Psychiatry Research 273 (2019) 767–769

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Short communication

Updating verbal fluency analysis for the 21st century: Applications for psychiatry

Terje B. Holmlund[a,*], Jian Cheng[b], Peter W. Foltz[c,d], Alex S. Cohen[e], Brita Elvevåg[a,f]

[a] Department of Clinical Medicine, University of Tromsø, UNN Åsgård, Postbox 6124, 9291 Tromsø, Norway
[b] Analytic Measures Inc., Palo Alto, CA, USA
[c] Institute of Cognitive Science, University of Colorado Boulder, USA
[d] Pearson PLC, London, England
[e] Department of Psychology, Louisiana State University, USA
[f] Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway

ABSTRACT

Evaluating patients' verbal fluency by counting the number of unique words (e.g., animals) produced in a short-period (e.g., 1–3 min) is one of the most widely employed cognitive tests in psychiatric research. We introduce new methods to analyze fluency output that leverage modern computational language technology. This enables moving beyond simple word counts to charting the temporal dynamics of speech and objectively quantifying the semantic relationship of the utterances. These metrics can greatly expand the current psychiatric research toolkit and can help refine clinical theories regarding the nature of putative language differences in patients.

1. Introduction

Language is affected by a large number of cortical disorders, and in psychiatric disorders specifically it is the medium through which many symptoms are expressed and thus measured. Skilled clinicians can detect and assess such symptoms intuitively in their daily practice, but numerous tests have also been developed to be easily administered. Although these tests were not designed to understand serious mental illness *per se*, due to the ease of administration they are frequently employed to shed additional light on the nature of the presenting language anomalies. We investigated how emerging technology could be leveraged for new opportunities in both the administration and analysis of the verbal fluency task, and how descriptions of high temporal resolution could inspire new insights into the dynamical nature of verbal fluency task performance.

The category verbal fluency task is one of the most widely used language tasks in psychiatric research. In this task, participants are asked to produce as many exemplars to a few noun cues (e.g., animals) for a specified duration (e.g., one minute) for each category. The experimenter then typically writes down all the exemplars and assigns a point for each unique exemplar produced. Such an operationalization ignores the fundamental fact that even on such a simple task there is a remarkable amount of structure and temporal information (Bousfield and Sedgewick, 1944; Bousfield et al., 1954). However, the exact timing between utterances has rarely been formally examined.

1.1. Moving beyond stopwatches and pencils to automatic speech transcription with accurate timing measurements

By collecting speech output digitally it is possible to use automatic transcription methods which work because of the statistical properties of language (e.g., word frequencies) derived from large scale language corpora. Although currently available tools are not designed specifically to analyze the verbal fluency task, when we calibrate the statistical model on relevant task words (i.e., what words are likely to occur, such as "elephant", "squirrel" and "giraffe") we have found that the automatic speech to text transcription accuracy is significantly improved (with a mere 6% word error rate - Holmlund et al., 2019). Additionally, it is possible to time-stamp each word utterance (using forced alignment tools). Although the current temporal resolution of publicly available speech recognition services are on the order of ± 100 ms and thus not really adequate for charting the flow of thought, a higher temporal resolution on the order of ± 10 ms is possible to develop in-house for specific tasks. Therefore, the use of automatic speech recognition combined with accurate temporal markers of the utterances can radically transform the manner in which verbal fluency data are analyzed.

[*] Corresponding author.
E-mail address: terje.holmlund@uit.no (T.B. Holmlund).

https://doi.org/10.1016/j.psychres.2019.02.014
Received 12 January 2019; Received in revised form 6 February 2019; Accepted 6 February 2019
Available online 07 February 2019

