Dissociating Semantic and Phonemic Search Strategies in the Phonemic Verbal Fluency Task in early Dementia

Hali Lindsay and **Philipp Mueller**
German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
{hali.lindsay, philipp.mueller}@dfki.de

Nicklas Linz and **Mario Mina**
ki elements, Saarbrücken, Germany
{nicklas, mario.mina}@ki-elements.de

Radia Zeghari
The CoBTeK, Université Cote d'Azur (UCA), Nice, France
radia.zeghari@gmail.com

Alexandra König
Stars Team, Institut National de Recherche en Informatique
et en Automatique (INRIA), Valbonne, France
alexandra.konig@inria.fr

Johannes Tröger
German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
johannes.troeger@dfki.de

Abstract

Effective management of dementia hinges on timely detection and precise diagnosis of the underlying cause of the syndrome at an early mild cognitive impairment (MCI) stage. Verbal fluency tasks are among the most often applied tests for early dementia detection due to their efficiency and ease of use. In these tasks, participants are asked to produce as many words as possible belonging to either a semantic category (SVF task) or a phonemic category (PVF task). Even though both SVF and PVF share neurocognitive function profiles, the PVF is typically believed to be less sensitive to measure MCI-related cognitive impairment and recent research on fine-grained automatic evaluation of VF tasks has mainly focused on the SVF. Contrary to this belief, we show that by applying state-of-the-art semantic and phonemic distance metrics in automatic analysis of PVF word productions, in-depth conclusions about production strategy of MCI patients are possible. Our results reveal a dissociation between semantically- and phonemically-guided search processes in the PVF. Specifically, we show that subjects with MCI rely less on semantic- and more on phonemic processes to guide their word production as compared to healthy controls (HC). We further show that semantic similarity-based features improve automatic MCI versus HC classification by 29% over previous approaches for the PVF. As such, these results point towards the yet underexplored utility of the PVF for in-depth assessment of cognition in MCI.

1 Introduction

Dementia is a syndrome primarily presenting with broad cognitive impairments. There are multiple underlying causes that result in dementia such as Alzheimer's Disease (AD) or fronto-temporal lobar degeneration or focal lesions (MacPherson et al., 2016). These sub-forms have different neurocognitive profiles. The most-common Alzheimer's Disease (AD)-related dementia is typically driven by an amnestic cognitive impairment (Kidd, 2008) whereas the fronto-temporal dementia is often associated with executive function impairment (Huey et al., 2009).

Early identification of dementia as well as precise differentiation between dementia sub-forms is crucial for effective management of the syndrome (Thyrian et al., 2016). Pairing high diagnostic sensitivity with ease of use, verbal fluency tests (VF) are amongst the most-applied tests in cognitive assessment of dementia (Troyer et al., 1997). In these tests, participants are asked to produce as many words from a specific category as they can in a fixed time. The two main variants of VF tests are the semantic verbal fluency (SVF) and the phonemic verbal fluency (PVF). In the SVF,

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 32–44

the word category is defined by semantics (e.g. all animal words), whereas in the PVF participants need to produce words starting with a specific letter (e.g. "S"). Traditionally, test scores are computed by counting the number of correctly named words within the given time (Gomez and White, 2006). Although both VF variants are quite similar in the way they engage different neurocognitive functions, the cognitive strategies of the task can indicate different patterns of the underlying neuropathology. For instance, an SVF impairment is often only regarded as evidence for amnestic dementia (Vaughan et al., 2016; Teng et al., 2013) whereas a PVF impairment is almost exclusively regarded as evidence for fronto-temporal dementias (Dubois et al., 2000).

Recently, advanced Natural Language Processing (NLP) techniques have been applied to allow for in-depth analysis of the produced word sequence in VF tasks, particularly for the SVF (Linz et al., 2017a; Kim et al., 2019; Diaz-Orueta et al., 2020; Zemla et al., 2020). By extracting clusters from the produced word sequence and by modelling the semantic relationships between- and within these clusters, it is possible to disentangle the effects of memory impairment from effects of executive function impairment (Tröger et al., 2019). Despite the success of these qualitative features in the SVF, their utility for automatic analysis of the PVF remains underexplored.

In this paper, we investigate both phonemic and semantic motivations for the underlying strategy of the phonemic verbal fluency task, and thereby reduce the gap between clinical theory and computational approaches to evaluating cognitive speech tasks. By contrasting semantic and phonemic distance measure in an analysis based on time bins, we show a dissociation between semantically- and phonemically-guided search processes: Subjects with mild cognitive impairment (MCI) exhibit significantly less semantic similarity in their productions as compared to healthy controls (HC). Finally, in experiments on automatic classification of MCI vs. HC from PVF word productions, we show that semantic features improve over previous approaches by 29%. Taken together, our results pave the way towards more fine-grained analysis of the PVF task that can help to improve clinical decision processes.

2 Clinical Background

2.1 Cognitive Processes in VF

Verbal Fluency tasks (VF) require a network of cognitive processes activating—-a region associated with language (Vigneau et al., 2006)—-the frontal lobe (Coslett et al., 1991; Miller, 1984), specifically the left hemisphere (Birn et al., 2010; Troyer et al., 1998; Mueller et al., 2015), as well as the temporal lobe (Newcombe, 1969; Cerhan et al., 2002).

VF are used to assess semantic memory and executive functions as a good VF performance hinges on intact semantic memory stores as well as the ability to access these memory stores (Chertkow and Bub, 1990; Hodges et al., 1992; Mueller et al., 2015). Executive functioning, specifically, working memory is thought to allow a person to effectively search through phonological and semantic stores while regulating and adapting the search strategy to produce more words over the task (Faust, 2012; Rende et al., 2002; Troyer et al., 1997; Rosen, 1980).Both PVF and SVF are hypothesised to span multiple overlapping cognitive abilities; executive, verbal, and attention abilities (Mueller et al., 2015; Li et al., 2017; Shao et al., 2014; Schmidt et al., 2017). However, there is evidence that each task measures a set of distinct cognitive processes.

