Vector space models for evaluating semantic fluency in autism

Emily Prud'hommeaux[†], **Jan van Santen**[°], **Douglas Gliner**[†]
[†]Rochester Institute of Technology, [°]Oregon Health & Science University
{emilypx,dgg5503}@rit.edu, vansantj@ohsu.edu

Abstract

A common test administered during neurological examination is the semantic fluency test, in which the patient must list as many examples of a given semantic category as possible under timed conditions. Poor performance is associated with neurological conditions characterized by impairments in executive function, such as dementia, schizophrenia, and autism spectrum disorder (ASD). Methods for analyzing semantic fluency responses at the level of detail necessary to uncover these differences have typically relied on subjective manual annotation. In this paper, we explore automated approaches for scoring semantic fluency responses that leverage ontological resources and distributional semantic models to characterize the semantic fluency responses produced by young children with and without ASD. Using these methods, we find significant differences in the semantic fluency responses of children with ASD, demonstrating the utility of using objective methods for clinical language analysis.

1 Introduction

Semantic fluency tasks, in which patients undergoing neuropsychological evaluation must list as many items as possible in a particular semantic category in a fixed, brief period of time, are widely used by clinicians to evaluate language, development, and cognition. Performance on such tasks is usually measured in terms of the raw number of appropriate items produced. A more detailed analysis of these lists, however, can reveal patterns associated with a variety of neurological conditions, including autism, dementia, and schizophrenia.

Semantic fluency responses hold particular promise for shedding light on the language of children with autism spectrum disorder (ASD). ASD has been associated with atypical semantics and pragmatic expression since the condition was was first identified over 70 years ago (Kanner, 1943). One linguistic feature of ASD, referenced in many of the diagnostic instruments for the disorder, is the use of words that are meaningful but unexpected (Lord et al., 2002; Rutter et al., 2003), a phenomenon that could play an important role in the production of semantically related words.

In this paper, we present NLP-informed approaches for automatically approximating the subjective manual methods described in the psychology literature for analyzing semantic fluency responses. Applying these methods to data collected from young children with and without ASD, we find that none of the standard manual measures of semantic fluency are able to distinguish children with ASD from those without. Several computationally derived measures, however, are significantly different between diagnostic groups. These results indicate that computationally derived measures of semantic fluency tap into subtle differences that would be difficult to detect using standard manual metrics, lending support for the clinical utility of computational linguistic analysis.

2 Background

The semantic fluency task is a subtype of a more general word-generation task commonly referred to as verbal fluency. In such tasks, a participant must verbally produce a list of words belonging to some category (e.g., animals) within a predetermined amount of time, usually 60 seconds. Performance on verbal fluency tasks has been correlated with executive function, and differences in verbal fluency scores have been noted in a variety of neurological conditions including dementia

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 32–37
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2006

(Henry et al., 2004), schizophrenia (Frith et al., 1995), and autism (Turner, 1999; Geurts et al., 2004; Spek et al., 2009; Begeer et al., 2014).

The rate at which speakers generate words in a semantic fluency response has been observed to vary throughout the timed period, typically with several related words being produced in close succession followed by a pause before a new burst of related words (Bousfield et al., 1954). Troyer et al. (1997) proposed two cognitive processes underlying this pattern: clustering and switching. Clustering refers to the tendency of speakers to list words in clusters according to their membership in a particular subcategory of the larger semantic category (e.g., *pets* for the larger category of *animals*). Switching is the decision made by the speaker to abandon a subcategory when it has been exhausted and to list items in a new subcategory.

Autism is associated with deficits in executive function, and thus we should expect to see consistent patterns demonstrating deficits in semantic fluency performance in the ASD population. Several studies have found overall weaker performance, in terms of raw item count, in individuals with ASD (Turner, 1999; Geurts et al., 2004; Spek et al., 2009); other more recent studies, however, have not been able to replicate this finding (Lopez et al., 2005; Inokuchi and Kamio, 2013; Begeer et al., 2014). Similarly conflicting results have been reported when evaluating the semantic relatedness of adjacent words, with some finding smaller clusters in ASD (Turner, 1999), some finding larger clusters (Begeer et al., 2014), and still others finding no differences (Spek et al., 2009).

