HHS Public Access

Author manuscript

Cortex. Author manuscript; available in PMC 2015 June 01.

Published in final edited form as:

Cortex. 2014 June ; 55: 97–106. doi:10.1016/j.cortex.2013.05.009.

A Computational Linguistic Measure of Clustering Behavior on Semantic Verbal Fluency Task Predicts Risk of Future Dementia in the Nun Study

Serguei V.S. Pakhomov[1] and **Laura S. Hemmy**[2,3]

[1]University of Minnesota Center for Clinical and Cognitive Neuropharmacology

[2]Department of Psychiatry, University of Minnesota, MN

[3]Geriatric Research Education and Clinical Center (GRECC), Minneapolis VA System

Abstract

Generative semantic verbal fluency (SVF) tests show early and disproportionate decline relative to other abilities in individuals developing Alzheimer's disease. Optimal performance on SVF tests depends on the efficiency of using clustered organization of semantically related items and the ability to switch between clusters. Traditional approaches to clustering and switching have relied on manual determination of clusters. We evaluated a novel automated computational linguistic approach for quantifying clustering behavior. Our approach is based on Latent Semantic Analysis (LSA) for computing strength of semantic relatedness between pairs of words produced in response to SVF test. The mean size of semantic clusters (MCS) and semantic chains (MChS) are calculated based on pairwise relatedness values between words. We evaluated the predictive validity of these measures on a set of 239 participants in the Nun Study, a longitudinal study of aging. All were cognitively intact at baseline assessment, measured with the CERAD battery, and were followed in 18 month waves for up to 20 years. The onset of either dementia or memory impairment were used as outcomes in Cox proportional hazards models adjusted for age and education and censored at follow up waves 5 (6.3 years) and 13 (16.96 years). Higher MCS was associated with 38% reduction in dementia risk at wave 5 and 26% reduction at wave 13, but not with the onset of memory impairment. Higher (+1 SD) MChS was associated with 39% dementia risk reduction at wave 5 but not wave 13, and association with memory impairment was not significant. Higher traditional SVF scores were associated with 22–29% memory impairment and 35–40% dementia risk reduction. SVF scores were not correlated with either MCS or MChS. Our study suggests that an automated approach to measuring clustering behavior can be used to estimate dementia risk in cognitively normal individuals.

Corresponding Author Information: Serguei V.S. Pakhomov, 7-125F Weaver Densford Hall, 308 Harvard Street SE, Minneapolis, MN 55455, Tel: (00+1) 612-624-1198, Fax: (00+1) 612-625-9931, pakh0002@umn.edu.

Keywords

semantic verbal fluency; dementia; latent semantic analysis; clustering; Alzheimer's disease

Introduction

Tests of phonemic and semantic verbal fluency are widely used in the assessment of individuals with memory complaints and in the clinical diagnosis of Alzheimer's disease (AD). Patients are asked to generate as many words as they can either starting with a certain letter of the alphabet (phonemic fluency) or belonging to a certain semantic category (semantic fluency). The performance on the semantic verbal fluency (SVF) test is typically measured by counting the number of correct words spoken by the patient in one minute. Although individual performance on both the phonemic and semantic fluency tests is impaired in individuals with AD, semantic performance is significantly more affected than phonemic early in the disease course. Deficits evident on SVF have been shown to be sensitive and specific in differentiating between healthy controls and patients with AD (Canning et al., 2004) and have been found to be predictive of the development of AD in people with memory complaints prior to a clinical diagnosis (Fagundo et al., 2008), as well as progression of healthy individuals to mild cognitive impairment (MCI) and dementia (Loewenstein et al., 2012). In addition to clinical use, these relatively simple to administer tests of verbal fluency have been used extensively as part of standard neuropsychological test batteries to study cognitive effects of dementia (Gorno-Tempini et al., 2004; Hodges et al., 2004; Knopman et al., 2008; Libon et al., 2007). In particular, SVF deficits have been widely reported in patients with various stages of AD and MCI (Chan et al., 2001; Ober et al., 1986; Rosen, 1980; Troyer et al., 1998) and often show early and disproportionate decline relative to other language, attention, and executive abilities (see Lezak, (2004), and Henry et al., (2004) for review).

Originally proposed by Troyer et al. (1997), additional qualitative metrics that examine the degree to which SVF responses are organized into groups of semantically related words (semantic clusters) and the frequency of transitions between these groups (switches) have also been extensively studied. Successful overall performance on the SVF test depends to a large extent on how well semantic information is organized into conceptually related clusters and whether one is able to use an efficient strategy that accesses these clusters during the test (Estes, 1974; Hodges and Patterson, 1995; Laine, 1988). The size of semantic clusters and the efficiency of switching from one cluster (after it has been exhausted) to another have been found to have different neuroanatomical correlates (Rich et al., 1999; Troyer, 2000; Troyer et al., 1997). Semantic cluster size was found to be associated with the left temporal lobe, whereas the processing associated with switching was associated more with the function of the frontal lobe. These studies suggested that cluster size and the amount of switching between clusters may index the strength of associations in the patient's lexical-semantic networks.

Studies of clustering and switching in AD also found smaller and fewer clusters produced on this task by people clinically diagnosed with AD (Ober et al., 1986; Rosen, 1980; Troster et

al., 1998; Troyer et al., 1998) and significantly fewer switches (Raoux et al., 2008) than by healthy controls. The latter finding is not surprising as the number of switches in a time-limited task such as the SVF test is likely to be highly correlated with the total number of words produced on the test. This is particularly true of approaches such as the original Troyer et al. (1997) methodology in which single words unrelated to adjacent words are counted as clusters of zero size. This approach results in discounting these "singleton" clusters in the computation of the mean cluster size but not the number of switches for any given SVF sample. When participants tend to produce small clusters, and a switch is defined as a transition between any two clusters (even those consisting of a single word), the number of switches is likely to be highly correlated with the total number of words produced, and thus not very informative. An alternative measure that normalizes the raw number of switches by the total number of words produced was proposed but was not found to add meaningful information (Troster et al., 1998).

