Mind your models! Distributional semantic models for the analysis of verbal fluency tasks in Schizophrenia Spectrum Disorders

Chiara Barattieri di San Pietro (University of Milano-Bicocca)
Giovanni de Girolamo (IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli)
Claudio Luzzatti (University of Milano-Bicocca)
Marco Marelli (University of Milano-Bicocca)
chiara.barattieridisanpietro@unimib.it

Verbal fluency (VF) tasks are widely employed to assess executive dysfunctions and impaired integrity of the semantic store in Schizophrenia Spectrum Disorders (SSDs). Studies aimed at defining the impact of these two neurocognitive factors with fine-grained measures of VF (Troyer et al., 1997) yielded heterogeneous results. Estimates from computational semantic models showed promising results, but consideration should be given to the kind of "meaning" these models can capture, in relation to the type of semantic associations elicited through different VF tasks.

This study aimed at assessing the classification performance of semantic measures from different computational architectures in discriminating people with and without SSDs, against standard manual annotation.

Two VF tasks (semantic – SVF, and generative associative naming – GAN) were administered to a group of people with SSDs (N = 37) and a matched group of healthy participants (HPs). In an SVF, the participant's performance is rated by the number of unique correct words produced in 1 min within a given semantic category (i.e., "animals"). In a GAN task, participants are asked to produce as many words as possible related to four words stimuli (i.e., "cat", "shoe", "rain", "strike") in 2 min; performance is rated by counting the number of correct answers (as per reference list: Spinnler & Tognoni, 1987; Bandera et al., 1991).

Outputs were analyzed by computing size of semantic clusters, number of switches between clusters, and coherence between responses. Measures were computed i) manually by a human rater following standard procedure (Troyer et al., 1997) and ii) by a set of algorithms relying on semantic representations derived from three semantic spaces: two were generated by a neural network (word2vec – Mikolov et al., 2013) and one by Latent Semantic Analysis model (LSA – Landauer & Dumais, 1997). All three semantic models used the untagged itWac corpus (about 1.9 billion tokens; Baroni et al., 2009), previously converted to UTF8 lower case and tokenized, having removed all special characters (except for vowels with orthographically marked stresses). The neural network models (Word-Embeddings Italian Semantic Space 1 and 2 – "WEISS1" and "WEISS2", http://meshugga.ugent.be/snaut-italian) are based on the word2vec CBOW model and consider words with a minimum frequency of 100. WEISS1 is built with 400 dimensions on a 9-word window; WEISS2 with 200 dimensions on a 5-word window. For the LSA model, we randomly extracted an untagged set of 91,058 documents from the itWac corpus (so as to match the TASA (http://lsa.colorado.edu) settings), comprising the same set of words (N = 180,080) of WEISS. The creation of a matrix of co-occurrences was carried out using the DISSECT toolkit (Dinu et al., 2013), and applying a Positive Pointwise Mutual Information weighting scheme, followed by dimensionality reduction by Singular Value Decomposition. We set the number of dimensions at 300.

To compute the number of switches and the mean size of clusters automatically, we created an algorithm that retrieves from the semantic space the vector corresponding to a given word produced and calculates the cosine proximity between it (w_n) and the next word (w_n+1) in the verbal fluency output. It then compares this value to a pre-specified threshold (the 10^{th}, 30^{th}, 50^{th}, 70^{th}, and 90^{th} quantile of the mean cosine proximity distribution of the whole dataset) and, if found equal or above such value, the two words are considered as part of a cluster. In this case, a new vector, representing the cluster meaning, is computed by averaging the vectors of all the words currently considered as part of the cluster. The algorithm then compares the next word in the list against this mean vector and repeats the last two steps (computing mean vector and comparing it to the next word) until the cosine proximity to the next word eventually falls below the threshold. In this case, the function registers a switch

and re-starts the process with a new cluster. Coherence of responses was calculated as the mean cosine distance between words at different distances (w_n+1, w_n+3, w_n+5, w_n+7).

A set of classifiers (logistic regressions models) was created by entering fluency measures (number of switches and mean size of clusters as interacting variables, and coherence as covariate) as predictors, and group membership as categorical dependent variable. Area Under the Curve (AUC, Table 1) values from Receiver Operating (ROC; Figure 1) graphs were compared to identify the best performing classifiers.

Compared to control participants, in both tasks people with SSD produced fewer switches and showed higher between-word coherence. In the GAN task, HPs also produced smaller clusters than SSDs participants.

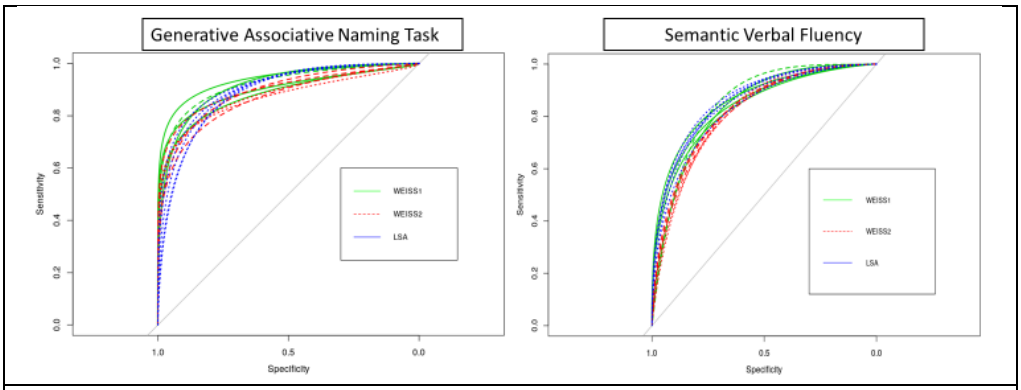

Figure 1. ROC curves of classifiers based on VF measures in a GAN and an SVF task to classify people with/without SSD.

	SVF	GAN
Semantic representation	**AUC**	**AUC**
No. of switches * Mean Size of clusters - Manual scoring	.76	.82
No. of switches * Mean Size of clusters + coherence - WEISS1	.86	**.94**
No. of switches * Mean Size of clusters + coherence - WEISS2	.84	.91
No. of switches * Mean Size of clusters + coherence - LSA	**.87**	.92

Table 1. Best AUC values from ROC analysis of classifiers for participants with/without SSD

We observed a dissociation of performance of the semantic models between the tasks. The LSA semantic representations appeared best equipped to inform a classifier based on the SVF output. During this task, participants may choose to produce syntagmatically-related words, that are known to be best represented by words-by-document co-occurrences. Inversely, WEISS1 outperformed other models in predicting participants' class from fluency values derived from a GAN task. Here, participants may have produced words entailing principally paradigmatic relations. Semantic representations based on narrow contextual windows are known to be best suited to capture this kind of relation. In this sense, the type of semantic relations that a VF can prompt should be taken into consideration when choosing the semantic representation to employ. Most importantly, scores based on computational representations outperformed manual scoring, indicating that a taxonomic-based approach to identify semantic cluster is limited by the model of knowledge it entails.

References

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. In *Psychological Review* (Vol. 1).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. Retrieved from http://arxiv.org/abs/1310.4546

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. doi.org/10.1037/0894-4105.11.1.138