Published in final edited form as:

Neuropsychologia. 2016 August ; 89: 42–56. doi:10.1016/j.neuropsychologia.2016.05.031.

Characterizing Cognitive Performance in a Large Longitudinal Study of Aging with Computerized Semantic Indices of Verbal Fluency

Serguei VS Pakhomov¹, Lynn Eberly², and David Knopman³

¹University of Minnesota Center for Clinical and Cognitive Neuropharmacology, Minneapolis, MN, USA

²Division of Biostatistics, Minneapolis, MN, USA

³Mayo Clinic, Department of Neurology, Rochester, MN, USA

Abstract

A computational approach for estimating several indices of performance on the animal category verbal fluency task was validated, and examined in a large longitudinal study of aging. The performance indices included the traditional verbal fluency score, size of semantic clusters, density of repeated words, as well as measures of semantic and lexical diversity. Change over time in these measures was modeled using mixed effects regression in several groups of participants, including those that remained cognitively normal throughout the study (CN) and those that were diagnosed with mild cognitive impairment (MCI) or Alzheimer's disease (AD) dementia at some point subsequent to the baseline visit. The results of the study show that, with the exception of mean cluster size, the indices showed significantly greater declines in the MCI and AD dementia groups as compared to CN participants. Examination of associations between the indices and cognitive domains of memory, attention and visuospatial functioning showed that the traditional verbal fluency scores were associated with declines in all three domains, whereas semantic and lexical diversity measures were associated with declines only in the visuospatial domain. Baseline repetition density was associated with declines in memory and visuospatial domains. Examination of lexical and semantic diversity measures in subgroups with high vs. low attention scores (but normal functioning in other domains) showed that the performance of individuals with low attention was influenced more by word frequency rather than strength of semantic relatedness between words. These findings suggest that various automatically semantic indices may be used to examine various aspects of cognitive performance affected by dementia.

Keywords

semantic verbal fluency; clustering; memory; attention; dementia; semantic relatedness; word frequency

Corresponding author: Serguei VS Pakhomov, 308 Harvard St. SE, Minneapolis, MN 55108, Tel: 612-624-1198, pakh0002@umn.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Semantic verbal fluency (SVF) tests are widely used as part of cognitive batteries in a variety of clinical and research settings including the assessment of semantic memory impairment in individuals with Alzheimer's disease (AD) dementia. SVF is assessed by asking the participant to name as many things as possible that belong to a semantic category (e.g., animals) in a limited amount of time (usually 60 seconds). The number of valid words produced on this test, after excluding errors and repetitions, is used as the standard quantitative measure of generative semantic verbal fluency (a.k.a., category fluency). A number of studies have demonstrated that individuals diagnosed with AD dementia produce significantly fewer words on SVF tests than do cognitively normal individuals (Epker, Lacritz, & Munro Cullum, 1999; Henry, Crawford, & Phillips, 2004; McDowd et al., 2011; Monsch et al., 1992; Price et al., 2012; Rosen, 1980; Weintraub, Wicklund, & Salmon, 2012).

SVF is a complex language task that is served by multiple cognitive mechanisms including semantic memory and executive function (Mayr, 2002). In order to maximize the performance on this task, the participant has to be able to successfully complete a number of cognitive subtasks including retrieval of category exemplars from the semantic store; matching of retrieved exemplars to appropriate lexical and phonological representations; compiling, initiating and carrying out an articulatory plan, continuously monitoring the speech being produced, as well as the items that have already been spoken, and inhibiting inappropriate responses (Henry et al., 2004; Levelt, 1989).

Furthermore, SVF test responses contain evidence that speakers also typically attempt to use a variety of strategies such as naming things in alphabetical order or, more commonly with naming animals, picturing themselves in a zoo or using taxonomic and visual/functional similarity relations within a category to access exemplars of subcategories (e.g., bear, polar bear, brown bear, koala, etc.). The ability to use a strategy typically manifests itself through producing measurably larger clusters of semantically related items. A number of studies have demonstrated that words on SVF tests are not produced at random but rather tend to be delivered in semantic groups (clusters) consisting on average of 2–3 semantically related words (e.g., horse, sheep, goat) (Raoux et al., 2008; Troyer, 2000; Troyer, Moscovitch, & Winocur, 1997). However, mixed results have been reported in the literature with respect to the association between clustering behavior on SVF tests and cognitive impairment resulting from neurodegeneration due to conditions such as AD dementia and mild cognitive impairment (MCI). Some studies have shown that the mean size of semantic clusters is significantly larger in cognitively normal individuals as compared to patients with MCI and AD dementia (Price et al., 2012; Troster et al., 1998; Troyer et al., 1997), whereas others have not found any significant differences (Epker et al., 1999; Raoux et al., 2008). Explanations offered for these conflicting results have to do with differences in dementia severity, demographic variables, sampling differences and study design differences (Raoux et al., 2008).

However, conflicting results reported in these studies may also have to do with the fact that most studies that examined the relationship between clustering and dementia have relied on a manual process for semantic cluster determination. Manual clustering assessments are by definition subjective even if they use the same guidelines (e.g., Troyer's methodology is used most frequently). The guidelines do not cover the full range of category exemplars and raters in different studies may interpret the guidelines differently resulting in systematic differences in cluster size assessments across studies. Thus, more objective and reproducible approaches to measuring clusters may be useful in reconciling these differences.

Several prior studies have demonstrated the utility of using computerized tools for assessing clustering behavior on verbal fluency tasks in patients with dementia (Chan, Butters, Salmon, & McGuire, 1993; Goni et al., 2011; Pakhomov & Hemmy, 2013; Pakhomov, Hemmy, & Lim, 2012; Pakhomov, Jones, & Knopman, 2015; Taler, Johns, Young, Sheppard, & Jones, 2013). In addition to dementia, use of computational semantic analysis of verbal fluency test responses has been examined in other clinical domains such as schizophrenia and bipolar disorder (Holshausen, Harvey, Elvevåg, Foltz, & Bowie, 2014; Rosenstein, Foltz, Vaskin, & Elvevag, 2015; Voorspoels et al., 2014). Much of this prior work examining the relationship between semantic clustering behavior on SVF tests and cognitive impairment has focused on either cross-sectional studies of differences between impaired and unimpaired groups or on longitudinal studies examining various aspects of semantic clustering behavior as a risk factor for subsequent onset of dementia.

In addition to examining these cross-sectional associations and longitudinal predictions, it is also important to examine how verbal behavior measured through various semantic indices changes over time in cognitively normal (CN) individuals as compared with individuals affected by dementia. Semantic indices with different trajectories in these groups may potentially be useful for early detection of cognitive decline. In our previous work (Pakhomov & Hemmy, 2013; Pakhomov et al., 2012, 2015), we have developed several computerized semantic indices based on the animal category fluency tests including mean size of semantic clusters and a measure of semantic diversity represented by the mean semantic relatedness averaged across all pairs of words spoken in response to the verbal fluency task. We also demonstrated that these semantic indices are associated with distinct brain mechanisms and thus measure different aspects of the verbal fluency performance (Pakhomov et al., 2015).

The objective of the current study was to compare the trajectories of change in computerized semantic indices across several diagnostic groups and to characterize the relationship between the changes in semantic indices and changes in memory, attention, and visuospatial cognitive domains in a large longitudinal study of aging. At the outset of the study, we anticipated finding some decline in computerized measures of mean cluster size, semantic diversity and increase in perseveration across all participants due to aging; however, we expected to find progressively steeper changes in individuals diagnosed with MCI and AD dementia as compared to those who remained cognitively normal throughout the study. Individuals with better semantic memory that utilize a semantic clustering strategy measured with computerized indices discussed further in this manuscript are expected to be able to find more exemplars for each cluster, thus resulting in larger cluster sizes. Producing

semantically diverse responses, on the other hand, requires the individual to maintain a balance between searching within categories and deciding when to abandon search to switch to a new category. For example, producing a response consisting of 20 animal names that are completely semantically unrelated would arguably require more executive control needed to switch between the names than producing 5 clusters of 4 semantically related animal names per cluster. In our previous work, we found some, albeit indirect, evidence that the size of semantic clusters is associated with language-related networks in the left medial temporal lobe involved in semantic memory storage and retrieval, while semantic diversity is associated with the left angular gyrus and the left inferior frontal gyrus involved in working memory and language production (Pakhomov et al., 2015). Thus, following prior work on clustering behavior, we also hypothesized that changes in the memory domain scores would be associated with the size of semantic clusters, whereas changes in the attention domain scores would be associated with measures of semantic diversity and perseveration.

Furthermore, in prior work on bilingual English-French speakers Taler et al. (2013) reported results demonstrating that SVF performance relies on several different information sources including word frequency and semantic similarity between words produced on the SVF test. This work also demonstrated that word frequency and similarity play competing roles, with decreasing importance of semantic information in the presence of increased demands on executive function induced with a language switching SVF task. In the current study, we wanted to determine if similar effects could be observed when the attention aspect of executive function capacity is decreased due to cognitive impairment. Thus, while we expect trajectories in attention domain scores to be associated with trajectories in measures of semantic diversity, when comparing individuals with low vs. high attention domain scores, changes in word frequency measures would be sensitive to the magnitude of the attention domain scores, whereas changes in semantic relatedness¹ measures would not.

