DISTRIBUTIONAL SEMANTICS, NLP, AND MACHINE LEARNING: A COMBINED APPROACH TO LANGUAGE ANALYSIS IN SCHIZOPHRENIA

Chiara Barattieri di San Pietro[1], Biagio Scalingi[1], Federico Frau[1], Giulia Agostoni[2], Margherita Bechi[2], Roberto Cavallaro[2,3], Marta Bosia[2,3], Nicholas Bianchini[4], Flavio Bertini[4], Valentina Bambini[1]

[1] Laboratory of Neurolinguistics and Experimental Pragmatics (NEP), University School for Advanced Studies IUSS, Pavia
[2] Schizophrenia Research and Clinical Unit, IRCCS San Raffaele Hospital, Milan, Italy
[3] School of Medicine, Vita-Salute San Raffaele University, Milan, Italy
[4] Department of Mathematical, Physical and Computer Sciences, University of Parma

Language is prone to modification in cognitive and mood disorders, and specific semantic and pragmatic features can be associated with clinical symptoms. To overcome the limit of qualitative analysis, applying Natural Language Processing (NLP) techniques and Distributional Semantics (DS) approaches to extract quantitative features from the language data could help identify reliable digital biomarkers for symptom severity assessment and treatment response. Here we present four studies employing different NLP techniques and DS models to capture latent and meaningful semantic associations in the linguistic production of people with schizophrenia, to be used with diagnostic and treatment response aims.

AUTOMATED CLUSTERING AND SWITCHING ALGORITHMS APPLIED TO SEMANTIC VERBAL FLUENCY DATA IN SSD

LANGUAGE, COGNITION AND NEUROSCIENCE
https://doi.org/10.1080/23273798.2023.2178662

Automated clustering and switching algorithms applied to semantic verbal fluency data in schizophrenia spectrum disorders

Chiara Barattieri di San Pietro[a], Claudio Luzzatti[a], Elisabetta Ferrari[b], Giovanni de Girolamo[b] and Marco Marelli[a]

[a]Department of Psychology, University of Milano-Bicocca, Milan, Italy; [b]Psychiatric Epidemiology and Evaluation Unit, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy

ABSTRACT
In the cognitive assessment of Schizophrenia Spectrum Disorders (SSD), the standard scoring method for Verbal Fluency (VF) tasks is the number of correct words produced. Finer-grained measures, such as the size of semantic clusters and the number of transitions between them, have been proposed to characterise the cognitive functions involved, but results based on human ratings are heterogeneous. The objective of this study was to develop a computational procedure based on Vector Space Models (VSMs) to assess the predictive ability of these fine-grained measures for class membership in SSD. A semantic VF task was administered to thirty-five people with SSD and a matched group of healthy participants, and their VF productions were characterised manually and using a set of ad-hoc algorithms. Computational estimates consistently showed higher predictive accuracy than models built on VF measures computed by a human rater and models built on the sole total number of words.

ARTICLE HISTORY
Received 28 June 2022
Accepted 1 February 2023

KEYWORDS
Verbal fluency; schizophrenia; latent semantic analysis; word2vec; vector space models

Abbreviations: AIC: Akaike Information Criterion; AUC: Area Under the Curve; CBOW: Continuous Bag-Of-Words; DSM-5: Diagnostic and Statistical Manual of Mental Disorders Fifth Edition; HP: Healthy Participants; LSA: Latent Semantic Analysis; NLP: Natural Language Processing; ROC: Receiving Operating Curve; SSD: Schizophrenia Spectrum Disorders; VF: Verbal Fluency; VSM: Vector Space Model