One likely source of these discrepancies is the subjectivity inherent in the cluster assignment task. Troyer et al. (1997) provide examples of common clusters and their member animals, but they note the difficulty in assigning items to subcategories, explaining that their proposed subcategories were not generated using any existing taxonomy but instead grew organically out of the patterns observed in the data. An additional complication is that a word's subcategory membership is dependent on its context. The word *camel*, for instance, could be assigned to any number of categories (e.g., *desert animal*, *zoo animal*), depending on the nearby words. This is particularly problematic when analyzing the responses of children, whose semantic categories might not align with those of an adult annotator.

In response to these challenges, some recent work has focused on modeling the cluster-switch behavior using computational linguistic methods, in particular, using latent semantic analysis to calculate the semantic similarity between adjacent words. Mean scores over these similarity values can capture a individual's tendency to use a naming strategy relying on similarity (Nicodemus et al., 2014; Rosenstein et al., 2015). Other work has focused on setting thresholds over these similarity values in order to delineate the boundaries between clusters or chains of related words (Rosenstein et al., 2015; Pakhomov and Hemmy, 2014). None of these studies, however, has compared the output of the automated methods to manual annotations in order to determine their accuracy. Furthermore, the thresholds used for cluster boundary identification in these studies were set by "rule of thumb" rather than empirically or probabilistically.

To our knowledge, this is the first attempt to use distributional semantic models to analyze semantic fluency responses in children with autism spectrum disorder. More importantly, it is the first study that uses machine learning to validate the utility of these models for replicating and, perhaps improving upon, human annotation methods of semantic fluency responses.

3 Data

The participants in this study were 22 children with typical development (TD) and 22 high-functioning children with ASD, ranging in age from 4 to 9 years. ASD was diagnosed via clinical consensus according to the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV-TR) criteria for Autistic Disorder (American Psychiatric Association, 2000) and the established thresholds on two commonly used diagnostic instruments: the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002) and the Social Communication Questionnaire (SCQ) (Rutter et al., 2003). None of the participants analyzed here met the criteria for language impairment, and the two groups were selected so that there were no statistically significant differences (via two-tailed t-test) between the groups in chronological age, verbal IQ, and full scale IQ. In addition to the experimental corpus, we had access to a development set of 55 semantic fluency responses that were discarded after the groups were matched on these three criteria.

During administration of the task, the clinician asked the child to name as many animals as he could as quickly as possible. The children's responses were timed and recorded. The audio was then transcribed by a speech-language pathologist, and the transcripts were reviewed to remove extraneous dialogue and to standardize spelling. Two manual annotations were performed: (1) semantic clusters (Troyer et al., 1997), in which a cluster consists of two or more animals belonging to same subcategory (*giraffe, elephant, lion*); and (2) semantic chains (Pakhomov and Hemmy, 2014), in which each animal shares something in common at least with the immediately preceding animal (*elephant, lion, cat*). Inter-annotator agreement for labeling cluster boundaries according to the Troyer criteria was low (Cohen's $\kappa < 0.4$); we therefore limit our discussion to semantic chains, whose boundaries were labeled with more substantial agreement ($\kappa = 0.71$).

4 Features

4.1 Manually derived measures

Performance on a verbal fluency task is normally evaluated by counting the number of unique items produced in the designated time period. Credit is given both to a general category such as *fish* and to examples of that category, such as *salmon*; however, a morphological or descriptive variation of another item (e.g., *doggy* for *dog*) is considered a repetition and does not contribute to the total. We report this count, along with the number of semantic chains and mean length of semantic chain.

4.2 Semantic similarity measures

There are a number of ways to measure the semantic similarity between two words, some relying on manually curated knowledge bases and other derived distributionally from large text corpora. A high mean similarity between adjacent word pairs in a list of words might suggest that the list contains a small number of large clusters of strongly related words (a cluster-and-switch strategy) or a sequence of items each of which is closely related to the previous item but not necessarily to the items before that (a chaining strategy). In either case, the participant is tapping into semantic subcategories when producing his response. A lower mean similarity should indicate that a participant has produced a large number of small clusters or has selected items from the larger category seemingly at random.

One possible way to capture relatedness is by using a manually curated lexical ontology that implicitly encodes the similarity between pairs of words, such as WordNet (Fellbaum, 1998). Various algorithms have been proposed for assigning similarities scores for two synsets in WordNet by traversing the hierarchical trees connecting those synsets. Here we calculate the mean path similarity for each adjacent word pair in a participant's generated wordlist. Words not appearing in WordNet were manually replaced with equivalent synsets (e.g., *puppy dog* was replaced with *puppy*). When multiple synsets were associated with a given item, we used the first synset whose hypernym included the synset for *animal* or *imaginary being* (e.g., *pegasus*).