In the current study, we performed a cross-sectional and a longitudinal analysis on two independent datasets. The purpose of the cross-sectional analysis was to validate the proposed approaches to computing semantic indices (detailed in Methods) that were subsequently applied to the longitudinal analysis. We also used the cross-sectional analysis to ensure that semantic indices computed in the current study were comparable to those that we reported on previously (Pakhomov & Hemmy, 2013; Pakhomov et al., 2012, 2015). By doing the longitudinal analysis we sought to answer the following questions for each semantic index:

1. Sensitivity to dementia: Is the trajectory of change in the index significantly different for different diagnostic groups?
2. Cognitive mechanism(s): Is there a significant association in the cognitively normal group between the baseline index score and changes in memory, attention, and visuospatial cognitive abilities?

¹We use a more general term “semantic relatedness” rather than the more specific “semantic similarity” in order to include distributional methods for computing the strength of relatedness between words (e.g., mouse and cat may not be semantically similar but are functionally related in many contexts).

3. Mechanism(s) driving lexical access: Do the trajectories of change in measures of semantic diversity and word frequency differ by baseline measure of attention?

In the remainder of this article, we present the methods used to calculate semantic indices and the results and discussion of the cross-sectional and longitudinal analyses. The methods and results pertaining to validation of semantic indices are reported in the Appendix.

2 Materials and methods

2.1 Participants

In this study, we used two independent cohorts of participants. One cohort was used for the cross-sectional analysis and the other was used for the longitudinal analysis.

The cross-sectional analysis cohort ($N = 170$) was obtained from the Mayo Clinic Alzheimer's Disease Research Center registry (ADRC cohort) and consisted of 53 patients with AD dementia, 71 patients with MCI (amnesic and multi-domain) and 46 persons diagnosed as cognitively normal (CN). For the purposes of the analysis performed in the current study, we excluded participants who produced fewer than 6 words on the SVF test. This reduced the cross-sectional cohort to 167 participants (50 with AD dementia, 71 patients with MCI and 46 CN).

The longitudinal analysis cohort consisted of 1,376 participants in the Mayo Clinic Study of Ageing (MCSA cohort) who had participated in at least two MCSA assessments as of 12/01/2013. The participants had 3.26 (SD 1.76) assessments on average with a mean follow-up period of 5.46 (SD 2.08) years. These participants were all CN at the baseline assessment in year 2005 and were subsequently assessed at approximately 15-month intervals for up to 6 additional study visits. During follow up, 968 of the 1,376 participants remained CN throughout the study. MCI was diagnosed in 336 and AD dementia was diagnosed in 72 of the remaining participants during at least at one of the follow up assessments. For this study we did not try to account for instances in which participants reverted from MCI to CN, or from AD dementia to CN or MCI status. The majority of people reverting to CN from a cognitively impaired status ($n=81$) did so temporarily and were subsequently again diagnosed with either MCI or AD dementia (Roberts et al., 2014). Thus, once a participant was diagnosed with either MCI or AD dementia, he/she was assigned to the corresponding diagnostic group even if he/she was diagnosed as cognitively normal on a subsequent visit. We further divided the cognitively normal at all visits group into two randomly selected subsets. One of these subsets (referred to below as CN-Train) was used to estimate strength of semantic relatedness between words produced on The SVF test ($N = 624$). The other subset ($N = 344$) was used for longitudinal analysis in conjunction with data from the MCI and AD dementia participants.

2.2 Cognitive and clinical assessments

At each study visit, all MCSA and ADRC participants underwent a clinical evaluation and neuropsychological testing. Clinical evaluation included a complete neurological examination, the Clinical Dementia Rating scale (Morris, 1993), and the Short Test of

Mental Status (STMS; Kokmen, Smith, Petersen, Tangalos, & Ivnik, 1991) among other assessments previously published elsewhere (Roberts et al., 2008). We summarize diagnostic criteria here for convenience. The diagnoses of MCI and AD dementia were made during consensus conferences including neurologists, neuropsychologists and nurses and took into account the full neuropsychological test battery, the neurological assessment and the views of the family informants, as obtained by the examining neurologists and nurses. A diagnosis of MCI was established based on the following: a) cognitive concern was raised by the participant, informant, nurse, or physician; b) impairment in 1 or more of the 4 cognitive domains (see description of neuropsychological assessments below); c) normal functional activities; and d) absence of dementia. A diagnosis of AD dementia was based on *DSM-IV* criteria (American Psychiatric Association, 1994). Subjects were considered to be cognitively normal if they performed within the normative range and did not meet criteria for MCI or AD dementia.

In addition to the clinical assessments, a neuropsychological battery consisting of nine tests was used. The tests were selected to assess function in four domains: *memory* (delayed free recall percent retention scores for Wechsler Memory Scale-Revised Logical Memory and Visual Reproduction tasks (Wechsler, 1981), and the Auditory Verbal Learning test (Ivnik et al., 1992)); *language* (Boston Naming test (Kaplan, Goodglass, & Weintraub, 1983) and category fluency (Lucas et al., 1998)); *attention aspect of executive function* (Trail Making test B (Reitan, 1958) and Digit Symbol Substitution subtest from the Wechsler Adult Intelligence Scale-Revised (Wechsler, 1981)); and *visuospatial skills* (picture completion and block design (Wechsler, 1981)). The raw scores on each test were adjusted for age using normative data from the Mayo Clinic's Older American Normative Studies (Ivnik et al., 1992). The adjusted test scores within each domain were summed and scaled to obtain domain-specific and global z-scores (Roberts et al., 2008). A decision about cognitive impairment was based on a consensus agreement among the examining nurse, physician, and neuropsychologist. We used quantitative cut-scores based on normative data and extensive experience with the measurement of cognitive abilities in the target population on neuropsychological domain performance scores. Values worse than -1.5 SD were considered as typical for MCI. However, based on our long-used approach (Petersen, 2004; Petersen et al., 1999), cut-off scores for neuropsychological tests were used as a guide to reach consensus for diagnosis rather than an absolute criterion. Thus, individual test scores were not explicitly used for diagnosis. The clinical evaluations were performed without using information from imaging or other biomarkers acquired during the study.

In the current study, we relied on adjusted z-scores to test for relationships between semantic indices and cognitive domains. We only used the memory, attention and visuospatial domains. The language domain was excluded because its z-scores contained the scores from the SVF task. Global z-scores were also excluded for the same reason as with the language domain.

2.3 Computation of semantic relatedness strength

The terms “semantic relatedness” and “semantic similarity” tend to be treated somewhat differently by different disciplines, which can potentially introduce inconsistency and

ambiguity. Standardizing cross-disciplinary nomenclatures is beyond the scope of the current work; therefore, we present a brief explanation of how terms relevant to the work presented in this paper align across disciplines. A detailed discussion of the definitions of these terms and the relationship between them is available in a previous publication (Pakhomov et al., 2012). We summarize it here for convenience.

In psycholinguistics, a distinction is made between “semantic relatedness” and “associative relatedness”. Associative relatedness refers to the probability that one word calls to mind another word (e.g., needle-thread), while semantic relatedness reflects the degree of semantic feature overlap between words (e.g., whale-dolphin). In computational linguistics, a distinction is made between the terms “semantic similarity” and “semantic relatedness” that roughly correspond to the psycholinguistic terms “semantic relatedness” and “associative relatedness”, respectively. In this manuscript, we will use the computational linguistic terminology. We also treat semantic similarity as a special case of semantic relatedness. For example, the words “cat” and “tiger” refer to somewhat similar concepts that share many semantic features (e.g., feline, four legs, fur, tail, etc.), while the words “cat” and “bird” refer to semantically dissimilar but functionally related concepts (e.g., cats are known to chase birds). In addition to the distinction between the terms “semantic relatedness” and “semantic similarity”, we will also make a distinction between the terms “strength of semantic relatedness” and “strength of association.” In the rest of the manuscript, we will refer to values calculated with the log-likelihood approach as reflecting the “strength of semantic relatedness”, and to values obtained manually with a free association task as the “strength of association.”

In order to quantify clustering behavior on the SVF test (‘animals’ category), we needed a measure of the strength of semantic relatedness between pairs of animal names. Ideally, we would like to generate a set of all possible pairs of animal names and have each pair manually judged by several raters. An equally labor-intensive alternative to this approach is to measure semantic relatedness based on free associations by asking participants to name words that come to mind in response to a target stimulus. Intuitively, words produced in response to the free association task would likely consist of most closely semantically related words to the target and their order may be used as a measure of relatedness. Unfortunately, either of the two manual tasks of the required magnitude is not currently feasible. For example, there are approximately 5000 animal names that we have been able to identify so far, which would result in over 24M possible (unordered) pairs. Even if we restricted these pairs to those that were actually produced by the present study participants, we would still have over 10,000 pairs that would have to be manually evaluated and still would not be likely to provide sufficient coverage for future assessments of SVF responses not in the current dataset. Apart from feasibility, this direct and fully manual approach would also not generalize easily to other semantic categories.

As an alternative to manual approaches, we and other groups have developed several ways of automatically computing semantic relatedness. In our prior work, we have investigated a number of ontology-based and distributional semantic approaches for computing semantic relatedness between animal names (Pakhomov & Hemmy, 2014; Pakhomov et al., 2012, 2015). We found that ontology-based approaches tend to be problematic, as they are more

useful to measure semantic similarity rather than relatedness and are not easily scalable to words that are not in the ontology. Distributional semantic approaches (e.g., using latent semantic analysis based on co-occurrences in Wikipedia) are scalable and can be used to measure semantic relatedness more directly; however, their weakness is that they tend to be noisy. Less noisy co-occurrence statistics can be obtained directly from SVF responses themselves by making a reasonable assumption that words that occur close together in an SVF response tend to be semantically related (Goni et al., 2011); however, that assumption does not necessarily hold at semantic cluster boundaries. Also, one needs to have a relatively large set of SVF responses in order to have a sufficient volume of co-occurrence data.

