Temporal Analysis of the Semantic Verbal Fluency Task in Persons with Subjective and Mild Cognitive Impairment

Nicklas Linz[1], Kristina Lundholm Fors[2], Hali Lindsay[1], Marie Eckerström[2], Jan Alexandersson[1], Dimitrios Kokkinakis[2]
[1]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[2]University of Gothenburg, Gothenburg, Sweden
nicklas.linz@dfki.de, kristina.lundholmfors@gu.se, hali.lindsay@dfki.de
marie.eckerstrom@neuro.gu.se, jan.alexandersson@dfki.de, dimitrios.kokkinakis@gu.se

Abstract

The Semantic Verbal Fluency (SVF) task is a classical neuropsychological assessment where persons are asked to produce words belonging to a semantic category (e.g., animals) in a given time. This paper introduces a novel method of temporal analysis for SVF tasks utilizing time intervals and applies it to a corpus of elderly Swedish subjects (mild cognitive impairment, subjective cognitive impairment and healthy controls). A general decline in word count and lexical frequency over the course of the task is revealed, as well as an increase in word transition times. Persons with subjective cognitive impairment had a higher word count during the last intervals, but produced words of the same lexical frequencies. Persons with MCI had a steeper decline in both word count and lexical frequencies during the third interval. Additional correlations with neuropsychological scores suggest these findings are linked to a person's overall vocabulary size and processing speed, respectively. Classification results improved when adding the novel features ($AUC = 0.72$), supporting their diagnostic value.

1 Introduction

Verbal fluency is a widely adapted neuropsychological test. Historically, Schiller (1947) used the "spontaneous naming by free association"-test for the assessment of brain injuries, becoming one of the first recorded instances of what would later be referred to as "category fluency". Category fluency, or semantic verbal fluency (SVF), requires the verbal production of as many different items from a given category, e.g., animals, as possible within a given timeframe. A large body of evidence substantiates the discriminative power of semantic verbal fluency for dementia due to Alzheimers Disease (AD) and its precursor Mild Cognitive Impairment (MCI) (Henry et al., 2004; Auriacombe et al., 2006; Gomez and White, 2006; Raoux et al., 2008; Linz et al., 2017).

As there is currently no cure for AD, preventive medication labeled to delay the onset or worsening of symptoms is the primary course of action, with an emphasis on early intervention being a beneficial factor for effective treatment. Early identification of subtle symptoms is also valuable for drug trial screening programs and supports early behavioral interventions that can delay the onset of the disease (Ashford et al., 2007; Zucchella et al., 2018).

SVF has been used to identify the early stages of dementia through traditional crude measures, such as the total number of unique words produced. This may overlook persons with very subtle cognitive impairment because they lack statistically significant differences from healthy controls. Thus, additional sensitive measures of performance are needed. Further analysis of SVF has often looked at the production as a series of *clusters* and *switches*, where a cluster is a group of semantically similar words (e.g. pets such as 'cat', 'dog' and 'hamster') and a switch is the task of changing semantic focus from one group of animals to another (e.g. switching from enumerating pets to producing animals that live in Africa) (Troyer et al., 1997). Authors have also suggested approaches to clustering and switching that solely rely on temporal information (Troeger et al., 2019).

SVF has been shown to be a valid measure of executive function and verbal ability, specifically vocabulary size and lexical access speed (Shao et al., 2014). It has been sug-

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 103–113
Minneapolis, Minnesota, June 6, 2019. ©2019 Association for Computational Linguistics

gested that word production in SVF is moderated by different cognitive processes over time, where the initial process is a semi-automatic retrieval of commonly used and readily available words, whereas later stages demand more effortful processing (Demetriou and Holtzer, 2017).

In this paper, we examine SVF results of three groups of Swedish participants; those with Subjective Cognitive Impairment (SCI), with MCI and healthy controls (HC). By analysing the data temporally, we are able to reveal differences that are not evident when looking at the SVF as a whole. This paper is structured in the following way: An overview of related work is given, with a focus on performance on the SVF by persons with MCI and SCI. Then the dataset and methodology are described as well as the features that were extracted. Finally, the results of our analyses and machine learning experiments are presented and discussed in tandem with other relevant neuropsychological metrics.

2 Related work

Performance of SVF tasks in healthy older adults tends to decline with age, and is partially attributed to a decrease in processing speed, rather than a diminished verbal knowledge (Elgamal et al., 2011). In line with this reasoning, Tallberg et al. (2008) found that the performance of Swedish speakers on SVF is negatively correlated with age and positively correlated with years of education. Healthy participants in the age range 65-89 with $\leq$12 years of education produced a mean of 14.9±6.4 animals, whereas those in the same age range but with an education of >12 years produced 19.4±5.6 animals in the same task.

The deterioration of cognition in MCI, with impairment both in processing speed and switching attention (Ashendorf et al., 2008), results in persons with amnestic MCI (aMCI) producing smaller clusters and fewer switches than healthy controls (Peter et al., 2016). This reduction across strategy generalises to persons with aMCI producing significantly less categorical words (Price et al., 2012; Mueller et al., 2015).