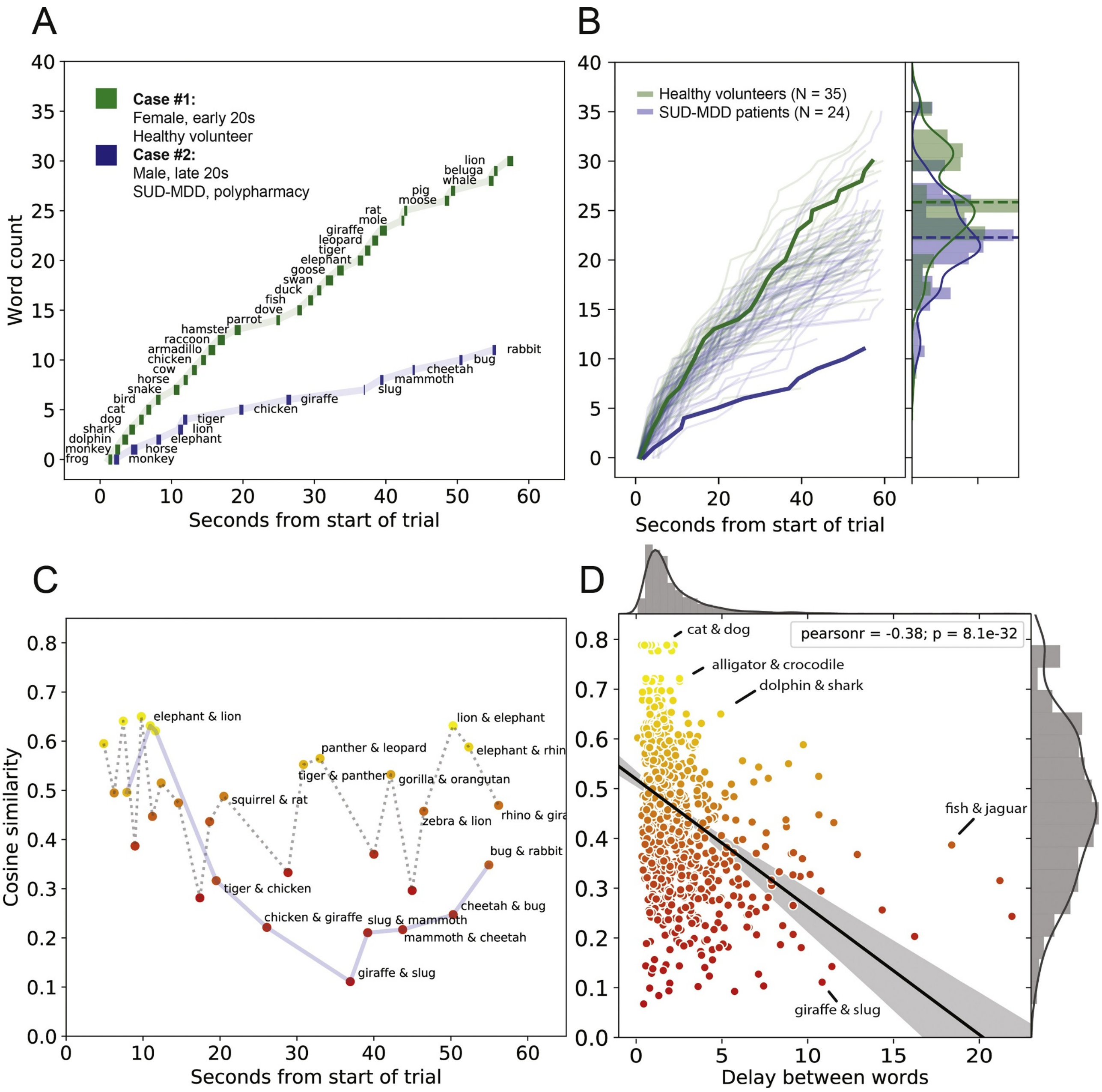

Fig. 1. The temporal sequence of utterances in example trials from the one-minute category fluency task. *Panel A:* Each individual response utterance is plotted on the timeline from the start of the trial (i.e., "0") to the end of the trial (i.e., "60"). The left and right margins of the colored boxes represent the duration of the respective onsets and offsets of the responses, demonstrating how a verbal response of "pig" is shorter than a response of "armadillo". The vertical axis represents the word count, and periods with a quick succession of words results in a steeper trajectory. *Panel B:* The two cases (Case #1 and Case #2) are plotted alongside the individual sequences from the other participants in the two groups (healthy, patients). The distribution of total word counts suggests the expected pattern where healthy participants produced more words (green, mean = 25) as compared to patients with substance use disorders and major depressive disorder comorbidity (SUD-MDD, blue, mean = 19). *Panel C:* The similarity between successive responses can fluctuate over time. The full blue line shows how the similarity between responses from Case #2 decreases over the course of the trial (e.g., "chicken and giraffe", "giraffe and slug"), but ends with words that more commonly occur together in sentences ("bug and rabbit"). The dotted line shows a trial from another participant with more consistently similar response words. *Panel D:* The semantic similarity between two successive words is related to the inter-word delay. All word pairs produced by the patients are presented as a point in the scatterplot, where word pairs that are more similar are higher up the vertical axis, and pairs produced with longer inter-word delays are further to the right on the horizontal axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1.2. Making sense of the semantic associations in free speech: measuring distances in semantic space

Traditionally psychiatry has concerned itself with analyzing the meaning of what people say, using a variety of theories and (arguably subjective) hand-coding methods. This is motivated by the notion that examining associations will provide some insight into the connection of ideas. The underlying theory is that well-organized and closely associated ideas and concepts will be generated faster (i.e., that meaning and temporal dynamics are intertwined). Today such conceptualizations of semantic associations can be formally examined by the use of large scale language corpora. These corpora can be leveraged to create vector representations of individual words and build semantic spaces which can then be used as tools to compute semantic distances between words and concepts. In psychiatry there has been a growing use of such computational methods to derive a variety of complex natural language processing assays, for example it has been shown that thought disorder can be associated with lower word to word coherence (Elvevåg et al., 2007). Critically, it is possible to explore whether such metrics are predictive of clinical state changes in a manner that is similar to human clinicians, or perhaps even more sensitive (Bedi et al., 2015; Rosenstein et al., 2015; Foltz et al., 2016; Corcoran et al., 2018).

