Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task

Alexandra König, Nicklas Linz, Johannes Tröger, Maria Wolters, Jan Alexandersson, Phillipe Robert, Alexandra Konig

▶ **To cite this version:**

Alexandra König, Nicklas Linz, Johannes Tröger, Maria Wolters, Jan Alexandersson, et al.. Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task. Dementia and Geriatric Cognitive Disorders, 2018, 45 (3-4), pp.198 - 209. 10.1159/000487852 . hal-01850408

HAL Id: hal-01850408
https://inria.hal.science/hal-01850408

Submitted on 30 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Fully automatic speech-based analysis of the semantic verbal fluency task

Alexandra König[a], Nicklas Linz[b], Johannes Tröger[b], Maria Wolters[c], Jan Alexandersson[b], Phillipe Robert[a]

[a]*Memory Clinic, Association IA, CoBTek Lab - CHU Université Côte d'Azur, France*

[b]*German Research Center for Artificial Intelligence (DFKI), Germany*
[c]*School of Informatics, University of Edinburgh, Scotland*

Running head: Automatic Analysis of Semantic Verbal Fluency

Abstract

Background: Semantic verbal fluency (SVF) tests are routinely used in screening for mild cognitive impairment (MCI). In this task, participants name as many items of a semantic category under a time constraint. Clinicians measure task performance manually by summing the number of correct words and errors. More fine-grained variables add valuable information to clinical assessment, but are time-consuming. Therefore, the aim of this study is to investigate whether automatic analysis of the SVF could provide these as accurate as manual and thus, support qualitative screening of neurocognitive impairment.

Methods: SVF data was collected from 95 older people with MCI (n=47), Alzheimer's or related dementias (ADRD; n=24) and healthy controls (HC; n=24). All data was an- notated manually and automatically with clusters and switches. The obtained metrics were validated using a classifier to distinguish HC, MCI and ADRD.

Results: Automatically extracted clusters and switches were highly correlated (r=0.9) with manually established values, and performed as well on the classification task separating healthy controls from persons with Alzheimer's (AUC=0.939) and MCI (AUC=0.758).

Conclusion: The results show that it is possible to automate fine-grained analyses of SVF data for the assessment of cognitive decline.

Keywords: Alzheimer's disease, dementia, Mild Cognitive Impairment (MCI), Neuropsychology, Assessment, Semantic Verbal Fluency, Speech recognition, Speech

processing, Machine Learning

Address of corresponding author:
Alexandra König
CoBTeK (Cognition Behaviour Technology) Research Lab - Université Côte d'azur ,

Centre Mémoire de Ressources et de Recherche- CHU de Nice, Institut Claude Pompidou, 10 rue Molière 06100 Nice, France; Tel.: +33 652021156, E-mail: alexandra.konig@inria.fr

1. Introduction

As life expectancy across the globe increases, the incidence of age-related cognitive impairment is soaring. Relevant current research focuses on early intervention to slow the progression of cognitive decline with a long-term goal of helping to find a cure for (reduce the occurrence of) Alzheimer's Disease and other dementias [1-3]. It has been demonstrated that prevention at prodromal stages targeting disease-modifying risk factors show promising results and are more likely to be effective [4].

While a full assessment of cognitive function requires a trained clinician, the increasing prevalence of dementia and milder forms of cognitive impairment warrant large-scale screening of the population. Even in high-income countries, as many as 50% of all relevant cases remain undiagnosed [5]. New approaches to screening and monitoring are needed [6,7].

In order to address this problem, we need new tools that are fast, do not need a laboratory, and can automatically indicate which patients might need to be referred to a specialist [8]. Such tools are highly scalable, and can be made accessible to health care professionals with little to no specialised training in old age psychiatry. Ideally, it should be possible to administer them remotely, and they should integrate easily with existing telehealth and telecare solutions for older patients. Automated analysis of speech, in particular speech that is produced during a standard clinical assessment, might be a prime candidate for such a tool [9-14]. Several research groups demonstrated the interest of adopting an automated approach to speech analysis for clinical assessment of older people [15-19]. Overall, reported work either uses speech from conversations, spontaneous speech tasks, reading or repetition tasks, and fluency tasks.