Nikolai et al. (2018) found categorical differences between naming animals and vegetables when comparing participants with SCI and HC on the SVF test. While the animal category revealed no differences, persons with SCI generated significantly fewer vegetables, specifically in the later 30 seconds. Participants with SCI produced smaller clusters and made more switches in the animal category. The groups did not differ significantly on any demographic variables (age, education, gender) or on the Mini-Mental State Examination (MMSE; Folstein et al. (1975)).

Throughout the SVF, word production rate decreases regardless of the presence of cognitive impairment. To further explore the performance of persons with MCI and healthy controls, Demetriou and Holtzer (2017) divided and analyzed the task into three 20-second sections with two substantial findings; both groups declined over time and generated more words in the first time span. However, persons with MCI performing within normal limits produced fewer words in the first time interval. Slow initiation of lexical search process suggests that MCI inhibits early semi-automatic word retrieval processes. This is in line with previous research showing that the last 30 seconds of the verbal fluency task does not differ between participants, whereas the first 30 seconds contain discriminating information (Fernaeus et al., 2008).

When performing an even finer-grained temporal analysis based on ten second intervals, Fernaeus et al. (2008) found that intervals 1 and 2 were useful in distinguishing persons with AD and MCI, and interval 3 made it possible to differentiate between persons with MCI and SCI, and MCI and AD respectively.

3 Methods

3.1 Recruitment and Data Acquisition

All the participants in the current study on ”Linguistic and extra-linguistic parameters for early detection of cognitive impairment” were recruited from the Gothenburg MCI study (Wallin et al., 2016). All participants were speakers of Swedish, selected according to detailed inclusion and exclusion criteria (Kokkinakis et al., 2017). Data collection took place in a quiet lab environment where participants were fitted with a lapel microphone (AudioTechnica ATR3350) and digitally recorded with a Zoom H4n Pro recorder (44.1 kHz sampling rate; 16bit resolution). The following instruction was given in Swedish: "Your task is to think of words. I want you to tell me all the different *animals* you can think of. You have 60 seconds. Do you have any questions? Are you ready? Go ahead and start." If the participant seemed unsure, they were told "any animals are okay: big ones, little ones, etc.". At the end of the 60 seconds, a timer would go off and the test leader would let the participant know that 60 seconds had passed. The resulting audio files were manually transcribed and manually time aligned in Praat (Boersma and Weenink, 2018). All animals named were transcribed on a separate tier.

A future follow-up visit at the memory clinic in 2019, after a second round of language tests, will include a renewed GDS (Global Deterioration Scale) classification and neuropsychological tests. The study was approved by local ethical committee (ref. number: 206-16, 2016 and T021-18, 2018).

3.2 Clinical Assessments

Participants in the Gothenburg MCI study were classified as having SCI, MCI, or dementia, and the controls were recruited separately and evaluated to ascertain that they were cognitively healthy. The classification is based on the Global Deterioration Scale (GDS), where level 1 codes for cognitively healthy, level 2 SCI, level 3 MCI and level 4 and above dementia (Auer and Reisberg, 1997; Wallin et al., 2016). Participants were further evaluated with neuropsychological tests, magnetic resonance imaging (MRI), blood samples, and spinal fluid samples (Wallin et al., 2016).

Compared to the other study participants, the persons with SCI were relatively young, had higher levels of education, higher prevalence of stress conditions and depressive symptoms as well as a family history of dementia (Eckerström et al., 2016).

3.3 Features

3.3.1 Traditional measures

From the manual transcripts, traditional SVF performance metrics were automatically extracted. The word count was determined as the number of unique, correctly named animals. Clusters and switches were determined based on a temporal metric proposed by Troeger et al. (2019). In this approach, the cluster structure is solely determined by the temporal position of words in the recording. Consecutive words are clustered if the transition time between them is shorter than then average transition time over the sample. This threshold is furthermore scaled over the process of the task to account for the decline in production speed. The mean number of clusters and the number of switches between them is extracted.

3.3.2 Temporally resolved measures

To explore different cognitive processes engaged over the course of the one minute task, SVF performance is examined in 10 second steps. Words in the transcript were assigned to a temporal interval based on their onset. Word count is determined for each interval, disregarding repetitions from earlier intervals. Lexical frequency of words were determined using the KORP collection of Swedish corpora (Borin et al., 2012). Transition times between consecutive words were defined as the difference between the end of the current word and the onset of the next. Word frequency and transition times are reported as the average over each interval.