In the present study, we experimented with a hybrid approach that was based on a relatively large dataset of SVF responses used to compute co-occurrence frequencies for a statistical measure of association, as well as a database of manually obtained free association norms. For the co-occurrence frequency based approach, we first extracted a vocabulary of animal names present in the SVF test responses of the CN-Train set (see Figure 1) resulting in a list of 600 names (these included multi-word names). Six hundred names can be paired in 179,400 different ways (ignoring the order of names in a pair and pairs consisting of the same name); however, only a fraction of these possible pairs are likely to be observed in actual participant responses. Thus, we then extracted all pairs of animal names that actually occurred in the CN-Train dataset resulting in a smaller set of 10,162 pairs. For each pair in this set, we were able to observe how frequently the names in the pair occurred with each other as well as how frequently they occurred with other names. Based on these co-occurrence frequency counts we were able to calculate the log likelihood ratio (G^2) as a way to quantify the strength of semantic relatedness between the names in the pair. The distribution of G^2 values computed from the CN-Train corpus is illustrated in Figure 2 and ranges between 10868.1 (dog – cat) and – 20.8 (deer – cat).

The notion of using G^2 to estimate the strength of semantic relatedness between lexical items was initially introduced to the computational linguistics field by Dunning (Dunning, 1993) and was shown to be a more robust alternative for smaller text corpora than other equivalent methods such as chi-square and mutual information. However, using SVF responses as a corpus of text to estimate G^2 presents a unique challenge. The existence of clustering behavior as a strategy in SVF responses has been clearly demonstrated (Troyer et al., 1997). It has also been shown that it is reasonable to assume that words spoken in close proximity to each other are likely to be semantically related (Goni et al., 2011). This observation has been successfully exploited by a number of investigations to determine the semantic structure of categories (such as animals, for example) based purely on the participants' responses (Chan et al., 1993; Goni et al., 2011; Taler et al., 2013). For example, Goni et al. (2011) rely on a binomial distribution to determine if the probability of any two animal names appearing within a small window of 5 words is greater than chance. If the probability is greater than 0.5, the words in the pair are said to be conceptually related. Other statistical measures of association such as G^2 and mutual information work in a similar fashion and result in a continuous measure that can be dichotomized into 'conceptually related' and 'unrelated' categories with a threshold.

While these distribution-dependent strategies work very well for the majority of cases where clustering behavior is evident, they do not take into account the fact that the distribution of word pairs at the borders of clusters (when the respondent switches from one cluster to the next) is not really random. For example, participants in an animal fluency test frequently switch between ‘pets’ and ‘farm animals’ or “African animals” categories. A typical response to the SVF task begins with “cat dog ...” or “dog cat ...” cluster followed by clusters that begin with words such as “horse”, “pig”, “chicken” or “rhinoceros”, “giraffe”, “hippopotamus.” This behavior may result in relatively high frequency of transitions between semantically weakly related words (e.g., ‘dog’ and ‘horse’ – $G^2 = 1344.4$), thus potentially inflating the strength of relatedness between these words. More importantly, due to data sparseness, some of the G^2 values for semantically closely related pairs containing words that do not occur often enough in the training data may be deflated. For example, the pair of very closely related words “cow – sow” has a G^2 value of 3.8, which places this pair towards the bottom of the relatedness range.

In the current study, we attempted to compensate for the limitations the G^2 values computed from the SVF data by introducing relatedness estimates based on the South Florida University Free Association Norms (FAN) database developed by Nelson et al. (Nelson, McEvoy, & Schreiber, 2004). Each participant in the FAN study was asked to produce a single word (target word) in response to a stimulus word (cue word) selected from a set of 5,019 cues, which resulted in a database of 72,000 common English word pairs along with the strength of forward and backward association between them collected from 6,000 participants. For the current study, we first identified those animal name pairs that exist in both the FAN database and in our dataset of SVF responses. Second, for each of these pairs, we calculated the strength of association as a ratio of the number of associates shared by the cue and target words (the “O+OMIA” fields in the FAN database) by the total size of the cue set (the “CSS” field). Third, we scaled the resulting strength of association value to match the range of G^2 values in our set of pairs according to the following formula in (1):

$$\text{Scaled_FAN} = \min(G^2) + \frac{(FAN - \min(FAN)) * (\max(G^2) - \min(G^2))}{\max(FAN) - \min(FAN)} \quad (1)$$

where FAN is the raw unscaled value obtained from the FAN dataset.

After that, we were able to merge the strength of association values found in the FAN dataset with the G^2 values by replacing the G^2 values with the scaled FAN values for the word pairs in the intersection between the two datasets ($N = 267$ pairs). Figure 2 illustrates the effects of introducing FAN associations in this way, the word pair ‘horse – dog’ had a relatedness value of 342.2 after merging FAN with G^2 , indicating that the words “horse” and “dog” are somewhat related but not as closely as the G^2 value of 1344.4 would suggest. Conversely, the pair “cow – sow” got a new value of 1140.7 after merging FAN with G^2 , which places this pair in the closely related part of the range (the G^2 value for “cow – pig” pair is 5424.9).

When a pair of animal names in the CN-Test corpus was not found in the dataset of merged FAN and G^2 values (3.4% of all pairs in CN-Test), we assigned a relatedness value of zero to

that pair, assuming that pairs that did not co-occur at all in the same SVF response in the SVF-Train data are more likely to be unrelated than closely related. An informal examination of a random selection of 100 of the missing pairs showed that only 16 of them were at least somewhat semantically related (e.g., “snake – gila_monster”, “salamander – python”, “osprey – crow”), the rest were unrelated (e.g., “zebra – chameleon”, “calf – chimpanzee”, “rat – mosquito”). Based on the relatively low prevalence of missing pairs and the majority of them being unrelated, we did not attempt to use any smoothing techniques.

2.4 Computation of word frequency

The frequency with which a word appears in any given language is associated with the ease of access and retrieval of this word from semantic memory (Murray & Forster, 2004). We followed the methodology introduced by Taler et al. (2013) to compute the frequencies of animal names used in SVF responses. We used the latest available English Wikipedia archive (created 06-02-2015) as the corpus for estimating word frequencies. In order to standardize all animal names in SVF responses to their singular form, we used a WordNet (Fellbaum, 1998) lemmatizer available through the Python Natural Language Processing Toolkit (NLTK) package Version 3.0 (Bird, Klein, & Loper, 2009). The lemmatizer was applied to all words in SVF responses as well as all words in the Wikipedia corpus. Following Taler et al. (2013), all animal name frequencies obtained from Wikipedia were converted to a natural logarithmic scale found to reflect human lexical access latency (Adelman and Brown, 2008) and known in computational linguistics literature as a measure of information content of the word (Resnik, 1999).

2.5 Study Design

The study design is illustrated in Figure 1. The longitudinal cohort contained 4 datasets: CN-Train, CN-Test, MCI and AD. The CN Train dataset was used to generate G^2 values for all available pairs of animal names that were stored in a matrix subsequently used to make clustering/switching decisions and generate semantic indices on the SVF samples contained in the CN-Test, MCI, and AD datasets. In addition to the traditional verbal fluency score (SVF), we defined the following computerized indices:

- Mean Cluster Size (MCS)
- Cumulative Relatedness/Semantic Diversity (CuRel)
- Repetition Density (RepD)
- Mean Word Frequency (MWF)

Clusters were defined as consecutive groups of semantically closely related animal names using an empirically determined threshold (see Appendix section A.3 for details on threshold calibration). Animal names that were not semantically related to their neighbors were assigned cluster size of one, two name clusters were assigned cluster size of two, and so on. This procedure constitutes a departure from the original Troyer et al. (1997) approach in which 1 was subtracted from each cluster size resulting in discounting items that were unrelated to their immediate neighbors. MCS was calculated by dividing the sum of cluster sizes by the total number of clusters in a given response.

We used the cross-sectional ADRC cohort to empirically determine the most optimal semantic relatedness threshold for grouping animal names in participants' responses into clusters. In order to do this, we experimented with a range of thresholds on semantic relatedness values and compared the resulting automatically determined number of clusters to the number of clusters found manually by human raters (described further in "Validation of automatic clustering" section). The threshold that resulted in the best correspondence between computerized and manual clustering procedures was subsequently used for automatically defining clusters in the CN-Train set responses.

CuRel was calculated as the mean of all pairwise G^2 values. Note that while this measure represents the degree of how semantically diverse the animal names are in a given SVF response, this is a fundamentally different notion from the semantic diversity of a single word proposed by Hoffman et al. (Hoffman, Meteyard, & Patterson, 2014). The latter is defined as the semantic distance between different contexts in which a given word occurs and reflects the specificity of the word's semantic content. RepD was calculated as the ratio of the count of repeated animal names to the total number of animal names in an SVF response. MWF was calculated by normalizing the sum of log-scaled Wikipedia word counts by the total number of words in the SVF response.