PVF burdens executive resources whereas the SVF demands linguistic-conceptual knowledge (Thompson-Schill et al., 1997; Vigneau et al., 2006; Shao et al., 2014; Mueller et al., 2015; Schmidt et al., 2017; Birn et al., 2010). SVF is theorized to engage the temporal lobe for lexical-semantic access and retrieval from semantic store (Newcombe, 1969; Mueller et al., 2015; Cerhan et al., 2002) where as the PVF is thought to rely on executive functioning and prefrontal lobe processes (Mueller et al., 2015) as well as phonological and orthographic cues for word retrieval (Li et al., 2017; Clark et al., 2013). Generally, it is hypothesised that SVF requires both semantic and retrieval processes whereas PVF relies only on retrieval processes (Fisher et al., 2004). However, there is conflicting research that PVF taps into the semantic network, although to a lesser extent than semantic fluency (Lezak et al., 2004; Mueller et al., 2015; Schmidt et al., 2017; Clark et al., 2013).

Bizzozero et al. (2013) investigated the extent to which SVF and PVF were related to semantic and attention processes and found evidence of semantic processes in both SVF and PVF. Nutter-Upham et al. (2008) observed a larger effect size

for the amnestic MCI (aMCI) group's deficit on semantic verbal fluency (Cohen's d=0.98) than for their deficit on phonemic verbal fluency (Cohen's d=0.66), due to greater variability in phonemic verbal fluency performance. Therefore, an alternative interpretation is that their findings actually do reflect a preferential deficit on semantic verbal fluency in aMCI. Supporting these findings, imaging studies combined with factor analysis have also suggested that the PVF task is relies on both semantic and phonemic processes (Schmidt et al., 2017; Clark et al., 2013).

2.2 VF for Diagnosis

Both the Phonemic and Semantic varieties of verbal fluency are commonly used to diagnosis and monitor cognitive decline such as mild cognitive impairment (MCI) and Alzheimer's Disease and Related Dementias (ADRD) (Marra et al., 2011; Clark et al., 2009; Gomez and White, 2006; Troyer et al., 1998).

SVF has been found to be more impaired than PVF in ADRD (Cerhan et al., 2002; Barr and Brandt, 1996; Zhao et al., 2013) and deficits in both semantic and phonemic memory have been reported. However there is conflicting research for PVF and SVF in the MCI group. For aMCI, only the SVF shows impairment (Hodges, 2006; Murphy et al., 2006; Teng et al., 2013). While other studies show decline on both the PVF and SVF task for MCI (Mueller et al., 2015; Vita et al., 2014; Nutter-Upham et al., 2008). Rinehardt et al. (2014) compared controls with aMCI, non-aMCI and AD and found that both MCI groups were less impaired on the SVF than the PVF, behaving more like controls than the AD group.

Clark et al. (2013) considered computationally-based phonemic and semantic measures when analyzing the PVF and SVF tasks in relation to gray matter correlates for HC, MCI and AD. They concluded that both tasks showed greater semantic motivations than phonemic motivation, even in the PVF task.

PVF may be a sensitive test for investigating phonemic and semantic processes but a global word count does not provide the in-depth information needed to understand the underlying cognitive processes (Gomez and White, 2006; Becker and Salles, 2016). In this paper, we apply recently developed automatic analysis techniques from computational linguistics to the PVF to obtain a better insight into the degradation of semantic and phonemic processes.

3 Previous Work

3.1 Analyzing Semantic and Phonemic Strategy for VF

Several modes of analysis have been proposed with the goal of observing the role that different cognitive strategies play throughout VF tasks.

Much work has been done on the semantic variety of verbal fluency, specifically for the animal category. Troyer et al. (1997) introduced a semantically-motivated hierarchical list of animals for determining semantic clusters. To overcome this time-intensive and subjective annotation process, previous research worked on automatically producing semantic clusters over SVF productions (Ryan, 2013; Pakhomov et al., 2015b, 2016; Linz et al., 2017b; König et al., 2018; Kim et al., 2019). For example, Pakhomov et al. (2015a) compared traditional and novel computational methods of evaluating SVF using medical imaging techniques between healthy and cognitively impaired individuals. The semantic relatedness of words was determined using latent semantic analysis of word co-occurrences from a large online corpora. This study showed that computational methods of evaluating the SVF were beneficial in understanding the relationships between the different cognitive processes.

Building off of this, Linz et al. (2017a) used neural word embeddings as a data-driven way to model semantic clustering in the SVF task. König et al. (2018) showed high correlations ($r = 0.9$) between automatically extracted clustering and switching features and clinical methods. From these clusters, several features including cluster size or number of switches between clusters were calculated to reflect cognitive processes (Linz et al., 2017a; König et al., 2018).

In addition to the SVF, Troyer et al. (1997) proposed a rule-based method for finding phonemically-related clusters of words in PVF productions. Lindsay et al. (2019) automated this rule-based method for determining phonemic clusters, and proposed three additional phonemic similarity metrics for evaluating the PVF task on healthy German students, namely the Levenshtein distance (LD), phonemically-weighted Levenshtein distance (PHON-LD), as well as position-weighted Levenshtein distance (POS-LD). Clark et al. (2013) pro-

	HC	MCI	p
N (#Female)	34(6)	48(22)	-
Age	73.56(6.74)	75.02(7.68)	0.40
Education	12.65(1.82)	10.71(4.01)	0.08
MMSE	28.76(1.28)	25.79(2.74)	<0.01

Table 1: Demographic information for the French population used. Age and Education are given in years. The Mini-Mental State Exam (MMSE) is a test to measure cognitive function (Max score 30). Means are given for the populations with standard deviation in parentheses. Significance testing between groups is reported in p column.

posed another phonemic distance measure using an English pronouncing dictionary and a formula for measuring string overlap to estimate phonemic-relatedness of adjacent words over the task.

Recently, (Linz et al., 2019) considered a binning-based approach (Fernaeus et al., 2008) for the automatic analysis of the SVF. In this approach, features were calculated separately on non-overlapping, 10-second time bins, which alloweda deeper investigation into the evolution of a participant's production strategy over time. Linz et al. (2019) used temporal binning to analyse at what points in time during SVF word production HC differed from MCI and AD patients with respect to word count, transition length, and word frequency.