3.4 Statistical analysis

Statistical analysis was performed using R (software version 3.4.0). For group comparisons of traditional measures, linear models with the measure as a function of diagnostic group were examined. Temporally resolved measures were examined with separate linear mixed effects analysis, one for each response variable –word count, lexical frequency and transition time– using the *lme4* (Bates et al., 2014) package. Each time interval is modelled as a single data point and with age and education level, as well as the interaction between the time interval (T) and diagnosis, as fixed effects. The participant identifier was modelled as a random intercept. Spearman correlations between the interval word count and neuropsychological scores were examined. Age and education were chosen as demographic variables. As neuropsychological correlates, the following scores were used: the Trail Making Test

Part A (TMT-A), as an indicator for processing speed; the Boston Naming Test (BNT; Kaplan et al. (1983)), which assess language ability with a spectrum of high to low frequency words as a proxy of vocabulary size; and the Wechsler Adult Intelligence Scale Similarities (WAIS-Similarities), which measures abstract thinking, concept formation and verbal reasoning (Wechsler, 1999).

3.5 Machine Learning

The predictive power of the proposed temporal and semantic features were validated with machine learning experiments for the HC and MCI populations. For each transcribed speech sample, the features described in Section 3.3.1 and 3.3.2 were extracted and label in accordance to their diagnostic category. Logistic Regression (LR) and Support Vector Machine (SVM) models, as implemented by the scikit-learn (Pedregosa et al., 2011) framework, were trained as binary classifiers to separate the groups. First, models were trained with only word count, to establish a baseline, and then, on the complete feature set, utilizing univariate feature selection.

Area under the Receiver-Operator curve (AUC) is reported as the evaluation parameter. Due to the small size of the dataset, we used leave-pair-out cross validation (LPO-CV), which has been shown to produce an unbiased estimate for AUC on small datasets (Airola et al., 2009). We also computed the standard deviation in AUC as described by Roark et al. (2011).

Feature scaling and hyper-parameter optimisation were done on the training set in each fold. Features were scaled using min-max scaling between 0 and 1. For both SVMs and LR, C was optimised between $C \in [10^{-4}, ..., 10^{4}]$ using a grid search. LR models were trained with both L1 and L2 loss; for SVM a linear and an rbf kernel were used.

For the extended feature set, feature selection based on χ^2-tests was applied to the training set in each fold. The number of selected features was scaled between 1 and the maximum of 30.

	HC	**SCI**	**MCI**
N	32	19	24
Sex (M/F)	12/20	8/11	11/13
Age (years)	68.1 (7.2)	66.0 (6.7)	70.8 (5.6)
Education (years)	13.2 (3.5)	16.0 (2.3)	13.8 (3.5)
MMSE (max 30)	29.7 (0.5)	29.6 (0.8)	28.5 (1.4)

Table 1: Demographic information; the MMSE (Mini Mental State Exam) is a general screening test of cognitive status and has a maximum score of 30.

4 Results

4.1 Demographic information

Demographic information by diagnostic group is reported in Table 1. The SCI group is slightly younger and has a higher education level than the other two groups. The MMSE, a general index of cognitive status with a maximum score of 30, is lower in the MCI group. With an average MMSE of 28.5, this MCI population is still quite functional in comparison to other MCI populations (mean MMSE score can vary between 23 and 29 in the MCI group) (Lonie et al., 2009). Note that cut-off points for MMSE may vary slightly: for Swedish, a cut-off value between 25 and 27 indicates possible cognitive impairment which should be further evaluated (Palmqvist et al., 2013) while other studies consider an "abnormal" MMSE score to be lower or equal to 25 (Zadikoff et al., 2008).

4.2 Traditional measures

A linear model of word count as a function of diagnosis revealed a significant main effect ($F(2, 72) = 8.57, p < 0.01$). Compared to the control group ($WC = 24.06 \pm 6.37$), the SCI group ($WC = 27.84 \pm 5.6$) had a significantly increased word count ($3.78 \pm 1.8, p < 0.5$); the MCI group ($WC = 20.12 \pm 6.08$) a significantly lowered one ($-3.94 \pm 1.6, p < 0.5$). No significant effects for the size of temporal clusters ($F(2, 72) = 2.59, p = 0.08$) or the number of temporal switches ($F(2, 72) = 1.64, p = 0.2$) as a function of diagnosis are found.

4.3 Temporally resolved measures

Word count, lexical word frequency and transition times by 10 second intervals is visualized in Figure 1 and the results of linear mixed random effects models are presented in Table 2.

