A System for Computerized Analysis of Verbal Fluency Tests

James O. Ryan

Institute for Health Informatics
Center for Cognitive Sciences
University of Minnesota Twin Cities

June 2013

Abstract

I present VF-Clust, a system for computerized analysis of psychological tests of *verbal fluency*, which are used in clinical settings to detect and assess neurological pathologies such as Alzheimer's disease. While the simple scoring of such tests may not be sensitive to underlying pathologies, more nuanced *clustering analyses* often are. Currently, clustering analyses on verbal fluency tests are conducted manually, by trained scorers, in a process that is labor-intensive and prone to human error and interrater variability. VF-Clust is a resource for automatically generating clustering analyses, on both semantic and phonemic verbal fluency test responses, by utilizing *latent semantic analysis* and computational methods for determining phonetic similarity. Results from a pilot study indicate that VF-Clust's automatically determined clustering measures are as useful as their manually determined equivalents. Additionally, in this study repetitions were more numerous in verbal fluency test responses from patients with dementia from Alzheimer's disease. In the case of phonemic verbal fluency, controls differed with patients with Alzheimer's diagnoses on the number of words repeated, but not on the number of words generated, which is the standard measure by which the test is scored. While repetitions are typically ignored in even the most in-depth analyses of verbal fluency test responses, these results indicate that future analyses should consider them. VF-Clust is freely available upon request.

Acknowledgments

I am deeply indebted to my advisor and friend, Dr. Serguei Pakhomov, whose mentorship has been absolutely invaluable not only in the writing of this thesis, but more broadly throughout three crucial formative years in my yet nascent academic and professional journey. Without happenstance presenting me the opportunity to work with him, and then without his teacherly disposition and strong intellectual vigor, I surely would be in some place far less befitting any goals or interests that I have ever had. It has been an honor, a privilege, and a delight to have worked with you, Serguei!

I would also like to thank another mentor, Dr. Genevieve Melton-Meaux, who was, at a time when I had no professional or continuing academic plans whatsoever, pivotal in my application to and acceptance into this program. Without her advocacy, I would likewise be on some other less fortunate path. Thank you, Genevieve!

Further, I would like to express warm gratitude to the additional members of my master's committee, Drs. Terrance Adam, Paul Johnson, and Brian Reese.

Lastly, I have several colleagues who have in various ways contributed to the present work: Mara Anderson Searls, Robert Bill, Aaron Free, Anja Laske, Kyle Marek-Spartz, and Hannah Sande. Thanks, guys!

Contents

1 Introduction

Alzheimer's disease (AD) is the most common cause of dementia, with an estimated prevalence of 30 million people worldwide (Holtzman et al., 2011). This number is expected to quadruple by 2050, at which time 1 in 85 people worldwide would be living with the disease (Holtzman et al., 2011; Roses et al., 2009). It has been estimated that the global disease burden could be decreased by over 9 million cases if disease onset were delayed by just one year (Brookmeyer et al., 2007). Reliable early detection of Alzheimer's, along with early intervention, could help contribute to delaying AD onset in individuals prone to the disease. Thankfully, there are objective measures that have shown utility in detecting early signs of impending dementia.

Neuropsychological tests of phonemic verbal fluency (PVF) and semantic verbal fluency (SVF) are commonly used as part of larger test batteries to study and assess cognitive impairment from AD, as well as from other neurological conditions, such as Parkinson's and Huntington's diseases and traumatic brain injury (Randolph et al., 1993; Butters et al., 1986; Henry et al., 2004). On these tests, the subject is asked to name as many words beginning with a specified letter (for PVF) or belonging to a specified category (usually *animals*; for SVF) as he or she can in one minute (Benton, 1968; Newcomb, 1969).

The standard measure by which these tests are scored is the total number of satisfactory words produced, which tends to be less in individuals with these conditions. The SVF test, in particular, has been shown to accurately discriminate between persons with AD dementia or mild AD dementia and demographically matched controls (e.g., Monsch et al., 1992; Troyer et al., 1998b). However, prior studies have found that impairment from these conditions also affects clustering and switching behavior on these tests (Troyer et al., 1998a,b; Raskin et al., 1992; Ho et al., 2002). *Clustering* refers to the contiguous grouping of semantically related or phonetically similar words in a test response, and *switching* denotes transitioning from one cluster to the next.

While phonetic and semantic clustering analyses are useful for early detection of cognitive impairment from AD, they must be conducted manually. Manual approaches to these analyses are laborious and necessarily rely on contrivances in defining clusters, due to the infeasibility of arguably more grounded procedures. My objective in the present work was to develop and pilot-test a system constituting an automated, computerized approach to this issue that would afford efficiency and scalability, along with more versatile methods for determining phonetic similarity and semantic relatedness.

VF-Clust is a piece of software that generates clustering analyses for both PVF and SVF test responses. Its phonetic clustering analysis (PCA) module uses phonetic representations for words to determine cluster spans, while the semantic clustering analysis (SCA) module utilizes semantic relatedness scores generated by latent semantic analysis, a computational method for determining distance in meaning between words.

I first discuss clustering analysis methodologies, and subsequently findings on the effect of Alzheimer's disease on phonetic and semantic clustering behavior on tests of verbal fluency. Next, the VF-Clust system architecture is outlined, with specific treatment given to both its PCA and SCA modules. Here, I explain the edit distance metric for determining the similarity of two strings − which is used by VF-Clust's PCA module to determine phonetic distance between words − and latent semantic analysis. Finally, I discuss the results of a pilot study, using data from persons with Alzheimer's disease and mild cognitive impairment, which demonstrate the utility of VF-Clust in detecting subtle signs of brain impairment due to dementia.

2 Background

It is well established that persons with Alzheimer's disease generate fewer words on both PVF and SVF tests (e.g., Chertkow & Bub, 1990; Martin & Fedio, 1983). Impairment on these tasks, particularly SVF, seems to be related to deficits in semantic memory, an early characteristic of AD (e.g., Hodges & Patterson, 1995; Martin & Fedio, 1983). This would appear to be supported by evidence that patients with AD perform worse on SVF tests than on PVF tests (e.g., Pasquier et al., 1995; Rosser & Hodges, 1994).

However, there is wide belief that performance on these tasks is multifactorial, as evidenced by several studies finding that fluency performance involves multiple brain regions (see discussion in Troyer et al., 1997). An indication from these findings is that closer examination should be given to the underlying cognitive components that contribute to performance on verbal fluency tests, rather than to just the total number of words generated. To this end, several investigations into clustering and switching on these tests have been undertaken.

A *cluster* is a group of contiguous words that are deemed by some metric to be either phonetically similar, for phonetic clusters, or semantically related, for semantic clusters. That is, each word in a cluster must be similar or related to each other word in the cluster, whether or not the two words are directly adjacent in the response. Lone words that do not belong to

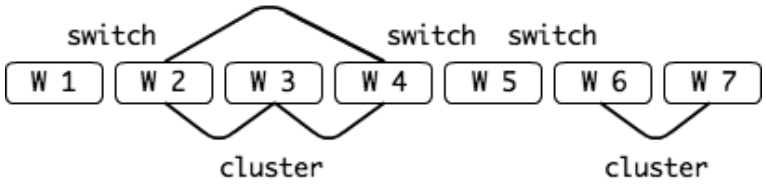

Figure 1: A series of words featuring two clusters and three switches, with connective lines indicating similarity or relatedness. `W1` and `W5` are singletons.

any cluster constitute *singleton* clusters. Transitions between clusters and singletons, or between clusters and other clusters or singletons and other singletons, are called *switches*. FIGURE 1 illustrates these concepts.

Clustering analysis critically relies on a methodology for determining whether any two words are similar or related. Examples of these are discussed in the next section.

2.1 Clustering analysis methodologies

Raskin et al. (1992) define phonetic clusters as comprising "successive words with the same second phoneme (e.g. fork, form) or two successive words which rhymed (e.g. fake, flake)" (96). It is unclear whether by 'phoneme' they actually mean orthographic letter, or if they indeed did compare phonemes. Because they conducted phonetic clustering analysis on not just PVF tests, but also SVF tests, they relax their first criterion for the latter task so that two words are considered phonetically similar if they share just an initial phoneme.

Troyer et al. (1997), whose method has become standard, likewise include in their similarity metric rhymes and words beginning with the same two letters. In the latter case, they explicitly specify using orthographic letters, presumably without special consideration for cases in which word spelling and pronunciation diverge. This is likely due to the impractical time burden of generating phonetic transcriptions for words prior to analysis, or alternatively the mental burden of considering words phonetically during analysis. Additionally, they allow for words that differ by only a vowel sound, such as *sat* and *seat*, and homophones, like *some* and *sum*.

For semantic clusters, Raskin et al. (1992) allow on PVF responses words that share "a semantic category (e.g. apple, apricot)" (96) or are inflectional variants of one another, such as *sing* and *sang*. They unfortunately do not

provide a list of semantic categories, if any, that were used during analysis. However, they report interrater reliability of 0.9 on SCA for PVF tests, so presumably they *did* rely on a reference list of categories, rather than subjective ones in the minds of individual raters. For SCA on SVF tests, the authors defined clusters as consisting of "successive words from the same category (e.g. lion, tiger)" (ibid.). Here, they presumably mean animal subcategories, as the category given to subjects for SVF was *animals*. Again, they do not report on what exactly their list of these, if there was one, may have included.[1]

Troyer et al. (1997) also used animal subcategories, which they specify, in defining what constitutes a semantic cluster. Their 22 subcategories include *African*, *Australian*, *Arctic*, *North American*, *Farm*, *Water*, *Beasts of burden*, *Animals used for fur*, *Pets*, *Birds*, *Bovine animals*, *Canine animals*, *Deer*, *Feline animals*, *Fish*, *Insects*, *Insectivores*, *Primates*, *Rabbits*, *Reptiles/Amphibians*, *Rodents*, *Weasels*.

For PVF and SVF, Troyer et al. (1997) calculate two clustering measures each: *mean cluster size* and *number of switches*. Each cluster is counted for size beginning with the second word in the cluster, so that a singleton is given size 0, a cluster of two words size 1, and so forth. Errors and repetitions are included in these counts, and mean cluster size is simply calculated as the mean of these size counts. The number of switches is equivalent to the number of transitions between clusters, in this case including singleton clusters.

In the following section, I briefly outline results from two studies employing verbal fluency clustering analysis. For both studies, the methodology of Troyer et al. (1997) was used.