However, if natural language is analysed, considerable effort has to be spent on preprocessing the data, e.g. annotating turns, or trimming the audio file, in order to prepare it for further computational learning which is not useful for an application in daily clinical practice. Moreover, in order to detect in speech early subtle changes of cognition, it seems crucial to induce a minimum of cognitive effort in a vocal task [15;

20].

Category fluency, or semantic verbal fluency (SVF) task, requires the verbal production of as many different items from a given category, e.g. animals, as possible in a given time period. The SVF task is one of the most widely used neuropsychological tests comprising both executive control and semantic memory retrieval processes. . They are relatively short and part of standard dementia screens such as the Addenbrookes Cognitive Examination- Revised (ACE-R) [21] and often used in assessing cognitive function in older people [22-24]. SVF performance can distinguish people with dementia from healthy controls and people with mild cognitive impairment [25-29], and additionally can be sensitive to early phases of neurodegenerative disease [30].

Most studies of SVF performance use the total number of correct words produced. However, in order to differentiate between multiple pathologies and gain detailed information on underlying cognitive processes, more elaborate measures have been established which serve as additional indicators [31,32].

One prominent approach, popularised by Troyer and collaborators [32,33], assumes two processes are involved in the production of SVF word sequences, the lexical search for a word from the category to be produced, and the retrieval of other lexical items that are semantically related to the original word. The temporal sequences of semantically related words are called clusters, and the executive search process between clusters is called switching. Typically, semantic clusters are determined using predefined semantic subcategories, often according to Troyer et al. (1997) [32]. After determining cluster boundaries, the mean size of clusters and the number of switches between clusters are computed. Various parameters related to the size of clusters and number of switches have been shown to be sensitive to cognitive decline and differentiate between different types of dementia.

Unfortunately, any analysis of SVF data that goes beyond word counts is too time consuming for daily clinical practice, especially for general practitioners and family physicians, who are typically the first point of contact for people who suspect that they or one of their loved ones has a cognitive impairment. In addition, any analysis strategy that is based on fixed, pre-defined categories is open to subjective judgement. This might explain some of the variation in cluster sizes and switch counts reported in the literature [33-36; 27-28].

While automatic analysis introduces its own systematic biases, it is objective, replicable and yields almost immediate results for clinicians to act on. Thus, computational approaches to analyse the SVF task have been proposed [37-39] for which statistical methods have been applied in order to obtain semantic clusters. Pakhomov et al. [39], as well as Ledoux et al. (2014) use Latent Semantic Analysis (LSA), to compute strength of semantic relations between pairs of words produced [40],. Woods et al. (2016) [37], use Explicit Semantic Analysis (ESA) [41] —a vector embedding trained on co-occurrence of words in Wikipedia articles—to identify chaining behaviour for different demographics based on pairwise cosine similarity. Clark et al. (2016) [38] propose novel semantic measure based on graph theory; most prominently they put forward graph-based coherence measures which compare the patient's created sequence/path of words with the "shortest" possible path through the fully connected weighted graph of all patient's words. Neural word embeddings based on large word2vec models [41] allow to directly measure the semantic distance between two given words using simple geometry in the created vector space [42]. In terms of scalability and feasibility for parallel version of categories, qualitative SVF analysis based on computational semantics may represent a promising step forward. However, before this method could make its way into daily clinical practice it should be demonstrated to provide reliable and valid data for a regular use.

For this, we set out in this research, to investigate whether fully automatic analysis of the SVF task can be (1) considered as reliable as the manual one, (2) can be used for automatic qualitative assessment of neurocognitive impairment within this task and the corresponding domain and (3) in the end could be used as a valid fast and scalable screening tool, based on a classification experiment.

2. Methods

2.1. Recruitment

Within the framework of a clinical study carried out for the European research project *Dem@care*, and the EIT-Digital project *ELEMENT*, speech recordings were conducted at the Memory Clinic located at the Institut Claude Pompidou and the University hospital in Nice, France. The Nice Ethics Committee approved the study.

Each participant gave informed consent before the assessment. Speech recordings of participants were collected using an automated recording app which was installed on an iPad. The application was provided by researchers from the University of Toronto, Canada and the company Winterlight Labs.