Subsequent studies that examined clustering and switching performance in less impaired individuals with memory complaints (e.g., MCI) consistently found significant differences between healthy controls and MCI participants on the traditional SVF score and switching, but not so with cluster size. For example, Raoux et al. (2008) examined a set of 153 participants including 51 incident cases of AD at 2 and 5 years prior to diagnosis. Significant cross-sectional differences between controls and the incident AD cases were found at all three time points in the traditional SVF score and the number of switches, but no difference was found in mean cluster size. In contrast, a recent study by Price et al. (2012) examined a group of 33 amnestic MCI participants matched on age, education and IQ with 33 healthy controls, and found significant differences in both the traditional SVF score and mean cluster size measures. Both of these studies calculated the cluster sizes according to Troyer et al. (1997) guidelines with one major difference, however. Rosen's study excluded repetitions from the calculation of cluster sizes, whereas Price's study followed Troyer's methodology more strictly and inlcuded all repetitions and errors in computation of clusters. It is unclear at this point if the inclusion/exclusion of repetitions in cluster size determination is responsible for the inconsistency in the results and conclusions; however, this is a good example where the qualitative nature of the methodology is a problem that we are attempting to address with automation in the present study.

All prior studies of clustering and switching in relation to dementia have relied on subjective assessments of semantic similarity between at least two (Rich et al., 1999; Troyer et al., 1997; Troyer et al., 1998) or three (Laine, 1988) adjacent words to define clusters. For example, the qualitative assessment proposed by Troyer et al. (1997) relies on manual determination if adjacent words belong to a top-level subcategory (e.g., zoological categories, human use, and living environment) with further more fine-grained subcategorizations (e.g. living environment category composed of African, Australian, Arctic/Far North, Farm, North American and Water Animals). In addition to their subjectivity, these manual approaches are time consuming and are difficult to implement and standardize, which may be responsible for some of the conflicting results obtained with these methods. Automated, computerized approaches to the assessment of clustering and switching behavior may help address some of the issues with qualitative approaches.

Independent of the efforts of neurologists and neuropsychologists, workers in the field of computational linguistics have developed a number of fully automated approaches to representing the degree to which any two words in a given language are semantically related (Pedersen et al., 2007; Rada et al., 1989; Resnik, 1998). Many of these approaches utilize variations on a technique called Latent Semantic Analysis (LSA: (Landauer, 2006; Landauer and Dumais, 1997)). This approach is described in detail in the Methods section. In brief, LSA relies on the co-occurrence of words in a large corpus of text consisting of newspaper articles, books, speeches and other sources of typical word usage, to represent the semantic content of a word or a term as a set of co-occurrence counts with other words used in the same context. These semantic representations can then be directly compared to each other to assign a numeric value indicative of the strength of semantic relatedness between them.

The objective of the current study was to determine if LSA-based automation of the assessment of clustering and switching behavior on SVF tests can produce measurements associated with the risk of developing dementia. We hypothesized that traditional manual approaches to determining the size of semantic clusters may be limited in their ability to capture the richness of associative semantic relations between concepts. Manual approaches are inherently constrained to relying on relatively small and artificial categorization schemes that are categorical rather than continuous in nature, and may not be granular enough to represent the nuances of associations between concepts in the mind of any given individual. Increasing the granularity of manual categorization schemes for natural language semantics inevitably leads to worsening of reliability of semantic relatedness judgments made based on these categorization schemes (Poesio and Vieira, 1998), thus imposing a practical limitation on the granularity of categorization. Automated approaches that utilize large text corpora of word usage in a given language are promising in that they can provide fine-grained and/or continuous scale measurements of the strength of semantic associations between concepts. Furthermore, apart from automation, our study is also different from prior attempts at examining the relationship between clustering and switching behavior and dementia in that we use a large longitudinal sample with longer follow up time than previously reported to determine if clustering measures predict future dementia and memory problems.

Methods

Participants

All data were obtained as part of a Human Studies IRB approved protocol for the University of Minnesota Nun Study. The Nun Study is a longitudinal study of aging in 678 U.S. School Sisters of Notre Dame aged 75+ years. The participants in the Nun Study underwent cognitive assessments at regular intervals (waves of approximately 18 months) for up to 20 years of follow up. Participants in the current study were limited to those sisters with two or more evaluation time points and intact cognitive function at baseline evaluation, resulting in a set of 239 subjects.

Cognitive Assessments

All participants underwent the standard Consortium to Establish a Registry for Alzheimer's Disease (CERAD) neuropsychological test battery and Mini-Mental State Examination

(MMSE). The CERAD battery included the semantic verbal fluency test ("animals" category). Participant responses on this test were recorded by the test examiner on a psychometric sheet that was subsequently used to calculate the traditional SVF score as the number of correct words excluding repetitions, intrusions and perseverations. However, all words recorded on psychometric sheets including errors and repetitions were used in calculating cluster size as detailed in the following section.

In addition to the SVF test, each CERAD assessment included a 10-item word list learning task and delayed free recall test designed to assess memory impairment (Fillenbaum *et al*., 2008). Performance on this test with a cutoff of 5 words on delayed free recall (WRCL) was used to define onset of future memory impairment. Risk of memory impairment was one of the two outcomes used in proportional hazards modeling. The other outcome was risk of Diagnostic and Statistical Manual (DSM-IV) dementia (American Psychiatric Association., 2000), operationalized in the Nun Study as the presence of memory impairment, impairment in at least one other cognitive domain, and impairment on performance-based measures of functional ability (activities and instrumental activities of daily living). For both of these outcomes, two censor variables were created – one at wave 5 and one at wave 13 – resulting in four censor variables total.