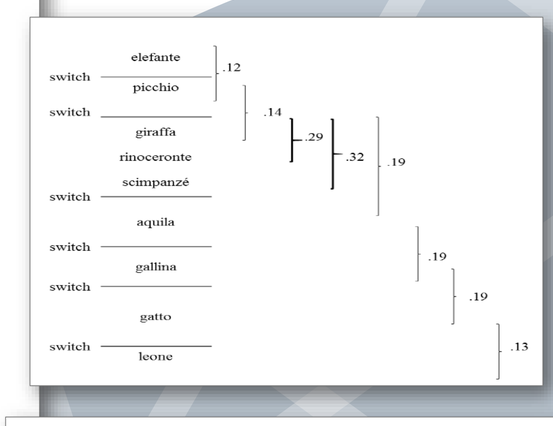
Introduction

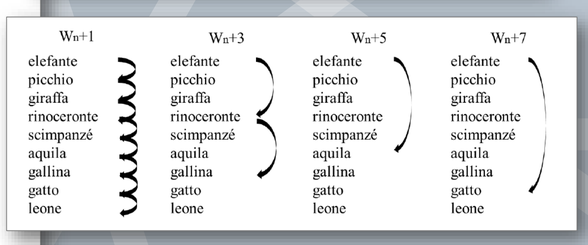
Schizophrenia Spectrum Disorders (SSD) are chronic psychotic disorders that can severely affect language and communication abilities (DeLisi, 2001; Kuperberg, 2010; Tavano et al., 2008). A commonly used method to study lexical retrieval is the Verbal Fluency (VF) task (Lezak et al., 2012; Neill et al., 2014). In its semantic version, the participant's performance is rated by the number of correct words produced in one minute for a given semantic category. Impairments in VF in people with SSD have been repeatedly reported in the literature (Bokat & Goldberg, 2003; Brébion et al., 2015; Elvevåg et al., 2010; Galaverna et al., 2016; Henry & Crawford, 2005; Juhasz et al., 2012; Ojeda et al., 2010; Sumiyoshi et al., 2005). Such impaired performance appears to be significantly related to the transition to psychosis in high-risk people (Fusar-Poli et al., 2012), and poor performance in VF tasks remains stable in individuals with SSD during the time course of the disorder (Robert et al., 1998).

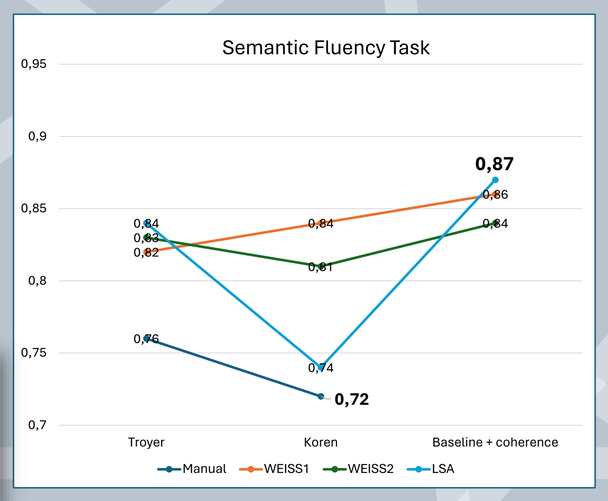
At least two main interacting cognitive functions are involved in VF: the semantic store system, supported by the left temporal lobe, and the executive control system, supported by left-hemisphere prefrontal regions. Approaches favouring the former over the latter in explaining VF deficits in people with SSD suggest the content of the semantic store to be deviant (Aloia et al., 1998; Bokat & Goldberg, 2003) and disorganised (Bozikas et al., 2005), with lack of conceptual knowledge (Goldberg et al., 1998; Paulsen et al., 1996; Rossell & David, 2006) and abnormal spreading activation (Moritz et al., 2002). In contrast, the increased number of perseverations (repetitions of earlier responses) and intrusions (the recollection of inappropriate information) suggests an increased susceptibility to interferences in recall strategies (Galaverna et al., 2016) and hence impaired executive functionality (Doughty & Done, 2009).

However, the number of correct words produced in a specific time unit, i.e. the standard score of a VF task,

CONTACT Chiara Barattieri di San Pietro chiarabarattieridisanpietro@unimib.it Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy
Supplemental data for this article can be accessed online at https://doi.org/10.1080/23273798.2023.2178662.
© 2023 Informa UK Limited, trading as Taylor & Francis Group







We created a set of algorithms based on distributional semantic models to compute the size of semantic clusters and the number of transitions between them, and we assessed their predictive ability for class membership in a sample of 35 people with SSD and a matched group of control participants.

Scores based on distributional semantic models **always outperformed manual scoring** in classifying people with and without schizophrenia, with semantic measures derived by **LSA** reaching an accuracy of AUC = .87 in the SF task.

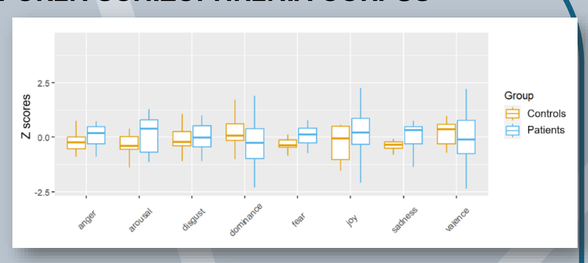
We hypothesize that the semantic fluency task prompts participants to produce words **paradigmatically related** and that the LSA space was the best fitted for capturing this kind of relationship, being its word vectors derived from a matrix that has all documents as co-occurrence contexts.

