

APPLYING NATURAL LANGUAGE PROCESSING TECHNIQUES TO SCORING THE SEMANTIC VERBAL FLUENCY TEST

Ken Anderson

CS501

Master's Non-Thesis Research

December 8, 2006

1. INTRODUCTION

Semantic verbal fluency is a commonly-used neuropsychological test which typically allows 60 seconds for a participant to generate as many words as possible belonging to a specified semantic category such as animals or vegetables [1]. Verbal fluency is commonly used for evaluating the integrity of memory and executive functions. Performance on the task is contingent on retrieval of words from long-term storage and executive control of the process. Evidence suggests that these tasks are sensitive to the presence of most forms of brain damage [2]. In addition, age-related differences in semantic fluency have been demonstrated, with older participants tending to generate fewer words than younger participants [1]. Researchers have also observed differences between older adults with dementia and those who do not have dementia [3, 4].

There are two fundamental types of scoring that can be applied to the verbal fluency test: quantitative and qualitative. The main quantitative measure is the number of words produced. Commonly used qualitative measures include clustering and switching. Clustering is defined as the production of a group of words in sequence that all belong to the same subcategory. A switch is simply a transition between clusters. Other qualitative measures could include any type of measure that analyzes the qualities of the words produced, for example, a measure of their semantic relatedness.

Another type of verbal of verbal fluency test is the phonemic verbal fluency test, in which participants are asked to generate words beginning with a specified letter. We will not consider the phonemic verbal fluency in depth, but mention it as some researchers have identified "task-discrepant clustering"—the use of phonemic clusters in a semantic test or vice versa—as potentially indicating an explicit use of strategy [2].

The original motivation for this research was to investigate whether semantic verbal fluency could be used to diagnose individuals with other brain disorders, in particular, Attention Deficit Hyperactivity Disorder (ADHD). We did not have access to data from ADHD patients during the scope of this project as we had hoped, so we focused on looking at various approaches to scoring the semantic verbal fluency test

using data from both healthy adults and adults with dementia. We investigated the use of NLP techniques to enable more sophisticated approaches toward scoring and interpreting the test. In particular, we seek to extend and increase the rigor of qualitative measures such as clustering and switching.

2. CURRENT SCORING METHODS

A number of methods of scoring verbal fluency test have been proposed. Traditionally, the test is scored by the number of acceptable responses within the allotted time. However, the fluency test provides qualitative information as well, including the characteristics of the words generated and patterns among the words [2]. There has been significant disagreement among researchers on how to best score and interpret the qualitative aspects of the test. Troyer et.al. [1] review evidence from lesion studies and functional brain imaging and conclude that the weight of evidence is that multiple brain regions are involved in the fluency task, and that semantic fluency seems to be more associated with temporal lesions. They also note that evidence from temporal and semantic analyses of word generation over extended time periods reveals that words tend to be produced in temporal clusters, with a short time interval between clusters and a longer pause between clusters. In addition, in the semantic fluency task, the words within the temporal clusters tend to be semantically related. For example, given the sequence *cat-dog-lion-elephant-zebra*, the subcategories might be labeled pets and African animals, with cats and dogs falling in the pet subcategory and the rest in the African animal subcategory. Troyer et. al. state that this response pattern has led to suggestions that performance on semantic fluency involves two processes: a) a search for subcategories and b) a search for and production of words within the subcategories. Troyer et. al. [1] propose an additional component of fluency, switching, that has not been emphasized previously. They define switching as the ability to shift efficiently to a new subcategory, and clustering as the production of words within semantic or phonemic subcategories. They propose a model of verbal fluency in which switching is associated with executive processes governed by

the frontal lobe and production of words within clusters is associated with semantic memory processes governed by the temporal lobes. Troyer [5] has also produced normative data that corrects for several demographic variables: age, education and sex. These normative data would be useful for any future work based on the Troyer system.

Troyer et. al. [1] suggest that phonemic clusters are rarely produced. In the data we examined, 2 out of 26 participants used an alphabetic strategy, generating exemplars in which each starts with a subsequent letter of the alphabet from the previous exemplar (e.g. *aardvark-bear-cat-dog-elephant*) and maintain this strategy through all or most of the allotted time period. The alphabetic strategy may not meet the strict definition of phonemic clustering used by some researchers, but it is clearly not semantic clustering, and 2/26 does not seem rare. Although we do not have enough data to analyze correlation between task discrepant clustering and overall performance, both participants who used an alphabetic strategy performed better than average which suggests that this question merits further research. Other researchers have suggested that the use of task-discrepant clustering may indicate intentional strategy use on both tasks [2]. It seems plausible that measuring intentional strategy use could have diagnostic value, so any automatic scoring system should be able to detect the use of alphabetic and other task-discrepant clustering.

Table 1 gives an example of an actual protocol from our dataset and illustrates scoring using the Troyer system as well as some additional measures used by other researchers or developed by us. Three measures are computed in the Troyer system: 1) Total number of correct words generated, defined as the sum of all words produced, excluding errors and repetitions; 2) Mean cluster size, where cluster size is counted starting with the 2nd word in a cluster, and includes errors and repetitions; and 3) Number of switches, defined as the total number of transitions between clusters, including single words as well as errors and repetitions. March and Pattison [6] also measure the raw number of subcategories and track the number of repetitions and category errors (e.g. naming a fruit when the semantic category is vegetables). In the example given, no repetitions or errors occurred, so these measures are not shown in the table. The mean latency within clusters and mean latency between clusters are measures that we added because we had time-aligned data. Additional measures used by other researcher or that would be enabled by techniques discussed in this paper will be discussed below.

A number of researchers have criticized various aspects of Troyer et. al.s [1] scoring system (referred to as the Troyer system from here on for simplicity). Others have adopted the Troyer system but have extended or refined it.