2.2 Prior findings

Several studies have investigated verbal fluency performance, as well as clustering and switching behaviors, in patients with AD dementia. While many of these have found patients with AD to generate fewer words on SVF (Binetti et al., 1998; Gomez & White, 2006; March & Pattison, 2006; Randolph et al., 1993; Raoux et al., 2008; Tröster et al., 1998; Troyer et al., 1998b), those that looked at PVF also found fewer words generated on that test (Gomez & White, 2006; Tröster et al., 1998; Troyer et al., 1998b). Additionally, patients with mild cognitive impairment have been found to generate

[1]The authors do, however, say that during SVF administration "categories were provided as cues (i.e. jungle animals, pets, farm animals, ocean animals)", though it is not clear if these are the subcategories they relied on for clustering analysis (96).

fewer words on SVF (Price et al., 2012). There has also been consistent reporting of AD patients generating smaller phonetic clusters on PVF (Gomez & White, 2006[2]; Tröster et al., 1998; Troyer et al., 1998b) and semantic clusters on SVF (Gomez & White, 2006; March & Pattison, 2006; Raoux et al., 2008; Tröster et al., 1998; Troyer et al., 1998b).

As for switching behavior on PVF and SVF, there has not been such consensus. While more than one study has found AD patients to switch less on PVF (Gomez & White, 2006; Tröster et al., 1998), Troyer et al. (1998b) found no such difference. On SVF, the majority of studies have reported less switching by AD patients (Gomez & White, 2006; Raoux et al., 2008; Tröster et al., 1998; Troyer et al., 1998b), though March & Pattison (2006) did not observe any difference. The latter study, however, included only patients with "mild to moderate AD" (549), so perhaps this explains the lack of impoverished switching behavior. Price et al. (2012) likewise found no difference in switching on SVF between mild cognitive impairment (MCI) patients and controls.

Troyer et al. (1998b) attribute the smaller cluster sizes generally observed on SVF among patients with AD to impoverished semantic memory, a characteristic of that population. They, however, were the study that did not find less switching on PVF among AD subjects, but they too studied patients with mild AD.

3 VF-Clust

VF-Clust is written in Python, and features both a phonetic and semantic clustering analysis module. Currently, it supports command-line interaction on Mac, Linux, and Windows operating systems. The following is abridged output for an example PVF response `fun fort friend fry fret fetch flip` (see appendix A for examples of full system output):

```
Number of permissible words:  7
Number of phonetic clusters:  1
Number of phonetic cluster switches:  3
```

VF-Clust's PCA module employs two methods for determining phonetic similarity, one of which is an edit distance metric based on a classic algorithm for measuring string similarity. For the SCA module, semantic relatedness is determined by latent semantic analysis. The modules, and the computational techniques they use, are described in the following sections.

[2]Though in this study no difference was found for PVF with specified letter *s*.

3.1 PCA module

VF-CLUST's PCA module generates values along several clustering measures by automatically identifying phonetic clusters through the use of two similarity metrics: *edit distance* and a *common-biphone check*. Both of these work on phonetic word representations, which come from a modified version of the CMU Pronouncing Dictionary (CMUDICT). If a word is not found in the dictionary, a phonetic representation is automatically generated for it, using a decision tree-based algorithm trained on CMUDICT. All of these aspects of the module, as well as its architecture and workflow, are described in more detail below.

3.1.1 Phonetic word representations

The PCA module uses phonetic word representations found in a modified version of CMUDICT, a pronunciation dictionary developed for speech recognition and synthesis applications at the Carnegie Mellon University (Weide, 1998). CMUDICT contains phonetic transcriptions, using a phone set based on ARPABET (Rabiner & Juang, 1993), for North American English word pronunciations. In particular, the latest version, *cmudict.0.7a*, which contains 133,746 entries, is used.

Each CMUDICT entry is a set of plaintext phone symbols, each separated by whitespace. Vowels that carry lexical stress have numerical indicators appended to their symbols, according to word pronunciation. Words with variant pronunciations may have multiple pronunciation entries in the dictionary. As an example, the entry for the word *phonetic* is `F AH0 N EH1 T IH0 K`. Because VF-CLUST's phonetic similarity metrics, described in the next section, require compact phonetic word representations, the system's CMUDICT was modified as follows.

From the full set of entries in CMUDICT, I removed alternative pronunciations for each word, leaving a single phonetic representation for each heteronymous set.[3] Additionally, all vowel symbols were stripped of numeric stress markings (e.g., `AH1` $\rightarrow$ `AH`), and all multicharacter phone symbols were converted to arbitrary single-character symbols, in lowercase to distinguish these symbols from any original single-character phone symbol (e.g., `AH` $\rightarrow$ `c`). Finally, all whitespace between phone symbols was removed, yielding compact phonetic-representation strings suitable for computing VF-CLUST's similarity metrics.

[3]For instance, the entry for *does* (n., pl., 'female deer') was removed in favor of that of its more prominent heteronym *does* (v., pres. of *do*).

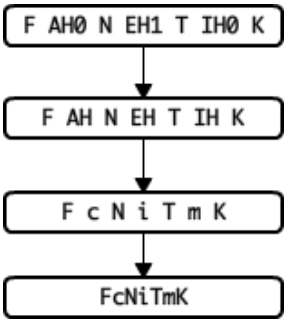

Figure 2: Modification of the default CMUDICT entry for *phonetic* into a compact phonetic representation.

The whole modification process is illustrated in FIGURE 2, which shows the steps by which the above-given default CMUDICT pronunciation entry for the word *phonetic*, `F AH0 N EH1 T IH0 K`, is rendered `FcNiTmK`.

3.1.2 Similarity metrics

As stated above, the PCA module uses two methods for determining phonetic similarity: edit distance and a common-biphone check. Each of these gives its own measure of similarity for a pair of phonetic representations, which I respectively call a *phonetic-similarity score* (PSS) and a *common-biphone score* (CBS).

The module's edit distance method, which gives a PSS, is based on the Levenshtein distance string metric (Levenshtein, 1966). Introduced by Vladimir Levenshtein in the mid-1960s, this metric quantifies the orthographic distance between two text strings by counting the minimum number of edit operations required to turn one of the strings into the other. The edit operations that can be used on the string being transformed are: INSERT, in which a character is inserted; DELETE, in which a character is deleted; and REPLACE, in which a character is replaced with another character.

As an example, the Levenshtein distance between the strings `phonetic` and `fanatic` is 4, because it takes four edit operations to transform the former into the latter: first, DELETE is used on `p`; then, REPLACE is used to change `h` to `f`; next, REPLACE is used again to switch out `o` for `a`; finally, after `n` is skipped, `e` is changed to `a` via REPLACE, and thus `phonetic` has become `fanatic`.

To compute the PSS for two words, the PCA module first computes the

Levenshtein distance between their compact phonetic-representation strings, described in the previous section. (Indeed, CMUDICT entries were modified *because* of how Levenshtein distance works.) This score is then normalized to the length of the longer string, and finally that value is subtracted from 1, giving the PSS. PSS values range from 0 to 1, with higher scores indicating greater phonetic similarity.

To illustrate PSS calculation, I will again use *phonetic* and *fanatic* as an example. These words are much more similar phonetically than they are orthographically. Indeed, their compact phonetic representations, respectively `FcNiTmK` and `FcNbTmK`, differ by only one character. Given this, the Levenshtein distance between the two representations is 1. Both representations have the same length, 7, so their Levenshtein distance is divided by that, which yields 0.143. Lastly, this value is subtracted from 1, giving a PSS score of 0.857, appropriately indicating very high phonetic similarity.

While PSSs are continuous, CBSs are categorical and binary. Specifically, a CBS of 1 is given for two words whose phonetic representations have a common initial or final biphone, and 0 for two strings that have neither in common. In the case of *phonetic* and *fanatic*, a CBS of 1 would be given, as their respective phonetic representations, `FcNiTmK` and `FcNbTmK`, have the same initial two characters (and also the same final two characters, though this does not further change the score).

3.1.3 Similarity classification

To identify phonetic clusters, the PCA module needs to classify contiguous word pairs as either phonetically similar or dissimilar, with pairs of similar words forming clusters. Using the common-biphone method, two words are considered phonetically similar simply if their CBS is 1. When using the edit distance method, however, the PCA module requires a PSS threshold for categorizing a word pair as similar or dissimilar. This is because PSSs are continuous, with values that in and of themselves do not lend to classification like binary CBS values do.

To remedy this, phonetic similarity thresholds were determined empirically for each letter in the alphabet. First, for each letter, I randomly sampled 1000 modified CMUDICT entries for words starting with that letter.[4] Then, I computed PSS scores for each sample's 499,500 pairwise combinations. Finally, the threshold for each letter was set as the value separating the upper quintile of the pairwise PSS scores for that letter's random sample.

[4]For *q*, *x*, *y*, and *z*, there were less than 1000 entries to sample from, in which case every entry was used.

Interestingly, some thresholds differed quite drastically from others, suggesting greater phonological variation among words beginning with certain letters relative to other letters. For instance, the threshold for q was found at 0.43, while both a and o had thresholds of 0.22.[5] This indicates less phonological variation across words beginning with q. This is not all that surprising, though, given that almost all English words beginning with q have `[kw]` for an initial biphone. Generally, however, there was not incredible variation in letter-specific thresholds, as 22 letters were given thresholds between 0.29 and 0.33.

Because the similarity method of Troyer et al. (1997) relies on orthography, words that intuitively should not be classified as phonetically similar sometimes will be. For instance, take the words *reconcile* and *reach*. Since these two words both have `re` as their first two letters, they would be manually determined to form a cluster. However, they really do not sound alike, as they only share an initial sound, which will be the case with nearly all words generated on any PVF response. With the common-biphone method, these words would not be deemed similar, as their phonetic representations show different initial biphones (`Ri` and `Rn`, respectively), as well as different final biphones (`ng` and `fL`, respectively). And by the edit-distance method, these words would receive a PSS of 0.125, far below the phonetic similarity threshold for r. These types of errors can also occur in the opposite direction, where words that are intuitively similar are not deemed to be so due to an orthographic quirk.

3.1.4 Module workflow

FIGURE 3 shows the high-level architecture and workflow of VF-CLUST's PCA module. The module accepts comma-separated transcriptions of PVF responses, and also expects the letter given for the test. As a pre-processing step, any words that do not begin with that letter are removed from the response, though these will be considered for count measures, described in the next section.

After pre-processing, all words are converted to compact phonetic representations, as described in SECTION 3.1.1. This is done by dictionary lookup into the modified CMUDICT. If a word is not found in the dictionary, a phonetic representation is automatically generated for it with a decision tree-based grapheme-to-phoneme algorithm, based on that of Pagel et al. (1998), that was trained on CMUDICT.[6] To be precise, the algorithm

[5] However, for q there were only 386 CMUDICT entries to use.