2.2. Clinical Assessments

Each participant underwent the standardised process used in French Memory clinics. After an initial medical consultation, with a geriatrician, neurologist or psychiatrist a neuropsychological assessment was performed.
Following this, participants were categorized into three groups: Control participants (HC) that were diagnosed as cognitively healthy after the clinical consultation, patients with mild cognitive impairment (MCI), and patients that were diagnosed as suffering from Alzheimer's Disease and related disorders (ADRD). For the ADRD and MCI group, the diagnosis was determined using the ICD-10 classification of mental and behavioural disorders [44]. Participants were excluded if they were not native speaker or had any major hearing or language problems, history of head trauma, loss of consciousness, addiction including alcoholism, psychotic or aberrant motor behaviour or prescribed medication influencing psycho-motor skills.
The cognitive assessment included (among others) the Mini-Mental State Examination (MMSE) [45], phonemic verbal fluency (letter 'f'), semantic verbal fluency (animals), and the Clinical Dementia Rating Scale [46].

Each participant performed the SVF task during a regular consultation with one of the Memory Center's clinicians who operated the mobile application. For the Dem@care data, the vocal tasks were recorded with an external microphone attached to the patients shirt and for the ELEMENT data, with the internal microphone. Instructions for the vocal tasks *("Pouvez-vous me dire le plus possible de noms d'animaux pendant une minute ?/Can you please give me in one minute as many animal names as you can think of?")* were pre-recorded by one of the psychologist of the center ensuring

standardised instruction over both experiments. Administration and recording were controlled by the application and facilitated the assessment procedure.

2.3. Speech data processing and transcription

Recordings of patients were analysed manually and automatically. For manual analysis, a group of trained speech pathology students transcribed the SVF performances following the CHAT protocol[47] and aligned the transcriptions with the speech signal using PRAAT [48]. For the automatic transcription, the speech signal was separated into sound and silent parts using a PRAAT script based on signal intensity. The sound segments were then analysed using Google's Automatic Speech Recognition (ASR) service, which returns several possible transcriptions for each segment together with a confidence score. The list of possible transcriptions was searched for the one with the maximum number of words that were in a predefined list of animals in French. In case of a tie, the transcription with the higher confidence score was chosen.

2.4. Features

Word count was defined as the number of animal names produced minus the number of repetitions.

Clusters were determined based on statistical word embeddings, a commonly used technique in computational linguistics, calculated with word2vec [42] based on the french FraWac corpus [49] as described in Linz et al. (2017) [43]. Let $a_1 \ldots a_n$ 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

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

with

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

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

. Mean cluster size was computed as the average number of words per cluster, and the number of switches was the number of clusters - 1.

2.5. Classification

In order to evaluate the feasibility of the automatic approach, we performed two

analyses that aimed to replicate existing results in the literature on differences in semantic verbal fluency performance between people with no impairment, mild neurocognitive impairment/MCI, and major neurocognitive impairment/AD [34,50]. The first used a staging approach using validated normative data provided by St-Hilaire et al. (2016) [50], and the second used machine learning (ML) classifiers.

2.5.1. Automatic norm-based neurocognitive evaluation

For simulation of a real world clinical application scenario, word counts from manual and automatic transcripts were compared using normative data for SVF. First, normative equations[50] were used to determine a z-value, based on manual word counts, age and education level, and people were staged in accordance with diagnostic categories of DSM-5 (z > -1 = no impairment, z > -2 = minor impairment, z <= -2
= major impairment). In a second step, people were staged using the normative equations, based on automatic word count, age and education level. The first staging was considered the ground truth and the second was compared to the first using classification metrics.

2.5.2. ML automatic diagnosis classification

To give an idea of how the collected features could be combined to make a diagnostic decision, a ML classifier was trained. Each person in the database was assigned to a label relating to his or her diagnosis (HC, MCI and ADRD). The features described in Section 2.4 were used, either calculated from automatic or manual transcripts, depending on the scenario. In all scenarios we use Support Vector Machines (SVMs) [51] implemented in the scikit-learn framework [52]. Leave-one-out cross validation was used for testing. In this procedure the data is split into 2 subsets (”folds”). One fold contains only one sample, the other contains all other samples. For each of the folds, the classifier is trained on the second fold and evaluated on the held out sample. 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. All features were normalised using z-standardisation, based on the training fold of each iteration.