One disadvantage inherent in the WordNet ontology of animal names is that it is derived from the biological taxonomy of the animal kingdom; that is, the degree to which two animals are semantically related within WordNet is determined primarily by their biological similarity and not by semantic features (e.g., region of origin, usual habitat) that a non-zoologist might use to organize animals names. In order to model multiple dimensions of similarity, we turn to the use of vector space models. We explore two vector-space representations: latent semantic analysis (LSA) (Landauer et al., 1998) and continuous space neural word embeddings (Bengio et al., 2003). Using the gensim Python library (Řehůřek and Sojka, 2010), we built an LSA model and a word2vec model, both with 400 dimensions but otherwise using default parameters settings, on the full text of Wikipedia downloaded in November, 2016. For each model, we take the mean of the set of cosine similarities between each adjacent pair of items in a participant's response. We also calculate the mean similarity over 100 random permutations of a participant's wordlist to capture "global coherence", as proposed by Nicodemus et al. (2014).

4.3 Measures of identifying semantic chains

Previous work in using word embeddings to model clustering relied on a simple cosine similarity threshold, determined heuristically (set arbitrarily 0.9 in Rosenstein et al. (2015), and at the 75th percentile in Pakhomov and Hemmy (2014)), in which a cluster boundary is inserted between any two adjacent words whose similarity did not exceed that threshold. We instead propose to empirically determine the optimal value of such a thresh-

Feature	TD	ASD	t
Raw count	12.0	10.2	1.043
Manual chain count	6.14	4.86	1.603
Manual chain length	2.0	2.13	-0.572
WordNet path similarity	0.169	0.1697	0.1721
LSA cosine similarity	0.365	0.308	1.636
LSA coherence	0.311	0.248	1.934∗
w2v cosine similarity	0.427	0.392	1.710∗
w2v coherence	0.409	0.375	1.530
LSA chain count	4.09	4.31	-0.316
LSA chain length	3.38	1.87	2.310∗
w2v chain count	4.14	4.41	-0.3800
w2v chain length	3.07	1.91	1.9265∗
SVM chain count	4.09	4.86	-1.0894
SVM chain length	3.66	2.19	2.4164∗

Table 1: Mean values by diagnostic group for semantic fluency metrics ($*p$ <0.05, one-tailed).

old. First, while leaving one subject out, we iteratively sweep through a range of possible values for the threshold to determine the value that maximizes the accuracy of semantic chain boundary identification for the rest of the participants. We then apply that threshold to the left-out subject.

In addition to thresholding over individual similarity metrics, we also use three similarity metrics (WordNet path similarity, LSA cosine similarity, and word2vec cosine similarity) as features within a support vector machine to classify any pair of adjacent words as either containing a semantic chain boundary or as belonging to the same semantic chain. Using all two-word sequences found in the children's responses and the manual indications of the locations of cluster boundaries, we perform leave-one-out cross validation to predict whether the second word in each word pair represents the start of a new chain or a continuation of the previous chain.

Although the methods all achieved reasonable boundary identification accuracy, with AUC ranging from 0.65 to 0.8, we note that the goal of determining cluster boundaries in this way is not to replicate human cluster boundary insertion, which we know to be subjective and difficult to perform reliably. Rather, we are attempting to develop an objective way to insert boundaries that does not rely on an annotator's ability to infer another individual's semantic organization of the world.

5 Results

Table 1 shows the mean value for each group and the t-statistic for each of the features. In contrast to some previous work (Turner, 1999; Geurts et al., 2004; Spek et al., 2009), we find no between-group differences in raw item count. These results, however, support other work that did not find such differences when comparing groups matched on verbal ability, as our groups are (Lopez et al., 2005; Inokuchi and Kamio, 2013).

Mean cosine similarity derived using the word2vec model is significantly different between the two groups, with the TD group showing a higher mean similarity between adjacent items. We also see that the global coherence measure, derived by taking the mean similarity over 100 random orderings of each list, is significantly higher in the TD group when derived using LSA.

Although there are no between-group differences in the manually derived measures of chain count and chain length, we find differences in chain length when derived using both thresholding over similarity measures and machine learning. In all three cases, children with typical development have longer semantic chains than children with ASD, suggesting that TD children employ the semantic chaining strategy that is reportedly preferred by neurotypical adults. In short, there are differences in the semantic fluency responses of young children with ASD, and these differences would be difficult to reliably detect without appealing to computational techniques.