[6] The autopronouncer used here was developed by Kyle Marek-Spartz and Serguei

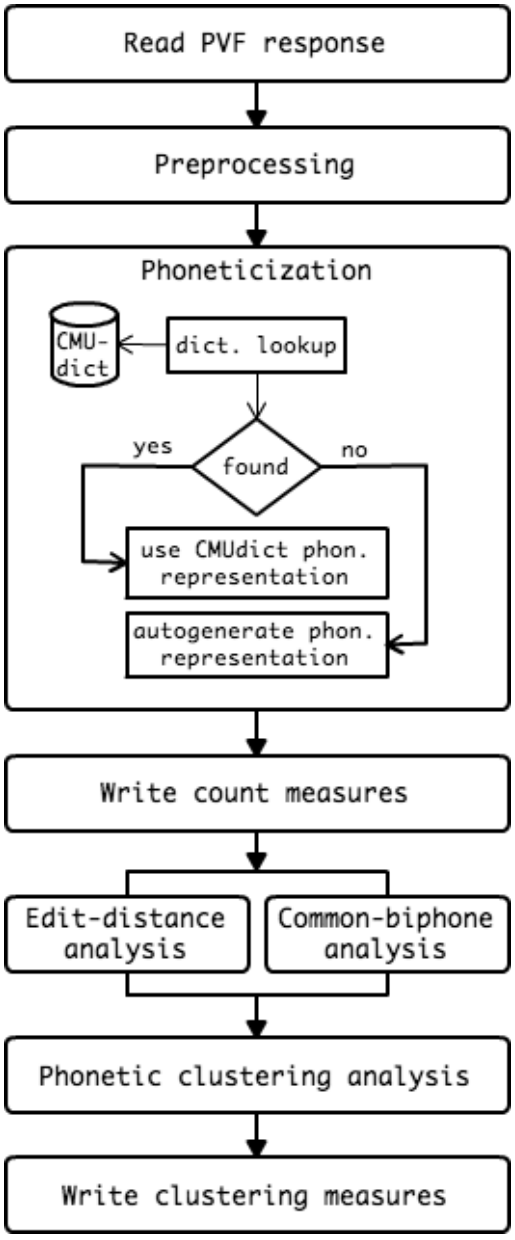

Figure 3: High-level PCA module architecture and workflow.

first generates a phonetic representation for the word in the format of a default CMUDICT entry, which is then compacted using the same method by which all native CMUDICT entries were modified.

Next, the module's count measures are totaled, and after this, PSSs and CBSs are computed sequentially for each pair of contiguous phonetic representations, and for all pairwise phonetic representations. Finally, these scores are used to identify clusters and generate values for VF-CLUST's phonetic clustering measures, which are described in the next section.

3.1.5 PVF measures

The PCA module generates values along several clustering and count measures. For count measures, it produces counts for total number of words (PHN-TOTAL), number of permissable words (PHN-WORDS), number of repetitions (PHN-REPS), and number of off-topic words (PHN-ASIDES). PHN-WORDS counts all words in a PVF response that begin with the letter specified for the test, while PHN-ASIDES is a count of all remaining words. PHN-REPS is a count of repeated permissible words only, and PHN-TOTAL is equivalent to PHN-WORDS + PHN-ASIDES.

The module's clustering measures are computed using both the edit-distance and common-biphone method, with two ways of defining a cluster. As explained in SECTION 2, a cluster customarily is composed of contiguous words that are each similar to or related to one another. In addition to clusters of this type, VF-CLUST also uses a more relaxed approach that defines clusters as *chains*.

Whereas each word in a cluster must be similar to each other word in the cluster, only directly adjacent words must be similar to one another in a chain. More formally, a chain comprises a sequence for which each word is similar to that of the word immediately prior to it in the chain (unless it is chain-initial) and the word immediately subsequent to it (unless it is chain-final). Chains based on the edit-distance method are called *phonetic chains*, and chains based on the common-biphone method are called *common-biphone chains*; both of these are illustrated in FIGURE 4. Similarly, clusters found via edit distance are called *phonetic clusters*, and those found with a common-biphone check are called *common-biphone clusters*.

Because there are two similarity classification methods that each use two cluster definitions, VF-CLUST generates values for a large number of phonetic clustering measures. For both the edit-distance and common-

Pakhomov.

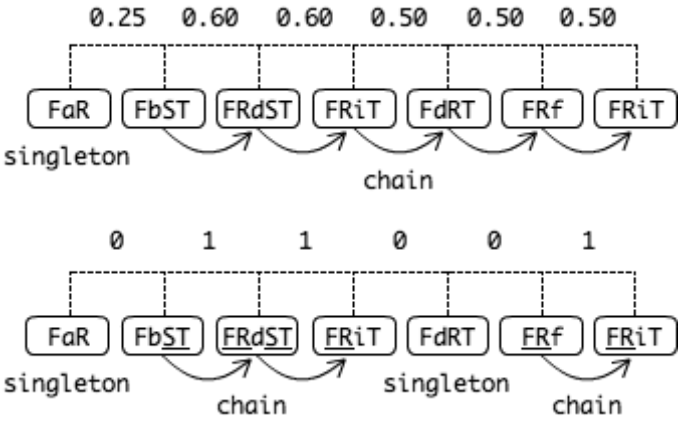

Figure 4: Phonetic chain and common-biphone chain (below) for an example PVF response.

biphone method, the following measures are computed: *number of clusters* (PHN-CLUSTERS, BPH-CLUSTERS); *number of clusters, excluding singletons* (PHN-CLUSTERS-NS, BPH-CLUSTERS-NS); *number of chains* (PHN-CHAINS, BPH-CHAINS); *number of chains, excluding singletons* (PHN-CHAINS-NS, BPH-CHAINS-NS); *mean cluster size* (PHN-MEAN-CLUSTER, BPH-MEAN-CLUSTER); *mean cluster size, excluding singletons*[7] (PHN-MEAN-CLUSTER-NS, BPH-MEAN-CLUSTER-NS); *mean chain length* (PHN-MEAN-CHAIN, BPH-MEAN-CHAIN); *mean chain length, excluding singletons* (PHN-MEAN-CHAIN-NS, BPH-MEAN-CHAIN-NS); *maximum cluster size*, i.e., size of the largest cluster (PHN-MAX-CLUSTER, BPH-MAX-CLUSTER); *maximum chain length*, i.e., length of the largest chain (PHN-MAX-CHAIN, BPH-MAX-CHAIN); *number of cluster switches*[8] (PHN-CL-SWITCHES, BPH-CL-SWITCHES); and *number of chain switches* (PHN-CH-SWITCHES, BPH-CH-SWITCHES).

Additionally, both similarity classification methods produce a mean pairwise similarity score, respectively called a *mean pairwise phonetic similarity score* (PHN-PAIRWISE) and a *mean pairwise common-biphone score* (BPH-PAIRWISE). The former is calculated as the mean of all pairwise PSSs for

[7]This measure is roughly an automatic equivalent to the mean cluster size measure of Troyer et al. (1997). However, in theirs, a cluster's size is calculated as the number of its constituents minus 1, whereas here merely the number of constituents is counted.

[8]NOTE: The number of switches for an attempt will always be equal to the number of corresponding clusters minus 1. VF-CLUST includes both as two different ways of conceptualizing the same phenomenon.

Table 1: VF-CLUST's PVF measures

BPH-CH-SWITCHES	maximum common-biphone chain switches
BPH-CHAINS	number of common-biphone chains
BPH-CHAINS-NS	number of common-biphone chains, not including singletons
BPH-CL-SWITCHES	maximum common-biphone cluster switches
BPH-CLUSTERS	number of common-biphone clusters
BPH-CLUSTERS-NS	number of common-biphone clusters, not including singletons
BPH-MAX-CHAIN	maximum common-biphone chain length
BPH-MAX-CLUSTER	maximum common-biphone cluster size
BPH-MEAN-CHAIN	mean common-biphone chain size
BPH-MEAN-CHAIN-NS	mean common-biphone chain size, not including singletons
BPH-MEAN-CLUSTER	mean common-biphone cluster size
BPH-MEAN-CLUSTER-NS	mean common-biphone cluster size, not including singletons
BPH-PAIRWISE	mean pairwise CBS
PHN-ASIDES	number of off-topic words
PHN-CH-SWITCHES	number of phonetic chain switches
PHN-CL-SWITCHES	number of phonetic cluster switches
PHN-CHAINS	number of phonetic chains
PHN-CHAINS-NS	number of phonetic chains, not including singletons
PHN-CLUSTERS	number of phonetic clusters
PHN-CLUSTERS-NS	number of phonetic clusters, not including singletons
PHN-MAX-CHAIN	maximum phonetic chain length
PHN-MAX-CLUSTER	maximum phonetic cluster size
PHN-MEAN-CHAIN	mean phonetic chain size
PHN-MEAN-CHAIN-NS	mean phonetic chain size, not including singletons
PHN-MEAN-CLUSTER	mean phonetic cluster size
PHN-MEAN-CLUSTER-NS	mean phonetic cluster size, not including singletons
PHN-PAIRWISE	mean pairwise PSS
PHN-REPS	number of repeated permissible words
PHN-TOTAL	total number of words
PHN-WORDS	number of permissible words

the words in a given PVF response, and the latter likewise, except using all pairwise CBSs.

TABLE 1 shows abbreviations for each of the VF-CLUST's PVF measures, along with corresponding definitions.

3.2 SCA module

VF-CLUST's SCA module produces values along its own clustering measures by identifying semantic clusters via relatedness scores generated by *latent semantic analysis*.

3.2.1 Latent semantic analysis

Latent semantic analysis (LSA) is a computational technique for representing the meanings of words according to their contextual distributions in a large corpus of text (Landauer & Dumais, 1997). LSA is typically used to measure relatedness between words or documents. It was originally developed as a method for automatic indexing and retrieval of documents in large databases (Deerwester et al., 1990), but has since been used in a variety of applications, such as automatic essay grading (Miller, 2003), intelligent tutor systems (Graesser et al., 2000; Wiemer-Hastings et al., 2004), and spell checking (Jones & Martin, 1997). LSA has been found to learn new words at a rate similar to that of schoolchildren, and to do as well on synonym knowledge tests as competent second-language speakers of English (Landauer & Dumais, 1997).

The method is built on the assumption that words close in meaning will occur in similar contexts. From a large collection of text, called a *corpus*, LSA collects each occurring word,[9] and each context in which those words occur. The words are referred to as *terms*, and their contexts as *documents*. What sort of context is treated as a document depends on the application, but documents should be coherent units of related information, e.g., encyclopedia entries or paragraphs. LSA then constructs a co-ocurrence matrix of the terms and documents, in which each row represents an individual term and each column an individual document. The cells of this term-document matrix are populated with frequency counts, such that each cell will have a count of the number of times the term of the corresponding row occurred in the document of the corresponding column. Since this matrix representation only takes into account term-document co-occurrence, word order in

[9]Stop words, e.g., *if* and *or*, are normally omitted. See SECTION 3.2.2 for more information.