2.6. Performance Measures

The performance of ASR systems is usually determined using Word Error Rate (WER) as a metric. WER is a combination of the mistakes made by ASR systems in the process of recognition. Mistakes are categorised into substitutions, deletions and intrusions. Let S, D and I be the count of these errors respectively, and N be the number of tokens in the ground truth. Then

$$WER = \frac{S + D + I}{N}$$

We only calculated WER for words describing animals, not for off-task speech, which also occurs in our data. We refer to this metric as VFER (Verbal Fluency Error Rate).

As performance measures for prediction of each class in the ML classification experiment, we report the receiver operator curve (ROC), as different trade-offs, between sensitivity and specificity are visible. We also report area under curve (AUC) as an overall performance metric.

3. Results

3.1. Participants characteristics

Relevant demographic characteristics of the HC group (n = 24, age 76.12 years, MMSE 28.21, CDR-SOB 0.46), the MCI group (n = 47, age 76.59 years, MMSE 26.02, CDR-SOB 1.68), and the ADRD group (n = 24, age 77.7 years, MMSE 18.83, CDR-SOB 7.5) are presented in Table 1. The total number of participants was 95. Excluding MMSE and CDR-SOB, no significant effects between the groups were found.

---- Table 1 ----

3.2. Automated Speech Recognition (ASR)

Evaluation of all samples in the corpus yielded a VFER of 20.01 %. Since not all types of errors might have the same impact on analysis (e.g. word count is not influenced by substitutions in all cases), the proportion of types of error made are considered. 50.3% of all errors were deletions, 29.8 % were substitutions and 19.9 % were insertions.

3.3. Correlation

The relationship between features extracted from automated transcripts and manual ones was examined.

Consequently, Spearman's correlation coefficient was computed. All relationships are reported in Figure 2. The correlation between manual and automatic SVF analysis was strong across all three relevant features with a correlation of $\rho = 0.921$ for the main clinical feature in this task, the word count.

3.4. Automatic norm-based neurocognitive evaluation

Neurocognitive disorder evaluations (no impairment, minor and major impairment) determined with the automatic word count, agree with labels based on the manual WC with an accuracy of 0.831, weighted precision of 0.83, weighted recall of 0.83 and F_1 of 0.83. When looking at sensitivity and specificity in a one versus all scenario, using HC as the negative class, the model achieves a sensitivity of 0.914 and a specificity of 0.833. A detailed confusion matrix is depicted in Figure 1.

3.5. ML automatic diagnosis classification

ROC curves for all scenarios are reported in Figure 3. Classifiers trained on automatic measures and manual ones perform comparably or better for two of three scenarios.

4. Discussion

In this paper, we describe an automated analysis method for the fine-grained analysis of SVF data in terms of clusters and switches and validate it for the category of animals. Clusters and switches, determined by the tool correlate well with clusters and switches that were determined manually using a strict annotation procedure. Both manually and automatically derived statistics were successful in distinguishing between healthy controls, people with mild cognitive impairment and people with Alzheimer's disease or related disorders.

4.1. Automated Speech Recognition (ASR)

Considering the reliability of the fully automated pipeline, the automated speech recognition (ASR) is often considered to be the main limiting factor [19]. Our results show an overall low error rate of 20.01 % for the automated system, compared to the manual transcripts. This in itself represents an improvement over results of other

authors using ASR systems for evaluating the SVF tasks [53,54]. In line with previous research, diagnostic groups differ significantly in the number of errors made by ASR (Kruskall-Wallis, χ^2= 13.7, df = 2, p < 0.001). More word errors are produced by the ASR for AD patients, compared to healthy subjects. Since persons with AD are expected to produce less words in an SVF task, this does not negatively affect further analysis. Closely looking at the types of errors, insertions and deletions are both problematic for further analysis. Both skew the raw word count, which still is the single most predictive performance indicator in SVF for dementia detection. Substitutions only affect qualitative measures such as the mean size of clusters and the number of switches between clusters, but do not effect the word count.

4.2. Automatic norm-based neurocognitive evaluation

Even though the ASR produced word errors, mainly deletions, which negatively affect the overall word count and thereby the main clinical measure of SVF, the correlation between the automated and manual systems is very strong, i.e. 0.921. This shows that although the ASR system introduces some errors, it does not greatly affect the overall clinical measure, since the errors are not correlated to cognitive status. In the first experiment, we benchmarked the automatic pipeline for a norm-based neurocognitive evaluation. The performed neurocognitive evaluation based on automatic word count agreed strongly with labels based on the manual word count. The confusion matrix (see Figure 1) shows that the automatic approach tends to systematically underestimate the performance of a person in the SVF task. This can be attributed to the deletions of the ASR. Thus, the automatic pipeline can be considered conservative, showing high sensitivity, which is of great importance to its use as a screening tool.