Abwender et. al. [2] performed a review and comparison of several scoring methods, including the Troyer system. They suggest that while Troyer et. al. [1] report a positive correlation between clustering and total productivity, this is not sufficient evidence that clustering necessarily leads to produc-

tion of more words or is a strategic process, since the more words that are produced, the more likely it is that what appear to be meaningful clusters will appear by chance. Abwender et. al. also state that the definition of switching used by Troyer et. al. is problematic with respect to their interpretation of how switching reflects underlying cognitive processes. Because Troyer et. al. define switching and clustering as mutually exclusive, a negative correlation must obtain for a given number of words generated. Also, by considering single words as clusters of size 0, smaller mean cluster sizes caused by many single words will be accompanied by a larger number of switches for a given number of words generated. These issues may limit the conceptual dissociability of these constructs, calling into question Troyer et. al.'s interpretation that switching is a product of strategic searching and mental flexibility. It is also possible that switching simply reflects a lack of ability to cluster. Also, positive correlations between switching and total word output do not prove that output is dependent on switching. Abwender et. al. point out that Epper et. al. [3] have suggested that the number of switches depends on total word output.

Ulrich Mayr [7] states that the Troyer system does not allow unambiguous inferences regarding the selective impairment of the switching process. This is because the number of switches depends on both the difficulty that a participant has in accessing a new cluster as well as the difficulty that a participant has in generating new words within clusters, because the more time a participant spends within a cluster, the less time remains for accessing clusters. Hence, the number of switches score does not differentiate between difficulties in accessing new clusters and difficulties in accessing new words within clusters. Mayr gives the following example: suppose a participant produces only three words from the same cluster after 15, 30, and 45s and no additional words in the remaining 15 s. In this case, the average cluster size is 2, and the number of switches is 0, but it is clear that the lack of switches is not due to a switching deficit, but rather a difficulty in retrieving words from memory.

Mayr and Kliegl [8] propose another model of how semantic and executive processes affect semantic fluency. They suggest that the time for every act of retrieval contains two additive components: one executive and the other semantic in nature. The executive component consists of tasks such as updating the current search criterion and stopping or starting single retrieval processes. It is assumed to be relatively constant, whether it occurs within or between clusters. The semantic component reflects the actual semantic search demands, such as spreading of activation in the semantic network. In this model, between-cluster retrieval (equivalent to switching in Troyer's model) should require more of the same semantic processing that occurs within cluster rather than separate executive processing. Mayr and Kliegl point out that like Troyer et. al. [1] this model assumes both semantic and executive components in semantic fluency but makes different assump-

exemplar	latency (ms)	label	word total	cluster size	no. of switches	no. of subcategories	mean latency within cluster	mean latency between clusters
cat	647.8	pet						
dog	0	pet						
bird	2395.9	pet		2				
monkey	3602.7	African						
elephant	0	African						
lion	541.3	African						
tiger	0	African						
giraffe	496.9	African		4				
mouse	2014.3	rodent						
mole	1144.7	rodent		1				
fish	11287.2	fish		0				
antelope	14019.2	deer						
deer	0	deer						
moose	6344.6	deer		2				
			14	1.8	4	5	1213.7	7748.9

Table 1. Scoring a semantic verbal fluency protocol

tions about where executive components vs. semantic retrieval come into play.

March and Pattison [6] use the Troyer system but add an additional variable: the raw number of subcategories produced. This additional variable is motivated by the observation that a participant can access as few as two subcategories and still demonstrate high rates of switching by spontaneously alternating between two. Hence, the switching variable does not account well for the variety of subcategories accessed during word generation. March and Pattison’s study also measures two error patterns: 1) number of repetitions and 2) number of categorical errors.

Epker et. al. [3] performed a study using the Troyer system, but concluded that switching is dependent on the number of words produced, most likely because as a participant produces more words, there is more opportunity to switch to different clusters. The study found no differences between groups in switching and clustering when normalized by word counts, but did find differences in word counts. Epker et. al. do suggest that qualitative measures could be useful for looking at certain disorders, as well as from a clinical standpoint.

Reverberi et. al. [9] suggest that several cognitive functions involved in verbal fluency could be attributed to the frontal lobes. They discuss compliance with a strategy as a new component to be analyzed and define two new constructs: 1) Loss of strategy as evidenced by lower average relatedness between successive words and higher frequency of switching between subcategories; and 2) Loss of switching evidenced by a lower switch rate, higher semantic relatedness between words, and few subcategories overall. They suggest that instead of counting switches, an index describing the structure of the word sequence would be more useful. They propose using the mean semantic proximity (i.e. relatedness) between

successive pairs of words to measure the degree of semantic organization of a sequence. Their study used fruits as the semantic category, so it was feasible to obtain human judgments of semantic relatedness as the number of fruits is smaller than the number of animals.

We found only one study that used the semantic verbal fluency to try to identify differences between participants with ADHD and healthy participants. Tucha et. al. [10] used the Troyer system, but also counted the number of labels produced (e.g. "bird" or "reptile" vs. "robin" or "chameleon") as well as clusters within clusters. They found a significant difference in word count and switching between patients with ADHD and healthy participants. We believe that to the extent that ADHD is associated with deficits in executive functions, a measure such as the one proposed by Reverberi may also be useful.

As computer scientists, we do not take a position on the implications of verbal fluency for understanding of brain function, but our research is informed by these debates, and we seek to both develop approaches to scoring that will allow for multiple perspectives and in particular, to expand the types of measures that can be obtained from verbal fluency data in ways that might enable future research. We agree with Mayr [7] that obtaining time-alignment data would help to overcome some of the issues with previous scoring methods discussed above, and we believe that it could also lead to the development of new measures that might be useful for diagnostic purposes. Also, we believe that a measure of semantic similarity between words within the semantic category being used for the test would allow for various kinds of analyses that do not depend on categorizing the production of each exemplar as either a switch or within the same cluster as the previous exemplar. This approach is motivated by the observation

that words can belong to more than one subcategory and thus it can be ambiguous whether a word is starting a new subcategory. For example, in the sequence *cat-dog-horse-zebra-lion-giraffe*, zebra could be considered as both equine and in a cluster with horse, as well as African and in a cluster with lion and giraffe. We suggest that with a measure of semantic similarity and latency data, the test can be evaluated without needing to categorize zebra as either equine or African.