Variable	Estimate	t	95% CI	p-Value
$\mathbf{WC_{T_1-T_2}}$	-0.456	-6.196	[-0.529, -0.382]	$< .01$
$\mathbf{WC_{T_1-T_3}}$	-0.698	-7.898	[-0.787, -0.61]	$< .01$
$\mathbf{WC_{T_1-T_4}}$	-0.937	-8.681	[-1.046, -0.83]	$< .01$
$\mathbf{WC_{T_1-T_5}}$	-1.301	-8.675	[-1.452, -1.152]	$< .01$
$\mathbf{WC_{T_1-T_6}}$	-1.290	-8.690	[-1.439, -1.142]	$< .01$
Age	-0.011	-3.294	[-0.014, -0.008]	$< .01$
Education	-0.003	-0.411	[-0.010, 0.004]	.68
SCI	-0.086	-1.128	[-0.164, -0.010]	.26
SCI x T				
SCI x $\mathbf{WC_{T_1-T_2}}$	0.247	2.161	[0.133, 0.361]	$< .03$
SCI x $WC_{T_1-T_3}$	0.155	1.102	[0.014, 0.296]	.27
SCI x $WC_{T_1-T_4}$	0.180	1.068	[0.012, 0.349]	.29
SCI x $\mathbf{WC_{T_1-T_5}}$	0.543	2.738	[0.345, 0.742]	$< .01$
SCI x $\mathbf{WC_{T_1-T_6}}$	0.575	2.959	[0.381, 0.770]	$< .01$
MCI	-0.041	-0.602	[-0.111, 0.028]	.55
MCI x T				
MCI x $WC_{T_1-T_2}$	-0.088	-0.724	[-0.210, 0.034]	.47
MCI x $\mathbf{WC_{T_1-T_3}}$	-0.383	-2.176	[-0.559, -0.207]	$< .05$
MCI x $WC_{T_1-T_4}$	-0.015	-0.089	[-0.189, 0.158]	.93
MCI x $WC_{T_1-T_5}$	-0.101	-0.396	[-0.354, 0.153]	.69
MCI x $WC_{T_1-T_6}$	-0.299	-1.046	[-0.585, -0.013]	.30

(a) Word Count

Variable	Estimate	t	95% CI	p-Value
$\mathbf{WF_{T_1-T_2}}$	-0.774	-2.558	[-1.077, -0.472]	$< .05$
$\mathbf{WF_{T_1-T_3}}$	-0.696	-2.298	[-0.999, -0.393]	$< .05$
$\mathbf{WF_{T_1-T_4}}$	-1.274	-4.208	[-1.577, -0.971]	$< .01$
$\mathbf{WF_{T_1-T_5}}$	-1.386	-4.578	[-1.689, -1.083]	$< .01$
$\mathbf{WF_{T_1-T_6}}$	-1.514	-5.000	[-1.816, -1.211]	$< .01$
Age	0.023	2.600	[0.014, 0.032]	$< .05$
Education	0.000	0.003	[-0.018, 0.018]	0.99
SCI	0.228	0.642	[-0.127, 0.582]	.52
SCI x T				
SCI x $WF_{T_1-T_2}$	-0.549	-1.108	[-1.045, -0.053]	.27
SCI x $WF_{T_1-T_3}$	-0.763	-1.539	[-1.259, -0.267]	.12
SCI x $WF_{T_1-T_4}$	-0.123	-0.248	[-0.619, 0.373]	.80
SCI x $WF_{T_1-T_5}$	-0.138	-0.279	[-0.634, 0.358]	.78
SCI x $WF_{T_1-T_6}$	-0.575	-1.159	[-1.071, -0.079]	.25
MCI	0.193	0.588	[-0.135, 0.521]	.56
MCI x T				
MCI x $WF_{T_1-T_2}$	-0.261	-0.564	[-0.723, 0.202]	.57
MCI x $\mathbf{WF_{T_1-T_3}}$	-0.936	-2.025	[-1.399, -0.474]	$< .05$
MCI x $WF_{T_1-T_4}$	-0.356	-0.769	[-0.818, 0.107]	.44
MCI x $WF_{T_1-T_5}$	-0.256	-0.554	[-0.719, 0.206]	.58
MCI x $WF_{T_1-T_6}$	-0.282	-0.610	[-0.745, 0.180]	.54

(b) Word frequency

Variable	Estimate	t	95% CI	p-Value
$L_{T_1-T_2}$	0.986	1.460	[0.311, 1.662]	.15
$\mathbf{L_{T_1-T_3}}$	2.557	3.786	[1.882, 3.233]	$<.01$
$\mathbf{L_{T_1-T_4}}$	2.641	3.911	[1.966, 3.317]	$<.01$
$\mathbf{L_{T_1-T_5}}$	5.245	7.766	[4.570, 5.921]	$<.01$
$\mathbf{L_{T_1-T_6}}$	5.641	8.352	[4.965, 6.316]	$<.01$
Age	0.028	1.029	[0.001, 0.055]	.31
Education	-0.074	-1.355	[-0.129, -0.019]	.18
SCI	0.311	0.365	[-0.541, 1.163]	.72
SCI x T				
SCI x $L_{T_1-T_2}$	-0.703	-0.635	[-1.81, 0.404]	.53
SCI x $L_{T_1-T_3}$	-1.429	-1.291	[-2.536, -0.322]	.20
SCI x $L_{T_1-T_4}$	-0.803	-0.726	[-1.910, 0.303]	.47
SCI x $\mathbf{L_{T_1-T_5}}$	-2.528	-2.284	[-3.634, -1.421]	$<.05$
SCI x $\mathbf{L_{T_1-T_6}}$	-2.384	-2.154	[-3.490, -1.277]	$<.05$
MCI	0.22	0.281	[-0.564, 1.004]	.78
MCI x T				
MCI x $L_{T_1-T_2}$	0.167	0.162	[-0.865, 1.198]	.87
MCI x $L_{T_1-T_3}$	0.510	0.494	[-0.522, 1.542]	.62
MCI x $L_{T_1-T_4}$	0.724	0.702	[-0.308, 1.756]	.48
MCI x $L_{T_1-T_5}$	-1.212	-1.175	[-2.244, -0.18]	.24
MCI x $L_{T_1-T_6}$	0.41	0.397	[-0.622, 1.441]	.69

(c) Transition Length

Table 2: Linear Mixed Random Effects model examining the effects of time interval, diagnosis, age and education on one of three variables, while controlling random effects per subject. Significant values ($p < .05$) are indicated in bold.