Automated Semantic Relatedness Computation

We followed the methodology described in Troyer et al. (1997) to determine the size of clusters, but substituted automated quantitative semantic relatedness assessments for qualitative judgments. To compute semantic relatedness between pairs of words, we relied on LSA. In order to compare the meanings of any given pair of words, we first represent the semantic content of each word as a set of other words found in the same context as the target word in a collection of texts. We chose Wikipedia entries as a convenient and relatively complete source of textual co-occurrence information in definitions of animal names. For example, the meaning of the word "tiger" is represented as a set of all other words that are found in the Wikipedia entry for "tiger" after exclusion of function words (e.g., the, he, she, it, on, at, etc.) including: "panthera", "tigris", "largest", "cat", "species", "most", "recognizable", "feature", "pattern", "dark", "vertical", "stripes", "reddish", "orange", "fur", "Russia", "Bangladesh", "India", "siberian", "asia", among others. Similarly, the meaning for the word "lion" contains words "panthera", "leo", "four", "big", "cats", "genus", "panthera", "second", "largest", "living", "cat", "tiger", "currently", "exist", "subsaharan", "Africa", "asia." Of course, the Wikepidia entries also contain many words irrelevant to comparing the meanings of the target words that need to be filtered out prior to the comparison. Furthermore, in some cases, it is useful to take advantage of the fact that a given pair of words may not appear in the same context but may still be linked through their co-occurrence with a third word. For example, the words, "tigris" and "leo" may never occur in the same Wikepedia entry but the word "tigris" co-occurs with "pantera" and so does the word "leo", thus forming a latent semantic association. LSA is a computational technique that operates by constructing a co-occurrence matrix for all words found in a given corpus of text (e.g., Wikipedia) and applying a variant of principal components analysis to filter out irrelevant words (dimensions) through singular value decomposition and to identify latent semantic associations between the words.

Applying LSA to Wikipedia entries for all animal names that were produced by the participants in our study resulted in representing each animal name as a vector in N-dimensional semantic space (described in more detail in the Results section). We then used the resulting vectors to compare pairs of animal names in this N-dimensional space by computing the cosine of the angle between the vectors. The cosine value ranges from -1 showing that the two vectors are at a 180 degree angle, (i.e., they are pointing in opposite directions), to zero corresponding to a 90 degree angle, (i.e., the two vectors are orthogonal), to 1 - zero degree angle (i.e., the two vectors are pointing in exactly the same direction). These cosine values between vectors in the N-dimensional semantic space can be interpreted in terms of semantic relatedness between the words that the vectors represent. We use the cosine values to compute measures of clustering behavior on responses to SVF test, as described in the next two sections.

Automatic Determination of Clusters in SVF responses

Based on the semantic relatedness tools described in the previous section, we developed an automated approach that follows the clustering and switching analysis introduced by Troyer et al. (1997) as closely as possible. Troyer and colleagues calculated cluster size by starting to increment the count of words in the cluster from the second word in the cluster, which makes the size of single-word clusters equal to zero. Thus, the size of two-word clusters was one, three-word clusters – two, and so on. Errors and repetitions were included. The mean clusters size was calculated by averaging the sizes of individual clusters. The lists of animals semantic in categories and subcategories used in the manual determination of clusters are too long to be repeated here but can be found in the Appendix to the original Troyer et al. (1997) publication. The only significant departure from Troyer's approach in the current implementation is that the qualitative human judgments of whether any given pair of words belongs to the same semantic category are replaced by quantitative semantic relatedness scores dichotomized with a threshold of 0.90 to identify very closely related pairs (i.e., their semantic vectors point almost in the same direction). Thus, if the relatedness value between two words exceeds this threshold, the words (or to be more precise, the animal senses of these words) are treated as belonging to the same cluster. The relatedness threshold of 0.9 was based on the desire to have a starting point which eliminates as much potential noise from clustering as possible and only retain those pairs of animals that are clearly highly related – i.e. maximize the specificity, perhaps at the cost of sensitivity. To arrive at this threshold, we considered the distribution of all LSA-derived relatedness values on all pairs of animals in the participants' responses. The threshold of 0.9 represents relatedness values in the top 10% of all computed values. The rest of clustering and switching computation was exactly the same as described by Troyer et al. (1997). It should be noted that we followed Troyer's methodology to make our results interpretable in light of the prior work that also relied on this methodology; however, automated computation of semantic relatedness enables further experimentation with modified approaches. Based on this automated approach to determining semantic clusters, we computed two measures: mean cluster size (MCS) and mean chain size (MChS). The former is the mean size of groups of words in which the value of LSA-based semantic relatedness between each word and all other words in the group is above the 0.9 threshold. The latter is the mean size of groups of words in which each word is closely related only to its immediate neighbors.

Validation of automated semantic relatedness measures

LSA relies on a number of parameters that can influence the resulting relatedness values between pairs of words. Two parameters are most important: a) choice of a corpus of textual data to obtain co-occurrence counts, and b) choice of the number of latent semantic dimensions (Landauer, 2006). For this study, we limited the choice of text corpus to Wikipedia entries for animals combined with a method for finding the optimal number of dimensions that relies on calculating the proportion (share) of the sum of singular values for the first N dimensions to the total sum of singular values for all dimensions. The number of dimensions for LSA computation is difficult to estimate a priori. As described in detail in Quesada ((2011), p. 82), unlike other related techniques such as multi-dimensional scaling, LSA does not currently have a internal criterion or a theoretical way of determining the most optimal number of dimensions. Therefore, we followed the recommended practice for empirical determination of dimensionality using an external criterion. To do this we compared the ratings between pairs of words produced by the LSA methods to manual assessments reported in an independent study conducted by Weber and colleagues (Weber *et al.*, 2009). In that study, manual ratings were produced by twelve participants on all possible pairs of nine animals: bear, camel, cougar, dolphin, elephant, giraffe, hippopotamus, horse, lion. The animal names were presented on 36 stimulus cards along with the photographic images of the animals but the participants were instructed to rank the pairs based on conceptual (i.e., semantic) rather than visual similarity. The study found that the mean similarity scores computed based on these ratings were correlated with the similarity in hemodynamic response patterns obtained with fMRI imaging using the photographs of the animals as stimuli ($r = 0.65$, $p < 0.001$; p. 863). Thus, this dataset represents a neurophysiologically validated reference standard that was used in the present study for the purpose of independent testing of the semantic similarity measurements.