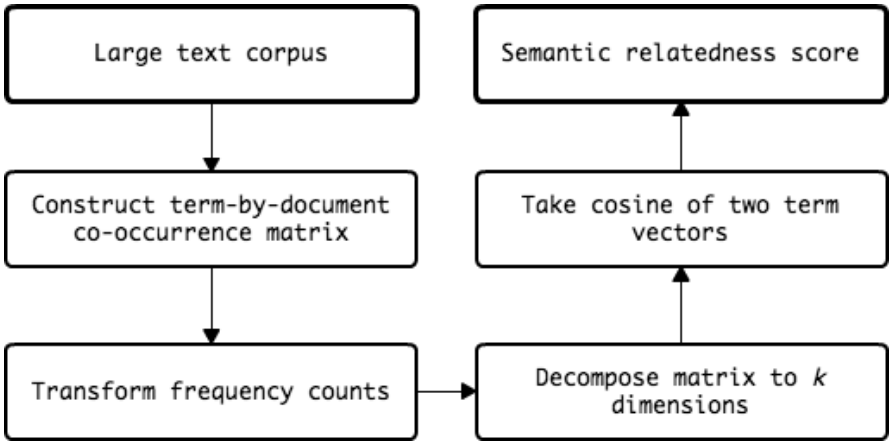

Figure 5: Steps in latent semantic analysis.

the documents is ignored − i.e., each document is represented as a 'bag of words'.

Rather than work with the raw term frequencies, the cell counts in the term-document matrix are often transformed. A common first step in this regard is to transform each cell count to the log of that cell count. As an effect of this log transformation, there will be a greater association between two words that both occur in two different contexts than if they had each appeared twice in the same context (Landauer & Dumais, 1997). Next, the cell entries generally are divided by the entropy for the event type. This inverse entropy measure captures how meaningful co-ocurrence with a particular word is. Consider that a word that occurs in many contexts transmits less information about words with which it co-occurs. This is because the more contexts a word occurs in, the less informative its occurrence in any single context becomes, and thus the less informative its co-occurrence with any word becomes.[10] The more documents a term occurs in, the larger its entropy will be, and so the inverse entropy transformation utilizes this effect.

At this point in the LSA derivation, each row in the matrix is a vector corresponding to a term, with values for that term's relatedness to each document. In a matrix of n terms by m documents, each row is an m-dimensional term vector. Given the number of documents in a typical corpus, these are likely to be *very* high-dimensional vectors.[11] LSA is differentiated from

[10]Extreme cases of this principal are frequently occurring functional words, like *the*.
[11]LSA is often trained on corpora comprising more than one million documents.

other vector space models by its dimensionality reduction, the result of a reduced-rank singular value decomposition (SVD) performed on the matrix (Dumais, 2005).

SVD is the general method for decomposition of a matrix into principle components. It is invoked with a parameter k, which specifies the desired number of dimensions. It is crucial to specify an appropriate number of dimensions for the SVD; typically, around 300 are chosen. Once the n x m matrix is submitted to SVD, the k dimensions with the largest singular values[12] are retained, with the remainder being set to 0. This causes the m-dimensional term vectors to become k-dimensional vectors in the space derived by the SVD.

It is not just a term's own distribution across documents that determines the values for its k-dimensional vector. Rather, the SVD uses all linear relations at its disposal to generate term vectors that best predict exactly which documents a term occurs in. Thus, a change in *any* cell of the original matrix usually will change *every* coefficient of *every* word vector in the reduced matrix (Landauer & Dumais, 1997).

Now, LSA has term vectors for each word that point outward in the semantic space derived by the SVD. With these, it is straightforward to compute relatedness between words. This is typically done by calculating the cosine between the two vectors for each term pair. The relatedness scores that LSA generates range from -1, where two words are not related at all, to 1, where two words have identical trajectories in LSA space.[13] FIGURE 5 recaps the intermediate procedures by which latent semantic analysis derives semantic relatedness scores for words in a corpus.

3.2.2 Constructing a corpus

Before latent semantic analysis could be used to generate semantic relatedness scores for word pairs in SVF responses, a corpus on which to train LSA had to be assembled. Since every word in an LSA corpus will usually affect the term vectors for every other word in the corpus, it is crucial not to include extraneous documents that may introduce noise into the model. In order to prevent such from happening, I constructed a corpus that would only contain documents pertinent to the SVF task domain. The category most commonly specified for SVF administration is *animals*, so I chose that

[12] I.e., the dimensions that capture the greatest variance in the original matrix. Thus, the SVD approximation is the best-possible k-dimensional representation of the original matrix.

[13] NOTE: Only a word and itself should have a relatedness score of 1.

for the domain to which documents for the corpus would pertain. Because it features an expansive array of articles written about animals, and moreover because its content is freely and easily available, I used Wikipedia to construct the corpus for LSA.

First, I compiled a list of 8903 animal common names, using several informal sources found on the web. This list included animals as ordinary as *cat* and *dog*, rare ones such as *acadian flycatcher* and *temple pit viper*, and everything between, including dinosaurs and insects.

From here, I used a Python script[14] to extract the text of any Wikipedia article that was returned by accessing each URL `http://en.wikipedia.org/wiki/[animalname]`, in which `[animal name]` was replaced with one of the 8903 names in my compilation. This process yielded texts from 5105 unique articles, of which 122 were excluded due to pertaining to a concept different from the animal with which they shared a name − e.g., *tailor*, *javelin*, *rifleman*.

Next, all non-ASCII characters were converted to their ASCII approximations, and each of the 4983 articles was stripped of any punctuation, symbols[15], or Wikipedia artifacts.

After the articles were cleaned, the final word of each instance of all 8903 animal names was lemmatized using the WordNet lemmatizer (Miller, 1993) available in the Natural Language Toolkit suite of Python libraries (Loper & Bird, 2002). *Lemmatization* refers to the conversion of inflected forms of a word to that word's canonical form, or *lemma*. For English nouns, this essentially means changing plurals to their singular forms − e.g., *dogs* → *dog* or *oxen* → *ox*. Animal names were lemmatized so that each name would correspond to only one LSA term vector, rather than multiple.

Consider this: if each instance of the word *dogs* was not lemmatized to `dog`, then LSA would consider these inflectional variants of the animal name completely different words, each with its own unique frequency distribution across corpus documents, and thus each with its own term vector in the derived semantic space. Because the end goal of using LSA here is to determine animal relatedness, and because LSA does this by taking the cosine of two words' respective term vectors, it is important that the meaning of each word − which LSA infers from that word's distribution in the corpus − is not divvied up among multiple term vectors. In the case that animal names would not have been lemmatized, one who wishes to test, via LSA, how related the pair *cat* and *dog* is would not be utilizing semantic information

[14]The core of this script was written by Anja Laske.

[15]Hyphens, underscores, slashes, and vertical bars were replaced with whitespace.

as to the meaning of the two animal names imparted by occurrences of the plurals *cats* and *dogs*. When all inflectional variants of a word are merged into its canonical form, LSA is better able to utilize *all* of that word's meaning, in as much as it can be inferred from the contexts of its occurrences in the corpus.

After lemmatization, I fused all multiword animal names into single-word tokens, so that, for the same reason I converted names to their lemmas, each animal name would have only a single term vector in the LSA model. Consider an animal such as the *arctic ground squirrel*. Because of how LSA works, this animal's name would be split into three separate words − *arctic*, *ground*, and *squirrel* − without concern for the fact that they should constitute a single term, in the LSA sense. As such, no term vector would exist for *arctic ground squirrel* that could be used to determine this animal's relatedness to any other animal, because every instance of the name would instead contribute to the term vectors for its constituent words. To remedy this, I replaced the interior whitespace of all multiword animal names with underscores so that, for example, each instance of `arctic ground squirrel` was contracted to the single token `arctic_ground_squirrel`.

Some longer multiword animal names actually contain shorter multiword names, as is the case with, e.g., *annulated sea snake* and *sea snake*. Because of this, I first contracted the largest animal names, made up of five words, before continuing onto four-word names, three-word names, and finally two-word names. Otherwise, contraction of instances of these embedded shorter names would have precluded contraction of the longer names they occurred within. For instance, if every occurrence of `sea snake` was tokenized to `sea_snake`, one would find `annulated sea_snake`, which would not be contracted to `annulated_sea_snake`, because only instances of `annulated sea snake`, without any underscores, are contracted so.

Lastly, I assembled a list of *stop words*, for which every instance of each word was removed from the corpus. Stop words are generally common or grammatically functional words that do not impart much in the way of meaning to words they co-occur with. Because they occur so frequently in the language, and thus in any given corpus, they are a classic source of noise for tasks such as LSA.

The stop list I used was a combination of a general list of 470 stop words[16] and a list assembled specifically for the *animals* domain. The tailored list was composed subjectively by examining the 500 most frequently occurring words in the corpus and adding those I deemed not to impart semantic value

[16]This list was prepared by Anja Laske.

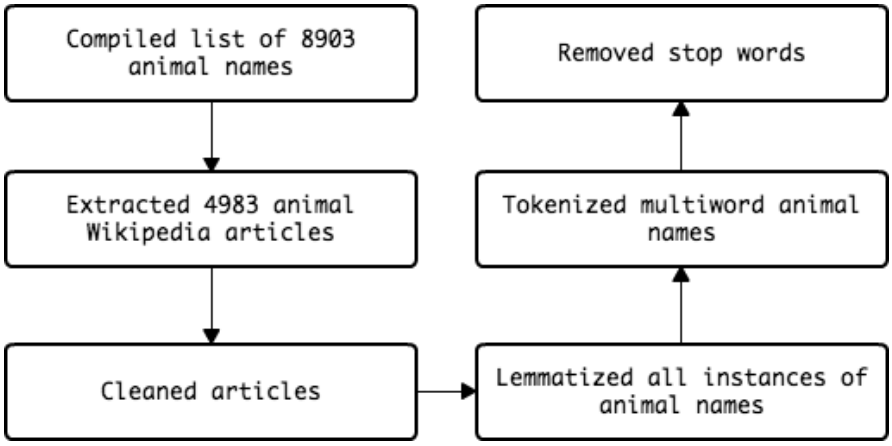

Figure 6: Steps in preparing the corpus for latent semantic analysis.

in the context of animal relatedness, due to an intuitively high likelihood of turning up in any encyclopedia article concerning animals (as evidenced by frequent appearance in the corpus). Examples of such words include *species*, *habitat*, and *family*. 652 words were included in the final stop list, which can be found in APPENDIX B.

At this point, the prepared corpus, consisting of nearly 5000 preprocessed documents, was suitable for latent semantic analysis. FIGURE 6 recaps the steps taken during preprocessing.

3.2.3 Deriving a semantic space

LSA models were derived from the corpus for every dimensionality k between 2 and 500 using GENSIM, a Python toolkit for distributional semantics that supports, among other models, LSA (Řehůřek & Sojka, 2010).[17]

In order to determine which dimensionality should be used to generate semantic relatedness scores for VF-CLUST's SCA module, a small informal study was conducted to elicit human relatedness judgements for use as a gold standard. For this, myself and six others[18] each indicated how semantically related, on a Likert scale from $0-3$, we believe sixty pairs of common animals to be. These included pairs such as *deer−elk*, *kangaroo−koala bear*,

[17] These models were built using computing resources granted by the Minnesota Supercomputing Institute.

[18] These were Mara Anderson Searls, Robert Bill, Nina Dylla, Aaron Free, Kyle Marek-Spartz, Serguei Pakhomov, and Hannah Sande.

elephant−cheetah, and *dolphin−parakeet*.