To conclude, while previous works introduced metrics for quantifying semantic as well as phonemic similarity in VF word productions, no comprehensive comparison of these metrics was performed on the PVF in a clinical setting. This leaves a gap between clinical theory of motivating cognitive strategies and computational methods as to how to automatically evaluate both phonemic and semantic strategy for the PVF task. To allow for a fine-grained analysis of production strategy over the course of the PVF task, we analyze semantic and phonemic distance metrics in the temporal binning framework.

3.2 PVF-based MCI Classification

Compared to the amount of work on HC versus MCI classification from the SVF (Linz et al., 2017a; König et al., 2018), considerably less studies have investigated this classification task using the PVF (Ryan, 2013; Lindsay et al., 2020). Ryan (2013) used logistic regression to classify between HC and MCI using only repetitions (AUC=0.53) and word count (AUC=0.5) from the PVF. Lindsay et al. (2020) reported a baseline PVF experiment between HC and MCI and reported an AUC of 0.75 using only word count on a very small dataset (8HC/19MCI). Additional temporal features lowered the classification (AUC=0.55). To the best of our knowledge, no study at the present time has investigated HC versus MCI classification with the PVF using phonemic and semantic measures.

4 Methods

4.1 Data

The data used in this research was collected during the Dem@Care (Karakostas et al., 2017) and ELEMENT (Tröger et al., 2017) projects. Participants were recruited through the Memory Clinic located in Nice University Hospital at the Institute Claude Pompidou in Nice, France. The study was approved by the Nice Ethics Committee. All participants were native speakers of French and asked to give informed consent before participating in the study. The French data was collected in the form of speech recordings via an automated recording application installed on a tablet computer. The recordings were manually transcribed in PRAAT (Boersma and Weenink, 2009) according to the CHAT protocol (MacWhinney, 1991). Participants were asked to complete a battery of cognitive tests, including a 60 second phonemic verbal fluency task for the letter category *F*. Demographics for the data used are displayed in Table 1. A Mann-Whitney U test was conducted between the HC and MCI populations to check for significant differences between age (W = 1106, p-value = 0.40) and education (W = 1492, p-value = 0.08) but none were found.

4.2 Binning, Clustering & Global Resolutions of VF Analysis

We look at three resolutions of the verbal fluency task that have been applied to the SVF task and consider them for the PVF task; temporal binning, clustering and switching and global features. Each method provides a different resolution for looking word retrieval strategy. Temporal binning (Linz et al., 2019; Fernaeus et al., 2008) gives the finest resolution of strategy. The clustering is motivated

by clinical theory to investigate the different cognitive processes (Troyer et al., 1998). Global features are what are the current norm in clinical practice (Troyer et al., 1998; Gomez and White, 2006).

4.2.1 Binning Methods

To produce temporal bins for the PVF, we follow the methodology in (Linz et al., 2019) that was previously used for SVF. The complete 60-second PVF response is split into into six 10-seconds bins. This produces a new resolution of the task from which we can then compute features. As done in (Linz et al., 2019), we include the word count as well as the average temporal distance(TD) between consecutive words. In addition, we include the average semantic distance between consecutive words as well as the averages of the three phonemic distance measures LD, PHON-LD, and POS-LD. This allows for a separate investigation of the phonemic, semantic and temporal measures that guide search processes during the span of the word production in the PVF task.

Semantic Distance (SD) We follow Linz et al. (2017a) who computed semantic similarity between two words as the cosine distance between their embedding vectors. To construct word embeddings, FastText models (Bojanowski et al., 2016) are used. For this paper, the cosine distance is used, where $Cosine_{distance} = 1 - Cosine_{similarity}$.

Levenshtein Distance (LD) Lindsay et al. (2019) used the Levenshtein distance as a measure of phonetic distance when evaluating the PVF task. They first phonetically transliterate the word using the python package epitran (Mortensen et al., 2018). They then proposed using the traditional levenshtein distance to measures the number of edits (insertions, substitutions and deletions) between consecutive words (Levenshtein, 1966). They also proposed two weighted measures of LD as described below.

Phonemic-weighted Levenshtein Distance (PHON-LD) In addition to LD, Lindsay et al. (2019) proposed a phonemically weighted version of levenshtein distance. Using the epitran package, each phoneme has a corresponding 21-length phonological vector to represents the characteristics of the sound (e.g. voice/unvoiced, front/back). When computing the levenshtein distance, they weighted substitutions as the cosine between the to phonological vectors. Insertions and deletion are still valued at 1.

Position-weighted Levenshtein Distance (POS-LD) Lindsay et al. (2019) also investigated a position weighted levenshtein distance as the distance between phonetic representations of consecutive words, weighted for position in the word. Deletions, insertions and substitutions are set weighted by exponential distribution (with $\lambda = 0.5$) at the position of the phoneme in the word.

Temporal Distance (TD) The temporal distance is defined as the time in seconds between the boundaries of consecutive words in the PVF production.

4.2.2 Clustering Methods

Clustering-based approaches for VF evaluation consist of two steps. First, the produced word sequence is partitioned into a set of clusters. Second, features (e.g. mean cluster size) are computed from the automatically produced clusters. In this study, we consider a rule-based phonemic clustering as well as an automated version of semantic clustering, and temporal clustering to investigate production. For each both phonemic and semantic clustering types, the mean cluster size and number of switches are computed.

Phonemic Clustering In the case of phonemic clustering features, we determine clusters in the word sequence following the phonemically-motivated, clinical approach from Troyer et al. (1997) that was automated by Lindsay et al. (2019). This approach uses phonemic similarity rules to determine whether subsequent words belong to the same cluster or not.

Semantic Clustering Semantic Clusters are determined as in Linz et al. (2017a). Using the semantic distance method described previously, a semantic threshold is determined for each participant by averaging the semantic distance between all words in the production. If the semantic distance between consecutive words is lower than the threshold, the words are said to be in a cluster. If the semantic distance between consecutive words is greater than the threshold, this introduces a cluster boundary.