2.6 Statistical Analysis

Inter-rater reliability was estimated with two-way, averaged intra-class correlation coefficient (ICC) for consistency in ratings. Tests for cross-sectional differences in semantic indices by diagnostic group were performed using ANOVA with Tukey-HSD correction for multiple comparisons. Further examination of the relationship between semantic indices and the scores obtained with the STMS clinical assessment instrument was conducted with linear regression models adjusted for age, sex, and years of education. Rates of change in semantic indices were examined using mixed effects regression modeling adjusting for age, sex, years of education, and the baseline SVF score when appropriate. We also used mixed effects modeling to test whether baseline values for the various semantic indices were associated with cognitive domain scores over time. All of these models included sex, education and age at baseline as nuisance covariates. Parametric Pearson correlation was used to measure associations in large sets of data with homogenous variable types – ADRC and MCSA groups (Table 1). Non-parametric Spearman rank correlation was used on smaller sets with heterogeneous variable types – Weber's dataset of 36 animal pairs.

To determine whether the trajectories of change in measures of semantic diversity and word frequency differ by baseline measure of attention, we followed work by Taler et al. (2013) in which the language switching SVF task introduced a greater executive load in the bilingual experiments. We mirrored these conditions in a monolingual SVF task by stratifying the participants into low and high attention group. Participants with baseline attention z-scores (z -scores < 0 and memory and visuospatial scores ≥ 0) were stratified into a low attention function (LOW_ATT) group ($n = 51$). Participants with baseline attention z-scores ≥ 0 and memory and visuospatial scores ≥ 0 were placed into a high attention function (HIGH_ATT) group ($n = 197$).

All statistical tests and models were conducted using the R statistical package (version 3.1). Semantic indices and the underlying distributional statistics including G^2 were calculated using a custom software package (VFClust) that was developed using the Python programming language and is freely available (<https://pypi.python.org/pypi/VFClust/0.1.1>). A web-based demo of the VFClust package is also freely available at (<http://gummo.pharmacy.umn.edu/salsa/vfclust/demo>).

3. Results

In this section we first report the results of the cross-sectional analysis followed by presenting the results of the longitudinal analyses. Validation results are presented in the Appendix.

3.1 Cross-sectional analysis of the ADRC cohort

To achieve the primary objective of the cross-sectional ADRC cohort analysis, we compared the means for manually and automatically computed MCS, automatically computed CuRel, and automatically computed RepD variables across diagnostic groups. The results of this comparison are presented in Figure 4 and show that MCS means are not different across the diagnostic groups, but CuRel, RepD, and MWF means for AD patients are significantly different from those for MCI patients and controls ($p < 0.05$).

We also found a significant positive association between SVF scores and STMS scores ($t = 7.24$; $df = 161$; $p < 0.001$). The regression models constructed to test for association of MCS, CuRel, MWF, and RepD with STMS scores were further adjusted by including the SVF score as a covariate. As a result of this modeling, we found that CuRel, RepD, and MWF were significantly negatively associated with STMS even after the adjustment for the SVF score ($t = -2.64$, $df = 160$, $p = 0.009$; $t = -5.94$, $df = 160$, $p < 0.001$; and $t = -2.67$, $df = 160$, $p < 0.001$, respectively). No significant association between STMS and MCS was found either before or after adjustment for SVF in addition to age, sex and education.

3.2 Longitudinal analysis of the MCSA cohort

First, we visualized the trajectories of change in the variables of interest over the five assessment cycles of the study. In order to illustrate group differences in trajectories and make the trajectories easily visually comparable, we only selected participants who entered the study on the initial assessment cycle and participated in each of the four subsequent cycles. Thus, these trajectories are shown only to illustrate the progression of change over the course of the study. It was not possible to calculate some of the scores for a small subset of participants in this group on various cycles due to missing data or insufficient output on SVF test (less than six words). The results of plotting these trajectories across the CN, MCI and AD groups are shown in Figure 5.

3.3 Trajectory of change in semantic indices over time

Mixed modeling was performed on the entire set of available data. The results of mixed modeling with SVF, MCS, CuRel, RepD, and MWF variables as the outcome and including the interaction between diagnostic category (NC, MCI, AD) with time, adjusted for age, sex,

years of education, and baseline SVF scores are summarized in Table 1. Since CuRel, MCS, and MWF variables may potentially be affected by the number of repeated words, we adjusted the corresponding models for the density of repetitions in addition to the other nuisance covariates. Linear mixed models for CuRel, RepD, and MWF trajectories were also repeated in an alternative analysis using covariate adjustment for each visit's updated SVF value (rather than just for the baseline SVF value). The results of this alternative analysis are summarized in Supplementary Table 1 in Appendix B.

Mixed modeling results confirm the differences in SVF, CuRel, RepD, and MWF trajectories and lack thereof in MCS trajectories apparent in visual comparisons across the groups presented in Figure 5. The trajectories for SVF show a decline (fewer words over time), while the trajectories for CuRel, RepD and MWF point in the opposite direction; however, the slopes of CuRel and RepD are only significant in the analysis that includes baseline SVF as a covariate (Table 1) and are not significant in the analysis that included SVF scores at each visit (Appendix B: Supplementary Table 1). Years of education turned out to be a significant predictor in all models – positive slope with SVF and negative with the rest of the indices. Being a male was significantly negatively associated with SVF, CuRel, and MWF but not MCS or RepD. Age at baseline was significantly positively associated only with RepD variable in models that included baseline SVF – older age at baseline associated with higher density of repetitions. The alternative analysis that included SVF scores at each visit also showed an association between age at baseline and CuRel (in addition to RepD).

3.4 Association between semantic indices and cognitive domains among participants that remained cognitively normal throughout the study

We tested for associations between semantic indices and three cognitive domains (memory, attention and visuospatial) after excluding all visits by incident MCI and AD participants. Following an earlier study by Mielke et al. (2013), we anticipated that incident MCI and AD cases may drive any associations between cognitive domains and semantic indices simply due to cognitive impairment associated with these conditions. Thus, in order to simplify the interpretation of associations, this analysis was performed only on cognitively normal controls who were normal at baseline and remained cognitively normal throughout the study. By only focusing on participants that remained CN throughout the study, we can reasonably factor out confounding effects of MCI and dementia on both semantic processing and cognitive domains, which would result in a significant association but may not necessarily show us whether a semantic measure is showing a true association with a cognitive domain or whether both the semantic measure and the domain are being modified by the disease but are not necessarily associated with each other. The results of this analysis are summarized in Table 2, which shows change over time in cognitive domain z-scores as well as z-scores of individual neuropsychological tests used to compose the domains (response variables) as a function of the baseline values of SVF, MCS, CuRel, RepD, and MWF (predictor variables) adjusted for age, sex and education.

Significant associations were found between the baseline SVF score and all cognitive domain scores – higher SVF scores at baseline were associated with better performance over

time in all three cognitive domains. No associations were found between MCS scores and any of the cognitive domains. CuRel at baseline was significantly associated with scores only for the visuospatial domain – greater semantic diversity indexed by lower CuRel scores was associated with better performance on visuospatial domain tasks. Baseline RepD scores were negatively associated with memory and visuospatial scores but not with attention scores – greater repetition density at baseline was associated with lower cognitive domain scores. Similarly to CuRel, the baseline MWF scores were associated only with the visuospatial domain, but not memory or attention; however, results for individual tests show a potentially interesting dissociation between these two indices (CuRel and MWF). As evident from Table 2, CuRel's negative association with the visuospatial domain is driven primarily by the performance on the Block Design task (although its association with the Picture Completion task was in the same direction but only approached significance), whereas MWF's also negative association with the same domain is driven primarily by the Picture Completion task (although its association with Block Design was also evident). Furthermore, while MWF is not significantly associated with the attention domain, the lack of association is driven by the Digit-Symbol task that shows a non-significant and positive association with MWF. However, MWF is negatively and significantly associated with the Trail Making (Part B) component of the composite score for the attention domain.

The Time variable was significantly associated with all three domains showing that the scores in these domains changed over time. However, the association was positive for memory and negative for the other two domains indicating an improvement over time in the former and deterioration in the latter two domains. The scores for individual neuropsychological tests comprising the domains followed the same pattern of change as the domain scores, with the exception of performance on AVLT (1/2 hour delay) task that remained unchanged over the duration of the study.

Correlations between semantic indices and cognitive domains at the baseline visit in CN participants are shown in Table 3. These results show stronger correlation between SVF scores and memory and attention domains as well as between RepD and memory and visuospatial domains. Weak correlations or no correlation were found between MCS or CuRel and any of the three cognitive domains. Similarly to the results of the cross-sectional validation study, the measures of CuRel and MCS showed weak or no correlation with the traditional SVF scores.

3.5 Trajectory of change in semantic indices over time in HIGH vs. LOW attention function groups

The results of mixed modeling in groups with high and low attention function scores are summarized in Table 4. These results show that the trajectory of change in the MWF measure is positive over time in individuals with low attention domain scores and negative in individuals with high attention scores. The trajectories of change in CuRel scores in the low and high attention groups are not significantly different from each other.

4. Discussion

At the outset of the study, we formulated hypotheses related to longitudinal effects of dementia on computerized semantic indices and cognitive mechanisms associated with these indices. We hypothesized that trajectories of deleterious change in computerized semantic indices of cluster size (MCS), semantic diversity (CuRel), repetition density (RepD) and word frequency (MWF) are sensitive to dementia and are thus significantly steeper in MCI and AD groups as compared to CN participants. With respect to cognitive mechanisms, we hypothesized that changes in memory would be primarily associated with changes in MCS and that changes in attention would be associated with CuRel and RepD measures. With also expected to find differences in lexical access reflected in MWF measure between individuals with high and low baseline attention scores.