While the LSA-based approach used in the current study is designed to produce associative relatedness rather than similarity measurements, semantic similarity can be viewed as a special case of relatedness. For example, in the mind of a person making the judgments, animals that are related by association may not necessarily be viewed as similar (e.g., cat and mouse); however, animals that are considered similar are highly likely to form an associative relationship (through their physical appearance, if nothing else). Thus, we believe that, with respect to the animals category, Weber's dataset is well suited to estimate how well the LSA-based methodology represents human relatedness judgments without having to make a formal distinction between similarity and relatedness. The entire dataset with manually obtained semantic similarity scores is available as an Appendix in Weber et al.'s (2009) publication.

Statistical Analyses

Cox proportional hazards models were used to examine the effect of baseline verbal fluency performance variables on the relative risk of memory impairment and dementia outcomes. All models were adjusted for baseline age and years of education. The predictor variables (traditional SVF score, MCS, and MChS) were converted to z-scores in order to facilitate interpretation of the results and comparisons between measurements produced on different scales. Cross-sectional associations between variables at the baseline assessment were tested

using Pearson correlation. Adjustments for multiple testing were made using the Holm method (Holm, 1979). Statistical analyses were performed using SPSS statistical software.

Results

Demographic and Baseline Characteristics

The demographic characteristics of the 239 study participants are presented in Table 1. In general, participants were older (M age 80.73, SD 3.98) and more educated (M years of education 16.96, SD 1.62) than the general population. The mean baseline MMSE (M 28.39, SD 1.60), WRCL (M 7.06, SD 1.39) and traditional SVF (M 18.05, SD 4.75) scores were all within normal limits according to published age- and education-based normative data (Beeri *et al.*, 2006; Crum et al., 1993). The mean follow up times elapsed between the baseline assessment and waves 5 and 13, along with the censor variables used in Cox modeling are shown in Table 2.

Memory Impairment and Dementia Risk

The results of modeling of memory impairment and dementia risk using Cox proportional hazards models adjusted for age and education are presented graphically in Figure 1 and in Table 3. Overall, memory impairment was detected in 24.4% of the participants by 5 waves of follow up and 43% by 13 waves. Dementia was diagnosed in 17% of the participants by 5 waves of follow up and 35% by 13 waves of follow up. Prior to adjustment for multiple testing, the standard SVF score was found to be a significant predictor of both future memory impairment and dementia at 5 and 13 waves of follow up. The MCS and MChS measures were also significant predictors of risk, but only for dementia, not memory impairment. Including MCS or MChS as a covariate along with the SVF score, age and education did not show these variables (MCS or MChS) to be significant predictors – only the raw SVF score and age were significant in these models. Testing MCS and MChS for interaction with the SVF score did not produce significant results either. Only SVF, MCS and MChS variables included in models predicting onset of dementia at 5 and 13 waves survived the adjustment for multiple testing.

Correlations Between Assessments

The correlation between the SVF score and both MCS and MChS was small and non-significant ($r = 0.08$, $p = 0.19$ for SVF vs. MCS and $r = 0.04$, $p = 0.54$ for SVF vs. MChS). MCS was weakly but significantly correlated with WRCL scores at baseline assessment ($r = 0.14$, $p = 0.04$). The correlation between MChS and WRCL scores was also weak and not significant ($r = 0.12$, $p = 0.07$). The SVF score was more strongly correlated with WRCL ($r = 0.27$, $p < 0.001$). No other correlations were found between SVF test–related variables and either MMSE or WRCL.

LSA Dimensionality Evaluation

The number of dimensions for LSA was calculated using a threshold at the first position in the descending sequence of singular values where the ratio of their sum to the total sum of all singular values met or exceeded 0.2 – roughly 20% of all possible dimensions available during singular value decomposition step of LSA. In our case, this thesholding method

resulted in 30 dimensions. A scatter plot illustrating the relationship between manual and LSA-based values obtained with 30-dimensional LSA is shown in Figure 2. The correlation between LSA-based semantic relatedness values manually determined similarity values in the Weber's dataset was in the moderate range ($r = 0.65$, $p < 0.01$).

Discussion

The current study is one of the first to demonstrate the use of a fully automated computational linguistic technique to assess semantic clusters in SVF responses predictive of future cognitive impairment. We tested the hypothesis that clustering of responses produced on the SVF test based on automatically computed pairwise semantic relatedness measures is associated with risk of dementia and memory impairment. We tested this hypothesis in a large longitudinal sample of participants in the Nun Study that were cognitively intact at the baseline assessment. Our results show that the size of semantic clusters and chains identified with an automated technique are significantly associated with the risk of dementia but are not significantly associated with risk of memory impairment. For every positive standard deviation change in the size of semantic clusters, the risk of developing dementia in initially intact individuals is reduced by 38% in the next 6 years (wave 5 model) and 26% in the next 17 years (wave 13 model). Similarly, for every positive standard deviation change in the traditional SVF score, the risk of developing dementia is reduced by 40% in 6 years and 34% at 17 years. Although, the traditional SVF score as a predictor of dementia did not survive adjustment for multiple testing, we believe this finding is still meaningful in light of the fact that the variables chosen for analysis are well motivated by prior work and are unlikely to be spurious. Thus even with borderline statistical significance, from a clinical standpoint, the traditional SVF score and the measurements based on semantic relatedness are similar in terms of their ability to predict future dementia but different in their ability to predict future memory impairment. It is also worth pointing out that despite the similar relative risk estimates, the traditional SVF score is not correlated with either MCS or MChS suggesting that these measures reflect different underlying cognitive mechanisms or, perhaps, different aspects of the same cognitive mechanism that eventually becomes impaired in individuals that develop dementia and/or memory impairment.