These ratings were then compared to relatedness scores generated by LSA at each dimensionality k from 2 to 500 using Spearman rank correlation.[19] In the end, dimensionality 91 was found to correlate best to the human judgements. This is rather few compared to dimensionalities commonly chosen for LSA, but the corpus used, with its 4983 documents, is likewise small.

3.2.4 Relatedness classification

Like its counterpart PCA module, VF-CLUST's SCA module cannot identify clusters without having a method for classifying word pairs as semantically related or unrelated. To this end, the module uses semantic relatedness scores (SRS) generated by latent semantic analysis of the corpus described above, with 91 dimensions by default. As with the phonetic similarity scores produced by the PCA module's edit-distance method, SRSs generated by LSA are continuous, and thus do not easily lend to binary relatedness classification.

To resolve this, I empirically determined SRS thresholds for dimensionalities 50 through 100.[20] First, I compiled a list of 314 unique animal names produced in SVF responses by participants in the study described below, in SECTION 4, as well as a preliminary sample from the Professional Fighters Brain Health Study (Bernick et al., in press; Ryan et al., in press); this list was intended as a fairly large sample representative of animals commonly named during SVF testing. Then, for each of the 51 dimensionalities for which thresholds were to be determined, SRSs were computed for all 49,141 unique pairs found among all possible combinations of the 314 animal names, excluding pairs of names and themselves.

As with the PCA module's edit-distance method, the threshold for each dimensionality was initially set as the upper quintile of the pairwise SRSs computed for it. However, upon inspection, this threshold appeared to be too inclusive of animal pairs that were not substantially related. So, thresholds were raised to the upper *deciles* of the pairwise SRSs for each dimensionality, which appeared to give much better classification results. The SRS threshold for dimensionality 91 was set at 0.14, while the remaining dimensionalities were given thresholds ranging from 0.135 to 0.23, with values increasing with smaller dimensionalities.

[19] This analysis was conducted by Serguei Pakhomov.

[20] While VF-CLUST's SCA module uses 91 for its default dimensionality, it also supports dimensionalities 50 through 100, as explained in SECTION 3.2.5.

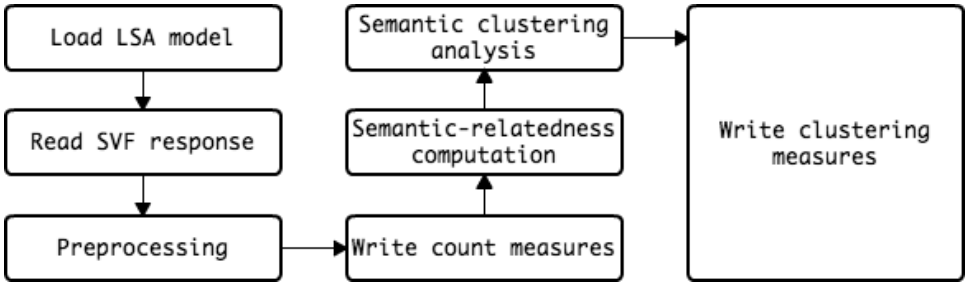

Figure 7: High-level SCA module workflow.

Now, with a method for quantifying semantic relatedness, and a threshold for determining relatedness classification, the SCA module is suitably equipped for semantic clustering analysis.

3.2.5 Module workflow

FIGURE 7 shows the high-level workflow of VF-CLUST's SCA module. This module accepts comma-separated transcriptions of SVF responses, and also expects an LSA dimensionality to use for SRS computation. If a dimensionality is not specified, the default of 91 is used.

Prior to preprocessing of the SVF response, the LSA model for the specified dimensionality is loaded. VF-CLUST comes preloaded with models for each dimensionality k in the range 50 through 100. (During determination of the ideal dimensionality for the corpus and task domain, dimensionalities in this range were found to correlate better to human relatedness judgements than the rest of dimensionalities 2 through 500.) Each LSA model has indexed term vectors for every term found in the corpus, so an SRS for any two words is easily computed by finding the cosine between the words' term vectors.

During preprocessing, multiword animal names produced in the SVF response are contracted to single tokens by replacing whitespace with underscores, as described in SECTION 3.2.2. Additionally, any word or token not included in VF-CLUST's animal names compilation − i.e., the list of 8903 names, described above − is removed, though these will be considered for the module's count measures, described below.

Next, the module's count measures are totaled, and then SRSs are computed sequentially for each contiguous pair of animal names, and also for all pairwise animal names. Finally, these scores are used to identify clusters

and to generate values for VF-CLUST's semantic clustering measures, which are described in the next section.

3.2.6 SVF measures

Like its counterpart module, VF-CLUST's SCA module generates values along several clustering and count measures. Its count measures are equivalent to the PCA module's. They are: *total number of words* (SEM-TOTAL), *number of permissable words* (SEM-WORDS), *number of repetitions* (SEM-REPS), and *number of off-topic words* (SEM-ASIDES). SEM-WORDS counts all words in a SVF response that are found in VF-CLUST's animal names list, while SEM-ASIDES is a count of all remaining words. SEM-REPS is a count of repeated permissible words only, and SEM-TOTAL is equivalent to SEM-WORDS + SEM-ASIDES.

Again like the PCA module, the SCA module identifies both clusters and chains, which are predictably called *semantic clusters* and *semantic chains*. However, because this module only uses one method for determining semantic relatedness, it generates fewer clustering measures than does the PCA module. These are: *number of clusters* (SEM-CLUSTERS); *number of clusters, excluding singletons* (SEM-CLUSTERS-NS); *number of chains* (SEM-CHAINS); *number of chains, excluding singletons* (SEM-CHAINS-NS); *mean cluster size* (SEM-MEAN-CLUSTER); *mean cluster size, excluding singletons* (SEM-MEAN-CLUSTER-NS); *mean chain length* (SEM-MEAN-CHAIN); *mean chain length, excluding singletons* (SEM-MEAN-CHAIN-NS); *maximum cluster size*, i.e., size of the largest cluster (SEM-MAX-CLUSTER); *maximum chain length*, i.e., length of the largest chain (SEM-MAX-CHAIN); *number of cluster switches* (SEM-CL-SWITCHES); and *number of chain switches* (SEM-CH-SWITCHES).

Additionally, the module produces a *mean pairwise semantic relatedness score* (SEM-PAIRWISE). This is calculated as the mean of all pairwise SRSs for the animal names in a given SVF response.

TABLE 2 shows abbreviations for each of the VF-CLUST's SVF measures, along with corresponding definitions.

In the next section, VF-CLUST's utility in detecting subtle indicators of cognitive impairment due to dementia is demonstrated with the results from a pilot study using data from patients with mild cognitive impairment and dementia from Alzheimer's disease.

Table 2: VF-CLUST's SVF measures

SEM-ASIDES	number of off-topic words
SEM-CH-SWITCHES	maximum common-biphone chain switches
SEM-CHAINS	number of common-biphone chains
SEM-CHAINS-NS	number of common-biphone chains, not including singletons
SEM-CL-SWITCHES	maximum common-biphone cluster switches
SEM-CLUSTERS	number of common-biphone clusters
SEM-CLUSTERS-NS	number of common-biphone clusters, not including singletons
SEM-MAX-CHAIN	maximum common-biphone chain length
SEM-MAX-CLUSTER	maximum common-biphone cluster size
SEM-MEAN-CHAIN	mean common-biphone chain size
SEM-MEAN-CHAIN-NS	mean common-biphone chain size, not including singletons
SEM-MEAN-CLUSTER	mean common-biphone cluster size
SEM-MEAN-CLUSTER-NS	mean common-biphone cluster size, not including singletons
SEM-PAIRWISE	mean pairwise CBS
SEM-REPS	number of repeated permissible words
SEM-TOTAL	total number of words
SEM-WORDS	number of permissible words

4 Pilot Study

A pilot study, using verbal fluency test responses from patients with mild cognitive impairment (MCI) and dementia from Alzheimer's disease, was conducted to test the utility of VF-Clust in assessing the effect of impairment from these conditions on verbal fluency performance and clustering behavior.

4.1 Participants

The verbal fluency test responses used for this study came from participants included in a random sample obtained from the Mayo Clinic Alzheimers Disease Research Center and the Mayo Clinic Study of Aging. This sample included 133 subjects, matched for age ($\bar{x} = 66.98$, $\sigma = 11.3$) and sex. At the time of neuropsychological testing, 37 participants had a clinical diagnosis of probable AD, by DSM-IV (American Psychiatric Association, 1994) and NINCDS-ADRDA (McKhann et al., 1984) criteria; 58 had a clinical diagnosis of MCI with an amnestic component (Peterson, 2004); and 38 were healthy controls. These clinical diagnoses were decided upon during consensus conferences among neurologists, neuropsychologists, and nurses, taking into consideration neurological assessment, performance on a neuropsychological test battery, and the views of family informants.

Positron emission tomography with amyloid tracer Pittsburgh Compound B (PiB-PET; Jack et al., 2008) − a technique used to measure amyloid load, elevated levels of which are a pathologic hallmark of Alzheimer's disease − was available for 58 of the 60 participants. Such imaging showed PiB-PET SUVR values greater than +1.4, consistent with AD-specific pathology, in all participants clinically diagnosed with AD.

4.2 Methods

During data collection, all participants underwent a neuropsychological test battery that included PVF tests for letters *c*, *f*, and *l*, and an SVF test for category *animals*. Additionally, all subjects were given Clinical Dementia Ratings (CDRs), which are used to stage dementia symptoms and have the following levels: 0, *none*; 0.5, *very mild*; 1, *mild*; 2, *moderate*; and 3, *severe* (Morris, 1997). By these ratings, patients with diagnoses of AD and MCI were generally quite mildly impaired (AD: $\bar{x} = 0.72$, $\sigma = 0.32$; MCI: $\bar{x} = 0.48$, $\sigma = 0.09$).

All verbal fluency test responses were audio-recorded during administration and later transcribed. Transcriptions were submitted to VF-CLUST for automated clustering analysis, and each SVF response also underwent manual clustering analysis, following the methodology of Troyer et al. (1997), by myself and another scorer.[21] Interrater reliabilities, determined by Pearson correlation coefficients, were high for both number of cluster switches ($r(131) = 0.93$) and mean cluster size ($r(131) = 0.84$).

For each participant, VF-CLUST output was averaged across the three PVF tests to give one value for each measure. Likewise, values for manually determined measures were averaged across raters to give one score for each subject for *manually determined mean semantic cluster size* (SEM-MEAN-CLUSTER-MAN) and *manually determined number of semantic cluster switches* (SEM-CL-SWITCHES-MAN). Transcriptions did not include off-topic speech, so no analysis of PHN-ASIDES or SEM-ASIDES was conducted.