2. Methods

We sought to chart the temporal dynamics of speech production and objectively quantify the semantic relationship of response words in a category fluency task, collected as part of a large project that developed and implemented a mobile software application for remote, frequent and self-administered psychological assessment to monitor mental states using smart devices (Holmlund et al., 2019). Participants were given the following vocal prompt from the smart device: "Name as many animals as you can. Any kind of animal, as many different animals as you can think of. You have up to 1 min; start now". The task was administered between one and three times to a subset of participants. We analyzed responses from 24 male patients recruited from an inpatient population with substance use disorders and major depressive disorder comorbidity (mean age = 39.1 years, SD = 10.7). For comparison, we analyzed responses from 35 presumed healthy volunteer participants (who were undergraduate students recruited from a large public university), substantially younger (mean age = 19.5, SD = 1.5), and predominantly female (87.9%).

One minute long recordings of responses were made via the microphone in smart devices at a sample rate of 16,000 Hz and saved in a .flac-format. These recordings were transcribed and response-words timestamped with a forced temporal alignment procedure using the Kaldi speech recognition toolkit (Povey et al., 2011). Non-animal words (e.g., "Let's", "see", "what", "else") were removed, and the remaining words were lemmatized (i.e., converted to their stem, e.g., "cats" to "cat") with the Natural Language Toolkit (Loper and Bird, 2002).

We derived an index of semantic associations between word pairs using a set of publicly available GloVe word vectors (Pennington et al., 2014), computing relationships between vector representations of the individual responses. This method provides a quantified measure of the degree of semantic association between words, based on how they co-occur in similar contexts in a given corpus. To base the analysis on a corpus with a wide variety of animal-word sources, we used a set of pre-trained word vectors calculated from approximately 42 billion tokens from the entire internet, courtesy of the Common Crawl project (Pennington et al., 2014). The GloVe word vectors were imported in a word2vec-format (Mikolov et al., 2013), and word-pair cosine similarity was derived using the Gensim python package (Řehůřek and Sojka, 2010). The range of the measure was 0–1, such that words that often co-occur such as "lion and tiger" got a score closer to one, while words that seldom occur together such as "giraffe and slug" got a score closer to zero.