A general decline in the word count for each time interval is visible and reflected in the model, regardless of diagnostic group. A significant effect for age is present, implicating that higher age leads to a reduced word count. For the SCI group, there is a significant interaction between the diagnostic group and the decline in WC_{T2},WC_{T5} and WC_{T6}. In these intervals, the decline of the SCI group is less severe. The MCI diagnostic group shows a significant interaction with the decline in WC_{T3}, with a stronger decline in word count than the other groups.

For lexical word frequency, again, a significant decline over time is visible, regardless of diagnostic group, which means that participants produce more common words at the start of the task, and less common words towards the end. Older participants produce words that are significantly more frequent. The MCI group has a significant interaction with WF_{T_3}, indicating this group uses lower frequency words in this time interval.

Starting from the third interval, a significant increase in word transition times is visible. A significant interaction between the SCI group and the fifth and sixth interval, indicates the SCI group shows significantly lower transition times in these intervals.

4.4 Correlation analysis

Spearman correlations between the word count by time interval, neuropsychological scores and demographic information is displayed in Figure 2. Only significant correlations are displayed.

Significant positive correlations between the BNT score and the word count in the last three time intervals are observed. The WAIS Similarity score shows positive correlations with the word count of the last two intervals. Negative correlations are observed between TMT A and the second and third interval, as well as between age and these two intervals (for the TMT A a lower score indicates a better per-

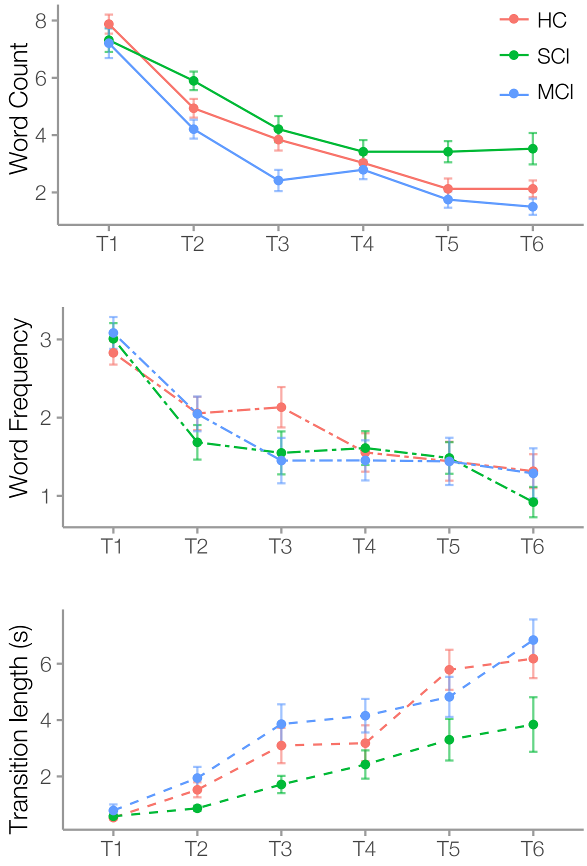

Figure 1: Word Count, Word Frequency and Transition length by time interval and for each group separately. Error bars display standard error.

formance).

4.5 Machine Learning

Figure 3 displays the results of the machine learning experiments. AUC is plotted, while varying the number of features chosen in feature selection, using different classifiers.

The baseline performances of models using just the word count is $AUC = 0.64$ for LR, both with $L1$ and $L2$ loss, and the linear SVM. The SVM with an rbf kernel only achieves $AUC = 0.62$ with the word count feature. Generally, the models using all features outperform the baseline. The best performance of $AUC = 0.72$ is observed for a linear SVM with 20 features. Generally, the linear and rbf SVM and the LR with $L1$ loss show similar performance patterns, across all number of features. The LR with $L2$ shows steadily increasing performance. The SVM with rbf kernel outperforms the other models with a lower number of features.