Consistent with the findings of Troyer et al. (1997), SEM-WORDS correlated much more strongly with age ($r(131) = -0.31$, $p < 0.001$) than did PHN-WORDS ($r(131) = 0.02$, $p = 0.86$). Because controls were considerable younger ($\bar{x} = 61.6$) than subjects with MCI ($\bar{x} = 69.0$) and AD ($\bar{x} = 69.33$), group differences in SVF measures were examined with participants younger than 65 and older than 80 excluded. For this age-constricted subset, there were 13 subjects with AD, 40 subjects with MCI, and 16 controls.

For comparisons of group means, t-tests were used. Logistic regression was used to compare the predictive value of various measures. Here, receiver operator characteristic area under the curve (AUC) is reported, along with Nagelkerke's r^2.

4.3 Results

The results of comparisons of group differences on PVF and SVF measures, and of manual and automatic clustering methods, as well as logistic regression using various predictors, are all reported below.

4.3.1 Differences in group means for PVF measures

These results are shown in TABLE 3. Surprisingly, no significant differences were observed in PHN-WORDS between controls and subjects with AD ($p = 0.235$), or between controls and subjects with MCI ($p = 0.235$). Interestingly, however, PHN-REPS for subjects with AD was twice that of controls ($\bar{x} = 0.76$ vs. 0.37, $p < 0.004$), while the difference between subjects with

[21] This was Mara Anderson Searls.

Table 3: Group means for PVF measures

Measure	**AD** ($n = 37$)	**MCI** ($n = 58$)	**Controls** ($n = 38$)
PHN-WORDS	11.17	12.24	12.33
PHN-REPS	0.76**	0.52	0.37
PHN-CLUSTERS	7.98	8.29	8.15
PHN-CLUSTERS-NS	2.84	3.24	3.25
BPH-CLUSTERS	9.92	10.26	10.43
BPH-CLUSTERS-NS	1.05*	1.56	1.50
PHN-CHAINS	7.65	7.71	7.49
PHN-CHAINS-NS	2.34	2.67	2.57
BPH-CHAINS	10.14	10.29	10.49
BPH-CHAINS-NS	1.02*	1.48	1.47
PHN-CL-SWITCHES	6.98	7.29	7.15
BPH-CL-SWITCHES	8.92	9.26	9.43
PHN-CH-SWITCHES	6.65	6.71	6.49
BPH-CH-SWITCHES	9.14	9.29	9.49
PHN-MEAN-CLUSTER	1.56	1.62	1.68
PHN-MEAN-CLUSTER-NS	2.23	2.35	2.40
BPH-MEAN-CLUSTER	1.15	1.21	1.20
BPH-MEAN-CLUSTER-NS	1.42	1.82	1.62
PHN-MEAN-CHAIN	1.63	1.71	1.86
PHN-MEAN-CHAIN-NS	2.56	2.67	2.90
BPH-MEAN-CHAIN	1.15	1.22	1.21
BPH-MEAN-CHAIN-NS	1.40	1.84	1.64
PHN-MAX-CLUSTER	2.61*	2.95	3.07
BPH-MAX-CLUSTER	1.84	2.20	2.02
PHN-MAX-CHAIN	3.05*	3.40	3.79
BPH-MAX-CHAIN	1.83	2.26	2.07
PHN-PAIRWISE	0.28	0.28	0.28
BPH-PAIRWISE	0.14	0.14	0.13

Note: Asterisk(s) indicates significant difference between group mean and that of controls: * $p < 0.05$, ** $p < 0.01$.

MCI and controls was not significant ($p = 0.18$). Unlike other PVF measures, PHN-REPS *was* found to correlate to age ($r(131) = 0.17$, $p < 0.05$). However, even with participant age restricted to $65-80$ years, PHN-REPS was significantly higher among subjects with AD relative to controls ($\bar{x} = 0.97$ vs. 0.46, $p < 0.05$).

While group means for PHN-CLUSTERS-NS, PHN-CHAINS-NS showed no significant differences, subjects with AD were found to generate significantly fewer BPH-CLUSTERS-NS ($\bar{x} = 1.05$ vs. 1.50, $p = 0.037$), and likewise fewer BPH-CHAINS-NS ($\bar{x} = 1.02$ vs. 1.47, $p = 0.027$). However, when singletons were included in analysis, common-biphone cluster and chain counts were equivalent between controls and subjects with AD. This was the same with the equivalent phonetic cluster and chain counts. Because these measures directly correspond to switching measures, no significant differences were found with those either. With or without singletons being included, controls and subjects with MCI generated equivalent amounts of common-biphone clusters and chains.

Subjects with AD and MCI did not significantly differ from controls on any mean cluster size or mean chain length measures, though differences between subjects with AD and controls were nearly significant for PHN-MEAN-CHAIN ($p = 0.087$) and PHN-MEAN-CHAIN-NS ($p = 0.085$). Significant differences *were* observed between subjects with AD and controls on both PHN-MAX-CLUSTER ($\bar{x} = 2.61$ vs. 3.07, $p < 0.05$) and PHN-MAX-CHAIN ($\bar{x} = 3.05$ vs. 3.79, $p = 0.032$), but not on their common-biphone equivalents. Maximum phonetic and common-biphone cluster sizes and chain lengths were comparable across controls and subjects with MCI. Likewise, PHN-PAIRWISE means were roughly equivalent across all groups.

4.3.2 Differences in group means for SVF measures

For SVF analysis, the age-constricted subset was used; these results are shown in TABLE 4. Group differences in SEM-WORDS were highly significant for both controls versus subjects with AD ($p = 0.003$) and controls versus subjects with MCI ($p = 0.016$). Comparisons between groups in repetitions on SVF showed similar results to those of repetitions on PVF ($\bar{x} = 2.54$ (AD) vs. 0.69 (controls), $p < 0.003$), though the difference between controls and subjects with MCI was nearly significant here ($\bar{x} = 1.35$ vs. 0.69, $p = 0.062$). Unlike with PHN-REPS, SEM-REPS was not found to correlate to age ($r(131) = 0.28$).

Subjects with AD generated significantly fewer semantic clusters ($\bar{x} = 2.92$ vs. 5.0, $p < 0.004$) and semantic chains ($\bar{x} = 2.54$ vs. 4.19, $p < 0.004$),

Table 4: Age-constricted group means for SVF measures

Measure	**AD** ($n = 13$)	**MCI** ($n = 40$)	**Controls** ($n = 16$)
SEM-WORDS	14.62**	16.08*	20.06
SEM-REPS	2.54**	1.35	0.69
SEM-CLUSTERS	10.92	11.23*	13.63
SEM-CLUSTERS-NS	2.92**	3.95	5.00
SEM-CHAINS	10.62	10.80	12.63
SEM-CHAINS-NS	2.54**	3.18	4.19
SEM-CL-SWITCHES	9.92	10.23*	12.63
SEM-CH-SWITCHES	9.62	9.80	9.63
SEM-MEAN-CLUSTER	1.42	1.53	1.61
SEM-MEAN-CLUSTER-NS	2.80	2.39	2.54
SEM-MEAN-CHAIN	1.48	1.57	1.67
SEM-MEAN-CHAIN-NS	3.15	2.68	2.85
SEM-MAX-CLUSTER	3.08	3.00	3.44
SEM-MAX-CHAIN	3.69	3.48	4.06
SEM-PAIRWISE	0.078	0.066	0.065

Note: Only participants aged $65 - 80$ were included in this analysis. Asterisk(s) indicates significant difference between group mean and that of controls: * $p < 0.05$, ** $p < 0.01$.

not including singletons, than did controls; when singletons were included, there were no significant differences. Due to direct correlation with these, switching measures gave the same results. Subjects with MCI, however, also had smaller SEM-CHAINS-NS than controls − though the difference was not quite significant ($\bar{x} = 3.18$ vs. 4.19, $p = 0.055$) − but differed more on SEM-CLUSTERS ($\bar{x} = 11.23$ vs. 13.63, $p < 0.05$) than on SEM-CLUSTERS-NS ($\bar{x} = 3.95$ vs. 5.0, $p = 0.1$).

Controls did not differ significantly from subjects with AD or MCI on SEM-MEAN-CLUSTER(-NS) or SEM-MEAN-CHAIN(-NS), though the largest difference was observed between subjects with AD and controls on SEM-MEAN-CLUSTER ($\bar{x} = 1.42$ vs. 1.61, $p < 0.11$). Likewise, no significant differences were found with SEM-MAX-CLUSTER or SEM-MAX-CHAIN.

As with PHN-PAIRWISE, differences between controls and subjects with AD and MCI for SEM-PAIRWISE were not significant.

No significant differences with controls were found on manually determined mean cluster size ($\bar{x} = 0.85$ (AD) vs. 1.10 (controls), $p = 0.133$; $\bar{x} = 1.07$ (MCI) vs. 1.10 (controls), $p = 0.77$), but differences in manually determined number of cluster switches were significant for subjects with MCI versus controls ($\bar{x} = 7.05$ vs. 8.91, $p = 0.03$), though not for subjects with AD against controls ($\bar{x} = 7.12$ vs. 8.91, $p = 0.07$). These results directly correspond to those for the equivalent automatically determined measures.

4.3.3 Classification by logistic regression

TABLE 5 shows the performance of several logistic regression models. Because there are only 13 AD cases in this subset, the full data set, in which there are 37 AD cases, was used for this analysis to allow for multiple regression.

Logistic regression with SEM-MEAN-CLUSTER-MAN and SEM-CL-SWITCHES-MAN as predictors differentiated age-constricted controls and subjects with AD with an AUC of 0.867 ($r^2 = 0.5$), and subjects with MCI with an AUC of 0.690 ($r^2 = 0.155$). When SEM-WORDS, which is fairly easy to determine manually, is included in this model, these AUCs respectively increase to 0.873 ($r^2 = 0.537$) and 0.713 ($r^2 = 0.154$). The best-performing model with the comparable automatic methods − one with SEM-WORDS, SEM-MEAN-CHAIN, and SEM-CH-SWITCHES − performs slightly better in classifying subjects with AD, with 0.878 AUC ($r^2 = 0.552$), but worse in classifying subjects with MCI, with 0.676 AUC ($r^2 = 0.126$).

