LingBaW. Linguistics Beyond and Within 11 (2025), 253–269

The CIPP-TRS Corpus: Corpus Construction and Preliminary Analyses

Laura Tagliaferro[1], Ludovica Fiorentino[1], Raffaele Guarasci[2], Luigi Franzese[3], Viviana Maria Saia[3], Giancarlo Spennato[3], Felice Iasevoli[3], Roberto Vitelli[3], Andrea de Bartolomeis[3], Francesca M. Dovetto[1]

[1]DSU, University of Naples Federico II; [2]ICAR-CNR; [3]Department of Psychiatry, University of Naples Federico II, Italy

Abstract

Schizophrenia, a neurodevelopmental disorder, significantly affects cognitive and linguistic functions, often resulting in disorganized speech, reduced syntactic complexity, and impaired discourse cohesion. While previous corpora have explored linguistic disruptions in schizophrenia, no dataset has systematically distinguished between treatment-resistant schizophrenia (TRS) and non-treatment-resistant (non-TRS) speech patterns. This study presents the CIPP-TRS Corpus, an annotated collection of transcribed speech from 20 individuals with schizophrenia (10 TRS, 10 non-TRS), alongside a control group of 10 neurotypical speakers. By analyzing peri-linguistic (e.g., interjections, pauses) and paralinguistic (e.g., breath patterns, output modalities) features, we investigate the linguistic manifestations of schizophrenia across these subgroups. Our preliminary findings suggest that TRS patients exhibit richer peri-linguistic markers, and increased hesitation phenomena, while non-TRS patients demonstrate greater lexical retrieval difficulties. Moreover, TRS individuals struggle more with temporal processing, particularly when recalling past events or engaging with past retellings, reinforcing theories on Theory of Mind (ToM) impairments and lived-time disturbances in schizophrenia. The CIPP-TRS Corpus represents a crucial step toward identifying linguistic biomarkers of schizophrenia and its treatment-resistant subtype. Future research will expand the dataset and incorporate prosodic, syntactic, and pragmatic analyses to refine our understanding of speech pathology in schizophrenia, with potential applications in clinical diagnostics and therapeutic interventions.

Keywords: Schizophrenia; treatment-resistant; corpus linguistics; disfluencies; lived time

1. Introduction

In recent years, research on schizophrenia has increasingly focused on the complex relationship between language symptoms and illness itself. Language disturbances, including disorganized speech and difficulties in coherence, have raised important questions about their

LingBaW. Linguistics Beyond and Within, e-ISSN: 2450-5188
DOI: 10.31743/lingbaw.19391

role in diagnosis, symptom progression, and underlying cognitive mechanisms (de Boer et al. 2020).

Despite this, research focusing on Italian remains limited (Dovetto & Gemelli 2013; Pennisi 1998, 2022). To address this gap, this study presents and analyzes the CIPP-TRS Corpus (*Italian Corpus of Patients with Treatment-Resistant Schizophrenia*), which comprises spoken interviews with both treatment-resistant (TRS) and non-treatment-resistant (non-TRS) patients. Language abnormalities affect up to 80% of individuals with schizophrenia, impacting multiple levels of linguistic structure — semantics, syntax, phonology — and ultimately leading to significant pragmatic difficulties. These challenges manifest as alogia, poverty of speech, increased pausing, monotone speech, tangentiality, derailment, and reduced coherence. The resulting communicative impairments can hinder the ability to establish and maintain social relationships, often contributing to social withdrawal and isolation.

TRS accounts for about 30% of schizophrenia cases, significantly impacting both patients and caregivers (de Bartolomeis et al. 2022). In this regard, the CIPP-TRS breaks new ground by being the first to systematically distinguish between TRS and non-TRS speech patterns. This distinction is not just a methodological novelty; it is a crucial step toward understanding the linguistic markers of TRS, paving the way for improved diagnostics, tailored therapeutic strategies, and deeper insights into the cognitive and communicative disruptions unique to this population. As in many areas of theoretical and computational linguistics, the scarcity of resources poses a significant challenge, particularly in specialized domains. To bridge this gap, we developed the CIPP-TRS Corpus from scratch, ensuring a structured, annotated, and linguistically rich dataset for both clinical and computational applications. Notably, this corpus is the first to systematically differentiate between TRS and non-TRS speech patterns—a methodological advancement with critical implications. By identifying the linguistic markers specific to TRS, this work lays the foundation for improved diagnostics, tailored therapeutic strategies, and a deeper understanding of the cognitive and communicative disruptions characteristic of this patient population. Among the linguistic features, in this study we particularly took into account traits denoting a difficulty in speech programming, such as paralinguistic and peri-linguistic indicators (De Mauro 2008), and the presence of pauses.