4.3. Automated ML diagnosis classification

For both the HC vs. AD and HC vs. MCI scenarios the performances of models trained on automatic and manual features have comparable AUC (0.723 vs. 0.758 and 0.953 vs. 0.939). In the MCI vs. AD scenario the AUC of models trained on automated features deteriorated (0.859 vs. 0.774). The difference of the previous experiment can be explained by the flexibility of ML models to learn decision boundaries, in contrast to pre-determined diagnostic norms. ML models are also able to accommodate the previously mentioned systematic errors of ASR.

A similar approach has been suggested by Clark et al (2016) [38], studying the utility of an automatic SVF score for the prediction of conversion with the result that higher prediction accuracy was obtained with the classifiers trained on all scores, rather than on manual scores. Overall, it can be stated that using automatic analysis of the SVF task allows immediate access to reliable and clinically relevant measures such as the word count, switches and clusters. This is potentially useful for differentiating between deficits in either executive or semantic processing. The automation of recording, transcription and analysis streamlines test administration and ultimately leads to more robust, reproducible data.

In addition to the assessment of cognitive decline, these qualitative measures extracted from the SVF performances may be of great interest as well for other neurocognitive disorders affecting verbal ability and executive control such as fronto-temporal dementia or primary progressive aphasia[55].

Costa et al. [30] states that we are far from having available reliable tools for the assessment of dementias, since one of the main problems is the heterogeneity of the tools used across different countries. Therefore, a working group of experts recently published recommendations for the harmonisation of neuropsychological assessment of neurodegenerative dementias with the aim to achieve more reliable data on the cognitive-behavioural examination. Automated speech analysis of the SVF could be one potential tool to assist in harmonising test procedures and outcomes. It also provides additional quantitative measurements extracted from speech signals for cognitive screening without increasing time, costs or even workload for the clinician. Such a tool could be used as an endpoint measurement in clinical trials to assess intervention outcome and monitor disease progress, even remotely over the phone.

4.4. Limitations

A few limitations of this study should be considered. We did not recruit healthy participants from the general elderly population, but were limited to include persons who came for clinical consultation to the memory clinic cognitively healthy but with some subjective complaints It should be further noted that the data set for this study is only in French, thus, limiting transferability of its results to other languages. A major goal for future work is the collection of SVF recordings in multiple languages and within the framework of longitudinal studies..

4.5. Conclusion

To conclude, the study demonstrates the feasibility of automatic analysis of SVF performances in elderly people to assess and monitor cognitive impairment. Furthermore, new measures beyond simple word counts such as word frequencies could be investigated in the future, possibly giving way to a deeper understanding of underlying cognitive functions and changes due to neurodegenerative disease.

Acknowledgements

This research 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.

Conflicts of Interests

The authors have no conflict of interest to declare.

References

1. Bateman, RJ, Xiong, C, Benzinger, TL, Fagan, AM, Goate, A, Fox, NC, Marcus, DS, Cairns, NJ, Xie, X, Blazey, TM, Holtzman, DM, Santacruz, A, Buckles, V, Oliver, A, Moulder, K, Aisen, PS, Ghetti, B, Klunk, WE, McDade, E, Martins, RN, Masters, CL, Mayeux, R, Ringman, JM, Rossor, MN, Schofield, PR, Sperling, RA, Salloway, S, Morris, JC: Clinical and biomarker changes in dominantly inherited Alzheimer's Disease. N Engl J Med 2012; 367(9): 795–804.
2. Langbaum, JB, Fleisher, AS, Chen, K, Ayutyanont, N, Lopera, F, Quiroz, YT, Caselli, RJ, Tariot, PN, Reiman, EM: Ushering in the study and treatment of preclinical Alzheimer Disease. Nat Rev Neurol 2013; 9(7): 371–381.
3. Sperling, RA, Aisen, PS, Beckett, LA, Bennett, DA, Craft, S, Fagan, M, Iwatsubo, T, Jack, CR, Kaye, J, Montine, TJ, Park, DC, Reiman, EM, Rowe, CC, Siemers, E, Stern, Y, Yaffe, K, Carrillo, MC, Thies, B, Morrison-Bogorad, M, Wagster, MV, Phelps, CH: 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 2011; 7(3): 280–292.