4.1 Sensitivity of semantic indices to dementia

The main findings of the current study are that trajectories of decline in some but not all computerized semantic indices differ significantly across diagnostic groups. We found steeper declines for the MCI and AD dementia groups in the traditional SVF scores (SVF), CuRel, RepD, MWF, but not in MCS. Note that positive slopes in CuRel reflect deteriorating semantic diversity, positive slopes in RepD reflect increased number of repetitions, and positive slopes in MWF reflect reliance on higher frequency words, thus all three trajectories reflect declining cognitive performance. Our results with respect to computerized measures of SVF and MCS performance are consistent with those obtained manually as part of the PAQUID study by Raoux et al. (Raoux et al., 2008) in a similar longitudinal study of aging in which participants who progressed to AD dementia also showed declines in SVF but not MCS measures.

4.2 Cognitive mechanisms associated with semantic indices

Contrary to our hypothesis, we did not find any significant associations between MCS and any of the cognitive domains. Also contrary to initial expectations, CuRel was found to be associated only with the visuospatial domain but not memory or attention, and RepD was found to be associated with memory and visuospatial domains but not attention. The finding that the CuRel measure of semantic diversity is associated with the visuospatial domain in CN participants may be due to the fact that the tests comprising the visuospatial domain in this study consisted of WAIS-R Picture Completion and Block Design tests. Both of these tests have been shown to correlate strongly with other predominantly verbal WAIS-R tests of Information, Vocabulary, and Similarities (Lezak, 2012). The Block Design and Picture Completion tasks are also part of the Perceptual Organization Index, one of several reliable factors identified in WAIS-R battery that indexes how well an individual can organize, process and reason with visual and spatial information. In this context, the fact that greater baseline semantic diversity (lower CuRel scores) is associated with better performance on the visuospatial domain tasks in CN individuals suggests that the degree to which an individual is able to draw on a broader variety of objects (animals in this case) may be related to their ability to categorize and retrieve objects from semantic memory based on their visual and perceptual features. These findings are also interesting in light of the fact that decreased function in the visuospatial cognitive domain as measured by the Block

Design task is an independent predictor of Alzheimer's disease, particularly in the early stages when diagnosis is difficult to establish (Arnáiz et al., 2001).

The fact that the traditional SVF scores in CN participants at baseline are only weakly correlated with both the visuospatial domain scores and with CuRel scores (see Table 2) suggests that the CuRel measure indexes an aspect of the task performance that is different from the traditional SVF score. Additional evidence for this hypothesis was provided in a prior neuroimaging study showing that the CuRel measure was associated with a different language neural network in CN participants from the network associated with the traditional SVF score (Pakhomov et al., 2015). Thus, it is possible that the CuRel measure of semantic diversity provides a link between the semantic memory and the visuospatial cognitive domains, which is not readily apparent with either the traditional measures of SVF performance, or measures based on semantic clusters such as the MCS. It is possible that some of the same cognitive mechanisms are involved in responding to an SVF task and the visuospatial tasks. Semantic relatedness between concrete objects such as animals may involve mechanisms that underlie perceptual processing of various visual features of these objects. We found some indirect evidence for this in a previous neuroimaging study (Pakhomov et al., 2015) in which we found an association between a measure of semantic diversity and brain networks that are typically associated with visual categorization (e.g., the fusiform gyrus). At the individual test level, the greater association of CuRel with Block Design vs. Picture Completion performance contrasted with the association of MWF with both Block Design and Picture Completion makes sense given that the Block Design task is focused more on testing perceptual organization and Picture Completion is focused more on testing visual recognition and is correlated with performance on tests of memory, information and vocabulary (Lezak, 2012).

Intuitively, it would make sense that perceptual organization abilities would enable strategies that would produce more semantically diverse responses. It is also plausible that individuals that perform better on tests of memory, information and vocabulary would also be better able to produce more low frequency words in their SVF responses. CuRel is also correlated reasonably well with MWF ($r = 0.65$), which suggests that the notions of semantic relatedness and word frequency are not completely independent even though they are measuring fundamentally different phenomena. This is not a surprising finding, given the context of prior work on using information content, a measure derived from word frequency, to estimate the amount of information shared between any given pair of concepts in an ontology as an approximation of their semantic similarity (Resnik, 1999).

Another interesting observation arising from longitudinal analyses of semantic indices in the CN group is the presence of a positive association between time and the memory domain while the association between time and other cognitive domains was negative. This finding suggests that CN participants' attention and visuospatial abilities are getting slightly worse over time (likely due to normal aging); however, memory appears to be improving. This seemingly counterintuitive result is likely due to the practice effects associated with the neuropsychological tests of delayed free recall comprising the memory domain scores (Lezak, 2012). This finding may have implications for using global cognitive functioning scores that include the memory domain as one of their components in longitudinal studies,

as the memory domain scores may ameliorate any differences between diagnostic groups in overall cognitive functioning.

With respect to the RepD index, the results of the current study also show that the frequency with which participants repeated words that had already been spoken by them was strongly associated with the slope of decline in the memory and visuospatial domains and less so in the attention domain. This is not a surprising finding because the scores for the memory domain are derived from tests of delayed free recall and verbal learning that rely heavily on episodic memory. One would expect that repeating words that were spoken earlier in response to this task would be sensitive to decreased episodic memory capacity.

Nonetheless, this finding is interesting in that it is the only index (of the ones that were examined in this study apart from SVF) that was associated with the memory domain making it potentially useful for differentiating types of cognitive impairment due to various forms of dementia. This is particularly important in light of the fact that the SVF test has been recommended as a quick, sensitive and specific tool for assessment of early signs of dementia (Canning, Leach, Stuss, Ngo, & Black, 2004). However, the traditional SVF score that is currently the primary measure in SVF testing is strongly associated with trajectories of change in all three domains making it a sensitive but not a particularly selective tool with respect to underlying cognitive mechanisms.

With respect to MVF, the frequency of animal names that the participants chose to use in their SVF responses, our results indicate that use of higher frequency words had a detrimental association with performance in the visuospatial domain and the Trail Making task in the attention domain. As expected, the scores on both tests decline with age (see Time variable in Table 2); however, only the Trail Making scores show a negative association with baseline word frequency. Both the Trail Making Part B and the Digit Symbol tests index attention and psychomotor speed components (Lezak, 2012); however, the Trail Making test has also been shown to correlate with other aspects of executive function such as mental flexibility in particular (Bowie & Harvey, 2006). Thus, the finding that word frequency was associated with the Trail Making but not the Digit Symbol task suggests that word frequency on The SVF test may be more sensitive to cognitive flexibility, but not processing speed. Further investigation is needed to examine this disassociation; however, the results of the present study show that using additional characteristics of SVF performance in addition to traditional SVF scores can help examine multiple cognitive mechanisms without having to administer further tests.

4.3 Effects of attention on lexical access

Frequency of a lexical item reflects how easy it is to access during production. The ease of access of frequent words that are not semantically related may directly compete with the ease of access to less frequent words that are strongly related. For example, the infrequent animal name “tasmanian devil” (122 mentions in Wikipedia) may follow the relatively frequent name “kangaroo” (8,373 mentions in Wikipedia) in an SVF response if the access mechanism is relying primarily on semantic relatedness rather than frequency. In contrast, if frequency is the primary mechanism, the name “kangaroo” may be followed by a highly

frequent word from another animal category such as “beaver” (8,123 mentions in Wikipedia).

Taler et al. (2013) hypothesized that the source of information that is used during the SVF task depends on how much executive capacity is available – the less of this capacity is available to the speaker, the more likely they are to rely on frequency rather than semantic relatedness. The results of the current study with respect to the associations between indices that represent semantic relatedness and word frequency information sources are consistent with results previously reported by Taler et al. (2013). In the current study, we found that longitudinal change in word frequency information is sensitive to decreased attention aspect of executive function, a condition that we treated as analogous to increased executive demands imposed by the language-switching task in Taler’s study. In contrast, the absence of a significant difference in the effect of decreased attention on a measure based on semantic relatedness suggests that lexical access mechanism(s) tend to be driven more by word frequency information and less so by semantic relatedness information under executive load. This result confirms the Taler et al. (2013) finding that semantic similarity (a special case of semantic relatedness) is a strong factor in a monolingual task but becomes less important than frequency in a higher executive demand language switching task.

4.4 Comparison to previous work on computerized semantic indices

This study demonstrates the utility of using computerized semantic indices derived from individual responses to the SVF task using log-likelihood estimation as a way to determine the degree of semantic relatedness between words. In previous studies, we and other investigators have examined similar computerized techniques based on SVF responses as well as large corpora of text such as Wikipedia to measure semantic relations between words produced on the SVF task in order to model semantic organization of various categories (e.g., animals) and use these models to detect and characterize cognitive impairment in individuals with neurodegenerative and psychiatric conditions based on their responses to the SVF task. To our knowledge, the current study is the first to apply computerized semantic indices to SVF responses obtained in a large longitudinal study of aging.

In previous work, we have shown on the Nun Study dataset that computerized semantic indices related to semantic clustering in SVF responses from initially cognitively normal individuals were associated with the risk of developing dementia later in life (Pakhomov & Hemmy, 2013). We also found, in a different study of elderly Veterans Administration (VA) patients, that computerized semantic indices significantly differed in groups with AD dementia and MCI and were strongly associated with neuropsychological measures of executive function as well as the rate of cognitive decline (Pakhomov et al., 2012). These previous studies; however, had a number of limitations including biased populations (all women in the Nun Study and predominantly men in the VA study), as well as limited availability of neuropsychological measures of executive function in the Nun Study and lack of steady longitudinal follow-up in the VA study. These limitations made it impossible at the time to compare trajectories of cognitive decline in any aspect of executive function and semantic indices.