3. METHODS AND EXPERIMENTAL RESULTS

Our research had two main goals: 1) to develop an algorithm for automatically labeling data and detect the use of task-discrepant clustering; and 2) to develop a measure of semantic relatedness. We also intended to evaluate possible uses of time-alignment data. We did not have sufficient data to do the evaluation, but we do identify several time-based measures that may be useful in future research.

3.1. Automatic Labeling and Detection of Task-Discrepant Clustering

We investigated using a Hidden Markov Model (HMM) tagger to automatically label semantic fluency data, in order to facilitate computation of measures based on clustering and switching. We built a model for a set of hand-labeled data consisting of 11 protocols obtained from elderly adult participants, including both patients with dementia as well as healthy older adults. The protocols were hand-labeled using the Troyer system, with one exception. In the Troyer system, more than one tag can be assigned to an exemplar. For example, in the sequence *mule-horse-zebra-lion-giraffe*, zebra could be tagged as both "equine" and "African". We decided to assign only one tag to an exemplar because allowing multiple tags seems to overly complicate the system without adding much in terms of explanatory power and because Troyer is not explicit on whether an exemplar labeled with two tags should be counted as a switch. Furthermore, we believe that latency can be used in judging whether a switch has occurred.

We then built a tagger using the Viterbi algorithm. We applied the model to a set of 16 protocols from a similar population that were first hand-labeled, in order to be able to evaluate the performance of the tagger. The tagger achieved 69% accuracy at the exemplar level. This is not very high compared to the performance of HMM taggers on other tasks, such as part-of-speech tagging, but there were several issues that could be remedied fairly easily which affected the performance. The first is that there were some discrepancies between some of the default labels used (i.e. the label assigned to an item if it is not part of a cluster) in labeling the first set and second set of data. This could be remedied by explicitly defining default labels for each animal prior to doing the hand labeling. The second issue is that the initial model was based on a very small dataset. It would be relatively easy to obtain additional data

and either hand-label it, or use the initial model as a starting point and use Expectation Maximization to incorporate additional unlabeled data into the model. Third, we used a very simple method of dealing with unseen data. For an unseen animal, we assumed that the animal was equally probable given any tag and relied only on tag transition probabilities to calculate the probability of each tag given that animal. For an unseen tag transition, we gave it an arbitrarily low probability of .01. Incorporating more data into the model would reduce the amount of unseen data. Also, a more sophisticated smoothing approach such as the backoff or deleted interpolation algorithms described by Jurafsky and Martin [11] could be used. Finally, we used a first-order Markov model. A second-order Markov model would probably perform better, assuming that the data sparsity issue could be overcome. This is a particularly important issue in this case since the tagging decision is heavily dependent on the tags that follow the current word, as well as the tags that precede it.

In addition, our HMM tagger produces a set of labelings that deviates from the Troyer approach in that each word is labeled with one and only one tag, whereas the Troyer system allows two tags to be assigned to a word if it belongs to more than one subcategory. We chose not to follow the Troyer system exactly, because we believe that timing data can be used to make an unambiguous labeling decision in almost all cases. We assume, based on our analysis of time-aligned data that the longer the latency of a word, the more likely the word is to represent a switch. In hand labeling our data we used the following heuristic: if word latency is greater than 1 s and it belongs to a subcategory that is different from the previous label and is the same as a subcategory that the following word belongs to, it represents a switch and is labeled accordingly. For example given the sequence *tiger-lion-cat-dog*, the labeling of cat is dependent on its latency. If it is greater than 1 s, we label it pet, otherwise, we would label it feline. In the Troyer method, cat would be labeled as both feline and pet, and presumably counted as a switch. We believe our approach provides a more unambiguous measure of switching and clustering than the Troyer approach. Our approach is also consistent with the model of verbal fluency proposed by Mayr and Kliegl [8] which, as a reminder, suggests that switching and producing words within a cluster both reflect the same type of semantic processing in the brain, and what is termed switching by Troyer represents greater distance in the semantic memory network and is not associated with a specific executive function.

Our HMM tagger was not successful in detecting the alphabetic strategy. It is not clear whether this is due to data sparsity or whether a probabilistic approach is just not suitable for this task, given that the majority of participants do not use task-discrepant strategies. In the latter case, it would be relatively straightforward to develop an algorithm that explicitly detects the use of task-discrepant strategies.

3.2. Semantic Relatedness

3.2.1. Introduction

Early in our research, we had the sense that having an objective measure of semantic similarity between exemplars would be useful. In particular, we thought that there might be significant differences between patients with ADHD and healthy controls on an index of overall semantic similarity in a test protocol. Our hypothesis was that patients with ADHD would exhibit less semantic similarity between words than would a healthy control group. Although we arrived at this viewpoint independently, our ideas are supported by the work of Reverberi et. al [9]. These researchers determined the mean semantic proximity between each successive pair of words in a protocol. They determined semantic proximity scores for word pairs by having 78 healthy subjects rate the semantic proximity between each pair of fruit in a set of 32 fruits most frequently reported in a semantic fluency task. In addition, eight healthy participants rated 420 additional pairs of fruit items in order to produce an objective similarity rating between all possible pairs of fruits that occurred in their experiment. One approach to doing this for animals would be to use human ratings, but this would not be feasible. There are roughly 200 animals that might reasonably be expected to occur in a fluency test using animals as the semantic category, resulting in 19,900 possible pairs. It would not be practical to rate that many pairs using human subjects. Therefore, we investigated the possibility of using other methods of estimating semantic relatedness, namely, deriving a measure from statistics obtained from a large corpus, and from a taxonomy such as WordNet. Both approaches have been extensively researched.

The terms semantic similarity, semantic relatedness and semantic distance are used somewhat interchangeably in the literature. Budanitsky and Hirst [12] propose the following definitions of the terms. Semantic relatedness is the broader terms and encompasses similarity as well as meronymy, antonymy, functional association, and other non-classical relations. Similarity is based on shared features, whereas relatedness is based on the other types of relatedness referred to above. For example, "dog" and "wolf" are similar, because they are closely phylogenetically related. "Dog" and "cat" are related because they are both commonly kept as pets. Budanitsky and Hirst suggest that the concept of semantic distance is inherently ambiguous, as it can be used as the inverse of either similarity or relatedness. These two uses are often consistent, but not always. For example, antonymous concepts are dissimilar and distant in that sense, but also semantically related, and not distant in that sense. Budanitsky and Hirst also point out that semantic relationships are between concepts, or particular senses of words, not between words themselves. Hence, semantic similarity is not the same as the similarity of distributional or co-occurrence behavior of the words. However, some researchers have investigated the use of distributional or

co-occurrence similarity as a proxy for semantic similarity or relatedness, with varying degrees of success (e.g. Mohammad and Hirst [13]).