The fact that cluster and chain size variables predict dementia risk but not risk of memory impairment, and the SVF score predicts both, may be due to the influence of an executive component of clustering behavior. Due to limitations of the early waves of cognitive testing in the Nun Study that did not include tests of executive function, we were not able to confirm this hypothesis using the present study sample. However, prior work on clustering and switching behavior in SVF tasks provides indirect evidence in support of this hypothesis. The initial formulation of clustering and switching measurements by Troyer et al. (1997) was proposed to disassociate the semantic from executive cognitive mechanisms involved in the SVF task. This two-component model has been subsequently criticized by Mayr and Kliegl (Mayr, 2002; Mayr and Kliegl, 2000). These authors maintain that the executive component is involved in both clustering and switching, and that clustering behavior without precise response latency information cannot be attributed solely to semantic memory function. Furthermore, in our own previous work, we used a similar

approach to computing semantic relatedness between words produced in response to the SVF task in a clinical AD sample for which a more extensive cognitive battery including assessment of executive function were available (Pakhomov et al., 2012). In that study, we showed that a measure of cumulative semantic relatedness, computed as the mean of pairwise relatedness values between all pairs of words produced in response to the SVF test, was strongly associated with executive function performance and less so with memory performance. Although the measure of cumulative semantic relatedness is not the same as the MCS and MChS measures tested in the current study, they are closely related by the nature of their computation. All other things being equal, the presence of larger clusters of closely related words in an SVF output would contribute to larger cumulative relatedness scores for that response set. Therefore, we believe the Pakhomov et al. (2012) study, in the context of other prior studies, provides indirect evidence in favor of linking MCS and MChS scores with executive function. Our findings are also consistent with prior work suggesting that impairment in verbal fluency in the preclinical phase of AD may reflect early deficits in both semantic memory and executive processing (Nutter-Upham et al., 2008; Raoux et al., 2008). Multiple other studies also reported impairment in episodic memory and executive function in the early stages of AD dementia and MCI (Amieva et al., 2005; Backman et al., 2005; Backman et al., 2004; Backman et al., 2001). Thus, a possible explanation for the pattern of results in our current study is that the association between risk of dementia and the size of the clusters produced in response to the SVF task reflects early stages of impairment in executive function and access to semantic representations more so than impairment in semantic representations themselves. In this context, the fact that the SVF scores predict both dementia and memory impairment suggests that the SVF score may better reflect both executive and semantic memory deficits, as would be expected based on prior work showing that optimal performance on this task depends on one’s ability to efficiently retrieve items from within semantically related groups of concepts as well as switch between these groups. In the current study, we found that MCS and MChS were significantly but weakly correlated with a memory task (word list delayed free recall), and the SVF score was more strongly correlated with this measure cross-sectionally at baseline. These results suggest that although memory is involved in the SVF task, its involvement may be captured by the different measures to different extents.

The MChS measure was found to be a significant predictor of dementia onset censored at wave 5 but not at wave 13; however, the relative risk estimates at wave 13 were consistent with those at wave 5. This dissociation is not likely to be due to the tendency towards producing single word “clusters” in SVF responses because single words unrelated to surrounding words would have the same negative effect on clusters and chains. This disassociation is more likely to be due to the fact that the MChS measure reflects an inherently “easier” behavior to accomplish as compared to MCS. To form a cluster, all words in that cluster have to be semantically closely related, whereas to form a chain, only the adjacent pairs of words have to be related. Thus, for example, a cluster of size 3 indicates a much more homogeneous group than a chain of the same size. Therefore, MChS is likely to show more variability and be less powerful at discriminating between healthy aging individuals and those who developed dementia at a later stage.

Our approach to clustering of animal names produced in response to the SVF task is based on extracting distributional statistics of words and their context in a corpus of text. As such, this approach may have a tendency towards quantifying the degree of associative relatedness between words, which is different from previous approaches to analysis of SVF tests that were more focused on assessing semantic similarity. Following Collins and Quillian's model (Collins and Quillian, 1969), if determination of semantic similarity tends to involve a comparison between various properties or semantic features of the concepts denoted by the words, then it is likely to rely on semantic rather than episodic memory cognitive mechanisms, as defined by Tulving (1972). This does not necessarily mean, however, that associative semantic relatedness relies more on episodic rather than semantic memory. The relationship between semantic relatedness and similarity is that of unidirectional entailment – concepts that are similar are necessarily semantically related but not vice versa (Pakhomov et al., 2010; Resnik, 1999). Thus it is possible that while semantic similarity involves predominantly semantic memory, semantic relatedness may involve both cognitive systems. Clustering measures based on semantic relatedness may reflect early deficits in executive function supported by frontal brain regions that are involved in retrieval from both episodic and semantic memory (Buckner and Wheeler, 2001). This hypothesis offers an explanation for the differential pattern of predictive and cross-sectional relationships observed among the SVF performance metrics evaluated in our previous work and the current study.

Limitations and Strengths

Certain limitations must be acknowledged to facilitate the interpretation of the results obtained in this study. One weakness of our data set and participant definition is incomplete coverage of cognitive domains. The Nun Study baseline battery lacks a test of executive function. This limits our ability to rule out all possible cognitive impairments at baseline, and makes it difficult to correctly attribute the performance on the automated variables. The study sample also consists exclusively of Caucasian women; therefore, the results may not be readily generalizable to other populations. The current study evaluates performance on a single semantic category (i.e., animals) in the generative verbal fluency test. These results may not generalize to other commonly used categories such as fruits/vegetables and supermarket items and need to be tested further in longitudinal samples that contain SVF tests with more categories. Due to the large size of the dataset, we were not able to conduct a direct comparison between manual and LSA-based semantic relatedness and clustering computation using all possible pairs of animal names. Consequently, we did not examine any differences in predictive power of manual and automated methods. Another limitation is that, currently, we could not assess the test-retest reliability of clustering measures on SVF tests administered to the same person multiple times. The current study only focused on predicting future dementia and memory impairment based on a single baseline assessment; however, prior work on investigating the reliability of semantic and phonemic verbal fluency (Ho et al., 2002) has found that manual clustering and switching measurements remain stable over time (five annual assessments in a sample of patients with Huntington's disease).