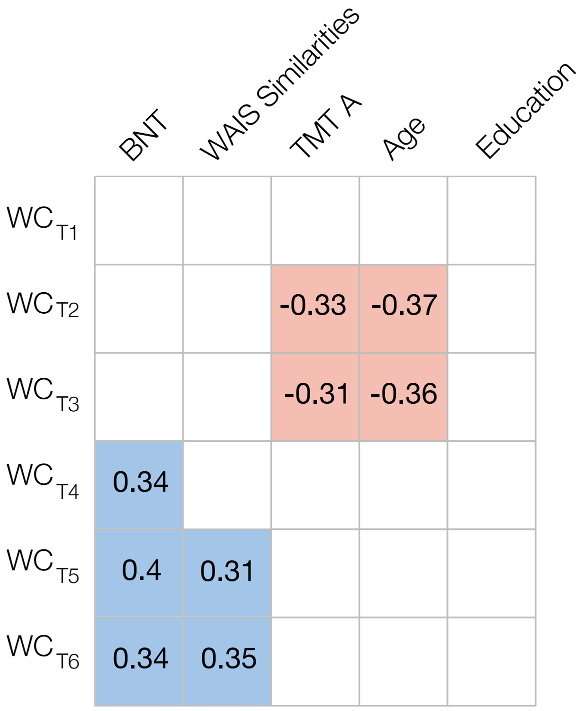

Figure 2: Spearman correlation between 10 second word count (WC) intervals and neuropsychological test scores. Only significant correlations are shown. Positive correlations in blue, negative ones in red.

5 Discussion

Reviewing the overall performance on the SVF, a significant difference in word count was found between the groups, but no differences in cluster size or number of temporal clusters. The temporally resolved measures showed that the MCI, SCI and HC group follow similar trends with regard to word count, word frequency and transition length: word count and word frequency generally decrease over time, while average transition times increase. Significant differences between the MCI group and the other two groups were found mainly for the third interval, where the participants in the MCI group produce fewer and less frequent words. For the word count, this is in line with previous findings from Fernaeus et al. (2008), and the lower word frequency in the third interval indicates that persons with MCI have to resort to low frequency words earlier in the task, switching from semi-automatic retrieval of more common words to effortful retrieval at an earlier point than the other groups.

The persons with SCI showed an increased

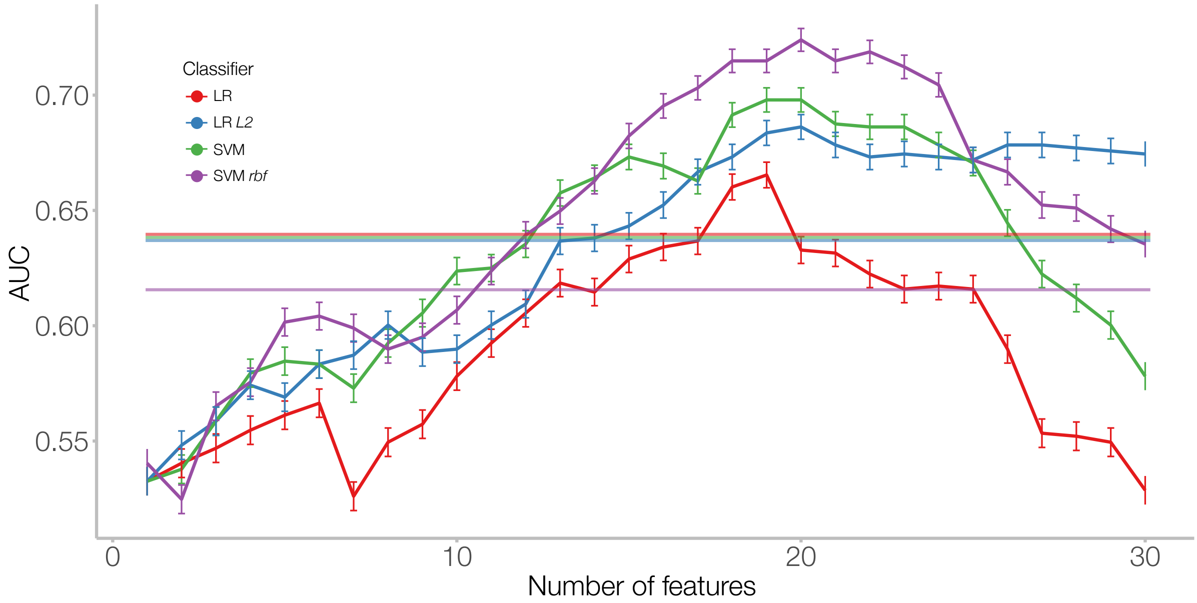

Figure 3: Area under the curve (AUC) of different classification models separating HC and MCI, plotted against number of features selected through univariate feature selection. Horizontal lines show the performance of models solely trained on the word count. Error bars indicate standard deviation of performance.

word count in the second, fifth and sixth interval, and reduced transition times in the fifth and the sixth interval. This suggests that they were able to sustain a continuous production for longer. The words they produced in the last intervals did not differ in frequency from the other groups, but the persons with SCI seemed to have access a larger store of words. Participants in the SCI group had a longer education than the general population, and one possibility is that the participants with SCI in the Gothenburg MCI study perform better because of higher premorbid functioning (Eckerström et al., 2016).