Logistic regression on the full data set with only SEM-WORDS classifies controls and subjects with AD with an AUC of 0.867 ($r^2 = 0.535$), and

Table 5: Performance of logistic regression models

Predictors	**Area under curve ($r2$)**	
	AD vs. control	**MCI vs. control**
SEM-WORDS SEM-REPS SEM-CLUSTERS BPH-MAX-CLUSTER	0.921 (0.667)	0.761 (0.28)
SEM-WORDS SEM-REPS	0.912 (0.65)	0.728 (0.204)
SEM-WORDS SEM-MEAN-CLUSTER-MAN SEM-CL-SWITCHES-MAN	0.873 (0.537)	0.713 (0.154)
SEM-WORDS SEM-MEAN-CHAIN SEM-CH-SWITCHES	0.878 (0.552)	0.676 (0.126)
SEM-MEAN-CLUSTER-MAN SEM-CL-SWITCHES-MAN	0.867 (0.5)	0.69 (0.155)
SEM-WORDS	0.867 (0.535)	0.672 (0.126)
PHN-REPS	0.69 (0.16)	0.526 (0.022)
PHN-WORDS	0.582 (0.026)	0.5 (0.001)

Note: All participants were included in this analysis. Values shown are receiver operator characteristic area under curve, with Nagelkerke's r^2. Models are listed in order of total performance in classifying subjects with AD and MCI vs. controls.

controls and subjects with MCI with an AUC of 0.672 ($r^2 = 0.126$). A model with only PHN-WORDS does much worse, with respective AUCs of 0.582 ($r^2 = 0.026$) and 0.5 ($r^2 < 0.001$). Using only PHN-REPS actually gives a much better model, classifying controls and subjects with AD with an AUC of 0.69 ($r^2 = 0.16$). Combining SEM-WORDS and SEM-REPS differentiates controls and subjects with AD with a very high AUC of 0.912 ($r^2 = 0.65$), and controls and subjects with MCI with an AUC of 0.728 ($r^2 = 0.204$).

Since there are just 37 AD cases in the full data set, it is only safe to include at most three to four independent variables in a logistic regression model for these data (Harrell, 2001). With all of VF-CLUST's count and clustering measures available, the best-performing model uses SEM-WORDS, SEM-REPS, SEM-CLUSTERS, and BPH-MAX-CLUSTER, differentiating controls and subjects with AD with an AUC of 0.921 ($r^2 = 0.667$), and controls and subjects with MCI with an AUC of 0.761 ($r^2 = 0.28$).

4.4 Discussion

While significant differences were found between controls and subjects with AD and MCI in the number of words generated on SVF, surprisingly no significant differences were found in the number of words generated on PVF, as has been observed in several studies. It is likely that this could be explained by the patients with MCI and AD in this study having only mild dementia symptoms at the time of data collection, as evidenced by their fairly low CDRs. This may also explain why mean phonetic and semantic cluster size were not found to significantly differ between groups by any manually- or automatically determined metric, and why subjects with AD and MCI were not found to switch less on either PVF or SVF.

Interestingly, when singletons were excluded, subjects with AD produced much smaller common-biphone clusters and chains, as well as semantic clusters and chains. Because these measures do not include singletons in their counts, they are unique to this study. The alternative cluster counts, which do include singletons, directly correspond to the number of switches in an attempt, which is a commonly reported measure. That the measures not including singletons were more sensitive to differences between controls and patients with very mild AD may suggest that singletons dilute cluster counts that include them. In the case of semantic clusters, subjects with MCI were actually significantly different than controls − which was not observed when singletons were excluded − while subjects with AD were not. Here, subjects with AD actually produced less clusters than subjects with MCI, but due to the age-constricted data set used for SVF analysis, there were only 13

AD cases compared to 40 MCI cases, giving the latter's comparison with controls more degrees of freedom.

Also unique to this study, and also showing significant differences between patients with AD and controls, are values for maximum phonetic cluster size and maximum phonetic chain length. While subjects with AD did not produce smaller phonetic clusters, their respective largest clusters were smaller than those of controls. This would seem to indicate that while subjects with AD produced only slightly smaller phonetic clusters, they showed less variability on these measures. This explanation is supported by much smaller standard deviations among patients with AD on both mean phonetic cluster size − 0.27 (AD) vs. 0.44 (controls) − and mean phonetic chain length − 0.34 (AD) vs. 0.72 (controls). Interestingly, the corresponding SVF measures for maximum cluster size and maximum chain length did not show significant differences between controls and subjects with AD, and likewise did not yield remarkably different standard deviations.

Perhaps the strongest result in this pilot study pertains to the significance of repetitions on verbal fluency tests. While most studies have ignored them, a few have reported on repetitions, which are also called perseverative errors, but only in SVF. Binetti et al. (1995) found few to no repetitions among AD patients, while March & Pattison (2007) found significantly more relative to controls; contrasting both, Raoux et al. (2008) reported no difference. In the present study, subjects with AD generated many more repetitions on both PVF and SVF. With PVF, this set these subjects apart from controls while the number of words generated did not. Further, logistic regression models that included repetition counts performed extremely well, particularly a model with only the number of words and repetitions generated on SVF as predictors. These results strongly suggest that repetitions should be considered in future analyses of verbal fluency test responses.

While this simple logistic regression model, including only the number of words and repetitions generated on SVF, did perform very well in classifying both subjects with AD and subjects with MCI, it was improved by the inclusion of one or two of several other predictors. This would seem to indicate that clustering measures can add classificational value beyond what the more basic measures − counts of words generated and of repetitions − give on their own.

In a study of the effects of repetitive head trauma in professional fighters, Ryan et al. (in press) compared some of VF-Clust's PVF measures with manually determined equivalents, and found the automatically determined measures − particularly those computed by the common-biphone method − to be more sensitive and to produce less variability. Here, both

by comparisons of group means and by logistic regression, VF-Clust's SVF measures were found to perform about equivalently to their manually determined counterparts. A comparison of automatically computed number of semantic cluster switches and mean semantic cluster size with the manually determined equivalents showed fairly high correlation for the former ($r(131) = 0.8$), and low correlation for the latter ($r(131) = 0.35$). In the case of mean cluster size, some discrepancy may be due to differing methods for determining cluster size, as described in SECTION 3.1.5.

5 Conclusion

VF-Clust is a publicly available system for computerized analysis of verbal fluency tests. Currently, clustering analyses on verbal fluency tests are conducted manually, by trained scorers, in a process that is labor-intensive and prone to human error and interrater variability. The objective in developing VF-Clust was to give a method for *automatically* generating clustering analyses on both semantic and phonemic verbal fluency test responses, which it does by using latent semantic analysis and computational methods for determining phonetic similarity.

The results from the pilot study described above show that, at worst, VF-Clust's automatically determined clustering measures seem to be as useful as their manually determined equivalents. Manual methods, however, are far more labor-intensive – VF-Clust can process several hundred test responses in a matter of seconds. As such, this automatic approach is much more scalable to larger numbers of responses. Moreover, automatic clustering analysis is not prone to human error, or to interrater variability, as manual methods are.

Additionally, a strong finding from the pilot study is that word repetitions were more numerous in verbal fluency test responses from patients with dementia from Alzheimer's disease. In the case of phonemic verbal fluency, controls differed with patients with Alzheimer's diagnoses on the number of words repeated, but not on the number of words generated, which is the standard measure by which the test is scored. While repetitions are typically ignored in even the most in-depth analyses of verbal fluency test responses, these results strongly indicate that future analyses should consider them.

In the future, VF-Clust's semantic clustering analysis module should be expanded to support analysis of categories besides *animals* that are also commonly specified in semantic verbal fluency test administration, such as *fruits*, *vegetables*, *furniture*, and *supermarket items*. LSA models for these

domains could be derived by the same methods described above. Beyond latent semantic analysis, other distributional-semantic techniques, such as *random indexing* or *latent Dirichlet allocation*, could be supported as well. While the SCA module, using LSA, does not appear to give results significantly better than those from manual clustering analysis, it possibly could using other computational techniques for determining semantic relatedness.

Further, both the PCA and SCA modules could be improved by supporting different methods by which phonetic and semantic clusters are identified. As it is, semantic clusters and phonetic clusters determined by edit distance use semantic relatedness score- and phonetic similarity score thresholds, respectively, to determine whether two words should form a cluster. An alternative, perhaps a better one, would be to use a clustering algorithm, such as *k-means clustering*, to determine a set of actual clusters to which words will belong. Then, rather than determining whether two words are similar enough, the system would check whether they belong to the same cluster. At the very least, this approach should be carried out in order to compare it to the current methods.

While there are system improvements to be made, and though further clinical validation is still needed, VF-Clust already shows utility as − if nothing else − an efficient alternative to a common manual approach.

References

[1] American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders: Diagnostic criteria from DSM-IV. American Psychiatric Association.

[2] Binetti, G., Magni, E., Cappa, S. F., Padovani, A., Bianchetti, A., & Trabucchi, M. (1995). Semantic memory in Alzheimer's disease: An analysis of category fluency. *Journal of Clinical and Experimental Neuropsychology, 17*(1), 82-89.

[3] Benton, A. L. (1968). Differential behavioral effects in frontal lobe disease. *Neuropsychologia, 6*(1), 53-60.

[4] Bernick, C., Banks, S.J., Jones, S., Shin, W., Phillips, M., Lowe, M., Modic, M. (In press). Professional fighters brain health study: Rationale and methods. *American Journal of Epidemiology.*

[5] Brookmeyer, R., Johnson, E., Ziegler-Graham, K., & Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's and Dementia, 3*(3), 186-191.

[6] Butters, N., Wolfe, J., Granholm, E., & Martone, M. (1986). An assessment of verbal recall, recognition and fluency abilities in patients with Huntington's disease. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior.*

[7] Chertkow, H., & Bub, D. (1990). Semantic memory loss in dementia of Alzheimer's type: What do various measures measure. *Brain, 113*(2), 397-417.

[8] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

[9] Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188-230.

[10] Gomez, R. G., & White, D. A. (2006). Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology, 21*(8), 771-775.

[11] Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group Tutoring Research Group, & Person, N. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8*(2), 129-147.

[12] Harrell, F. E. (2001). Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag.

[13] Henry, J. D., & Crawford, J. R. (2004). A meta-analytic review of verbal fluency performance in patients with traumatic brain injury. *Neuropsychology, 18*, 621-628.

[14] Ho, A. K., Sahakian, B. J., Robbins, T. W., Barker, R. A., Rosser, A. E., & Hodges, J. R. (2002). Verbal fluency in Huntingtons disease: A longitudinal analysis of phonemic and semantic clustering and switching. *Neuropsychologia, 40*(8), 1277-1284.

[15] Hodges, J. R., & Patterson, K. (1995). Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. *Neuropsychologia, 33*(4), 441-459.

[16] Holtzman, D. M., Morris, J. C., & Goate, A. M. (2011). Alzheimer's disease: The challenge of the second century. *Science Translational Medicine, 3*(77).

[17] Jack, C. R., Lowe, V. J., Senjem, M. L., Weigand, S. D., Kemp, B. J., Shiung, M. M., ... & Petersen, R. C. (2008). 11C PiB and structural MRI provide complementary information in imaging of Alzheimer's disease and amnestic mild cognitive impairment. *Brain, 131*(3), 665-680.

[18] Jones, M. P., & Martin, J. H. (1997). Contextual spelling correction using latent semantic analysis. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 166-173. Association for Computational Linguistics.

[19] Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

[20] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*, 707-710.

[21] Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics-Volume 1*, 63-70. Association for Computational Linguistics.

[22] March, E. G., & Pattison, P. (2006). Semantic verbal fluency in Alzheimer's disease: Approaches beyond the traditional scoring system. *Journal of Clinical and Experimental Neuropsychology, 28*(4), 549-566.