Our study also has a number of distinct strengths. The baseline sample is relatively large – 239 participants. To our knowledge, the only other large longitudinal study comparable to

ours was conducted by Raoux et al. (2008) using the PAQUID[1] study to derive a sample of 153 participants. In addition to the large sample size, our study also has the advantage of longer follow up. Raoux et al. (2008) were able to perform retrospective analysis spanning 5 years, while our study was based on 6 (wave 5) and 17 (wave 13) year follow up periods. However, unlike Raoux et al. (2008), our current data sample did not allow us to study the trajectory of change over time across the various assessment time points as we only had digital transcriptions of the SVF tests available from the baseline assessment. Going forward, we intend to convert the SVF test responses for all assessment periods to digital format enabling us to evaluate automated clustering analysis performance over time. Another strength is that, despite the lack of executive testing, our sample is well defined as cognitively normal (as opposed to simply "not demented"). The baseline cognitive performance of the participants in our study is within one standard deviation of ageappropriate normative data for all baseline measures available (verbal memory, constructional praxis, verbal fluency, confrontation naming, MMSE) and all participants were unimpaired on performance-based assessment of instrumental activities of daily living. Since LSA-based clustering assessments are computerized, there is no instrument variability associated with this approach; however, test-retest reliability of measuring clustering behavior of individuals in multiple assessments over time remains to be examined in future work.

Acknowledgments

The work on this study was supported in part by the National Institutes of Health National Library of Medicine Grant [LM00962301 - S.P.] and the Nun Study data collection was supported by a grant from the National Institute of Aging (R01AG09862). The authors also wish to thank Heather Hoecker for helping with digitization of the semantic verbal fluency samples.

References

American Psychiatric Association. Diagnostic criteria from DSM-IV-TR. Washington, D.C: American Psychiatric Association; 2000.

Amieva H, Jacqmin-Gadda H, Orgogozo JM, Le Carret N, Helmer C, Letenneur L, et al. The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. Brain. 2005; 128:1093–101. [PubMed: 15774508]

Backman L, Jones S, Berger AK, Laukka EJ, Small BJ. Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. Neuropsychology. 2005; 19:520–31. [PubMed: 16060827]

Backman L, Jones S, Berger AK, Laukka EJ, Small BJ. Multiple cognitive deficits during the transition to Alzheimer's disease. J Intern Med. 2004; 256:195–204. [PubMed: 15324363]

Backman L, Small BJ, Fratiglioni L. Stability of the preclinical episodic memory deficit in Alzheimer's disease. Brain. 2001; 124:96–102. [PubMed: 11133790]

Beeri MS, Schmeidler J, Sano M, Wang J, Lally R, Grossman H, et al. Age, gender, and education norms on the CERAD neuropsychological battery in the oldest old. Neurology. 2006; 67:1006–10. [PubMed: 17000969]

Buckner RL, Wheeler ME. The cognitive neuroscience of remembering. Nat Rev Neurosci. 2001; 2:624–34. [PubMed: 11533730]

Canning SJ, Leach L, Stuss D, Ngo L, Black SE. Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. Neurology. 2004; 62:556–62. [PubMed: 14981170]

[1]PAQUID is a cohort study of normal and pathological brain aging in 4,000 elderly subjects living at home. These subjects were randomly chosen in the general population of 75 communities of South- Western France.

Chan AS, Salmon DP, De La Pena J. Abnormal semantic network for “animals” but not “tools” in patients with Alzheimer’s disease. Cortex. 2001; 37:197–217. [PubMed: 11394721]

Collins AM, Quillian MR. Retrieval Time from Semantic Memory. Journal of Verbal Learning and Verbal Behavior. 1969; 8:240–7.

Crum RM, Anthony JC, Bassett SS, Folstein MF. Population-based norms for the Mini-Mental State Examination by age and educational level. JAMA. 1993; 269:2386–91. [PubMed: 8479064]

Estes WK. Learning theory and intelligence. Am Psychol. 1974; 29:740–9.

Fagundo AB, Lopez S, Romero M, Guarch J, Marcos T, Salamero M. Clustering and switching in semantic fluency: predictors of the development of Alzheimer’s disease. Int J Geriatr Psychiatry. 2008; 23:1007–13. [PubMed: 18416452]

Fillenbaum GG, van Belle G, Morris JC, Mohs RC, Mirra SS, Davis PC, et al. Consortium to Establish a Registry for Alzheimer’s Disease (CERAD): the first twenty years. Alzheimer’s Dement. 2008; 4:96–109. [PubMed: 18631955]

Gorno-Tempini ML, Dronkers NF, Rankin KP, Ogar JM, Phengrasamy L, Rosen HJ, et al. Cognition and anatomy in three variants of primary progressive aphasia. Ann Neurol. 2004; 55:335–46. [PubMed: 14991811]

Henry JD, Crawford JR, Phillips LH. Verbal fluency performance in dementia of the Alzheimer’s type: a meta-analysis. Neuropsychologia. 2004; 42:1212–22. [PubMed: 15178173]

Ho AK, Sahakian BJ, Robbins TW, Barker RA, Rosser AE, Hodges JR. Verbal fluency in Huntington’s disease: a longitudinal analysis of phonemic and semantic clustering and switching. Neuropsychologia. 2002; 40:1277–84. [PubMed: 11931930]

Hodges JR, Davies RR, Xuereb JH, Casey B, Broe M, Bak TH, et al. Clinicopathological Correlates in Frontotemporal Dementia. Ann Neurol. 2004; 56:399–406. [PubMed: 15349867]

Hodges JR, Patterson H. Is semantic memory consistently impaired early in the course of Alzheimer’s disease. Neuroanatomical and diagnostic implications. Neuropsychologia. 1995; 33:441–59. [PubMed: 7617154]

Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979:6.

Knopman DS, Kramer JH, Boeve BF, Caselli RJ, Graff-Radford NR, Mendez MF, et al. Development of methodology for conducting clinical trials in frontotemporal lobar degeneration. Brain. 2008; 131:2957–68. [PubMed: 18829698]

Laine, M. Correlates of word fluency performance. In: Koivuselka-Sallinen, P.; Sarajarvi, L., editors. Studies in Languages. Joensuu, Finnland: University of Joensuu; 1988.

Landauer, TK. Handbook of latent semantic analysis. Mahwah, N.J: Lawrence Erlbaum Associates; 2006.

Landauer TK, Dumais ST. A solution to Plato’s problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review. 1997; 104:211–40.

Lezak, MD. Neuropsychological Assessment. 4. Oxford: Oxford University Press; 2004.