Correlation analysis with additional psychometric data lends a deeper understanding of the results, and significant correlations showed that higher BNT and WAIS similarities scores were associated with a higher word count in the latter part of the SVF. This suggests that having a broader vocabulary, as measured by the BNT, predicts a higher word count in the second half of the SVF. When reviewing the word count graph in Figure 1 and comparing the groups, it is evident that the ability of participants with SCI to sustain performance in the later time intervals can be explained by the access to a larger vocabulary as measured by the BNT. Age and TMT-A both show significant negative correlation with the second and third time intervals of the SVF. TMT-A is a measure of processing speed, and it decreases with increasing age. A decrease in processing speed seems to specifically inhibit production in the second and third interval. Demetriou and Holtzer (2017) suggested a semi-automatic retrieval phase at the beginning and a more effortful retrieval at the end of the task. Our findings support the notion of these phases occurring over the course of task, where the first phase is more influenced by processing speed and the later benefits more strongly from a larger vocabulary.

The benefits of temporal analysis were apparent in the increase of the ability to correctly classify participants as HC or MCI, compared to a classification based solely on word count. In the best case, the performance of the SVM with rbf kernel improved from $AUC = 0.62$ to $AUC = 0.72$ with temporal analysis. While this study was based on manually transcribed data, previous research shows that this type of analysis can be done fully automatically including ASR, which allows for easy scaling of the task (König et al., 2018).

6 Conclusion

This paper introduced a novel, interval-based temporal analysis method for SVF tasks. The resulting outcome revealed distinct patterns that differentiated the groups: persons with SCI had a higher word count and sustained lexical frequency level during the last intervals, while persons with MCI had a steeper decline in both word count and lexical frequencies during the third interval. Correlations with neuropsychological scores suggested that the superior performance of the SCI group could be attributed to vocabulary size. Classification results improved when adding the novel features ($AUC = 0.72$), supporting their diagnostic value. This increase over the baseline performance underlines the value of using novel methods in addition to clinical standards.

The results of group comparisons and correlations are in line with previous findings about phases of production in SVF. The special role of the third time interval in discriminating MCI patients is also supported by previous research. Future research should strive to validate these findings on larger data sets, for other languages and other semantic categories.

Based on our findings, we suggest that temporal analysis of the SVF may be useful as a screening tool, when assessing persons with self-perceived memory problem, as this type of analysis seems to highlight the subtle differences between the groups. We see it as a strength that instead of adding new tasks, we are using an already clinically validated tool in an innovative and new manner.

7 Acknowledgements

This research has been funded by Riksbankens Jubileumsfond - The Swedish Foundation for Humanities and Social Sciences, through the grant agreement no: NHS 14-1761:1. We would also like to thank the Gothenburg University Centre for Ageing and Health (AgeCap) and Sahlgrenska University Hospital for supporting this study.

References

Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. 2009. A comparison of AUC estimators in small-sample studies. In *Machine Learning in Systems Biology*, pages 3–13.

Lee Ashendorf, Angela L Jefferson, Maureen K O'Connor, Christine Chaisson, Robert C Green, and Robert A Stern. 2008. Trail Making Test errors in normal aging, mild cognitive impairment, and dementia. *Archives of Clinical Neuropsychology*, 23:129–137.

J Wesson Ashford, Soo Borson, Ruth O'Hara, Paul Dash, Lori Frank, Philippe Robert, William R Shankle, Mary C Tierney, Henry Brodaty, Frederick A Schmitt, Helena C Kraemer, Herman Buschke, and Howard Fillit. 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.

Stefanie Auer and Barry Reisberg. 1997. The GDS/FAST staging system. *International Psychogeriatrics*, 9(SUPPL. 1):167–171.

Sophie Auriacombe, Nathalie Lechevallier, Hélène Amieva, Sandrine Harston, Nadine Raoux, and J-F Dartigues. 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.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Paul Boersma and David Weenink. 2018. Praat: doing phonetics by computer. version 6.0.40. Computer program.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp the corpus infrastructure of Språkbanken. In *The 8th international conference on Language Resources and Evaluation (LREC)*, pages 474–478, Istanbul, Turkey.

Eleni Demetriou and Roee Holtzer. 2017. Mild Cognitive Impairments Moderate the Effect of Time on Verbal Fluency Performance. *Journal of the International Neuropsychological Society*, 23:44–55.

Marie Eckerström, Anne Ingeborg Berg, Arto Nordlund, Sindre Rolstad, Simona Sacuiu, and Anders Wallin. 2016. High Prevalence of Stress and Low Prevalence of Alzheimer Disease CSF Biomarkers in a Clinical Sample with Subjective Cognitive Impairment. *Dementia and Geriatric Cognitive Disorders*, 42(1-2):93–105.

Safa A. Elgamal, Eric A. Roy, and Michael T. Sharratt. 2011. Age and Verbal Fluency: The Mediating Effect of Speed of Processing. *Canadian Geriatrics Journal*, 14(3).

Sven Erik Fernaeus, Per Östberg, Åke Hellström, and Lars Olof Wahlund. 2008. Cut the coda: Early fluency intervals predict diagnoses. *Cortex*, 44(2):161–169.