Figure 1 shows two semantic fluency responses, one produced by a child with ASD and one by a child with TD, with plots indicating the cosine similarities between adjacent words derived from both the LSA and word2vec models. Semantic chain boundaries proposed by the SVM are indicated with vertical dashed lines. Note that LSA and word2vec similarity values are only somewhat correlated, underscoring the potential utility of combining the two scores for chain boundary identification. As expected given the results in Table 1, the child with ASD has generally lower cosine similarity scores and many more chain boundaries than the typically developing child.

6 Discussion and future work

One problem with applying the chaining and clustering paradigms to children is that the semantic features linking animals for a child might be very different those of adults. Well over half of the children in this study included the sequence *cat, bear* or *bear, cat*, despite the lack of clear relation between the two words from an adult's perspective. We found, however, that our automated methods usually grouped these two words together, recognizing some similarity that adults seem to miss. At

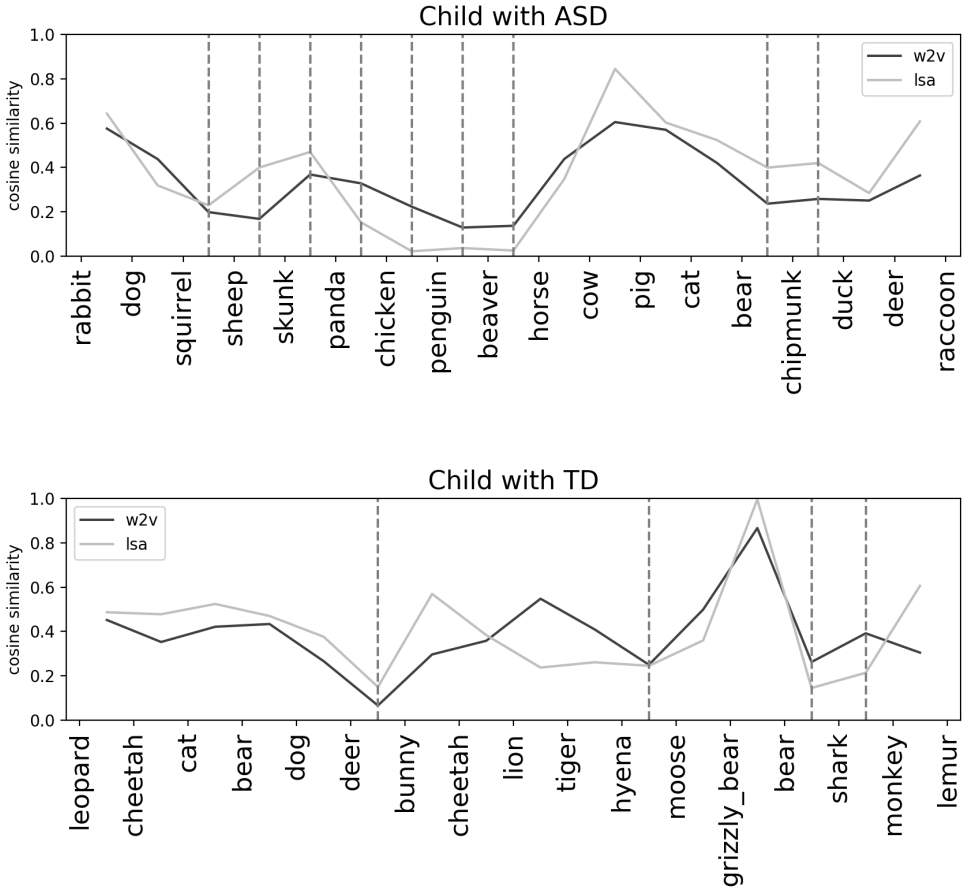

Figure 1: Plots of successive word-pair cosine similarity values derived using LSA and word2vec models for a child with ASD (upper panel) and a child with TD (lower panel). Vertical dashed lines indicate semantic chain boundaries proposed by the SVM.

the same time, relying on large corpora of adult-focused texts may introduce problems: the lowest similarity values found in our data set involved the word *turkey*, suggesting a preponderance in the data of the country rather than the bird. More sensitive text normalization methods could likely resolve this problem, but we also plan to build LSA and neural word embedding models using child language data (e.g., the CHILDES corpus (MacWhinney, 2000)) and child-oriented texts in the public domain.