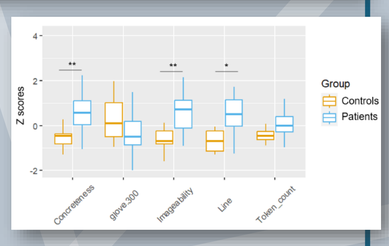
COMPUTATIONAL INSIGHTS INTO A SPOKEN SCHIZOPHRENIA CORPUS

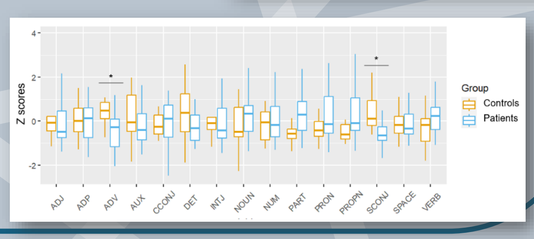
We analyzed the "Discussion Abstract Ideas in Schizophrenia Corpus " (DAIS-C, Delgaram Nejad et al., 2023), from which we extracted a set of lexical, semantic, discourse, sentiment and emotion variables.

People with schizophrenia produced significantly more sentences (U = 32, p < .05) and used words more characterized in terms of Concreteness (t(22.55) = 3.71, p < .01) and Imageability (t(23.91) = 4.16, p < .01) than controls. In contrast, patients used less adverbs (t(25.82) = 3.19, p < .05) and less subordinating conjunctions (U = -157.5, p < .05) than controls.

Results point to a **distinct linguistic profile** of people with schizophrenia, characterized by **verbose expressions**, albeit with **simple constructions**. A tendency among people with schizophrenia to employ **concrete and imageable words** is noted, possibly reflecting semantic alteration and **concrete thinking** in this clinical population. The absence of emotional connotations appears to be task-dependent



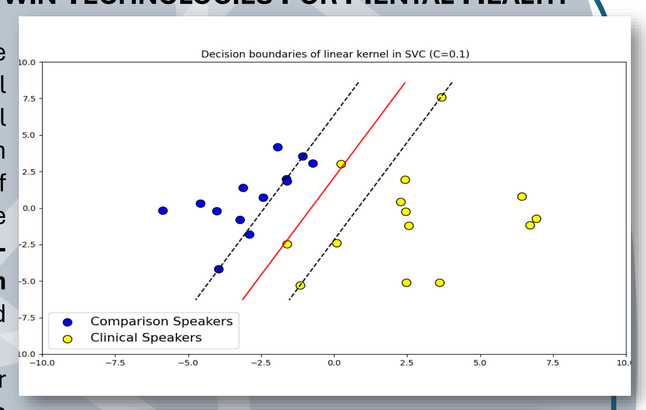


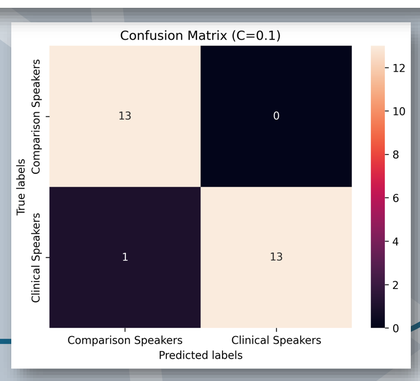


TOWARDS AN IN-SILICO MIND: DIGITAL TWIN TECHNOLOGIES FOR MENTAL HEALTH

Digital Twin (DT) technologies allow for the creation of digital models that mirror physical objects and processes. The application to mental health research has only recently started in medicine, thanks to the untapped potentialities of conversational agents based on Large Language Models (LLM). In the future, **LLM-based DT fine-tuned to the linguistic profile of the person** could help **personalize treatments** and therapeutic approaches.

A mandatory step toward creating DTs for psychiatric care is identifying **tools to evaluate the value** of such technologies. Here, we present an **SVM classifier** to distinguish people with and without schizophrenia trained on 43 language indexes derived from the DAIS corpus (14 clinical and 13 controls). The model has been cross-validated (Leave-One-Out), resulting in an overall **accuracy of 96.30%** in classifying patients vs controls.





STABILITY OF COSINE SIMILARITY MEASURES ACROSS TASKS

Measures of words' association, such as the cosine similarity derived from vector space models, are often linked to cognitive performances in clinical populations. However, the **relationship between cosine similarity and standard test scoring** has not been fully assessed.

We computed the cosine similarity between words (1, 3, 5, and 7 words of distance) produced by 68 persons with schizophrenia (Bambini et al., 2022) in i) semantic and phonological verbal fluency tests and ii) a set of semi-structured interviews. Answers were analyzed using three differently trained word2vec semantic space models and an LSA model.