The current study was designed to avoid the limitations of previous studies in a number of important ways. First, we selected a large and heterogeneous population of participants in the Mayo Clinic Study of Aging (MCSA) that participated in longitudinal neurocognitive assessments. The large size of this dataset allowed us to split the participants who remained cognitively normal throughout the study into two groups – one used to derive measures of semantic relatedness based on participant responses and the other to compare semantic indices based on these measures to other groups of participants diagnosed with MCI and AD dementia. The results of the cross-sectional analysis in the present study show that semantic indices computed using the combination of FAN and G² approaches are comparable to those developed and tested in prior studies.

5. Limitations

This study has a number of limitations that affect the interpretation of results. First, we only examined SVF responses to the ‘animals’ category. The results obtained with this category may not translate directly to alternative and commonly used in SVF tests categories such as fruits/vegetables, tools and supermarket items. Second, in the current study we used experiment-based modeling of semantic relations (i.e., the computation of the strength of relatedness was based on participants’ responses that were withheld from both cross-sectional and longitudinal evaluations). Thus, the computerized semantic indices resulting from this approach are different from those based on distributional models (i.e., those that rely on large corpora of text to compute semantic relatedness measures). Distributional approaches may introduce additional noise into the computation of semantic indices resulting in more variability in measurements and, thus, would need to be further validated. The analysis presented in this manuscript does not take into account depression as a concomitant condition with dementia that may also affect semantic memory and SVF performance (Brunet et al., 2011; Callahan et al., 2015; Klumpp et al., 2010). We plan to investigate effects of depression in future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by grants from the Alzheimer’s Association (DNCFI-12-242985) and the National Institutes of Health (P50 AG16574, U01 AG06786). We also would like to thank Jennifer Strommen, Benjamin Eischens, Mara Anderson, the University of Minnesota students that worked on digitizing thousands of hand-written SVF tests, and James Ryan and Thomas Christie that worked on programming the VFClust package. We also would like to thank Dana Swenson-Dravis and Dorla Burton for extracting and collating psychometric information at the Mayo Clinic for the ADRC and MCSA study participants. Last but not least, we extend special thanks to the study participants, their families, and caregivers.

Appendix A. Semantic relatedness and clustering approaches calibration and validation results

A.1 Validation of semantic relatedness measures

Approach

To validate the G^2 values combined with scaled FAN values as a way to measure the strength of semantic relatedness between pairs of animal names, we compared these values to manual semantic relatedness judgments. The manual judgments were obtained on a set of 36 pairs originating from a different study by Weber et al., (2009). This set consists of nine animal names and all possible pairs ($n=36$) of these nine animals. In the original study by Weber et al. (2009), these pairs were presented to 12 participants in a series of behavioral and imaging experiments in which the participants were asked to judge photographic images of these pairs of animals for similarity. The current study relies on the notion of semantic relatedness rather than similarity for clustering animal names. Therefore, rather than using the similarity scores obtained in Weber's study we performed a separate study in which we asked three raters (graduate and undergraduate students in linguistics at the University of Minnesota) to judge these 36 pairs for semantic relatedness. The raters registered their judgments by pressing a solid color bar (representing a continuous relatedness scale) on a touchscreen computer after each pair of animal names was presented to them for 4 seconds. The left side of the bar was labeled as "completely unrelated" and the right side of the bar was labeled as "closely related." The raters were asked to rate the 36 animal pairs after receiving instruction and a brief practice session on a small independent set of animal pairs. The x-coordinate position of the place where the raters touched the screen was used as a continuous measure of relatedness, with larger values (towards the right-side of the screen) representing greater relatedness values.

Results

After calculating inter-rater agreement, we compared the mean relatedness values obtained through manual assessments with the automatically computed G^2 scores combined with the scaled FAN scores. The three human raters that judged the 36 animal pairs for semantic relatedness achieved a good level of agreement with each other ($ICC = 0.65$; 95% CI 0.40 – 0.81). When compared to the ratings of visual similarity obtained from 12 participants in Weber's study, the two types of ratings were highly correlated ($\rho = 0.83$). The disagreements between the two sets of ratings had to do with the distinction between relatedness and similarity. For example, in Weber's similarity dataset, the pairs cougar-bear, hippopotamus-horse, and hippopotamus-bear were rated as similar (> 0.5 on a scale of 0 to 1.0). These pairs were rated as somewhat unrelated (x-coordinate < 1000 on a scale of 0 to 2000) in our relatedness dataset. In contrast, our raters judged giraffe-hippopotamus as somewhat related, whereas this pair was judged as dissimilar in Weber's dataset. These results support the distinction between similarity and relatedness judgments; however, they also indicate that while re-annotating this dataset for relatedness likely improved its representation of the phenomenon of semantic relatedness in a few cases, the difference between similarity and relatedness judgments is not very large. Nonetheless, we used the

dataset annotated for relatedness in further validation of G^2 and FAN based semantic associations.

We found that manual scores were highly correlated with the combined G^2 and FAN scores ($\rho = 0.78$). We also found that the correlation between manual scores and G^2 scores without replacing the G^2 with FAN scores where available resulted in lower correlation ($\rho = 0.73$). FAN scores are not available for all pairs; therefore, we could not correlate manual scores to FAN scores alone (without the influence of G^2 scores).

A.2 Validation of automatic clustering

Approach

To validate the application of G^2 scores combined with the scaled FAN scores to define clusters in the SVF responses, we created a reference standard by manually identifying clusters in the responses from the participants in the cross-sectional ADRC cohort. Following a brief training session, manual cluster identification was performed independently by two raters. Both raters were native speakers of English. The raters were instructed to use their intuition to define cluster boundaries in SVF responses and use Troyer's clustering guidelines only as a reference for category membership. However, we departed from Troyer's methodology by instructing the raters to rely more on their intuition of semantic associations rather than strictly using the categorization scheme. Thus, we introduced the possibility that the raters would place functionally related animal names into the same cluster even if they belonged to different animal categories (e.g. "wolf" and "sheep").

Results

After two human raters judged 167 SVF responses from the cross-sectional dataset, we measured inter-rater agreement using two different methods to compute MCS: one included singleton clusters, and the other excluded them. Both methods resulted in good-to-excellent agreement ($ICC = 0.86$ and 0.95 , respectively). We could not directly correlate manual and automated MCS measures because they are not truly independent for the following reasons. The count of clusters for both measures is a function of the total number of words – the correlation between the word count and automatic cluster count is 0.91 and the correlation between word count and the manual cluster count is 0.77 . Manual and automatic cluster counts are also highly correlated ($r = 0.82$). Thus, the cluster count is in the denominator of both manual and automatic MCS measures and is very highly correlated to the total number of words. Directly correlating ratios that share any of the terms would result in spurious and invalid correlations (Pearson, 1896). Therefore, instead of correlating the ratios that represent manual and automatic MCS measures, we correlated the numerators and denominators of the ratios separately resulting in correlation coefficients of 0.97 and 0.82 , respectively. In addition to correlations, we also computed the mean and median absolute differences between manual and automatic MCS measures resulting in a small mean cluster size difference of 0.45 words (95% CI $0.38 - 0.52$) and median difference of 0.39 words (95% CI $0.33 - 0.47$).

A.3 Calibration of relatedness threshold for automatic clustering

In addition to validating automatic clustering measures, we used the manual clustering assessments to find the optimal threshold for computerized semantic relatedness values. The range of thresholds with which we experimented and the corresponding evaluation metrics are shown in Figure 3. The relatedness score threshold of 410 is the point at which the correlations between the manual and automatic MCS numerators and the denominators is the highest and the mean absolute difference between MCS measures is the lowest. We used this threshold to compute automatic MCS measures in all subsequent analyses reported in this article.

References

- Adelman JS, Brown GDA. Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*. 2008; 115(1):214–227. <http://doi.org/10.1037/0033-295X.115.1.214>. [PubMed: 18211194]
- American Psychiatric Association. DSM-IV: Diagnostic and Statistical Manual of Mental Disorders. 4th. Washington, DC: American Psychiatric Association; 1994.
- Arnáiz E, Jelic V, Almkvist O, Wahlund LO, Winblad B, Valind S, Nordberg A. Impaired cerebral glucose metabolism and cognitive functioning predict deterioration in mild cognitive impairment. *Neuroreport*. 2001; 12(4):851–855. [PubMed: 11277595]
- Bird, S.; Klein, E.; Loper, E. Natural language processing with Python. 1st. Beijing: Cambridge [Mass.]: O'Reilly; 2009.
- Bowie CR, Harvey PD. Administration and interpretation of the Trail Making Test. *Nature Protocols*. 2006; 1(5):2277–2281. <http://doi.org/10.1038/nprot.2006.390>. [PubMed: 17406468]
- Brunet J, Hudon C, Macoir J, Belleville S, Rousseau F, Bouchard RW, Joubert S. The relation between depressive symptoms and semantic memory in amnesic mild cognitive impairment and in late-life depression. *Journal of the International Neuropsychological Society: JINS*. 2011; 17(5):865–874. <http://doi.org/10.1017/S1355617711000877>. [PubMed: 21729399]
- Callahan BL, Joubert S, Tremblay MP, Macoir J, Belleville S, Rousseau F, Hudon C. Semantic Memory Impairment for Biological and Man-Made Objects in Individuals With Amnesic Mild Cognitive Impairment or Late-Life Depression. *Journal of Geriatric Psychiatry and Neurology*. 2015; 28(2):108–116. <http://doi.org/10.1177/0891988714554708>. [PubMed: 25344480]
- Canning SJD, Leach L, Stuss D, Ngo L, Black SE. Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. *Neurology*. 2004; 62(4):556–562. <http://doi.org/10.1212/WNL.62.4.556>. [PubMed: 14981170]
- Chan AS, Butters N, Salmon DP, McGuire KA. Dimensionality and clustering in the semantic network of patients with Alzheimer's disease. *Psychol Aging*. 1993; 8:411–9. [PubMed: 8216961]
- Dunning T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*. 1993; 19(1):61–74.
- Epker MO, Lacritz LH, Munro Cullum C. Comparative analysis of qualitative verbal fluency performance in normal elderly and demented populations. *J Clin Exp Neuropsychol*. 1999; 21:425–34. [PubMed: 10550803]
- Fellbaum, C. WordNet: an electronic lexical database. Cambridge, Mass: MIT Press; 1998.
- Goni J, Arrondo G, Sepulcre J, Martincorena I, Velez de Mendizabal N, Corominas-Murtra B, Villoslada P. The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cognitive Processing*. 2011; 12:183–96. [PubMed: 20938799]
- Henry JD, Crawford JR, Phillips LH. Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia*. 2004; 42(9):1212–1222. <http://doi.org/10.1016/j.neuropsychologia.2004.02.001>. [PubMed: 15178173]
- Hoffman P, Meteyard L, Patterson K. Broadly speaking: Vocabulary in semantic dementia shifts towards general, semantically diverse words. *Cortex*. 2014; 55:30–42. <http://doi.org/10.1016/j.cortex.2012.11.004>. [PubMed: 23261549]