In the case of animal pairs that occur in verbal fluency tests, we see examples of both semantic relatedness and similarity. In the sequence lion-zebra-giraffe, the concepts are related because they all are found in Africa. Yet, they are not very similar to each other. On the other hand, a sequence like deer-elk-antelope might be considered as a subcategory of deer, and these concepts are similar. Because the term semantic relatedness encompasses semantic similarity in the Budanitsky and Hirst definition, we will use the term semantic relatedness in general, and use the term semantic similarity only when we specifically need to refer to a case of a similarity relationship.

There are some challenges to using distributional similarity and co-occurrence behavior as a proxy for semantic relatedness. For example, a large portion of our corpus consisted of Wall Street Journal text, in which bull and bear occur frequently, and probably more often in their financial senses than in the animal senses. This could distort a measure of relatedness between their animal senses and other animals.

3.2.2. WordNet-based Measures

The simplest method of computing semantic relatedness is to use lexical resources, such as WordNet. Several researchers have found that WordNet based methods work better than corpus-based methods when evaluated against human-rated sets of word pairs such as the Rubenstein-Goodenough set or the Miller-Charles set [12, 14, 15]. At first, we did not think a WordNet based approach would be suitable for computing semantic relatedness of word pairs in the verbal fluency task, because the majority of the relations in WordNet are hyponymy/hypernymy, whereas many of the word pairs in the verbal fluency tests are based on other relations. For example, if the word dog appears in a protocol, there is a very high probability that cat will follow it. In any WordNet based measure, however, dog would be more related to wolf than to cat because wolf is closer to dog than cat in a network based on hyponymy/hypernymy relations.

Later in our research, we came back to WordNet and discovered that for some animal pairs, a WordNet based relatedness score appears to perform better than co-occurrence based statistics. We used a simple WordNet-based approach, computing the path length between two word senses by counting the number of edges between them. More sophisticated measures exist, which tend to perform better than the simple measure that we used, at least on some tasks (see [12, 16] for definitions and evaluations of other WordNet-based measures).

3.2.3. Corpus-based Measures

Next, we turned to corpus-based measures of relatedness. As mentioned earlier, distributional similarity of words is not the

same as semantic relatedness of word senses. A key question is the extent to which distributional similarity can act as a proxy for semantic relatedness. Budanitsky and Hirst [12] define the distributional similarity of two words w_1 and w_2 as 1) the extent to which they occur in similar contexts, for some definition of context; 2) the extent that the set of words that w_1 tends to occur with is the same as the set that w_2 tends to occur with; or 3) the extent to which if w_1 is substituted for w_2 in a context, its plausibility is unchanged. A context may be a small or large window around the word, an entire document, or a syntactic relationship. An example of distributional similarity based on syntactic relationships would be using verbs as a context for nouns in the object position, so that nouns would be similar to the extent that they occur as direct objects of the same set of verb. We hypothesized that sentential co-occurrence would be a useful approach because many of the subcategories we observed in our data were based on geography or living environment. For example, one would expect that zebra and giraffe to co-occur with Africa, and possibly specific African countries.

A number of measures of distributional similarity have been proposed (see Mohammad and Hirst [13] for an in-depth survey). We evaluated three, using a very large corpus of over 300 million words drawn from multiple sources including the Switchboard Corpus and the Wall Street Journal. We used the Porter Stemmer [17] to reduce words to their linguistic roots, and a stop list to filter out closed-class words. Our initial approach was to look at sentential co-occurrence based on the log-likelihood statistic, for a set of animals. We identified the set of words that co-occurred with any of the animals in the set with a log-likelihood ratio of greater than 10.83. We used this set as the components of feature vector and measured the cosine of the angle between the vectors representing two animals, similar to the approach used by McDonald [18]. The cutoff of 10.83 was motivated by the fact that if λ is a likelihood ratio, the quantity $-2\log\lambda$ is asymptotically χ^2 distributed [19]. Therefore, we compute the log-likelihood statistic as $-2\log\lambda$ and use 10.83 as a threshold, which represents significance at the $p = 0.001$ level. As an example, suppose that cat and dog have the following co-occurrences, with the log-likelihood score shown in parentheses (note that these log-likelihood scores are made up for illustrative purposes):

cat	kitten(21.4)	cute(38.7)	meow(59.0)	scratch(14.3)	pet(44.8)
dog	puppy(33.7)	cute(29.8)	bark(40.0)	bite(15.8)	pet(55.8)

The feature vectors would be

	kitten	cute	meow	scratch	pet	puppy	bark	bite
cat	21.4	38.7	59.0	14.3	44.8	0	0	0
dog	0	29.8	0	0	55.8	33.7	40.0	15.8

The cosine of the angle between vectors is computed by normalizing the vectors and taking their dot product.

We also used the vector space approach using pointwise mutual information (PMI) as the vector components instead

of log-likelihood. According to Manning and Schütze [19], the log-likelihood ratio is considered to be a better indicator of co-occurrence than pointwise mutual information for two key reasons. The first is that PMI tends to artificially increase scores of infrequently occurring word pairs. The second is that they measure different things. Given a word pair $w_1 w_2$, log-likelihood measures how likely it is that w_1 is dependent on w_2 . PMI measures how much information about the occurrence of w_2 is provided by the occurrence of w_1 , or how well w_1 predicts w_2 . For most tasks, the reduction in uncertainty measured by PMI does not correspond to collocation or co-occurrence as well as the dependency measured by log-likelihood.