Marshal F Folstein, Susan E Folstein, and Paul R. McHugh. 1975. Mini-mental status. a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

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

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

Edith Kaplan, Harold Goodglass, Sandra Weintraub, and Osa Segal. 1983. Boston naming test. In *Psychological Corporation*, Philadelphia: Lea & Febiger.

Dimitrios Kokkinakis, Kristina Lundholm Fors, Eva Björkner, and Arto Nordlund. 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 of Computational Linguistics*, volume 75, pages 172–182. Linköping University Electronic Press.

Alexandra König, Nicklas Linz, Johannes Tröger, Maria Wolters, Jan Alexandersson, and Phillipe Robert. 2018. Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and geriatric cognitive disorders*, 45(3-4):198–209.

Nicklas Linz, Johannes Tröger, Jan Alexandersson, and Alexandra Konig. 2017. Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *IWCS 2017 - 12th International Conference on Computational Semantics*, pages 1–7, Montpellier, France.

Jane A Lonie, Kevin M Tierney, and Klaus P Ebmeier. 2009. Screening for mild cognitive impairment: A systematic review. *International Journal of Geriatric Psychiatry*, 24(9):902–915.

Kimberly Diggle Mueller, Rebecca L. Koscik, Asenath LaRue, Lindsay R. Clark, Bruce Hermann, Sterling C. Johnson, and Mark A. Sager. 2015. Verbal fluency and early memory decline: Results from the wisconsin registry for alzheimer's prevention. *Archives of Clinical Neuropsychology*, 30(5):448–457.

Tomas Nikolai, Ondrej Bezdicek, Hana Markova, Hana Stepankova, Jiri Michalec, Miloslav Kopecek, Monika Dokoupilova, Jakub Hort, and Martin Vyhnalek. 2018. Semantic verbal fluency impairment is detectable in patients with subjective cognitive decline. *Applied Neuropsychology:Adult*, 25(5):448–457.

Sebastian Palmqvist, B Terzis, C Strobel, and Anders Wallin. 2013. Mmse-sr: Mini mental state examination - swedish revision, version 2.

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. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jessica Peter, Jannis Kaiser, Verena Landerer, Lena Köstering, Christoph P Kaller, Bernhard Heimbach, Michael Hüll, Tobias Bormann, and Stefan Klöppel. 2016. Category and design fluency in mild cognitive impairment: Performance, strategy use, and neural correlates. *Neuropsychologia*, 93:21–29.

Sarah E. Price, Glynda J. Kinsella, Ben Ong, Elsdon Storey, Elizabeth Mullaly, Margaret Phillips, Lanki Pangnadasa-Fox, and Diana Perre. 2012. Semantic verbal fluency strategies in amnestic mild cognitive impairment. *Neuropsychology*, 26(4):490–497.

Nadine Raoux, Hélène Amieva, Mélanie Le Goff, Sophie Auriacombe, Laure Carcaillon, Luc Letenneur, and Jean-François Dartigues. 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.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffery Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090.

F Schiller. 1947. Aphasia studied in patients with missile wounds. *J Neurol Neurosurg Psychiatry*, 10(4):183–197.

Zeshu Shao, Esther Janse, Karina Visser, and Antje S. Meyer. 2014. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, 5(JUL):1–10.

Ing Mari Tallberg, E. Ivachova, K. Jones Tinghag, and Per Östberg. 2008. Swedish norms for word fluency tests: FAS, animals and verbs. *Scandinavian Journal of Psychology*, 49(5):479–485.

Johannes Troeger, Nicklas Linz, Alexandra Koenig, Jessica Peter, Philippe Robert, and Jan Alexandersson. 2019. Exploitation vs. ExplorationComputational Temporal and Semantic Analysis Explains Semantic Verbal Fluency Impairment in Alzheimers Disease. *Neuropsychologia*. Submitted.

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

Anders Wallin, Arto Nordlund, Michael Jonsson, Karin Lind, Åke Edman, Mattias Göthlin, Jacob Stålhammar, Marie Eckerström, Silke Kern, Anne Börjesson-Hanson, Mårten Carlsson, Erik Olsson, Henrik Zetterberg, Kaj Blennow, Johan Svensson, Annika Öhrfelt, Maria Bjerke, Sindre Rolstad, and Carl Eckerström. 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.

D Wechsler. 1999. Wechsler abbreviated intelligence scale. In *Psychological Corporation*, San Antonio, TX, USA.

Cindy Zadikoff, Susan H. Fox, David F. TangWai, Teri Thomsen, Rob M.A. de Bie, Pettarusup Wadia, Janis Miyasaki, Sarah DuffCanning, Anthony E. Lang, and Connie Marras. 2008. A comparison of the mini mental state exam to the montreal cognitive assessment in identifying cognitive deficits in parkinsons disease. *Movement disorders*, 23(2):297–299.

Chiara Zucchella, Elena Sinforiani, Stefano Tamburin, Angela Federico, Elisa Mantovani, Sara Bernini, Casale Roberto, and Michelangelo Bartolo. 2018. The multidisciplinary approach to alzheimer's disease and dementia. a narrative review of non-pharmacological treatment. *Front. Neurol.*, 9(1058).