- Holshausen K, Harvey PD, Elvevåg B, Foltz PW, Bowie CR. Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex*. 2014; 55:88–96. <http://doi.org/10.1016/j.cortex.2013.02.006>. [PubMed: 23510635]
- Ivnik RJ, Malec JF, Smith GE, Tangalos EG, Petersen RC, Kokmen E, Kurland LT. Mayo's older americans normative studies: Updated AVLT norms for ages 56 to 97. *Clinical Neuropsychologist*. 1992; 6(sup001):83–104. <http://doi.org/10.1080/13854049208401880>.
- Kaplan, E.; Goodglass, H.; Weintraub, S. Boston Naming Test. Philadelphia: Lea & Febiger; 1983.
- Klumpp H, Keller J, Miller GA, Casas BR, Best JL, Deldin PJ. Semantic processing of emotional words in depression and schizophrenia. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*. 2010; 75(2):211–215. <http://doi.org/10.1016/j.ijpsycho.2009.12.004>. [PubMed: 20006969]
- Kokmen E, Smith GE, Petersen RC, Tangalos E, Ivnik RC. The short test of mental status. Correlations with standardized psychometric testing. *Arch Neurol*. 1991; 48:725–8. [PubMed: 1859300]
- Levelt, WJ. Speaking: from intention to articulation. Cambridge, MA: MIT Press; 1989.
- Lezak, MD., editor. Neuropsychological assessment. 5th. Oxford; New York: Oxford University Press; 2012.
- Lucas JA, Ivnik RJ, Smith GE, Bohac DL, Tangalos EG, Graff-Radford NR, Petersen RC. Mayo's Older Americans Normative Studies: Category Fluency Norms. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A)*. 1998; 20(2):194–200. <http://doi.org/10.1076/jcen.20.2.194.1173>.
- Mayr U. On the dissociation between clustering and switching in verbal fluency: comment on Troyer, Moscovitch, Winocur, Alexander and Stuss. *Neuropsychologia*. 2002; 40(5):562–566. [http://doi.org/10.1016/S0028-3932\(01\)00132-4](http://doi.org/10.1016/S0028-3932(01)00132-4). [PubMed: 11749985]
- McDowd J, Hoffman L, Rozek E, Lyons KE, Pahwa R, Burns J, Kemper S. Understanding verbal fluency in healthy aging, Alzheimer's disease, and Parkinson's disease. *Neuropsychology*. 2011; 25(2):210–225. <http://doi.org/10.1037/a0021531>. [PubMed: 21381827]
- Mielke MM, Roberts RO, Savica R, Cha R, Drubach DI, Christianson T, Petersen RC. Assessing the temporal relationship between cognition and gait: slow gait predicts cognitive decline in the Mayo Clinic Study of Aging. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*. 2013; 68(8):929–937. <http://doi.org/10.1093/gerona/gls256>.
- Monsch AU, Bondi MW, Butters N, Salmon DP, Katzman R, Thal LJ. Comparisons of Verbal Fluency Tasks in the Detection of Dementia of the Alzheimer Type. *Archives of Neurology*. 1992; 49(12):1253–1258. <http://doi.org/10.1001/archneur.1992.00530360051017>. [PubMed: 1449404]
- Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*. 1993; 43(11):2412–2412. <http://doi.org/10.1212/WNL.43.11.2412-a>. [PubMed: 8232972]
- Murray WS, Forster KI. Serial Mechanisms in Lexical Access: The Rank Hypothesis. *Psychological Review*. 2004; 111(3):721–756. <http://doi.org/10.1037/0033-295X.111.3.721>. [PubMed: 15250781]
- Nelson DL, McEvoy CL, Schreiber TA. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*. 2004; 36(3):402–407. <http://doi.org/10.3758/BF03195588>.
- Pakhomov S, Hemmy Laur. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex*. 2013 (in press), (epub ahead of print).
- Pakhomov S, Hemmy LS. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex*. 2014; 55:97–106. <http://doi.org/10.1016/j.cortex.2013.05.009>. [PubMed: 23845236]
- Pakhomov S, Hemmy LS, Lim KO. Automated semantic indices related to cognitive function and rate of cognitive decline. *Neuropsychologia*. 2012; 50:2165–75. <http://doi.org/10.1016/j.neuropsychologia.2012.05.016>. [PubMed: 22659109]
- Pakhomov S, Jones DT, Knopman DS. Language networks associated with computerized semantic indices. *NeuroImage*. 2015; 104:125–137. <http://doi.org/10.1016/j.neuroimage.2014.10.008>. [PubMed: 25315785]

- Pearson K. Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. *Proceedings of the Royal Society of London* (1854–1905). 1896; 60(1):489–498. <http://doi.org/10.1098/rspl.1896.0076>.
- Petersen RC. Mild cognitive impairment as a diagnostic entity. *J Intern Med*. 2004; 256:183–194. [PubMed: 15324362]
- Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*. 1999; 56(3):303–308. [PubMed: 10190820]
- Price SE, Kinsella GJ, Ong B, Storey E, Mullaly E, Phillips M, Perre D. Semantic verbal fluency strategies in amnesic mild cognitive impairment. *Neuropsychology*. 2012; 26:490–7. <http://doi.org/10.1037/a0028567>. [PubMed: 22746308]
- Raoux N, Amieva H, Le Goff M, Auriacombe S, Carcaillon L, Letenneur L, Dartigues JF. Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: results from the PAQUID longitudinal study. *Cortex*. 2008; 44:1188–96. [http://doi.org/S0010-9452\(07\)00158-X](http://doi.org/S0010-9452(07)00158-X) [pii] 10.1016/j.cortex.2007.08.019. [PubMed: 18761132]
- Reitan RM. VALIDITY OF THE TRAIL MAKING TEST AS AN INDICATOR OF ORGANIC BRAIN DAMAGE. *Perceptual and Motor Skills*. 1958; 8(3):271–276. <http://doi.org/10.2466/pms.1958.8.3.271>.
- Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J of Artif Intell Res*. 1999; 11:95–130.
- Roberts RO, Geda YE, Knopman DS, Cha RH, Pankratz VS, Boeve BF, Rocca WA. The Mayo Clinic Study of Aging: Design and Sampling, Participation, Baseline Measures and Sample Characteristics. *Neuroepidemiology*. 2008; 30(1):58–69. <http://doi.org/10.1159/000115751>. [PubMed: 18259084]
- Roberts RO, Knopman DS, Mielke MM, Cha RH, Pankratz VS, Christianson TJH, Petersen RC. Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal. *Neurology*. 2014; 82(4):317–325. <http://doi.org/10.1212/WNL.000000000000055>. [PubMed: 24353333]
- Rosenstein, M.; Foltz, P.; Vaskin, A.; Elvevag, B. 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Denver, CO: 2015. Practical issues in developing semantic frameworks for the analysis of verbal fluency data: A Norwegian data case study; p. 124–133.
- Rosen WG. Verbal fluency in aging and dementia. *J Clin Neuropsychol*. 1980; 2:135–146.
- Taler V, Johns BT, Young K, Sheppard C, Jones MN. A computational analysis of semantic structure in bilingual verbal fluency performance. *Journal of Memory and Language*. 2013; 69(4):607–618. <http://doi.org/10.1016/j.jml.2013.08.004>.
- Troster AI, Fields JA, Testa JA, Paul RH, Blanco CR, Hames KA, Beatty WW. 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]
- Voorspoels W, Storms G, Longenecker J, Verheyen S, Weinberger DR, Elvevåg B. Deriving semantic structure from category fluency: Clustering techniques and their pitfalls. *Cortex*. 2014; 55:130–147. <http://doi.org/10.1016/j.cortex.2013.09.006>. [PubMed: 24275165]
- 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. [http://doi.org/S0028-3932\(08\)00504-6](http://doi.org/S0028-3932(08)00504-6) [pii] 10.1016/j.neuropsychologia.2008.12.029. [PubMed: 19162048]
- Wechsler, D. Wechsler Memory Scale Manual, revised. New York: Psychological Corporation; 1981.