Despite the argument advanced by Manning and Schütze, we decided to use pointwise mutual information, because the mix of similarity and other relations that occurs in the semantic verbal fluency task seems to set it apart from many other tasks where collocation might be used. Preliminary results suggest that PMI actually performs better in this case. It may be that for our task, the overestimating of low-frequency word pairs is beneficial, because it may highlight context words that are actually most strongly associated with a given animal. For example, kangaroo co-occurs with Australia, which is clearly a feature that one would expect it to have in common with other animals such as koala or wombat. However, kangaroo also co-occurs with a number of other words that don't seem particularly relevant. Because the frequency of kangaroo-Australia is low relative to the frequency of a number of other words that don't seem particularly relevant, PMI would do a better job of measuring the relatedness of a pair such as kangaroo-wombat compared to the log-likelihood statistic.

Our third measure was simply the log-likelihood statistic for the co-occurrence of two animals.

To get a rough idea of how well the vector space measures worked, we looked at the 50 pairs of animals with the highest cosines, using log-likelihood scores as the vector components. In looking at the top 50 based on the initial set of vectors, we observed that the vector space method did not seem to do a very good job at identifying semantically related animals. Boar and burro were rated as the most similar with a cosine of .557, which is not very intuitive, and dog and cat, which are clearly the most related based on observations of the test protocols, had a cosine of .208, although kitten and dog had a cosine of .550 and was ranked third. Kitten and cat had a cosine of only .201. To get a sense of what might be causing the overall poor results, we investigated two pairs, wombat-platypus and wombat-kangaroo. Intuition suggests that these pairs have a similar level of semantic relatedness, yet the vector approach yielded a score for wombat-platypus that was significantly higher than wombat-kangaroo. (.0557 for wombat-platypus vs. .225 for wombat-kangaroo). In investigating the words that each of these three animals co-occurred with, we noted that kangaroo was asso-

ciated with many more words than either wombat or platypus, which is why the cosine is lower for wombat-kangaroo than for wombat-platypus. Also the association of wombat with Australia was lost because the wombat occurred with an instance of Australia with a period at the end. It is clear that while the vector space method yielded plausible scores for many animal pairs, there were some significant issues. We experimented with a possible direction for improving the results by manually choosing certain words to remove from the vectors, such as the names of cities (assuming that they were associated with sports team names), and other words that did not seem likely to be good context words either because they seemed to be associated with other words senses or did not seem to be related to animals in ways that would be relevant for the task. Table 2 shows the top 50 pairs in each of 4 iterations. After the 1st iteration, we removed additional words in each subsequent iteration.

We did not attempt to formally evaluate the different sets of features, but in looking at the four iterations, it is clear that pruning the context words does have an impact, and it appears that in some cases, it is improving the scores. For example, cat and kitten show up in the 4th set (which is the most restricted,) but not in any of the others. By the fourth iteration, several pairs that seem related got added (e.g. lizard-snake and gerbil-hamster) and some that didnt seem very related got dropped (e.g. weasel-burro and parrot-squirrel). Boar-burro, which was at the top of the list in the first iteration, dropped off the list in the fourth iteration as well.

Based on these results, we assume that further reducing the dimensionality of the vectors would lead to even better results. We next attempted to find a more principled method of reducing the dimensionality, first investigating the use of Principal Components Analysis (PCA). PCA is a mathematical technique that reduces dimensionality by projecting a high-dimensional space onto a lower-dimensional space (see Smith [20] for a good overview of the technique). We performed PCA on the initial set of vectors, experimenting with a wide range of number of dimensions retained. Again lacking a good formal evaluation method, we used k-means clustering to cluster the animals into subcategories, before applying PCA and after performing PCA with various numbers of dimensions retained. Clustering at least gives an intuitive sense of how effective the method was. After the experiment, we concluded that PCA was not an effective approach to dimensionality reduction for this task. Table 3 shows some of the clusters that were produced. Note that after applying PCA, a number of small clusters and one very large cluster were formed. This result occurred, regardless of the number of dimensions retained and the number of clusters produced. While some of the clusters obtained after applying PCA make sense (e.g. *cat*, *dog*), many others do not (e.g. *frog*, *bear*). After examining the clusters, it does not appear that PCA is a viable approach for this task, although it might be worth correlating the vectors produced by PCA with an objective

standard, if one becomes available, to confirm this preliminary judgment.

As mentioned earlier, each measure does well with some animal pairs and not with others, and tends to do well on a different set of animal pairs. Several approaches could be taken to remedy this problem. One would be to refine the vector-based space measures by trying measures other than the cosine, and possibly interpolating between the log-likelihood and pointwise mutual information measures. In addition, incorporating the vector-based measures, the co-occurrence of animal pairs and a WordNet based measure into one overall measure might be effective, since the different measures seem to work well on different sets of animal pairs. The challenge here is that the measures use different units and have different ranges so there is no obvious way to combine them. In order to get a sense of whether combining the measures would improve performance, we scaled the co-occurrence values and then standardized the data by computing z-scores. The z-score expresses an observation in terms of the number of standard deviations away from the mean it is. Z-scores are appropriate when the distributions approximate the normal distribution. Some of these measures probably not normally distributed, but we proceeded with this assumption to determine whether the approach seemed feasible.

We scaled the co-occurrence values as follows:

Range	Scaled Value
0-10.83	0
10.84-99	1
100-199	2
200-299	3
300-399	4
400-499	5
500-599	6
600-699	7
700-799	8
800-899	9
>900	10

We then evaluated each of the measures in isolation as well as in the model

$score = \alpha * (\lambda * v_1 + (1 - \lambda) * v_2) + \beta * c - \gamma * t$ where
 v_1 = the cosine of angle between vectors using the log-likelihood statistic as the vector components
 v_2 = the cosine of angle between vectors using pointwise mutual information as the vector components
 c = the log-likelihood statistic for the co-occurrence of the two animals in a pair
 t = the WordNet-based measure
 λ = the interpolation factor, $0 \leq \lambda \leq 1$
 α = the weight of the interpolated cosines
 β = the weight of the co-occurrence measure
 γ = the weight of the WordNet-based measure