To obtain semantic word embeddings, the pretrained French fastText model is used. This model is trained on Common Crawl and Wikipedia corpora using the continuous bag of words (CBOW) algorithm with a negative sampling loss function. FastText models are trained at the character level using a character n-gram model. The 300-dimension

	HC		MCI		HC v. MCI	
	Mean	**SE**	**Mean**	**SE**	***W***	***p***
Average Over Bins						
Word Count	2.70	0.17	2.00	0.11	1145	**0.002**
Semantic Distance	0.54	0.12	0.57	0.12	584	**0.040**
Temporal Distance	4.25	0.29	5.96	0.36	496	**0.002**
LD	3.09	0.13	2.57	0.11	1125	**0.004**
PHON-LD	1.92	0.08	1.70	0.06	1016	0.060
POS-LD	1.66	0.05	1.49	0.04	1096	**0.008**
Rule-Based Phonemic Clustering						
Mean Cluster Size	4.63	1.74	4.02	1.57	1042	**0.033**
Number of Switches	2.51	1.17	2.19	0.98	947.5	0.195
Automatic Semantic Clustering						
Mean Cluster Size	2.81	0.79	2.63	0.83	928.5	0.287
Number of Switches	9.09	4.15	7.04	3.27	1077	**0.014**

Table 2: Significance testing results between HC and MCI for the binning and clustering methods with a Mann-Whitney U test. The p-value is reported and a significance level is set at 0.05. Significant values are shown in bold type face. Standard Error (SE). Means and SE are provided to understand relationship between the groups. The top half of the table reports values for the binning analysis. The bottom half of the table reports significance results for the clustering analysis.

model is used for this analysis. For specific numerical parameter values, or to download the models used in this research, please see the link in the footnote[1].

4.3 Global Features

In addition to the binning features and clustering features in (Section 4.2.2), we include the traditional way of evaluating verbal fluency tasks, which computes aggregate features for the whole 60 second long word production. For an overview of all features used, please see Appendix A. The most general and widely adopted measures of verbal fluency are the word count and repetition count (Spreen et al., 1991; Tombaugh et al., 1999). The word count is the count of all relevant words produced in (e.g. all words said start with the letter *F*), excluding repeated words. The repetition count is the number of words produce more than once.

4.4 Experiments

Statistical Analysis was done in R Studio (R Core Team, 2017). All coding experiments are implemented using python 3.7. For significance testing, a non-parametric Mann-Whitney U test for significance is always reported.

4.4.1 Comparing Strategic Processes With Binning Methods

To visualize what the strategic process over the duration of the PVF task, we plot the group averages of each feature across the bins. For overall performance, we plot the average word count and transition time by bin. To investigate semantic processes we plot the semantic distance between the words in each bin. To investigate the phonemic measures, we plot the LD, PHON-LD, and POS-LD.

In addition, we compute the bin average and standard error (se) for each group over all distance measures. A non-parametric Mann-Whitney U test for significance is reported to see if the bin averages differ between groups.

4.4.2 Classification Experiments

The classification models are created using the scikit-learn library[2] (Pedregosa et al., 2011).

For the classification application of these features, we focused on an early diagnostic scenario; distinguishing between healthy controls and mild cognitive impairment. To observe how age and education bias our classifier, we trained individual models on each potential bias (Nogueira et al., 2016; Petti et al., 2020). For the clinical baseline, a model was produced by training on only word count (word count) (Lindsay et al., 2020). To compare to previous work, a model was trained on number of repetitions (Ryan, 2013).

In addition to the baseline comparison experiments, we investigated individual and combined models. Four individual models were built using the features for semantic clustering, semantic binning, phonemic clustering or phonemic binning.

[1] https://fasttext.cc/docs/en/crawl-vectors.html

[2] sklearn version==0.24.0 for python 3.7

To investigate the proposed analysis modes and cognitive strategies, we built four combined models; all binning features (binning), all clustering features (clustering), all semantic features (semantic), and all phonemic features (phonemic).

Finally, we investigate a model using all features (All) and compare the models performance to the proposed baselines.

Classification Specifications To compare these methods, the extremely randomized trees (also known as extra trees) algorithm is used to train a classifier for each experimental scenario. This algorithm was chosen due to its ability to reduce variance and lesser likelihood of overfitting on a relatively small dataset with high dimensionality. Due to the limited amount of data available (34HC/48MCI), training-testing data splits were created using leave one out cross validation to maximize the amount of training data available, while still testing on every available data point. Due to the extreme randomness of the algorithm chosen, performance metrics can fluctuate between runs. To nullify the potential of the bias effects of random initialization, the experiment is repeated 50 times. For each model, the Area Under the Receiver Operator Curve (AUC) is averaged of the 50 iterations and reported.

5 Results

Results from the experiments to investigate strategic process as described in Section 4.4.1 are visualized in Figure 1. Significance testing between the HC and MCI groups are given in Table 2

5.1 Strategic Processes

For all binning features, excluding word count, a lower average bin distance represents a higher similarity between adjacent words. Compared to the HC group, the MCI group has a lower average word count, is less semantically motivated and more phonemically related. They also have longer transition times. The MCI group also show significantly smaller phonemic cluster (p=0.03) and lower number of semantic switches (p=0.01).

5.2 Classification results

To reduce the complexity of Figure 2, baseline and combined classifications are visualized with ROC-AUC curves and additional classification experiments are reported in the text of this section.

Both the age (AUC=0.41) and education (AUC=0.24) models perform below chance. The most common clinical evaluation, word count, performs at chance (AUC=0.50). The model trained using all features (AUC=0.71) proposed in this study improves over all baselines including the previous Ryan (2013) model (AUC=0.42) by 29 points.

Not shown in Figure 2, we compare each of the semantic and phonemic process in combination with the binning and clustering methods. Semantic clustering methods (AUC=0.61) achieve similar performance when used for binning (AUC=0.64) where as phonemic features are best when combined with the binning methods (AUC=0.70) but perform poorly for clustering (AUC=0.45).

