Can language detect different clinical profiles in schizophrenia? A semi-automated analysis on Italian-speaking patients

Frau F[1], Bischetti L[1], Cuoco F[2], Agostoni G[2,3], Cavallaro R[2,4], Bechi M[2], Buonocore M[2], Sapienza J[2], Bosia M[2,4], Bambini V[1] | [1]University School for Advanced Studies IUSS Pavia, Pavia, IT; [2]IRCCS S. Raffaele Scientific Institute, Milan, IT; [3]School of Psychology, S. Raffaele University, Milan, IT; [4]School of Medicine, S. Raffaele University, Milan, IT | E-mail: federico.frau@iusspavia.it

Background: Language disorders are a core symptom of schizophrenia[1,2]. Currently, computational approaches, which provide possibly quick and fine-grained quantitative linguistic analyses, represent a promising tool for research on language disturbances, with potential clinical impact. Previous studies using automated methods for linguistic analysis have mainly focused on diagnoses (e.g., to identify youth at risk of psychosis), using different speech and language measures, such as mean length of utterance and discourse coherence[3,4]. Instead, little research examined such characteristics in subjects with a long-term history of schizophrenia[5,6]. Using (semi-)automated linguistic analyses in this type of subjects might disclose important information to discriminate between individuals with different linguistic and clinical profiles. In turn, this might help structuring targeted rehabilitation interventions to improve functioning and quality of life. This study, the first of this type conducted on Italian-speakers, aimed at grouping chronic patients on different linguistic features in their speech, and to compare the resulting groups on functioning and symptomatology. Innovatively, we performed a multi-layered linguistic analysis targeting different domains within patients' speech.

Methods: We analyzed the speech of 67 people with schizophrenia (age: 39.8±11; education: 11.9±2.7; antipsychotic treatment (atypical/typical): 61/6), native Italian speakers, elicited with the interview task included in the APACS test[7]. All patients were assessed for psychopathology[8] and quality of life[9]. Linguistic features were selected from three domains (see Table1): a) Fluency (i.e., Mean Length of Utterance, pause duration, and pause-to-word ratio); b) Lexical Richness (i.e., Type-Token Ratio and Lexical Frequency); and c) frequency of personal Pronouns and Selected Semantic Classes (SSC) (i.e., affective and metacognitive words), as derived with the LIWC software[10]. We then performed a varimax rotated PCA on the linguistic features: four Principal Components were identified and used to feed a K-means algorithm, which returned a two-cluster solution supported by the Silhouette statistics and confirmed by a linear discriminant analysis (accuracy = 0.94). The two groups were compared on functioning and symptomatology, as well as on demographic and clinical variables.

Results: The algorithm assigned patients to Cluster 1 (n=47) if their speech was characterized by ↑Fluency, ↓Lexical Richness, ↑Pronouns, and ↑SSC, and to Cluster 2 (n=20) if their speech was characterized by ↓Fluency, ↑Lexical Richness, ↓Pronouns, and ↓SSC. Patients in the two groups did not differ for demographic measures nor illness duration (ts < .99, ps > .326). Conversely, patients in Cluster 1 had higher Quality of Life (Interpersonal Relations, Personal Autonomy and Total) and lower symptomatology (Table 2 and Figure 1).

Discussion: The novel aspect of this study is the identification of two linguistic profiles in patients with chronic schizophrenia, based on a semi-automated analysis of several speech characteristics. These profiles highlight distinguished performance on different domains of language and, importantly, are associated with different illness outcomes in terms of symptomatology and specific aspects of quality of life (Figure 2). Among all linguistic aspects, features related to fluency seem key in determining the profiles. Reduced fluency – in terms of shorter utterances and longer pauses – negatively impacts functioning, as well as the global severity of psychopathology, in line with previous literature[11,12]. Other features, such as lexical richness of speech and frequency of affective words and pronouns, contribute to discriminate groups of patients. The relation of these other characteristics with symptoms and functioning is a further novel aspect of this study. Overall, our findings suggest that automated speech analysis is a promising tool not only for providing early and differential diagnosis, as shown in previous literature, but also for predicting and monitoring response to treatment and ultimately improving quality of life.

References: 1. Covington et al. *Schizophr Res*, 77(1), 85-98 (2005). 2. Bambini et al. *Compr Psychiatry*, 71, 106-20 (2016). 3. Elvevåg B et al. *Schizophr Res*, 93(1-3), 304-16 (2007). 4. Cohen & Elvevåg *Curr Opin Psychiatry*, 27(3), 203-9 (2014). 5. de Boer et al. *NPJ Schizophr*, 6(1), 10 (2020). 6. Buck & Penn *J Nerv Ment Dis*, 203(9), 702-8 (2015). 7. Arcara & Bambini *Front Psychol*, 7, 70 (2016). 8. Kay et al. *Schizophr Bull*, 13, 261-76 (1987). 9. Heinrichs et al. *Schizophr Bull*, 10, 388-98 (1984). 10. Pennebaker et al., Austin, TX: liwc.net, (2015). 11. Bowie & Harvey *Schizophr Res*, 103, 240-7 (2008). 12. Parola et al. *Schizophr Res*, 216, 24-40 (2020).