Lacking a gold standard, we evaluated the measures by

1st Iteration			2nd Iteration			3rd Iteration			4th Iteration		
boar	burro	0.565	wombat	platypus	0.573	wombat	platypus	0.573	wombat	platypus	0.700
wombat	platypus	0.557	boar	burro	0.564	boar	burro	0.564	orangutan	panda	0.646
kitten	dog	0.550	kitten	dog	0.546	lion	otter	0.552	vicuna	alpaca	0.616
lion	otter	0.533	lion	otter	0.532	kitten	dog	0.550	moth	wasp	0.596
orangutan	panda	0.522	orangutan	panda	0.519	orangutan	panda	0.523	wombat	kangaroo	0.483
vicuna	alpaca	0.510	vicuna	alpaca	0.510	vicuna	alpaca	0.510	panda	gorilla	0.480
mink	chinchilla	0.491	mink	chinchilla	0.492	mink	chinchilla	0.500	koala	panda	0.475
wolf	zebra	0.478	boar	weasel	0.462	boar	weasel	0.467	hippo	panda	0.460
raccoon	opossum	0.473	dingo	cat	0.430	stallion	otter	0.436	boar	mule	0.457
boar	weasel	0.454	stallion	otter	0.423	dingo	cat	0.431	leopard	cheetah	0.452
ferret	cat	0.443	raccoon	opossum	0.421	ferret	cat	0.424	ant	wasp	0.434
dingo	cat	0.419	ferret	cat	0.417	boar	aardvark	0.422	mink	chinchilla	0.423
stallion	otter	0.414	goat	lamb	0.408	raccoon	opossum	0.421	vole	rat	0.423
goat	lamb	0.405	bison	elk	0.398	goat	lamb	0.408	gerbil	rat	0.421
bison	elk	0.404	dingo	boar	0.389	bison	elk	0.399	goat	cow	0.418
bird	fish	0.394	bird	fish	0.376	dingo	boar	0.389	goat	lamb	0.417
dingo	boar	0.387	whale	otter	0.369	whale	otter	0.377	llama	vicuna	0.416
deer	bison	0.374	boar	aardvark	0.363	bird	fish	0.374	opossum	squirrel	0.409
whale	otter	0.372	gerbil	cat	0.357	lamb	chicken	0.360	orangutan	gorilla	0.408
boar	aardvark	0.368	parakeet	cat	0.354	deer	bison	0.359	raccoon	opossum	0.404
gerbil	cat	0.366	lamb	chicken	0.351	gerbil	cat	0.358	deer	bison	0.402
opossum	squirrel	0.358	deer	bison	0.349	parakeet	cat	0.357	lamb	chicken	0.398
chipmunk	opossum	0.357	deer	elk	0.343	parrot	squirrel	0.351	elk	mule	0.395
lamb	chicken	0.350	moth	wasp	0.336	koala	kangaroo	0.350	chipmunk	rabbit	0.392
deer	elk	0.349	stallion	lion	0.333	stallion	lion	0.346	kangaroo	platypus	0.380
parakeet	cat	0.345	burro	mule	0.328	deer	elk	0.341	ant	moth	0.377
doe	zebra	0.343	llama	vicuna	0.328	moth	wasp	0.336	sheep	cow	0.373
bird	parrot	0.341	dingo	burro	0.323	bird	parrot	0.335	raccoon	mule	0.372
bird	squirrel	0.340	whale	fish	0.323	burro	mule	0.329	gerbil	hamster	0.368
burro	mule	0.339	elk	mule	0.320	llama	vicuna	0.328	bison	elk	0.365
moth	wasp	0.337	weasel	burro	0.319	mule	elk	0.325	chipmunk	opossum	0.365
whale	fish	0.337	chipmunk	opossum	0.318	finch	parakeet	0.324	orangutan	koala	0.353
sheep	goat	0.328	finch	parakeet	0.317	dingo	burro	0.323	opossum	mule	0.350
deer	squirrel	0.325	puma	stallion	0.313	whale	fish	0.323	llama	gnu	0.347
elk	mule	0.324	sheep	goat	0.312	puma	stallion	0.322	dolphin	fish	0.343
stallion	lion	0.324	doe	zebra	0.306	weasel	burro	0.322	cat	kitten	0.342
llama	vicuna	0.319	otter	fish	0.306	chipmunk	opossum	0.320	hippo	orangutan	0.342
raccoon	muskrat	0.317	hare	opossum	0.303	cougar	jaguar	0.313	hare	opossum	0.329
dingo	burro	0.316	cougar	jaguar	0.300	gerbil	hamster	0.313	hog	robin	0.327
finch	parakeet	0.310	opossum	squirrel	0.298	sheep	goat	0.311	koala	kangaroo	0.326
otter	fish	0.310	raccoon	muskrat	0.298	hare	opossum	0.304	leopard	lion	0.326
weasel	burro	0.308	koala	kangaroo	0.297	otter	fish	0.304	hamster	rat	0.323
gerbil	rat	0.307	gerbil	rat	0.296	parrot	parakeet	0.301	colt	ram	0.322
raccoon	mule	0.303	bird	parrot	0.295	opossum	squirrel	0.300	panda	cheetah	0.322
deer	bird	0.302	deer	squirrel	0.294	deer	squirrel	0.298	raccoon	muskrat	0.322
puma	stallion	0.302	bird	squirrel	0.292	gerbil	rat	0.298	kitten	dog	0.312
ferret	dog	0.301	raccoon	mule	0.289	hippo	panda	0.298	koala	impala	0.311
bird	otter	0.300	gerbil	hamster	0.288	raccoon	muskrat	0.298	lizard	snake	0.310
parrot	squirrel	0.298	bird	otter	0.287	bird	squirrel	0.295	boar	opossum	0.306
sheep	bison	0.298	whale	lion	0.285	whale	lion	0.295	sheep	robin	0.306