As shown in Figure 2, the combined binning methods (AUC=0.67) perform similarly to the combined clustering methods (AUC=0.64). The combined phonemic features (AUC=0.76) perform the best overall for the early diagnostic classification scenario.

6 Discussion

The phonemic verbal fluency task remains underexplored in its use for clinical assessment as well as research of MCI.

However, in this paper we show, that with state-of-the-art semantic as well as phonemic distance metrics, the PVF can reveal neurocognitive function involvement that is crucial to better assess MCI. Our data shows that with recent semantic and phonemic similarity metrics, we can capture MCI-related impairments, such as a general semantic impairment, that have also been reported in the SVF (Verma and Howard, 2012; Taler and Phillips, 2008) but not on the PVF. Our results show significantly lower semantic distance for HC responses when compared to the MCI group in the PVF task which is, by nature, phonemically motivated. In return, MCI patients show significantly lower phonemic distance. This could possibly be explained by the MCI group relying heavily on a phonemic strategy to guide their search rather than a utilizing a semantic strategy. The higher semantic distance for the MCI group could be interpreted as a structural deficit to access semantic memory efficiently as has been shown to be very prominent at all stages of AD-related dementia (Verma and Howard, 2012).

This is especially striking as one would expect

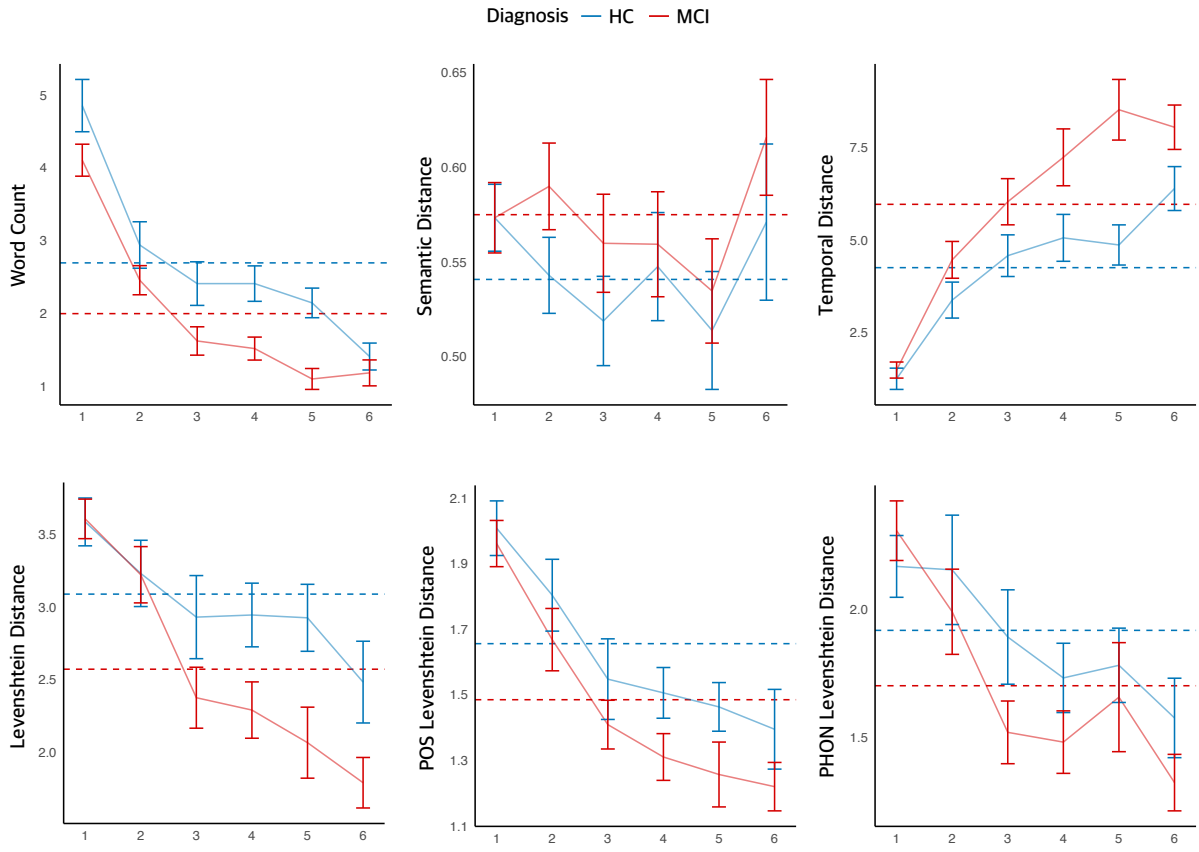

Figure 1: Graphical representation of binning results for each distance measure. Standard error bars are given for the HC and MCI groups at each bin. The dashed line represents the group average overall bins. For interpreting semantic and phonemic (LD, POS-LD, PHON-LD) distance metrics, a lower distance is interpreted as indicating a higher similarity.

the phonemic distance to increase as more words are produced (with a larger number of words per bin, the mean distance of adjacent words should be higher). Such an increase is the case for the phonemic distance where MCIs produce fewer words overall and are more phonemically related in comparison to HC, who produce more words and have a larger average phonemic distance over the bins. However, the exact opposite is the case for the semantic distance where MCIs produce fewer words while generating a list of less semantically related words in comparison to the HC group. This strongly points towards the conclusion that MCI patients struggle to exploit the associative network of their semantic memory.

By making neurocognitive processes visible in the PVF that are traditionally reserved for the SVF in clinical practice, the PVF becomes significantly more relevant to real-world MCI and dementia assessment. In order to support the diagnostic usage of the PVF for MCI assessment, we simulate a diagnostic decision scenario through downstream machine learning classification using the semantic as well as phonemic features in the PVF. Our results show that by using semantic and phonemic features we can improve classification results over previous clinical and automatic baselines. The all features model (AUC=0.71) out performs both the word count (AUC=0.50) and previous work of Ryan (2013) (AUC=0.42).

Both clustering (AUC=0.64) and binning (AUC=0.67) methods of analysis perform comparatively. Both the semantic (AUC=0.65) and phonemic (AUC=0.76) measures outperform the clinical baselines (0.50). The classification results support that while the task is overall a phonemic task, semantic investigation of the PVF is relevant for future research and capable of discriminating between HC and MCI better than the clinical baseline.