4. Sindi, S, Mangialasche, F, and Kivipelto, M: Advances in the prevention of Alzheimer's Disease. F1000Prime Rep 2015; 7: 50.

5. Prince, M, Comas-Herrera, A, Knapp, M, Guerchet, M, Karagiannidou, M: World Alzheimer Report 2016 Improving healthcare for people living with dementia. Coverage, Quality and Costs now and in the Future. Tech. rep 2016.

6. Laske, C, Sohrabi, HR, Frost, SM, Lopez-de Ipina, K, Garrard, P, Buscema, M, Dauwels, J, Soekadar, SR, Mueller, S, Linnemann, C, Bridenbaugh, SA, Kanagasingam, Y., Martins, RN, O'Bryant, SE: Innovative diagnostic tools for early detection of Alzheimer's Disease. Alzheimers Dement 2015; 11(5): 561–578.

7. Snyder, PJ, Kahle-Wrobleski, K, Brannan, S, Miller, DS, Schindler, RJ, DeSanti, S, Ryan, JM, Morrison, G, Grundman, M, Chandler, J, Caselli, RJ, Isaac, M, Bain, L, Carrillo, MC: Assessing cognition and function in Alzheimer's Disease clinical trials: do we have the right tools? Alzheimers Dement 2014; 10(6): 853–860.

8. Tröger, J, Linz, N, Alexandersson, J, König, A, Robert, P: 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, May 23–26, 2017, Barcelona, Spain.

9. König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., David, R., 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's Disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1 (1), 112–124.

10. Satt, A, Hoory, R, König, A, Aalten, P, Robert, PH: Speech-based automatic and robust detection of very early dementia. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTER-SPEECH. pp. 2538–2542. September 14-18, 2014, Singapore.

11. Hoffmann, I, Nemeth, D, Dye, CD, Pákáski, M, Irinyi, T, Kálmán, J: Temporal parameters of spontaneous speech in Alzheimer's Disease. Int J Speech Lang Pathol 2010; 12(1): 29–34.

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

13. Lopez-de Ipina, K, Martinez-de Lizarduy, U, Calvo, P, Mekyskac, J, Beitiad, B, Barroso, N, Estanga, A, Tainta, M, Ecay-Torres, M: Advances on automatic speech analysis for early detection of Alzheimer Disease: a non-linear multi-task approach. Curr Alzheimer Res 2017; 15(2): 139-148.

14. Tóth L, Hoffmann I, Gosztolyac G, Vinczec V, Szatlóczkid G, Bánrétib Z: A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Curr Alzheimer Res 2018; 15(2): 130-38.

15. König, A, Satt, A, Sorin, A, Hoory, R, Derreumaux, A, David, R, & Robert, PH: Use of speech analyses within a Mobile Application for the assessment of cognitive impairment in elderly people. Curr Alzheimer Res 2018; 15(2): 120–129.

16. Fraser, KC, Meltzer, JA, & Rudzicz, F: Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimers Dis 2015; 49(2): 407–422.

17. Lopez-de-Ipiña K, Alonso, JB, Travieso, CM, Solé-Casals, J, Egiraun, H, Faundez-Zanuy, M, Ezeiza, A, Barroso, N, Ecay, M, Martinez-Lage, P, Martinez-de-Lizardui, U: On the selection of non-invasive methods based on speech analysis oriented to Automatic Alzheimer Disease Diagnosis. Sensors 2013; 13(5): 6730-6745.

18. Meilan JJG, Martinez-Sanchez F, Carro J, Carcavilla N, Ivanova O. Voice markers of lexical access in mild cognitive impairment and Alzheimer's Disease. Curr Alzheimer Res. 2018; 15(2): 111-119.

19. Tóth L, Gosztolya G, VinczeV, Hoffmann I, Szatlóczki G, Biró E, *et al*. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In Proceedings: Interspeech 2015 Tutorials & Main Conference, Germany: Dresden.

20. Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M.: Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front Aging Neurosci 2015; 7: 195.