Libon DJ, Xie SX, Moore P, Farmer J, Antani S, McCawley G, et al. Patterns of neuropsychological impairment in frontotemporal dementia. Neurology. 2007; 68:369–75. [PubMed: 17261685]

Loewenstein DA, Greig MT, Schinka JA, Barker W, Shen Q, Potter E, et al. An investigation of PreMCI: subtypes and longitudinal outcomes. Alzheimers Dement. 2012; 8:172–9. [PubMed: 22546351]

Mayr U. On the dissociation between clustering and switching in verbal fluency: comment on Troyer, Moscovitch, Winocur, Alexander and Stuss. Neuropsychologia. 2002; 40:562–6. [PubMed: 11749985]

Mayr U, Kliegl R. Complex semantic processing in old age: does it stay or does it go? Psychology of Aging. 2000; 15:29–43.

Nutter-Upham KE, Saykin AJ, Rabin LA, Roth RM, Wishart HA, Pare N, et al. Verbal fluency performance in amnestic MCI and older adults with cognitive complaints. Archives of Clinical Neuropsychology. 2008; 23:229–41. [PubMed: 18339515]

Ober BA, Dronkers NF, Koss E, Delis DC, Friedland RP. Retrieval from semantic memory in Alzheimer-type dementia. J Clin Exp Neuropsychol. 1986; 8:75–92. [PubMed: 3944246]

Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. AMIA Annu Symp Proc. 2010:572–6. [PubMed: 21347043]

Pakhomov SV, Hemmy LS, Lim KO. Automated semantic indices related to cognitive function and rate of cognitive decline. Neuropsychologia. 2012; 50:2165–75. [PubMed: 22659109]

Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform. 2007; 40:288–99. [PubMed: 16875881]

Poesio M, Vieira R. A corpus-based investigation of definite description use. Computational Linguistics. 1998; 24:183–216.

Price SE, Kinsella GJ, Ong B, Storey E, Mullaly E, Phillips M, et al. Semantic verbal fluency strategies in amnestic mild cognitive impairment. Neuropsychology. 2012; 26:490–7. [PubMed: 22746308]

Quesada, J. Creating Your Own LSA Spaces. In: Landauer, T.; McNamara, D.; Dennis, S.; Kintsch, W., editors. Handbook of Latent Semantic Analysis. New York: Taylor and Francis Group; 2011. p. 71-89.

Rada R, Mili H, Bicknell E, Blettner M. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man and Cybernetics. 1989; 19:17–30.

Raoux N, Amieva H, Le Goff M, Auriacombe S, Carcaillon L, Letenneur L, et al. Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: results from the PAQUID longitudinal study. Cortex. 2008; 44:1188–96. [PubMed: 18761132]

Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research. 1999; 11:95–130.

Resnik, P. WordNet and Class-based Probabilities. In: Fellbaum, C., editor. WordNet: An electronic lexical database. Cambridge, MA: MIT Press; 1998. p. 239-63.

Rich JB, Troyer AK, Bylsma FW, Brandt J. Longitudinal analysis of phonemic clustering and switching during word-list generation in Huntington's disease. Neuropsychology. 1999; 13:525–31. [PubMed: 10527060]

Rosen WG. Verbal fluency in aging and dementia. J Clin Neuropsychol. 1980; 2:135–46.

Troster AI, Fields JA, Testa JA, Paul RH, Blanco CR, Hames KA, et al. Cortical and subcortical influences on clustering and switching in the performance of verbal fluency tasks. Neuropsychologia. 1998; 36:295–304. [PubMed: 9665640]

Troyer AK. Normative data for clustering and switching on verbal fluency tasks. J Clin Exp Neuropsychol. 2000; 22:370–8. [PubMed: 10855044]

Troyer AK, Moscovitch M, Winocur G. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology. 1997; 11:138–46. [PubMed: 9055277]

Troyer AK, Moscovitch M, Winocur G, Alexander MP, Stuss D. Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. Neuropsychologia. 1998; 36:499–504. [PubMed: 9705059]

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; 4:137–43. [PubMed: 9529823]

Tulving, E. Episodic and semantic memory. In: Tulving, E.; Donaldson, W., editors. Organization of Memory. New York: Academic Press; 1972. p. 381-402.

Weber M, Thompson-Schill SL, Osherson D, Haxby J, Parsons L. Predicting judged similarity of natural categories from their neural representations. Neuropsychologia. 2009; 47:859–68. [PubMed: 19162048]

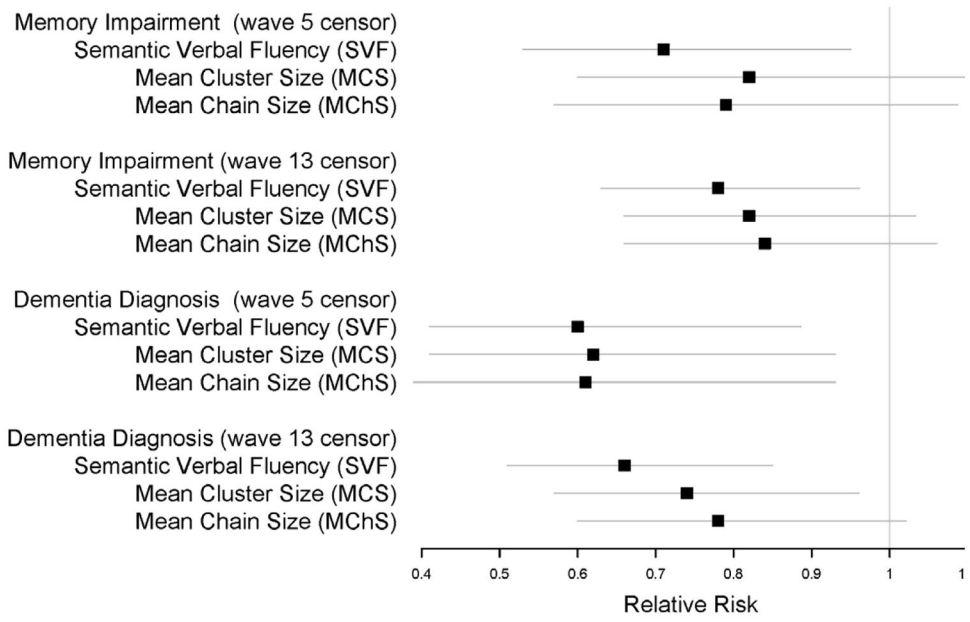

Figure 1.