As an additional finding, the machine learning task benefits from a combined binning and cluster-

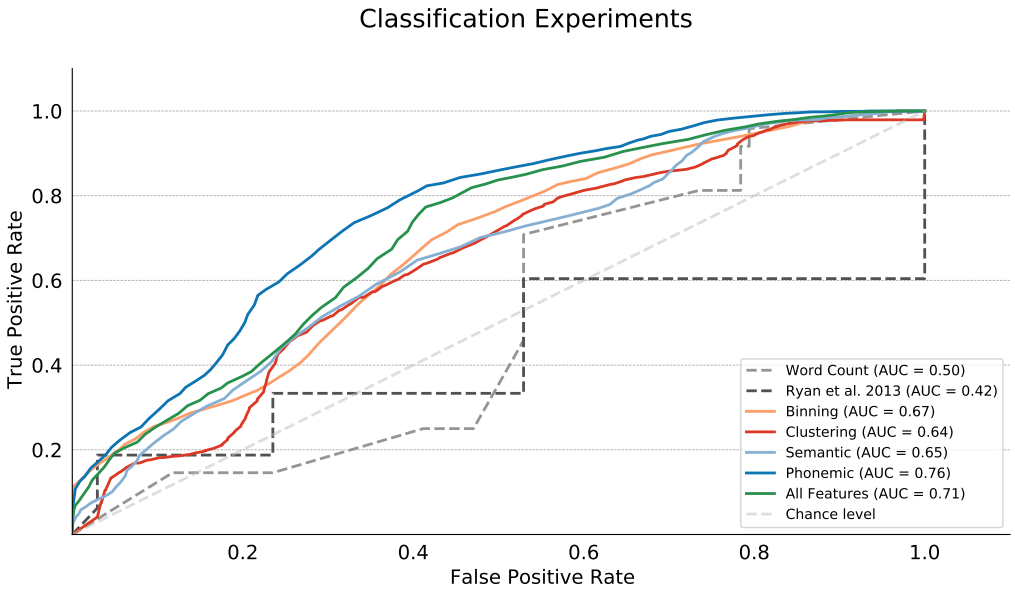

Figure 2: Visualzation of the ROC curve for the binary classification results between HC and MCI. Baseline methods are dashed in shades of gray. Ryan et al. 2013 is a previously published approach for comparison. Resolution modes are given in red. Strategy classifications are given in blue. The over all experiment is in green. AUC scores are given in the legend in the lower right corner. A perfect classification is 1.0. Chance is illustrated at 0.50.

ing approach when modelling the phonemic processes (AUC=0.76), increasing over only phonemic clustering (AUC=0.45) or phonemic binning methods (AUC=0.70) for classification.

7 Conclusion

This paper set out to investigate the ability of computational linguistic techniques for understanding phonemic and semantic cognitive processes of the under-explored phonemic verbal fluency task. Utilizing three resolutions of analysis, temporal binning, clustering and global measures, combined with semantic and phonemic distance measures, we found semantic impairment in a phonemic task as has been hypothesized in previous clinical research. In addition to giving a finer-resolution for understanding the PVF task, the additional phonemic and semantic features improved classification over previous clinical and automatic baselines for early dementia detection with the PVF task. Future work should investigate these measures in additional languages and possibly combine the features presented in this paper with medical imaging techniques to see if the findings can be replicated.

Acknowledgements

This research was funded by MEPHESTO project Q10 (BMBF Grant Number 01IS20075).

References

Amy Barr and Jason Brandt. 1996. Word-list generation deficits in dementia. *Journal of clinical and experimental neuropsychology*, 18(6):810–822.

Natalia Becker and Jerusa Salles. 2016. Methodological criteria for scoring clustering and switching in verbal fluency tasks. *Psico-USF*, 21:445–457.

Rasmus M Birn, Lauren Kenworthy, Laura Case, Rachel Caravella, Tyler B Jones, Peter A Bandettini, and Alex Martin. 2010. Neural systems supporting lexical search guided by letter and semantic category cues: a self-paced overt response fmri study of verbal fluency. *Neuroimage*, 49(1):1099–1107.

Ilaria Bizzozero, Stefania Scotti, Francesca Clerici, Simone Pomati, Marcella Laiacona, and Erminio Capitani. 2013. On which abilities are category fluency and letter fluency grounded a confirmatory factor analysis of 53 alzheimer's dementia patients. *Dementia and geriatric cognitive disorders extra*, 3(1):179–191.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (version 5.1.13).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jane Cerhan, Robert Ivnik, Glenn Smith, Eric Tangalos, Ronald Petersen, and Brad Boeve. 2002. Diagnostic utility of letter fluency, category fluency, and fluency

difference scores in alzheimer's disease. *The Clinical neuropsychologist*, 16:35–42.

Howard Chertkow and Daniel Bub. 1990. Semantic memory loss in dementia of alzheimer's type. *Brain : a journal of neurology*, 113 (Pt 2):397–417.

David Clark, Virginia Wadley, P. Kapur, Thomas DeRamus, Brandon Singletary, Anthony Nicholas, P.D. Blanton, K. Lokken, Hrishikesh Deshpande, D. Marson, and Georg Deutsch. 2013. Lexical factors and cerebral regions influencing verbal fluency performance in mci. *Neuropsychologia*, 54.

L. J. Clark, M. Gatz, L. Zheng, Y. L. Chen, C. McCleary, and W. J. Mack. 2009. Longitudinal Verbal Fluency in normal Aging, Preclinical, and Prevalent Alzheimer's Disease. *Am J Alzheimers Dis Other Demen*, 24(6):461–468.

H. Coslett, Dawn Bowers, Mieke Verfaellie, and K Heilman. 1991. Frontal verbal amnesia. phonological amnesia. *Archives of neurology*, 48:949–55.

Unai Diaz-Orueta, Alberto Blanco-Campal, Melissa Lamar, David J. Libon, and Teresa Burke. 2020. Marrying past and present neuropsychology: Is the future of the process-based approach technology-based? *Frontiers in Psychology*, 11:361.