Table 1. Description of linguistic features obtained from patients' speech.

Domain	Measure	Description
Fluency	Mean length of utterance	Mean number of words per utterance.[a]
	Mean pause duration	Mean duration of silent and filled pauses in milliseconds.[b]
	Pause-to-word ratio	Total number of pauses divided by the total number of words.
Lexical Richness	Type-Token Ratio	Number of unique words divided by the total number of words.
	Mean Lexical Frequency	Frequency value associated with the words used by the patient.[c]
Frequency of Personal Pronouns		Percentage of words in this word type category.[d]
Semantic Classes	Affective words	Percentage of words related to positive and negative emotions, and cognitive mechanisms.[d]
	Metacognitive words	

[a] utterances segmented following CHILDES-CHAT guidelines (MacWhinney, 2000); [b] pause duration extracted using the PRAAT software (Boersma & Weenink, 2020); [c] values from the Corpus and Frequency Lexicon of Written Italian (CoLFIS; Bertinetto et al., 2005); [d] frequency values derived using the Linguistic Inquiry Word Count Software (LIWC2015; Pennebaker et al., 2015).

Measures	Cluster 1	Cluster 2	t-statistics	p-value
QLS IRe	20.91±6.08	14.40±5.83	$t(64) = 4.05$	< .001[a]
QLS IRo	4.83±5.45	2.85±4.94	$t(64) = 1.39$	.169[a]
QLS PA	28.96±7.04	18.80±8.01	$t(64) = 5.17$	< .001[a]
QLS Total	54.70±14.08	36.05±14.42	$t(64) = 4.91$	< .001[a]
PANSS Pos	16.23±3.76	18.70±4.52	$t(65) = -2.31$	.024[a]
PANSS Neg	19.72±4.71	23.50±3.95	$t(65) = -3.14$	.008[a]
PANSS Gen	37.15±6.62	41.55±4.94	$t(65) = -2.67$	.014[a]

QLS = Quality of Life Scale; IRe = Interpersonal Relations; IRo = Instrumental Role; PA = Personal Autonomy; PANSS = Positive and Negative Syndrome Scale; Pos/Neg/Gen = Positive Scale Total Score/Negative Scale Total Score/General Scale Total Score; [a] FDR adjusted

Table 2. Clinical and functional descriptive measures and t-test comparisons between clusters.

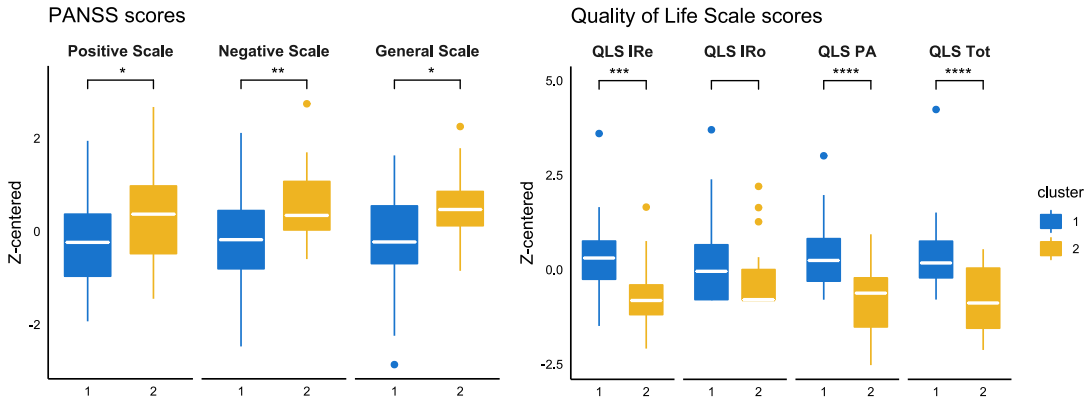

Figure 1. Cluster comparisons across symptomatology (Positive and Negative Syndrome Scale) and functioning (Quality of Life Scale).

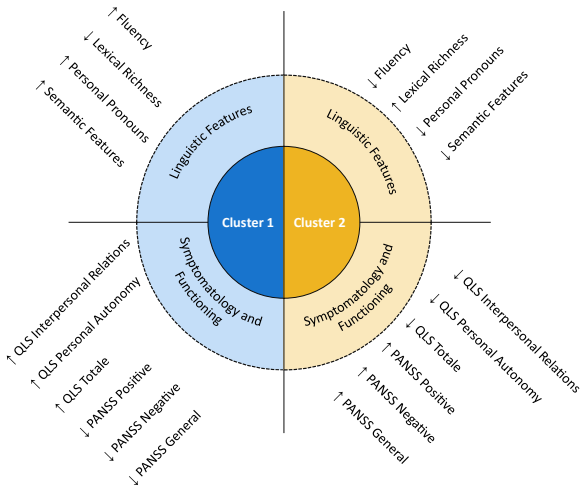

Figure 2. Summary of the linguistic, clinical and functional profiles associated with the patients belonging to the two clusters.