21. Mioshi, E, Dawson, K, Mitchell, J, Arnold, R, Hodges, JR: TheAddenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. Int J Geriatr Psychiatry 2006; 21(11): 1078–1085.

22. Peter, J, Kaiser, J, Landerer, V, Kostering, L, Kaller, CP, Heimbach, B, Hull, M, Bormann, T, Kloppel, S: Category and design fluency in mild cognitive impairment: performance, strategy use, and neural cognitive correlates. Neuropsychologia 2016; 93 (Pt A): 21–29.

23. Canning, SJ, Leach, L, Stuss, D, Ngo, L, Black, SE: Diagnostic utility of abbreviated fluency measures in Alzheimer Disease and vascular dementia. Neurology 2004; 62(4): 556–562.

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

25. Clark, LJ, Gatz, M, Zheng, L, Chen, YL, McCleary, C, Mack, WJ, Longitudinal

verbal fluency in normal aging, preclinical, and prevalent Alzheimer's Disease. Am J Alzheimers Dis Other Demen 2009; 24(6): 461–468.

26. Henry, JD, Crawford, JR, Phillips, LH: Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. Neuropsychologia 2004; 42(9): 1212–1222.
27. Pakhomov, SV, Eberly, L, Knopman, D: Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. Neuropsychologia 2016; 89: 42–56.
28. Raoux, N, Amieva, H, Le Goff, M, Auriacombe, S, Carcaillon, L, Letenneur, L, Dartigues, JF: Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study. Cortex 2008; 44(9): 1188–1196.
29. Auriacombe, S, Lechevallier, N, Amieva, H, Harston, S, Raoux, N, Dartigues, J.-F: A longitudinal study of quantitative and qualitative features of category verbal fluency in incident Alzheimer's Disease subjects: results from the PAQUID study. Dement Geriatr Cogn Disord 2006; 21(4): 260–266.
30. Costa, A, Bak, T, Caffarra, P, Caltagirone, C, Ceccaldi, M, Collette, F, Crutch, S, Della Sala, S, Demonet, JF, Dubois, B, Duzel, E, Nestor, P, Papageorgiou, SG, Salmon, E, Sikkes, S, Tiraboschi, P, van der Flier, WM, Visser, PJ, Cappa, SF: The need for harmonisation and innovation of neuropsychological assessment in neurodegenerative dementias in Europe: consensus document of the Joint Program for Neurodegenerative Diseases Working Group. Alzheimers Res Ther 2017; 9(1): 27.
31. Gruenewald, PJ, Lockhead, GR: The Free Recall of Category Examples. Journal of Experimental Psychology: Human Learning and Memory 1980; 6: 225–240.
32. Troyer, A. K., Moscovitch, M., Winocur, G., 1997. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology 11 (1), 138–146.
33. Troyer, AK, Moscovitch, M, Winocur, G, Leach, L, Freedman, M: Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. J Int Neuropsychol Soc 1998; 4(2), 137–143.
34. Murphy, KJ, Rich, JB, Troyer, AK: Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia. J Int Neuropsychol Soc 2006; 12(4): 570–4.
35. Gomez, RG, White, DA: Using verbal fluency to detect very mild dementia of the Alzheimer type. Arch Clin Neuropsychol 2006; 21(8): 771–775.
36. Mueller, KD, Koscik, RL, LaRue, A, Clark, LR, Hermann, B, Johnson, SC., Sager, MA: Verbal fluency and early memory decline: results from the Wisconsin registry for

Alzheimer's Prevention. Arch Clin Neuropsychol 2015; 30(5): 448.

37. Woods, DL, Wyma, JM, Herron, TJ, Yund, EW: Computerized analysis of verbal fluency: normative data and the effects of repeated testing, simulated malingering, and traumatic brain injury. PLOS ONE 2016; 11(12): 1–37.

38. Clark, DG, McLaughlin, PM, Woo, E, Hwang, K, Hurtz, S, Ramirez, L, Eastman, J, Dukes, RM, Kapur, P, DeRamus, TP, Apostolova, LG: Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. Alzheimers Dement (Amst) 2016; 2: 113–122.