B. Dubois, A. Slachevsky, I. Litvan, and B. Pillon. 2000. The fab. *Neurology*, 55(11):1621–1626.

Miriam Faust, editor. 2012. *The handbook of neuropsychology of language. Volume 1: Language processing in the brain: basic science. Volume 2: Language processing in the brain: clinical populations*. Wiley-Blackwell, Chichester. Bibtex: faust_handbook_2012.

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.

Nancy J. Fisher, Mary C. Tierney, Byron P. Rourke, and John P. Szalai. 2004. Verbal fluency patterns in two subgroups of patients with alzheimer's disease. *The Clinical Neuropsychologist*, 18(1):122–131. PMID: 15595364.

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

John R Hodges. 2006. Alzheimer's centennial legacy: origins, landmarks and the current status of knowledge concerning cognitive aspects. *Brain*, 129(11):2811–2822.

John R. Hodges, David P. Salmon, and Nelson Butters. 1992. Semantic memory impairment in alzheimer's disease: Failure of access or degraded knowledge? *Neuropsychologia*, 30(4):301–314.

E. Huey, E. Goveia, S. Paviol, M. Pardini, F. Krueger, G. Zamboni, M. Tierney, E. Wassermann, and J. Grafman. 2009. Executive dysfunction in frontotemporal dementia and corticobasal syndrome. *Neurology*, 72:453 – 459.

Anastasios Karakostas, Alexia Briassouli, Konstantinos Avgerinakis, Ioannis Kompatsiaris, and Magda Tsolaki. 2017. The dem@care experiments and datasets: a technical report. *CoRR*, abs/1701.01142.

Parris Kidd. 2008. Alzheimer's disease, amnestic mild cognitive impairment, and age-associated memory impairment: Current understanding and progress toward integrative prevention. *Alternative medicine review : a journal of clinical therapeutic*, 13:85–115.

Najoung Kim, Jung-Ho Kim, Maria K. Wolters, Sarah E. MacPherson, and Jong C. Park. 2019. Automatic scoring of semantic fluency. *Frontiers in Psychology*, 10:1020.

A. König, N. Linz, J. Töger, M. Wolters, J. Alexandersson, and P. Robert. 2018. Fully automatic analysis of semantic verbal fluency performance for the assessment of cognitive decline. *Dementia and Geriatric Cognitive Disorders*. Accepted.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals.

Muriel Deutsch Lezak, Diane B Howieson, David W Loring, Jill S Fischer, et al. 2004. *Neuropsychological assessment*. Oxford University Press, USA.

Yunqing Li, Ping Li, Qing X. Yang, Paul J. Eslinger, Chris T. Sica, and Prasanna Karunanayaka. 2017. Lexical-semantic search under different covert verbal fluency tasks: An fmri study. *Frontiers in Behavioral Neuroscience*, 11:131.

Hali Lindsay, Nicklas Linz, Johannes Tröger, and Jan Alexandersson. 2019. Automatic data-driven approaches for evaluating the phonemic verbal fluency task with healthy adults. In *ICNLSP*.

Hali Lindsay, Johannes Tröger, Jan Alexandersson, and Alexandra König. 2020. What difference does it make? early dementia detection using the semantic and phonemic verbal fluency task. In *LREC 2020 Workshop RaPID-3: Resources and ProcessIng of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments*.

Nicklas Linz, Kristina Lundholm Fors, Hali Lindsay, Marie Eckerström, Jan Alexandersson, and Dimitrios Kokkinakis. 2019. Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 103–113, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Nicklas Linz, Johannes Tröger, Jan Alexandersson, Maria Wolters, Alexandra König, and Philippe Robert. 2017b. Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728.

Sarah MacPherson, Colm Healy, Michael Allerhand, Barbara Spano, Carina Tudor-Sfetea, Mark White, Daniela Smirni, Tim Shallice, Edgar Chan, Marco Bozzali, and Lisa Cipolotti. 2016. Cognitive reserve and cognitive performance of patients with focal frontal lesions. *Neuropsychologia*, 96.

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

Camillo Marra, Monica Ferraccioli, Maria Vita, Davide Quaranta, and Guido Gainotti. 2011. Patterns of cognitive decline and rates of conversion to dementia in patients with degenerative and vascular forms of mci. *Current Alzheimer research*, 8:24–31.

Edgar Miller. 1984. Verbal fluency as a function of a measure of verbal intelligence and in relation to different types of cerebral pathology. *British Journal of Clinical Psychology*, 23(1):53–57.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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.

Kelly J. Murphy, Jill B. Rich, and Angela K. Troyer. 2006. Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of alzheimer's type dementia. *Journal of the International Neuropsychological Society*, 12(4):570–574.

Freda Newcombe. 1969. Missile wounds of the brain: A study of psychological deficits.

Dalia Santos Nogueira, Elizabeth Azevedo Reis, and Ana Vieira. 2016. Verbal fluency tasks: Effects of age, gender, and education. *Folia Phoniatrica et Logopaedica*, 68(3):124–133.

Katherine E Nutter-Upham, Andrew Saykin, Laura Rabin, Robert Roth, Heather Wishart, Nadia Pare, and Laura Flashman. 2008. Verbal fluency performance in amnestic mci and older adults with cognitive complaints. *Arch Clin Neuropsychol*, 23(3):229–41.

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

Serguei V.S. Pakhomov, David T. Jones, and David S. Knopman. 2015a. Language networks associated with computerized semantic indices. *NeuroImage*, 104:125–137.

Serguei V.S. Pakhomov, Susan E. Marino, Sarah Banks, and Charles Bernick. 2015b. Using Automatic Speech Recognition to Assess Spoken Responses to Cognitive Tests of Semantic Verbal Fluency. *Speech Communication*, 75:14–26.

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.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Barbara Rende, Gail Ramsberger, and Akira Miyake. 2002. Commonalities and differences in the working memory components underlying letter and category fluency tasks: A dual-task investigation. *Neuropsychology*, 16:309–21.

