund, "Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury," *PLOS ONE*, vol. 11, no. 12, pp. 1–37, 2016.

[31] N. Linz, J. Tröger, J. Alexandersson, and A. König, "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)*, 2017, in press.

[32] J. C. Morris, D. W. McKeel, K. Fulling, R. M. Torack, and L. Berg, "Validation of clinical diagnostic criteria for Alzheimer's disease," *Ann. Neurol.*, vol. 24, no. 1, pp. 17–22, 1988.

[33] G. G. Fillenbaum, B. Peterson, and J. C. Morris, "Estimating the validity of the clinical Dementia Rating Scale: the CERAD experience. Consortium to Establish a Registry for Alzheimer's Disease," *Aging (Milano)*, vol. 8, no. 6, pp. 379–385, 1996.

[34] S. E. O'Bryant, S. C. Waring, C. M. Cullum, J. Hall, L. Lacritz, P. J. Massman, P. J. Lupo, J. S. Reisch, R. Doody, and T. A. R. Texas

Alzheimer's Research Consortium, "Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer's research consortium study." *Arch. Neurol.*, vol. 65, no. 8, pp. 1091–1095, 2008.

[35] C. A. Lynch, C. Walsh, A. Blanco, M. Moran, R. F. Coen, J. B. Walsh, and B. A. Lawlor, "The Clinical Dementia Rating Sum of Box Score in Mild Dementia," *Dement Geriatr Cogn Disord.*, vol. 21, no. 1, pp. 40–43, 2006.

[36] B. Sheehan, "Assessment scales in dementia," *Ther Adv Neurol Disord*, vol. 5, no. 6, pp. 349–358, 2012.

[37] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-Mental State". A Practical Method for Grading the Cognitive State of Patients for the Clinician," *J Psychiatr Res*, vol. 12, no. 3, pp. 189–198, 1975.

[38] P. T. Trzepacz, H. Hochstetler, S. Wang, B. Walker, and A. J. Saykin, "Relationship between the Montreal Cognitive Assessment and Mini-mental State Examination for assessment of mild cognitive impairment in older adults," *BMC Geriatr.*, vol. 15, no. 1, pp. 107–115, 2015.

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

[40] B. MacWhinney, *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum Associates, Inc, 1991.

[41] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux *et al.*, "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*, vol. 7, no. 3, pp. 263–269, 2011.

[42] W. H. Organization, *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization, 1992.

[43] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild Cognitive Impairment: Clinical Characterization and Outcome," *Arch. Neurol.*, vol. 56, no. 3, pp. 303–308, 1999.

[44] S. T. Creavin, S. Wisniewski, A. H. Noel-Storr, C. M. Trevelyan, T. Hampton, D. Rayment, V. M. Thom, K. J. E. Nash, H. Elhamoui, R. Milligan, A. S. Patel, D. V. Tsivos, T. Wing, E. Phillips, S. M. Kellman, H. L. Shackleton, G. F. Singleton, B. E. Neale, M. E. Watton, and S. Cullum, "Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations," in *Cochrane Database of Systematic Reviews*, 2016.

[45] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 3111–3119.

[46] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta, "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora," *Language Resources and Evaluation*, vol. 43, no. 3, pp. 209–226, 2009.

[47] R. Speer, J. Chin, A. Lin, L. Nathan, and S. Jewett, "wordfreq: v1.5.1," 2016.

[48] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," http://www.fon.hum.uva.nl/praat/, accessed: 2017-03-19.

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

[50] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[51] M. Wolters, "Give me your data, and i will diagnose you," in *Data Power Conference 2017, Ottawa, CA*, 2017.

[52] B. Carpenter and J. Dave, "Disclosing a Dementia Diagnosis: A Review of Opinion and Practice, and a Proposed Research Agenda," *The Gerontologist*, vol. 44, no. 2, pp. 149–158, 2004.

[53] F. Aminzadeh, A. Byszewski, F. J. Molnar, and M. Eisner, "Emotional impact of dementia diagnosis: Exploring persons with dementia and caregivers' perspectives," *Aging & Mental Health*, vol. 11, no. 3, pp. 281–290, 2007.

[54] N. Mattsson, D. Brax, and H. Zetterberg, "To Know or Not to Know: Ethical Issues Related to Early Diagnosis of Alzheimer's Disease," *International Journal of Alzheimer's Disease*, 2010.

[55] S. V. Pakhomov, S. E. Marino, S. Banks, and C. Bernick, "Using Automatic Speech Recognition to Assess Spoken Responses to Cognitive Tests of Semantic Verbal Fluency," *Speech Communication*, vol. 75, pp. 14–26, 2015.