Table 2. Top 50 pairs using four different sets of features

cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
initial set – 3000 dimensions	camel goat pig llama lamb gnu vicuna alpaca burro ostrich	weasel elephant frog newt wolf woodchuck panther	cougar colt steer stallion mare ram jaguar	beaver walrus mink dolphin seal bird whale chinchilla otter fish	bulldog hog sheep robin deer bison yak cow reindeer	gerbil chipmunk hamster rabbit human vole parrot finch parakeet chameleon	puma leopard hyena warthog rhinoceros lion cheetah tiger aardvark zebra	gopher ant moth wasp crow	dingo koala wombat kangaroo platypus	hippo lemur orangutan baboon mole panda groundhog gorilla	duck fawn polarbear sloth cat kitten dog	bull hedgehog ox bat buffalo fox doe	armadillo dinosaur lizard grizzlybear toad iguana ferret snake	jackal shrew impala sealion eland bear mouse	hare boar raccoon elk opossum badger squirrel skunk mule muskrat
after PCA, 50 dimensions retained	chicken bull deer bison elk	bat tiger panther	dolphin whale fish	cougar dinosaur ant moth snake	colt steer ram jaguar doe	rat frog bear	ocelot cat dog	pig rabbit human rat	beaver mink seal otter	camel goat lamb cow	hippo dingo puma gopher bulldog gerbil lemur armadillo leopard walrus orangutan hare boar weasel stallion jackal fawn hedgehog ox mare elephant chipmunk hamster koala raccoon hyena baboon llama shrew impala lizard newt wombat yak polarbear sealion vole mole warthog grizzlybear rhinoceros parrot opossum groundhog gnu finch badger parakeet chameleon woodchuck eland vicuna chinchilla sloth kangaroo squirrel toad iguana alpaca wasp platypus cheetah burro ferret skunk mule ostrich kitten crow aardvark zebra mouse ocelot muskrat reindeer	panda lion gorilla	duck bird chicken	buffalo fox wolf	hog sheep robin

Table 3. Results of k-means clustering before and after principal components analysis

Measure	Correlation
Cosine, log-likelihood	-0.154
Cosine, pointwise mutual information	-0.193
Unscaled log-likelihood statistic	-0.127
Scaled log-likelihood statistic	-0.246
WordNet	0.189

Table 4. Correlations of individual measures

No. of pairs	λ	α	β	γ	Correlation
136	1	0.2	4.8	2.3	-0.264
69	1	0.2	4.8	2.3	-0.197
69	.05	3.0	5.5	1.5	-0.208

Table 5. Correlations of combined measures

correlating them with the mean latency, making the assumption that there is a correlation between mean latency and semantic relatedness. A better method would be to obtain human ratings of relatedness, and use them to both test the assumption of correlation between mean latency and relatedness, as well as to evaluate the various measures of semantic relatedness. However, we did not have the resources to obtain human ratings. We computed correlations based on the 136 pairs of words that occurred in our time-aligned dataset. We only computed pointwise mutual information for a subset of 69 pairs, so any measure that includes pointwise mutual information was correlated with only the subset. Because they were correlated with different sets of word-pairs, measures including pointwise information are not directly comparable with measures that exclude. Despite this issue, the data yielded some interesting insights. Table 4 shows the correlation for individual measures. Table 5 shows the correlations for measures combined according to the model described above. Parameter values were estimated through trial and error, by varying each parameter until a maximum correlation was achieved, and then varying each additional parameter in the same way. Note that when λ is set to 1, the pointwise mutual information measure is excluded.

The correlation of the combined measure including all four measures cannot be directly compared to the correlation of the combined measure including three measures for the set of 136 pairs, but given that it performs better than the three-measure combination for the subset, it would be worth obtaining the PMI measures for the set of 136. We would expect that a four-measure combination for the set of 136 pairs would perform better than the three-measure combination and thus achieve a correlation greater than -0.264. Interestingly, scaled co-occurrence by itself performs nearly as well as the combination, although the combination does achieve higher correlation. None of the measures has particularly high correlation, but there are a number of issues with the available data, and there are ways to refine the measures used. As discussed earlier, there are numerous alternate WordNet measures and

measures of vector distance, some of which may perform better than the ones we used. A more effective method of reducing the dimensionality of the vectors would probably also increase the correlations. Finally, our data for calculating mean latency were quite sparse for many pairs of words, there was only one observation. More data are required in order to determine whether there is a meaningful correlation between mean latency and semantic relatedness.

4. DISCUSSION

Multiple theories of verbal fluency, and related approaches to scoring verbal fluency tests, have been put forth by researchers. Two key theories are proposed by Troyer [1] and Mayr and Kliegl [8]. The main difference between them is their differing views of the role of executive functions in semantic verbal fluency. Both approaches have developed measures that show statistically significant differences between healthy participants and patients with various brain disorders including dementia and ADHD. Our work shows that it is possible to develop automated approaches to scoring that support both of the major theories.

We believe that incorporating latency data and an objective measure of semantic relatedness, whether derived from ratings by human subjects, or by methods using large corpora and/or taxonomies, will enable new ways of scoring and interpreting the test can be developed that may prove useful in future research. For example, looking at latency appears to be useful in determining whether a given production represents a switch when using the Troyer scoring system, which may help to achieve greater accuracy and objectivity in counting switches.

Given that the validity of Troyer’s model of verbal fluency performance is in question, and the various issues that have been raised about the constructs of clustering and switching, constructing an index of semantic relatedness provides an alternative method of evaluating performance on the test that is consistent with at least one of the major alternative theories of verbal fluency, that of Mayr and Kliegl.

Reverberi et. al. [9] did use an index of semantic relatedness. They did not find statistically significant differences between controls and frontal patients on this measure, but did find a robust trend. This result is promising and supports the idea that semantic relatedness could be useful in looking at other illnesses, such as ADHD.

Our measures also enable new possibilities for scoring and interpreting the verbal fluency test. For example, if participants produce results that have similar semantic relatedness but different timing characteristics that could be indicative of disturbances in brain functioning. Incorporating timing also allows for more sophisticated analysis such as looking at patterns of when exemplars are produced. For example, two participants might produce the same number of words, but one could produce most of them in the first half of the test period,

whereas the other might produce them such that they more evenly distributed throughout the allotted time period.

5. FUTURE WORK

Our results are promising, but much work remains to be done.

5.1. HMM Tagger

Following are areas for further work:

1. Refine the model by incorporating additional data and extending to a 2nd-order model.
2. If additional hand-labeling will be done, develop written guidelines for labeling, including establishing a default tag for each animal, to ensure consistent results.
3. Investigate the use of more sophisticated smoothing methods for handling unseen data.
4. Determine whether the HMM tagger can detect the use of task-discrepant clustering with a better model. If not, develop an algorithm to do so.