Relative risk estimates and 95% CIs obtained with Cox models for SVF score, MCS and MChS measures showing risk of future memory impairment and dementia at waves 5 and 13. (SVF – semantic verbal fleuncy, MChS – mean chain size, MCS – mean cluster size)

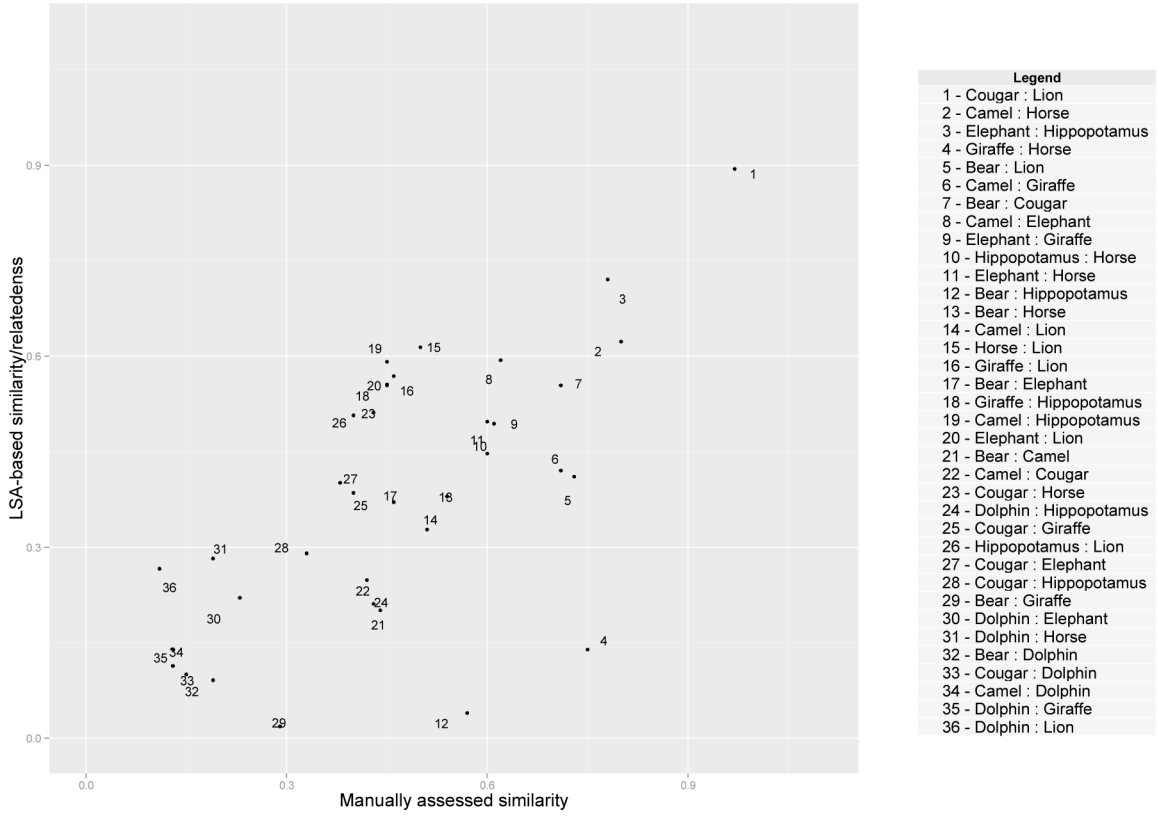

Figure 2.

Scatter-plot illustrating the correlation between LSA-based relatedness judgments and manual ratings defined by the similarity of pairs of exemplars. The manual ratings were obtained from an independent study by Weber et al. (2009)

Table 1

Baseline Assessment Characteristics for Study Participants

	School Sisters of Notre Dame (N=239; all female)	
	Mean	**SD**
Age	80.73	3.98
Years of education	16.96	1.62
Mini-Mental State Examination (MMSE) score	28.39	1.60
Delayed word recall score	7.06	1.39
Semantic verbal fluency (SVF) score	18.05	4.75
Mean Cluster Size (MCS)	0.66	0.33
Mean Chain Size (MChS)	0.67	0.35

Table 2

Elapsed Time Variables in Years

	School Sisters of Notre Dame (N=239; all female)	
Waves	**Mean**	**SD**
Time to wave 5 assessment	6.30	0.53
Time to wave 13 assessment	16.96	0.53
Censor Variables		
Time to Memory 5 censor	5.04	2.00
Time to Memory 13 censor	7.60	4.73
Time to Dementia 5 censor	5.27	1.88
Time to Dementia 13 censor	8.06	4.67

Table 3

Cox Regression Modeling Results

	Exp(B)	Unadjusted p-value	Adjusted p-value	95%CI Low	95%CI High
Predicting Memory Impairment in 5 Waves					
Semantic Verbal Fluency	0.71	0.022*	0.066	0.53	0.95
Mean Cluster Size	0.82	0.192	0.294	0.60	1.11
Mean Chain Size	0.79	0.147	0.294	0.57	1.09
Predicting Memory Impairment in 13 Waves					
Semantic Verbal Fluency	0.78	0.019*	0.057	0.63	0.96
Mean Cluster Size	0.82	0.093	0.186	0.66	1.03
Mean Chain Size	0.84	0.141	0.186	0.66	1.06
Predicting Dementia Onset in 5 Waves					
Semantic Verbal Fluency	0.60	0.010*	0.030*	0.41	0.89
Mean Cluster Size	0.62	0.022*	0.044*	0.41	0.93
Mean Chain Size	0.61	0.022*	0.044*	0.39	0.93
Predicting Dementia Onset in 13 Waves					
Semantic Verbal Fluency	0.66	0.001**	0.003*	0.51	0.85
Mean Cluster Size	0.74	0.023*	0.046*	0.57	0.96
Mean Chain Size	0.78	0.071	0.171	0.60	1.02

** indicates significance at 0.01 level

* indicates significance at 0.05 level