3. Results

Overall, healthy participants generated 25 words (range = 8–36) and patients with substance use disorder and major depressive disorder comorbidity generated on average slightly fewer (mean = 19 words; range = 11–36; $t(57) = -4.0$; $p < 0.001$). To illustrate how the temporal trajectories can differ between individuals we chose two noteworthy examples, namely a healthy person who generated 31 words versus a patient who generated only 12 words in the one minute period (Fig. 1A). We also illustrate how much variability there can be in individual data as compared to group means (Fig. 1B). Over the course of a minute the word to word similarity can fluctuate considerably (Fig. 1C) as a function of whether there is a high level of similarity between successive word pairs or not. Not surprisingly, the semantic coherence between two successive words was related to the speed of speech (Fig. 1D). We found a significant negative correlation (Entire sample: $r = -0.36$, $p < 0.001$; patients: $r = -0.38$, $p < 0.001$) indicating the tendency for longer pauses between semantically dissimilar words. Thus, word pairs with a high similarity index, such as "cat and dog", were spoken very quickly and thus seldom had inter-word delays longer than three seconds.

4. Discussion

We have introduced preliminary analytic methods that showcase how the currently untapped temporal and semantic information in a simple category fluency task can be formally extracted. While the results are encouraging, much remains to be explored and established before widespread implementation. As an increasing number of statistical and mathematical approaches to language emerge that are applied within psychiatry (Elvevåg et al., 2017), it remains extremely important to validate and calibrate these methods (Foltz et al., 2016). Most notably, it is critical to establish that these assays and analytic methods are consistent across data and analysis platforms, that the subtle nuances of the various natural language processing techniques are not affecting the results in a manner that is unexpected, and that the value of the extra technical effort is established to be worthwhile.

Acknowledgments

This research was funded by a grant from the Research Council of Norway to Brita Elvevåg (#231395).

References

Bedi, G., Carrillo, F., Cecchi, G., Slezak, D., Sigman, M., Mota, N., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, C., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 1 (1), 15030.

Bousfield, W.A., Sedgewick, H.W., 1944. An analysis of sequences of restricted associative responses. J. Gen. Psychol. 30, 149–165.

Bousfield, W.A., Sedgewick, H.W., Cohen, B.H., 1954. Certain temporal characteristics of the recall of verbal associates. Am. J. Psychol. 67, 111–118.

Corcoran, C., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D., Bearden, C.E., Cecchi, G., 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry 17 (1), 67–75.

Elvevåg, B, Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr. Res. 93 (1–3), 304–316.

Elvevåg, B, Foltz, P.W., Rosenstein, M, Ferrer-i-Cancho, R, De Deyne, S, Mizraji, E., Cohen, A.S., 2017. Thoughts about disordered thinking: measuring and quantifying the laws of order and disorder. Schizophr. Bull. 43, 509–513.

Foltz, P.W., Rosenstein, M., Elvevåg, B., 2016. Detecting clinically significant events through automated language analysis: quo imus. NPJ Schizophr. 2 (1), 15054.

Holmlund, T.B., Foltz, P.W., Cohen, A.S., Johansen, H., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E., Elvevåg, B., 2019. Moving psychological assessment out of the controlled laboratory setting: practical challenges. Psychol. Assess. 31 (3). https://doi.org/10.1037/pas0000647.

Loper, E., Bird, S., 2002. NLTK: the natural language toolkit. In: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, July, 2002.

Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient estimation of word representations in vector space. arXiv:1301.3781v3 [cs.CL], 2013.

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The KALDI speech recognition toolkit. In: Proceedings IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hawaii, USA, December 2011.

Řehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Witte, R., Cunningham, H., Patrick, J., Beisswanger, E., Buyko, E., Hahn, U., Verspoor, H., Coden, A. (Eds.), Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks Valletta, Malta, May, 2010, pp. 45–50.

Rosenstein, M., Foltz, P.W., DeLisi, L.E., Elvevåg, B., 2015. Language as a biomarker in those at high-risk for psychosis. Schizophr. Res. 165, 249–250.