Future work will focus on improving our methods for identifying semantic chains while accounting for different methods of semantic organization by combining information gained from the rich but out-of-domain data scenarios described here with in-domain experimental data. In addition to incorporating more child-oriented training data, we plan to use graph-based models to capture the ways in which speakers proceed through the semantic space (Abbott et al., 2015).

As the contradictory results in the literature indicate, the precise nature of the linguistic deficits associated with ASD is somewhat unclear. Many of the most widely reported linguistic deficits fail to obtain when participants are carefully matched, particularly on verbal IQ. The atypical language features that do persist under strict matching are usually semantic or pragmatic and, hence, more difficult to detect using easily scored standard language assessment instruments. Methods leveraging large corpora that reflect neurotypical language use may prove to be one of the more useful tools for identifying atypical language in ASD.

Acknowledgments

This work was supported in part by NIH grants R01DC013996, R01DC012033, and R01DC007129. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH.

References

Joshua T Abbott, Joseph L Austerweil, and Thomas L Griffiths. 2015. Random walks on semantic networks can resemble optimal foraging. *Psychological Review* 122(3):558–569.

American Psychiatric Association. 2000. *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, Washington, DC.

Sander Begeer, Marlies Wierda, Anke M Scheeren, Jan-Pieter Teunisse, Hans M Koot, and Hilde M Geurts. 2014. Verbal fluency in children with autism spectrum disorders: Clustering and switching strategies. *Autism* 18(8):1014–1018.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3:1137–1155.

Weston Ashmore Bousfield, CHW Sedgewick, and BH Cohen. 1954. Certain temporal characteristics of the recall of verbal associates. *The American Journal of Psychology* 67(1):111–118.

Christian Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

CD Frith, KJ Friston, S Herold, D Silbersweig, P Fletcher, C Cahill, RJ Dolan, RS Frackowiak, and PF Liddle. 1995. Regional brain activity in chronic schizophrenic patients during the performance of a verbal fluency task. *The British Journal of Psychiatry* 167(3):343–349.

Hilde M Geurts, Sylvie Verté, Jaap Oosterlaan, Herbert Roeyers, and Joseph A Sergeant. 2004. How specific are executive functioning deficits in attention deficit hyperactivity disorder and autism? *Journal of child psychology and psychiatry* 45(4):836–854.

Julie D Henry, John R Crawford, and Louise H Phillips. 2004. Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia* 42(9):1212–1222.

Eiko Inokuchi and Yoko Kamio. 2013. Qualitative analyses of verbal fluency in adolescents and young adults with high-functioning autism spectrum disorder. *Research in Autism Spectrum Disorders* 7:1403–1410.

Leo Kanner. 1943. Autistic disturbances of affective content. *Nervous Child* 2:217–250.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25(2-3):259–284.

Brian Lopez, Alan Lincoln, Sally Ozonoff, and Zona Lai. 2005. Examining the relationship between executive functions and restricted, repetitive symptoms of autistic disorder. *Journal of Autism and Developmental Disorders* 35(4).

Catherine Lord, Michael Rutter, Pamela DiLavore, and Susan Risi. 2002. *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services, Los Angeles.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Kristin K Nicodemus, Brita Elvevåg, Peter W Foltz, Mark Rosenstein, Catherine Diaz-Asper, and Daniel R Weinberger. 2014. Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. *Cortex* 55:182–191.

Serguei V.S. Pakhomov and Laura S. Hemmy. 2014. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex* 55:97–106.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. pages 45–50.

Mark Rosenstein, Peter W. Foltz, Anja Vaskinn, and Brita Elvevg. 2015. Practical issues in developing semantic frameworks for the analysis of verbal fluency data: A norwegian data case study. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*. pages 124–133.

Michael Rutter, Anthony Bailey, and Catherine Lord. 2003. *Social Communication Questionnaire (SCQ)*. Western Psychological Services, Los Angeles.

Annelies Spek, Tjeerd Schatorjé, Evert Scholte, and Ina van Berckelaer-Onnes. 2009. Verbal fluency in adults with high functioning autism or asperger syndrome. *Neuropsychologia* 47(3):652–656.

Angela K Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology* 11(1):138–146.

Michelle A Turner. 1999. Generating novel ideas: Fluency performance in high-functioning and learning disabled individuals with autism. *Journal of Child Psychology and Psychiatry* 40(2):189–201.