In the **verbal fluency** task (Fig.1), **higher cosine similarity** measures were positively related to a **higher number of words produced**, particularly the similarity of words at longer distances. The most sensitive model was the LSA semantic space, followed by the 9-word context window word2vec space. In the **Interview task** (Fig.2), **high similarity** between words was significantly correlated with **better pragmatic communication skills**. The difference between the semantic fluency measures might be related to the inability of the approach to capture the known clustering processes in these tasks. In phonological fluency, subjects seem to adopt a strategy to produce a **single large cluster** of semantically associated words. The fact that spaces trained on large windows are more sensitive indicates a selection of words on a paradigmatic basis. Further research will assess the relationships between the same measures and types of interpretations of nonliteral expressions (Fig. 3a,b, and c illustrate cosine similarity measures on the y-axis and, on the x-axis, types of interpretations – null, concrete, abstract – for metaphors, idioms, and proverbs, respectively).

Figure 1

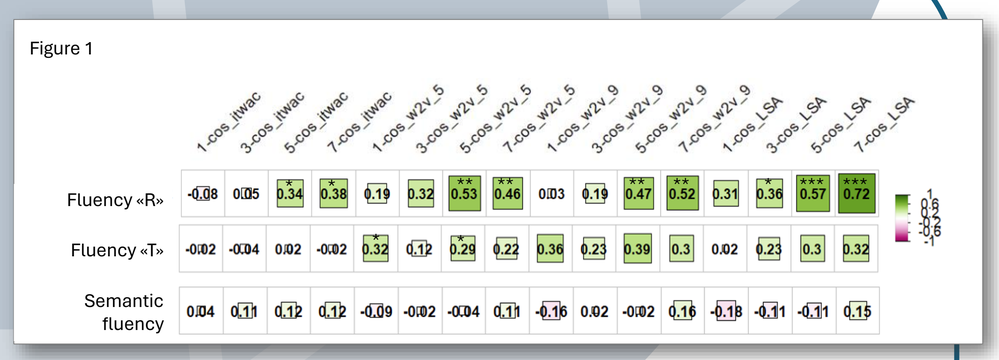
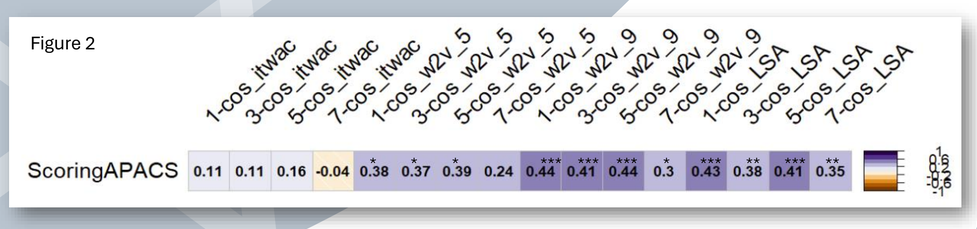


Figure 2



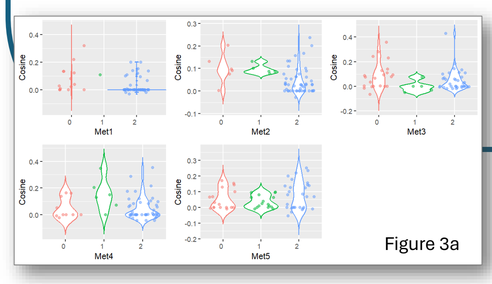


Figure 3a

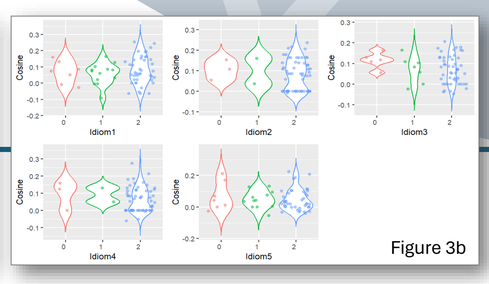


Figure 3b

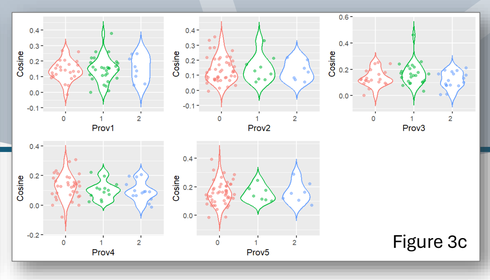


Figure 3c

IUSS Scuola Universitaria Superiore Pavia

NEP LAB Neurolinguistics and Experimental Pragmatics Lab

UniSR Università Vita-Salute San Raffaele