Eric Rinehardt, Katie Eichstaedt, John A Schinka, David A Loewenstein, Michelle Mattingly, Jean Fils, Ranjan Duara, and Mike R Schoenberg. 2014. Verbal fluency patterns in mild cognitive impairment and alzheimer's disease. *Dementia and geriatric cognitive disorders*, 38(1-2):1–9.

Wilma G Rosen. 1980. Verbal fluency in aging and dementia. *Journal of clinical and experimental neuropsychology*, 2(2):135–146.

James Ryan. 2013. *A System for Computerized Analysis of Verbal Fluency Tests*. Ph.D. thesis.

Charlotte SM Schmidt, Lena V Schumacher, Pia Römer, Rainer Leonhart, Lena Beume, Markus Martin, Andrea Dressing, Cornelius Weiller, and Christoph P Kaller. 2017. Are semantic and phonological fluency based on the same or distinct sets of

cognitive processes? insights from factor analyses in healthy adults and stroke patients. *Neuropsychologia*, 99:148–155.

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*, 5:772.

O. Spreen, P.P.O. Spreen, E. Strauss, and P.P.E. Strauss. 1991. *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. Maestría de neuropsicología. Oxford University Press.

Vanessa Taler and Natalie A Phillips. 2008. Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology*, 30(5):501–556.

Edmond Teng, Judith Leone-Friedman, Grace J. Lee, Stephanie Woo, Liana G. Apostolova, Shelly Harrell, John M. Ringman, and Po H. Lu. 2013. Similar Verbal Fluency Patterns in Amnestic Mild Cognitive Impairment and Alzheimer's Disease. *Archives of Clinical Neuropsychology*, 28(5):400–410.

Sharon L Thompson-Schill, Mark D'Esposito, Geoffrey K Aguirre, and Martha J Farah. 1997. Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences*, 94(26):14792–14797.

Jochen René Thyrian, Tilly Eichler, Andrea Pooch, Kerstin Albuerne, Adina Dreier, Bernhard Michalowsky, Wolfgang Hoffmann, and Diana Wucherer. 2016. Systematic, early identification of dementia and dementia care management are highly appreciated by general physicians in primary care - results within a cluster-randomized-controlled trial (delphi). *Journal of Multidisciplinary Healthcare*, 9:183.

Tom N Tombaugh, Jean Kozak, and Laura Rees. 1999. Normative data stratified by age and education for two measures of verbal fluency: Fas and animal naming. *Archives of Clinical Neuropsychology*, 14(2):167–177.

Johannes Tröger, Nicklas Linz, Jan Alexandersson, Alexandra König, and Philippe Robert. 2017. Automated Speech-based Screening for Alzheimer's Disease in a Care Service Scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*.

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.

Angela K Troyer, Morris Moscovitch, Gordon Winocur, Michael P Alexander, and Don Stuss. 1998. Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia*, 36(6):499 – 504.

Johannes Tröger, Nicklas Linz, Alexandra König, P. Robert, Jan Alexandersson, Jessica Peter, and Jutta Kray. 2019. Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in alzheimer's disease. *Neuropsychologia*, 131.

Roisin M. Vaughan, Robert F. Coen, RoseAnne Kenny, and Brian A. Lawlor. 2016. Preservation of the semantic verbal fluency advantage in a large population-based sample: Normative data from the tilda study. *Journal of the International Neuropsychological Society*, 22(5):570–576.

Malvika Verma and Robert J Howard. 2012. Semantic memory and language dysfunction in early alzheimer's disease: a review. *International journal of geriatric psychiatry*, 27(12):1209–1217.

Mathieu Vigneau, V Beaucousin, Pierre-Yves Hervé, Hugues Duffau, Fabrice Crivello, O Houdé, Bernard Mazoyer, and N Tzourio-Mazoyer. 2006. Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *NeuroImage*, 30:1414–32.

Maria Vita, Camillo Marra, Pietro Spinelli, Alessia Caprara, Eugenia Scaricamazza, Diana Castelli, Serena Canulli, Guido Gainotti, and Davide Quaranta. 2014. Typicality of words produced on a semantic fluency task in amnesic mild cognitive impairment: Linguistic analysis and risk of conversion to dementia. *Journal of Alzheimer's disease : JAD*, 42.

Jeffrey C Zemla, Kesong Cao, Kimberly D Mueller, and Joseph L Austerweil. 2020. Snafu: The semantic network and fluency utility. *Behavior research methods*, pages 1–19.

Qianhua Zhao, Qihao Guo, and Zhen Hong. 2013. Clustering and switching during a semantic verbal fluency test contribute to differential diagnosis of cognitive impairment. *Neuroscience bulletin*, 29(1):75–82.

A Appendix

Category	Feature Name	Description
Global Features	*Measures that span over the task as a whole*	
	Word Count	The total number of words excluding repetitions. Scoring system used in clinical practice
	Number of Repetitions	Number of repetitions said during the task. Previously suggested in Ryan (2013).
Phonemic Features	*Rule-based measures for phonemic clustering strategies proposed by Troyer et al. (1997) and automated by Lindsay et al. (2019)*	
	Mean Cluster Size	Average number of words in clinical phonemic clusters
	Number of Switches	Total number of switches between clinical phonemic clusters
Semantic Features	*Automatic data-driven methods for determining semantically motivated clusters as proposed in Linz et al. (2017a)* ...	
	Mean Cluster Size	Average number of words in a semantic cluster
	Number of Switches	Total number of switches between semantic clusters
Binning Features	*10-second binning approach for finer resolution of task proposed by Linz et al. (2019); The following features are computed for each of the six, 10-second bins.*	
	Word Count by Bin	The number of words per 10 second bin
	LD by Bin	Levenshtein distance per 10 second bin
	POS-LD by Bin	Position-weighted Levenshtein distance per 10 second bin
	PHON-LD by Bin	Phonemic-weighted Levenshtein distance per 10 second bin
	Semantic Distance by Bin	Semantic Distance between consecutive words per 10 second bin
	Mean Temporal Distance by Bin	The average transition time in seconds between the end of one word and the onset of the next word by 10 second bin

Table 3: The following features were extracted from the PVF task produced by the participants.