[23] Martin, A., & Fedio, P. (1983). Word production and comprehension in Alzheimer's disease: The breakdown of semantic knowledge. *Brain and Language, 19*(1), 124-141.

[24] McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34*(7), 939-944.

[25] Miller, G. A. (1993). Five papers on WordNet. *Technical Report CLS-Rep-43*. Cognitive Science Laboratory, Princeton University.

[26] Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research, 29*(4), 495-512.

[27] Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., & Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of Neurology, 49*(12), 1253-1258.

[28] Monsch, A. U., Bondi, M. W., Butters, N., Paulsen, J. S., Salmon, D. P., Brugger, P., & Swenson, M. R. (1994). A comparison of category and letter fluency in Alzheimer's disease and Huntington's disease. *Neuropsychology, 8*(1), 25-30.

[29] Morris, J. C. (1997). Clinical dementia rating: A reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International Psychogeriatrics, 9*(S1), 173-176.

[30] Newcombe, F. (1969). *Missile wounds of the brain: A study of psychological deficits*. London: Oxford University Press.

[31] Pagel, V., Lenzo, K., & Black, A. W. (1998). Letter to Sound Rules for Accented Lexicon Compression.

[32] Pasquier, F., Lebert, F., Grymonprez, L., & Petit, H. (1995). Verbal fluency in dementia of frontal lobe type and dementia of Alzheimer type. *Journal of Neurology, Neurosurgery & Psychiatry, 58*(1), 81-84.

[33] Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine, 256*(3), 183-194.

[34] Price, S. E., Ong, B., Mullaly, E., Pangnadasa-Fox, L., Kinsella, G. J., Storey, E., ... & Perre, D. (2012). Semantic verbal fluency strategies in amnestic mild cognitive impairment. *Neuropsychology, 26*(4), 490-497.

[35] Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition*.

[36] Randolph, C., Braun, A. R., Goldberg, T. E., & Chase, T. N. (1993). Semantic fluency in Alzheimer's, Parkinson's, and Huntington's disease: Dissociation of storage and retrieval failures. *Neuropsychology, 7*(1), 82-88.

[37] Raoux, N., Amieva, H., Le Goff, M., Auriacombe, S., Carcaillon, L., Letenneur, L., & Dartigues, J. F. (2008). Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study. *Cortex, 44*(9), 1188-1196.

[38] Raskin, S. A., Sliwinski, M., & Borod, J. C. (1992). Clustering strategies on tasks of verbal fluency in Parkinson's disease. *Neuropsychologia, 30*(1), 95-99.

[39] Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of LREC 2010 workshop on New Challenges for NLP Frameworks*, 46-50.

[40] Roses, A. D., Lutz, M. W., Amrine-Madsen, H., Saunders, A. M., Crenshaw, D. G., Sundseth, S. S., ... & Reiman, E. M. (2009). A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *The Pharmacogenomics Journal, 10*(5), 375-384.

[41] Rosser, A., & Hodges, J. R. (1994). Initial letter and semantic category fluency in Alzheimer's disease, Huntington's disease, and progressive supranuclear palsy. *Journal of Neurology, Neurosurgery & Psychiatry, 57*(11), 1389-1394.

[42] Ryan, J. O., Pakhomov, S., Marino, S., Bernick, C., Banks, S. (In press). Computerized analysis of a verbal fluency test. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Short Papers).*

[43] Trster, A. I., Fields, J. A., Testa, J. A., Paul, R. H., Blanco, C. R., Hames, K. A., ... & Beatty, W. W. (1998). Cortical and subcortical influences on clustering and switching in the performance of verbal fluency tasks. *Neuropsychologia, 36*(4), 295-304.

[44] Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology, 11*(1), 138-146.

[45] Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M. P., & Stuss, D. (1998a). Clustering and switching on verbal fluency: The effects of focal frontal-and temporal-lobe lesions. *Neuropsychologia*.

[46] Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., & Freedman, A. M. (1998b). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society, 4*(2), 137-143.

[47] Weide, R. L. (1998). The CMU pronouncing dictionary, v. 0.7a. *URL: http://www. speech. cs. cmu. edu/cgi-bin/cmudict*.

[48] Wiemer-Hastings, P., Allbritton, D., & Arnott, E. (2004). RMT: A dialog-based research methods tutor with or without a head. *Intelligent Tutoring Systems*, 141-229.

Appendix A: System output

VF-Clust also supports batch processing of several attempts, for which a comma-separated report is instead generated. The following is full VF-Clust output for a single example PVF response `fun fort friend fry um i don't know fret fetch flip`:

```
Total number of words:  11
Number of permissible words:  7
Number of repetitions:  0
Number of off-topic words:  4
Number of phonetic clusters:  4
Number of phonetic clusters (excl.  singletons):  1
Number of common-biphone clusters:  5
Number of common-biphone clusters (excl.  singletons):  1
Number of phonetic chains:  4
Number of phonetic chains (excl.  singletons):  1
Number of common-biphone chains:  5
Number of common-biphone chains (excl.  singletons):  1
Number of phonetic cluster switches:  3
Number of common-biphone cluster switches:  4
Number of phonetic chain switches:  3
Number of common-biphone chain switches:  4
Mean phonetic cluster size:  1.75
Mean phonetic cluster size (excl.  singletons):  4.0
Mean common-biphone cluster size:  1.4
Mean common-biphone cluster size (excl.  singletons):  3.0
Mean phonetic chain length:  1.75
Mean phonetic chain length (excl.  singletons):  4.0
Mean common-biphone chain length:  1.4
Mean common-biphone chain length (excl.  singletons):  3.0
Max.  phonetic cluster size:  4
Max.  common-biphone cluster size:  3
Max.  phonetic chain length:  4
Max.  common-biphone chain length:  3
Mean pairwise phonetic similarity score:  0.342857142857
Mean pairwise common-biphone score:  0.142857142857
```

The following is full output for an example SVF response `cat cougar spider bee fly dog wolf raccoon squirrel tiger zebra i cant think of any more`, with LSA dimensionality set to the default of 91:

```
Total number of words:  17
Number of permissible words:  11
Number of repetitions:  0
Number of off-topic words:  6
Number of semantic clusters:  8
Number of semantic clusters (excl.  singletons):  3
Number of semantic chains:  8
Number of semantic chains (excl.  singletons):  3
Number of semantic cluster switches:  7
Number of semantic chain switches:  7
Mean semantic cluster size:  1.375
Mean semantic cluster size (excl.  singletons):  2.0
Mean semantic chain length:  1.375
Mean semantic chain length (excl.  singletons):  2.0
Max.  semantic cluster size:  2
Max.  semantic chain length:  2
Mean pairwise semantic relatedness score:  0.00992223451022
```

Appendix B: Stop words

a, able, about, above, according, across, adult, adults, after, again, against, age, ago, al, all, almost, along, already, also, although, always, am, among, an, and, animal, animals, another, any, anybody, anyone, anything, anywhere, appearance, approximately, are, area, areas, around, as, ask, asked, asking, asks, at, average, away, b, back, backed, backing, backs, based, be, became, because, become, becomes, been, before, began, behind, being, beings, believed, below, best, better, between, big, black, body, both, breed, breeding, breeds, but, by, c, called, came, can, cannot, cant, case, cases, central, certain, certainly, citation, clear, clearly, close, closely, color, colour, come, common, commonly, complete, considered, consists, could, couldn't, currently, d, day, days, derived, described, developed, did, didn't, diet, differ, different, differently, distinct, do, does, doesn't, doing, done, dont, down, downed, downing, downs, due, during, e, each, early, eat, either, end, ended, ending, ends, enough, especially, estimated, et, even, evenly, ever, every, everybody, everyone, everything, everywhere, example, except, eye, f, fact, facts, family, far, feed, feeding, feet, felt, female, females, few, find, finds, first, five, following, food, for, form, formation, forms, found, four, from, front, full, fully, further, furthered, furthering, furthers, g, gave, genera, general, generally, genus, get, gets, give, given, gives, go, going, good, goods, got, great, greater, greatest, ground, group, grouped, grouping, groups, grow, h, habitat, habitats, had, hadnt, half, has, hasnt, have, havent, having, he, head, hed, hell, her, here, here's, hers, herself, hes, high, higher, highest, highly, him, himself, his, how, however, hows, i, if, i'll, i'm, important, in, include, includes, including, individual, individuals, interest, interested, interesting, interests, into, introduced, is, isn't, it, its, itself, i've, j, just, k, keep, keeps, kind, knew, know, known, knows, l, largely, larger, largest, last, late, later, latest, least, length, less, let, lets, life, light, like, likely, listed, live, lived, living, long, longer, longest, low, lower, m, made, main, mainly, make, making, male, males, man, many, material, mating, mature, may, me, meaning, means, member, members, mi, middle, might, mm, more, most, mostly, mr, mrs, much, must, mustn't, my, myself, n, name, named, national, native, natural, near, necessary, need, needed, needing, needs, nest, never, new, newer, newest, next, no, nobody, non, noone, nor, north, not, nothing, now, nowhere, number, numbers, o, occasionally, occur, occurs, of, off,

often, old, older, on, once, one, only, open, opened, opening, opens, or, order, ordered, ordering, orders, other, others, ought, our, ours, ourselves, out, over, own, p, part, parted, particularly, parting, parts, per, perhaps, period, place, placed, places, point, pointed, pointing, points, population, populations, possible, possibly, predators, present, presented, presenting, presents, prey, primarily, probably, problem, problems, put, puts, q, quite, r, range, ranges, rather, reach, really, recent, recognized, recorded, refer, referred, region, regions, related, relatively, remains, reported, right, room, rooms, s, said, same, saw, say, says, scale, season, second, seconds, see, seem, seemed, seeming, seems, seen, sees, separate, several, shall, shant, shaped, she, shed, shell, shes, should, shouldnt, show, showed, showing, shows, side, sides, similar, since, single, six, size, sized, slightly, smaller, so, some, somebody, someone, something, sometimes, somewhere, species, specific, specimen, specimens, stage, state, states, still, study, subspecies, such, sure, surface, t, take, taken, tend, th, than, that, thats, the, their, theirs, them, themselves, then, there, therefore, there's, these, they, they'd, they'll, they're, they've, thing, things, think, thinks, this, those, though, thought, thoughts, three, through, throughout, thus, time, times, to, today, together, too, took, total, toward, true, turn, turned, turning, turns, two, type, typical, typically, u, under, unlike, until, up, upon, upper, us, use, used, uses, using, usually, v, various, very, w, want, wanted, wanting, wants, was, wasn't, water, way, ways, we, wed, weeks, weigh, weight, well, wells, went, were, werent, weve, what, whats, when, whens, where, wheres, whether, which, while, who, whole, whom, whos, whose, why, whys, wide, wikimedia, wikipedia, wild, will, with, within, without, wont, word, work, worked, working, works, world, would, wouldn't, x, y, year, years, yes, yet, you, you'd, you'll, young, your, you're, yours, yourself, yourselves, you've, z.