Weintraub S, Wicklund AH, Salmon DP. The Neuropsychological Profile of Alzheimer Disease. Cold Spring Harbor Perspectives in Medicine. 2012; 2(4):a006171–a006171. <http://doi.org/10.1101/cshperspect.a006171>. [PubMed: 22474609]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Highlights

- a novel computerized approach to characterize semantic verbal fluency performance
- trajectories of decline in some but not all computerized semantic indices differ across diagnostic groups computerized
- longitudinal change in word frequency information is sensitive to decreased executive function
- lexical access mechanism(s) are driven word frequency and less so by semantic relatedness under executive function load

Figure 1.

Study Design.

Note: N_p – number of participants; N_s – number of samples.

Figure 2. Distribution of sorted relatedness values computed from CN-Train corpus and supplemented with FAN values. (Actual values have been log-transformed and shifted up by 30 for readability)

Figure 3.

Results of calibrating VFClust semantic relatedness threshold. The left y-axis reflects cluster count and cluster sum correlation coefficients. The right y-axis reflects the magnitude of the difference between manual and automatic cluster counts.

Figure 4. Differences in means across study variables by diagnostic group in the cross-sectional ADRC cohort.

Figure 5.
Illustration of differences in trajectories in study variables across diagnostic groups in the longitudinal MCSA cohort.

Table 1

Differences in trajectories of semantic indices over time by diagnostic group (adjusted for the baseline SVF score)

Predictor	Coef. β	SE (β)	t-value	p-value
Model 1 (SVF)				
Age at baseline	−0.37	0.05	−7.20	<0.001 ***
Sex (male)	−5.05	0.51	−9.97	<0.001 ***
Years of educ.	0.52	0.08	6.22	<0.001 ***
CN slope (per year)	−0.55	0.07	−7.99	<0.001 ***
Diff MCI slope from CN Slope	−0.82	0.10	−8.17	<0.001 ***
Diff AD slope from CN Slope	−2.53	0.18	−13.97	<0.001 ***
Model 2 (MCS)				
Age at baseline	−0.001	0.002	−0.88	0.377
Sex (male)	0.005	0.015	0.35	0.726
Years of educ.	−0.009	0.003	−3.50	<0.001 ***
Repeated words	0.013	0.005	2.60	0.001 **
Baseline SVF	−0.003	0.002	−1.82	0.069
CN slope (per year)	0.006	0.004	1.82	0.068
Diff MCI slope from CN slope	0.010	0.005	1.88	0.060
Diff AD slope from CN slope	−0.010	0.010	−1.02	0.307
Model 3 (log CuRel)				
Age at baseline	−0.002	0.002	−0.98	0.329
Sex (male)	−0.036	0.018	−2.07	0.039 *
Years of educ.	−0.015	0.003	−5.03	<0.001 ***
Repeated words	0.074	0.006	13.27	<0.001 ***
Baseline SVF	−0.038	0.002	−18.09	<0.001 ***
CN slope (per year)	0.018	0.004	4.16	<0.001 ***
Diff MCI slope from CN slope	0.024	0.006	3.82	<0.001 ***
Diff AD slope from CN slope	0.079	0.012	6.86	<0.001 ***
Model 4 (log RepD)				
Age at baseline	0.006	0.003	2.36	0.019 *
Sex (male)	−0.004	0.024	−0.18	0.859
Years of educ.	−0.014	0.004	−3.52	<0.005 ***
Baseline SVF	−0.017	0.003	−6.14	<0.001 ***
CN slope (per year)	0.015	0.004	4.12	0.012 *
Diff MCI slope from CN slope	0.023	0.008	2.86	<0.004 ***

Predictor	Coef. β	SE (β)	t-value	p-value
Diff AD slope from CN slope	0.103	0.015	6.95	<0.001 ***
Model 5 (log MWF)				
Age at baseline	0.000	0.002	-0.11	0.915
Sex (male)	-0.115	0.021	-5.40	<0.001 ***
Years of educ.	-0.016	0.004	-4.43	<0.001 ***
Repeated words	0.057	0.006	9.19	<0.001 ***
Baseline SVF	-0.031	0.003	-11.98	<0.001 ***
CN slope (per year)	0.023	0.005	4.59	<0.001 ***
Diff MCI slope from CN slope	0.026	0.007	3.62	<0.001 ***
Diff AD slope from CN slope	0.069	0.013	5.32	<0.001 ***

* significant at < 0.05;

** significant at < 0.01;

*** significant at < 0.001

Table 2

Associations between baseline semantic indices and change in memory, attention and visuospatial scores over time in the MCSA cohort. Values reported in the column labeled “Time” correspond to the slopes in the response variable (Memory, Attention, Visuospatial) for Model 1. The “Time” variable was included in each of the 5 models; however, its effects were almost identical (differed in the third decimal point) in all 5 models. For space considerations, so we only reported these effects for one of the 5 models but they apply to the other 4 models as well. Values in the columns marked “Baseline” indicate how much the slope in response variable (Memory, Attention, Visuospatial) shifted up or down based on the magnitude of the baseline measurement in the predictor variable (SVF, MCS, CuRel, RepD, MWFF).

	Change over time in response variable	Model 1	Model 2	Model 3	Model 4	Model 5
Response variable	Time (years) β (p-value)	Baseline SVF β (p-value)	Baseline MCS β (p-value)	Baseline log(CuRel) β (p-value)	Baseline log(RepD) β (p-value)	Baseline log(MWFF) β (p-value)
Memory	0.030 (<0.001) ***	0.023 (<0.001) ***	0.221 (0.121)	-0.144 (0.157)	-0.245 (<0.001) ***	-0.155 (0.112)
WMS-R Delayed Verbal Recall	0.050 (<0.001) ***	0.017 (0.002) **	0.344 (0.032) *	-0.014 (0.902)	-0.160 (0.040) *	-0.107 (0.350)
WMS-R Delayed Visual Reproduction	0.025 (<0.001) ***	0.016 (<0.001) ***	0.202 (0.136)	-0.010 (0.308)	-0.221 (0.001) **	-0.128 (0.179)
AVLT ½ hour delay	-0.003 (0.641)	0.020 (<0.001) ***	0.173 (0.230)	-0.174 (0.097)	-0.269 (<0.001) ***	-0.155 (0.064)
Attention	-0.050 (0.005) **	0.015 (<0.001) ***	0.124 (0.343)	-0.025 (0.787)	-0.024 (0.068)	-0.066 (0.475)
TMT-B	-0.050 (<0.001) ***	0.017 (<0.001) ***	0.039 (0.749)	-0.119 (0.168)	-0.060 (0.303)	-0.044 (<0.001) ***
WAIS-R Digit-Symbol	-0.044 (<0.001) ***	0.011 (0.030) *	0.182 (0.218)	0.059 (0.579)	0.008 (0.903)	0.016 (0.874)
Visuospatial	-0.015 (0.005) **	0.017 (<0.001) ***	-0.009 (0.949)	-0.267 (0.009) **	-0.212 (0.001) **	-0.328 (0.001) **
WAIS-R Picture Completion	0.027 (<0.001) ***	0.013 (0.003) **	-0.054 (0.678)	-0.177 (0.063)	-0.178 (0.008) **	-0.329 (<0.001) ***
WAIS-R Block Design	-0.054 (<0.001) ***	0.016 (0.002) **	0.012 (0.939)	-0.306 (0.007) ***	-0.210 (0.005) **	-0.267 (0.020) *

* significant at < 0.05;

** significant at < 0.01;

*** significant at < 0.001

Table 3
Correlations between semantic indices and cognitive domain scores found at the baseline visit in the MCSA CN group.

	Memory	Attention	Visuospatial	SVF	MCS	log(CuRel)	log(RepD)
Memory							
Attention	0.29**						
Visuospatial	0.30**	0.41**					
SVF	0.37**	0.36**	0.23**				
MCS	0.06	0.01	-0.02	0.00			
log(CuRel)	-0.09	-0.07	-0.17*	-0.41**	0.55**		
log(RepD)	-0.25	-0.10	-0.26**	-0.36**	0.09	0.41**	
log(MWF)	-0.09	-0.10	-0.21**	-0.32**	0.32**	0.60**	0.33**

* - significant at alpha < 0.05;
** - significant at alpha < 0.01

Table 4

Differences in trajectories of CuRel and MWF over time in HIGH vs. LOW attention groups

Predictor	Coef. β	SE (β)	t-value	p-value
Model 1 log(CuRel)				
Age at baseline	0.001	0.004	0.20	0.844
Sex (male)	-0.040	0.032	-1.26	0.210
Years of educ.	-0.017	0.006	-3.10	0.003**
Repeated words	2.788	0.207	13.49	<0.001***
LOW_ATT slope (per year)	0.025	0.010	2.43	0.015**
Diff HIGH_ATT slope from LOW_ATT slope	-0.008	0.011	-0.67	0.506
Model 2 log(MWF)				
Age at baseline	0.003	0.004	0.71	0.477
Sex (male)	-0.109	0.004	-2.89	0.004**
Years of educ.	-0.017	0.007	-2.54	0.012**
Repeated words	1.875	0.233	8.04	<0.001***
LOW_ATT slope (per year)	0.060	0.013	4.76	<0.001***
Diff HIGH_ATT slope from LOW_ATT slope	-0.039	0.014	-2.80	0.005**

* significant at alpha < 0.05;

** significant at alpha < 0.01;

*** significant at alpha < 0.001