5.2. Semantic Relatedness

There are a number of areas for further work in developing a semantic relatedness measure:

1. Improve the existing vector-space measures.
 - (a) Refine method of extracting sentential co-occurrence data from the corpus.
 - i. Enhance the Porter Stemmer to merge variants of words that the base version doesn't (e.g. Australia, Australian, Australia's).
 - ii. Expand the stop list to filter out additional words in order to reduce the dimensionality of the vectors.
 - iii. Enhance the algorithm used to extract data from the corpus to detect collocations such as "sea lion" or "grizzly bear".
 - (b) Principal Components Analysis does not appear to be a good approach to reducing dimensionality. Validate this preliminary conclusion and research other methods of reducing dimensionality. One possibility would be to remove one feature at a time, evaluate the remaining features using some fitness function, and drop the removed feature if the fitness function improves.
2. Test other measures of distance in vector space in addition the cosine of the angle between vectors.
3. Validate the use of z-scores to standardize data in order to combine measurements that are in different units.

4. Investigate incorporating measures based on syntactic relations in addition to sentential co-occurrence.
5. Obtain human judgments of semantic relatedness for at least a subset of the animal pairs observed in the initial dataset, so that the various measures of semantic relatedness can be evaluated against a gold standard.
6. Test additional WordNet based measures.

5.3. Time-based measures

Following are areas for further work relating to time-based measures:

1. Although we believe that time-based measures could be useful for classification, we did not have sufficient data to determine this, so this would be an important area for future work.
2. We used a latency threshold to determine whether a transition was a switch in ambiguous cases. Further work is needed to validate this approach and to develop a better method for determining the best value of the threshold.
3. Finally, we believe that there is a correlation between latency and semantic relatedness. It would be useful to obtain human judgments of relatedness of the set of animal pairs observed in a larger time-aligned dataset than was available to us, in order to test this hypothesis.

6. ACKNOWLEDGMENT

I wish to thank Professor Brian Roark for supervising me in this project and for providing extensive discussions and suggestions. Many of the ideas explored in the project were originally his, and I could not have completed the project without his support.

7. REFERENCES

- [1] Angela K. Troyer, Morris Moscovitch, and Gordon Winocur, "Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults," *Neuropsychology*, vol. 11, no. 1, pp. 138–146, 1997.
- [2] David A. Abwender, Jeffrey G. Swan, John T. Bowerman, and Sean W. Connolly, "Qualitative analysis of verbal fluency output: Review and comparison of several scoring methods," *Assessment*, vol. 8, no. 3, pp. 323–336, 2001.

- [3] Melissa O. Epker, Laura H. Lacritz, and C. Munro Cullum, "Comparative analysis of qualitative verbal fluency performance in normal elderly and demented populations," *Journal of Clinical and Experimental Neuropsychology*, vol. 21, no. 4, pp. 425–434, 1999.
- [4] Angela K. Troyer, Morris Moscovitch, Gordon Winocur, Larry Leach, and Morris Freedman, "Clustering and switching on verbal fluency tests in alzheimer's and parkinson's disease," *Journal of the International Neuropsychological Society*, vol. 4, no. 2, pp. 137–143, 1998.
- [5] Angela K. Troyer, "Normative data for clustering and switching on verbal fluency tasks," *Journal of Clinical and Experimental Neuropsychology*, vol. 22, no. 3, pp. 370–378, 2000.
- [6] Evrim Gocer March and Philippa Pattison, "Semantic verbal fluency in alzheimer's disease: Approaches beyond the traditional scoring system," *Journal of Clinical and Experimental Neuropsychology*, vol. 28, no. 4, pp. 549–566, 2006.
- [7] Ulrich Mayr, "On the dissociation between clustering and switching in verbal fluency: comment on troyer, moscovitch, winocur, alexander and stuss," *Neuropsychologia*, vol. 40, no. 5, pp. 562–566, 2002.
- [8] Ulrich Mayr and Reinhold Kliegel, "Complex semantic processing in old age: Does it stay or does it go?," *Psychology and Aging*, vol. 15, no. 1, pp. 29–43, 2000.
- [9] Carlo Reverberi, Marcella Laiacona, and Erminio Capitani, "Qualitative features of semantic fluency performance in mesial and lateral frontal patients," *Neuropsychologia*, vol. 44, no. 3, pp. 469–478, 2006.
- [10] Oliver Tucha, Lara Mecklinger, Rainer Laufkotter, Ivo Kaunzinger, Geraldine M. Paul, Helmfried E. Klein, and Klaus W. Lange, "Clustering and switching on verbal and figural fluency functions in adults with attention deficit hyperactivity disorder," *Cognitive Neuropsychiatry*, vol. 10, no. 3, pp. 231–248, 2005.
- [11] Daniel Jurafsky and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2000.
- [12] Alexander Budanitsky and Graeme Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.
- [13] Saif Mohammad and Graeme Hirst, "Distributional measures as proxies for semantic relatedness," 2005, submitted for publication.
- [14] Yuhua Li, Zuhair A. Bandar, and David McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on Knowledge Engineering and Data Engineering*, vol. 15, no. 4, pp. 871–882, 2003.
- [15] Julie Elizabeth Weeds, *Measures and Applications of Lexical Distributional Similarity*, Ph.D. thesis, University of Sussex, 2003.
- [16] Dongqiang Yang and David M. W. Powers, "Measuring semantic similarity in the taxonomy of wordnet," in *ACSC '05: Proceedings of the Twenty-eighth Australasian conference on Computer Science*. 2005, pp. 315–322, Australian Computer Society, Inc.
- [17] M.F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [18] Scott McDonald, "Exploring the validity of corpus-derived measures of semantic similarity," 1997, Presented at the 9th Annual CCH/HCRC Postgraduate Conference, University of Edinburgh, June 18-19, 1997.
- [19] Christopher D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.
- [20] Lindsay I. Smith, "A tutorial on principal components analysis," February 26 2002, Student Tutorial, University of Otago.