39. Pakhomov, SV, Hemmy, LS: A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. Cortex 2014, 55: 97-106.

40. Ledoux, K, Vannorsdall, TD, Pickett, EJ, Bosley, LV, Gordon, B, Schretlen, DJ: Capturing additional information about the organi-zation of entries in the lexicon from verbal fluency productions. J Clin Exp Neuropsychol 2014; 36(2): 205–220.

41. Gabrilovich, E, Markovitch, S: Wikipedia-based semantic interpretation for natural language processing. J Artif Int Res 2009; 34(1): 443– 498 (URL http://dl.acm.org/citation.cfm?id=1622716.162272).

42. Mikolov, T, Sutskever, I, Chen, K, Corrado, GS, Dean, J: Distributed representations of words and phrases and their compositionality. In: Ad- vances in Neural Information Processing Systems 2013; 26: 3111–3119.

43. Linz, N, Tröger, J, Alexandersson, J, König, A: 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). In press. September 19-22, 2017, Montpellier, France.

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

45. Folstein, MF, Folstein, SE, McHugh, PR: "Mini-Mental State". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975; 12(3): 189–198.

46. O'Bryant, SE, Lacritz, LH, Hall, J, Waring, SC, Chan, W, Khodr, ZG, Massman, PJ, Hobson, V, Cullum, CM: 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 2010; 67(6): 746–749.

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

48. Boersma, P, Weenink, D: Praat, a system for doing phonetics by computer. Glot international 2001; 5: 341–345.

49. Baroni, M, Bernardini, S, Ferraresi, A, Zanchetta, E: The WaCky Wide Web: A collection of very large linguistically processed web- crawled corpora. Language Resources and Evaluation 2009; 43(3): 209–226.

50. St-Hilaire, A, Hudon, C, Vallet, GT, Bherer, L, Lussier, M, Gagnon, JF, Simard, M, Gosselin, N, Escudier, F, Rouleau, I, Macoir, J: Normative data for phonemic and semantic verbal fluency test in the adult French-Quebec population and validation study in Alzheimer's disease and depression. Clin Neuropsychol 2016. 30(7): 1126–1150.

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

52. 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, Duchesnay, E: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12: 2825–2830.

53. Pakhomov, SV, Marino, SE, Banks, S, Bernick, C: Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency. Speech Communication 2015; 5:14–26.

54. Lehr, M, Prud'hommeaux, E, Shafran, I, Roark, B: Fully automated neuropsychological assessment for detecting Mild Cognitive Impairment. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2012; 1039–1042.

55. van den Berg, E, Jiskoot, L, Grosveld, JH, van Swieten, MC, Papma, J: 2017. Qualitative assessment of verbal fluency performance in frontotemporal dementia. Dement Geriatr Cogn Disord 2017; 44: 35–44.

Table 1: Demographic data and clinical scores by diagnostic group; mean (standard deviation); HC='Healthy control', MCI='Mild cognitive impairment', ADRD= 'Alzheimer's disease and related disorders'. Significant difference ($p < 0.05$) from the control population in a Wilcoxon-Mann-Whitney test are marked with ∗

	HC	MCI	ADRD
N	24	47	24
Age	76.12 (4.41)	76.59 (7.6)	77.7 (3.99)
Sex	5M/19F	23M/24F	8M/16F
Education	10.50 (4.05)	10.81 (3.6)	9.75 (4.69)
MMSE	28.21 (1.82)	26.02* (2.5)	18.83* (4.99)
CDR-SOB	0.46 (0.67)	1.68* (1.11)	7.5* (3.7)

F**igure 1:** Confusion matrix for diagnosis based on normative data, automatic WC and manual WC; (no='z > −1 no impairment', minor='z > −2 minor impairment', major='z <= −2 major impairment')

(attached as separate file)

Figure 2: Correlation matrix and scatter plots for features based on manual and automatic transcripts. Spearman's correlation coefficients are reported. WC = "Word Count", MCS = "Mean Cluster Size", NOS = "Number of Switches". Diagnostic groups are encoded on the scatter plot as HC = "blue triangles", MCI = "green circles", AD = "red squares".

(attached as separate file)

Figure 3: Receiver Operator Curve (ROC) of classification models for different scenarios. Models trained on manually extracted features are displayed as dotted lines, ones based on automatic features are displayed as solid lines. Color indicates the classification scenario, as coded in the legend. Area under the curve (AUC) reported in the legend for each scenario and feature set.

(attached as separate file)