Our hypotheses are that (a) TRS shows more fragmented discourse and richer peri-linguistic and paralinguistic phenomena, and that (b) pauses are more present in TRS patients due to the severity of their condition. Moreover, literature widely documents a difficulty in moving across the line of time (Minkowski 2004), which can directly signify that questions regarding future and/or hypothetical events could be deemed as challenging for patients.

The first part of the analysis will examine peri- and paralinguistic phenomena, which involve elements pertaining to oral language-systems with a focus on interjections ‘eh’ and ‘beh’ as expressions of discourse-related fillers. These elements may indicate programming difficulties in speech or function as emotionally driven markers reflecting heightened speaker involvement. Additionally, we provide a preliminary investigation into silent pauses as potential indicators of hesitation, particularly in relation to patients’ ability to envision themselves in past situations.

2. Aim and research questions

This study aims to identify potential differences between two patient groups, namely treatment resistant patients (TRS) and non-treatment resistant patients (non-TRS), by comparing their linguistic patterns to a healthy control group (HC). By incorporating both qualitative and quantitative analyses, the research seeks to provide a comprehensive understanding of how linguistic markers vary across these populations.

The research aims to answer different questions:

a) Are there substantial differences in the language production of the two groups of patients, namely TRS and non-TRS compared with the HC group? If so, are there substantial differences involving the production of peri-linguistic and paralinguistic phenomena in the three groups examined?
b) Is there a most frequent occurrence of specific type of interjection? Are there some recurrent patterns?
c) Are pauses, as important predictors of difficulties and dysfunctionality in speech programming, more frequent in the TRS group?

3. Background and related works

Schizophrenia is a chronic and complex brain disorder with a multifactorial etiology, affecting approximately 24 million people worldwide (WHO 2022). In particular, treatment-resistant schizophrenia (TRS) occurs in 30% of patients, persisting despite at least two adequate trials of antipsychotic medication (Potkin et al. 2020). TRS is associated with more severe symptoms, greater functional impairment, and a significant burden on patients and caregivers. In such cases, the second-generation antipsychotic clozapine is considered the gold standard treatment, especially for those experiencing neuromotor side effects from other medications.

Considering the two patient cohorts, only one similar study is currently available (de Boer et al. 2020). To date, no other freely available corpus of pathological schizophrenic speech—expertly labeled with linguistic annotation for automatic analysis—is known, except for two projects: the CIPPS Corpus (Dovetto & Gemelli 2013) and the C-ORAL-ESQ Corpus (*Corpus Oral de Esquizofrênicos*) (Raso et al. 2023). Particularly in relation to the CIPPS Corpus, various analyses were conducted considering pragmatical to interactional approaches (Dovetto & Gemelli 2013). Other studies like Pennisi (1998, 2022) offered an interdisciplinary approach that integrates linguistic, cognitive, and philosophical perspectives. Additionally previous studies focused on pausing patterns as a diagnostic tool for identifying language-related disturbances (Çokal et al. 2019). Furthermore, schizophrenia is linked to cognitive and social impairments, notably deficits in Theory of Mind (ToM) (Doody et al. 1998). Frith (1992) suggested ToM overlaps with pragmatics, which involves interpreting a speaker's intentions through various cues like language, gestures, and expressions. Impairments in these areas complicate understanding non-literal language, such as irony and metaphor (Abu-Akel 1999; Langdon et al. 2002). While the exact impact of ToM deficits on social competence is debated, factors like motivation, interest, and engagement also play a role.

4. Materials and methods

This section illustrates the criteria used for data collection and the annotation process. In particular, the construction of the corpus is described and details concerning the annotation process are motivated. The basic workflow is presented in Figure 1.

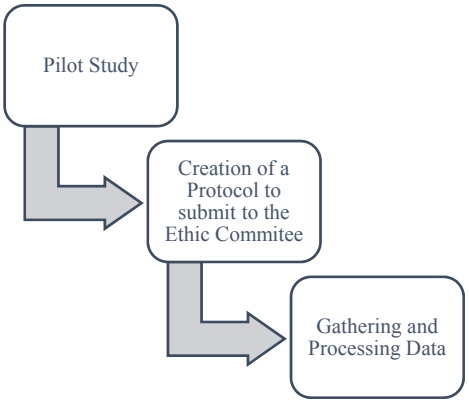


Figure 1: *Workflow*

4.1. Participants and selection parameters

The sample consisted of three groups of equal numbers of subjects affected with TRS, non-non-TRS, and Italian native-speaking healthy controls (HC). The TRS and non-TRS groups' members are currently being treated at the Department of Neuroscience, Reproductive Sciences, and Odontostomatology, Integrated Department of Clinical care, Section on Psychiatry and Clinical Unit of Psychiatry and Psychology of the Azienda Ospedaliera Universitaria of Federico II, Naples. The individuals who took part in the experiment were aged between 18 and 65 years old with an average of 11.6 years of schooling. All participants signed an informed consent for personal data protection and were assessed using the following assessment procedures:

- PANNS – Positive and Negative Scale
- BACS – Brief Assessment Cognition Schizophrenia Scale
- Mini International Neuropsychiatric Interview (MINI plus) – DSM 5

Patients were evaluated using the PANSS and BACS scales to explore correlations between linguistic abnormalities, symptomatology, and cognitive traits. To rule out underlying psychiatric conditions, the HC group was validated with the MINI Plus scale. Specific inclusion and exclusion criteria were established for both clinical and HC groups as listed below in Table 1.

Table 1: *Inclusion and exclusion criteria adopted to collect data*

Groups	**Criteria**	
	Inclusion	*Exclusion*
TRS non-TRS	1. Subjects aged between 18 and 65 years old. 2. Subjects who provided valid written informed consent. 3. Subjects affected by schizophrenia, both responsive (10 non-TRS) and non-responsive to treatment (10 TRS).	1. The presence of other psychiatric conditions, such as bipolar disorder or major depressive disorder. 2. The presence of organic pathologies that compromise language production and articulation. Suffering from intellectual disability, severe systemic disease (poor prognosis; or having severe disability), psychiatric condition secondary to general medical condition or substance abuse. 3. Failure to provide informed consent.
HC	1. Subjects aged between 18 and 65 years old. 2. Subjects who provided valid written informed consent. 3. Subjects who do not have specific psychopathological conditions.	1. The presence of bipolar disorder and depression or other major psychiatric disorders. 2. The presence of diseases that could affect articulation and speech production. 3. Persons with disabilities or suffering from systemic pathologies. 4. Being under the influence of drugs. 5. Failure to provide informed consent.

4.2. Assessing emotional valence in semi-structured interview

We organized a 15-minute semi-structured interview, in which a series of questions from a set of 61 were asked in a randomized order. This project utilized the same materials employed in de Boer and colleagues (2020), whose set of questions was translated and adapted into the Italian language. For their study, the Dutch group considered the set as 'neutral', as the questions concerned general life experiences, underlining that the topics with marked emotional valence such as 'quality of life' and 'health' were avoided.

Before recording, we assessed whether the set was perceived as neutral or emotionally connotated by randomly selected healthy participants from our target population. We created a Google survey where each translated question was followed by six randomized options corresponding to basic emotion, i.e., happiness, anger, surprise, sadness, fear, and no emotion (Ekman 1972), to minimize bias. The survey was completed by 50 students (24 humanities, 26 medical residents), and the results (see Figure 2) indicated that most questions were perceived as emotionally driven, likely due to their tendency to elicit positive memories.

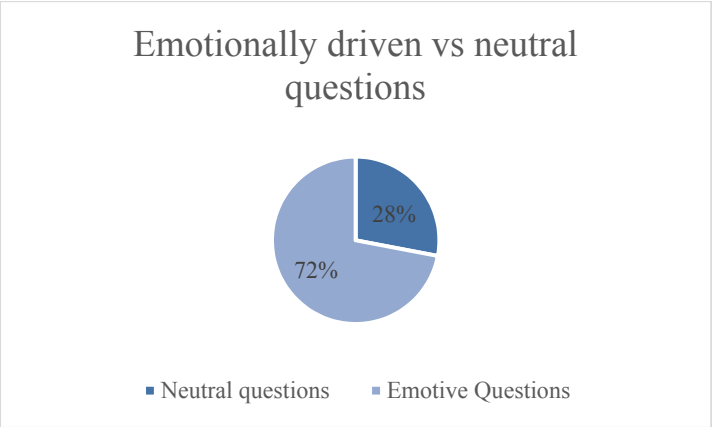


Figure 2: *Results on emotionally driven questions*

4.3. Question optimization for the interview protocol

To determine the optimal order for presenting questions to patients, we examined their nature and structure. Since the questions varied in content, we identified the following topics:

- Hypothetical questions (imaginative unrealistic events)
- Questions regarding the past (i.e., childhood and youth)
- Questions regarding the present with reference to vacation, free time, family, hobbies, work etc.

Since the groups of questions for each topic were different in numbers, we opted to present them in a randomized order for each participant in order to prevent potential biases related to question sequencing. The randomization was performed on excel and each question was treated as a path to follow.

When the interviewee seemed particularly engaged in responding to a given question, the interviewer deemed it necessary to ask for more information that could not be explicitly found in the questions' list. For instance, if the interviewee said that he practiced another sport like tennis instead of swimming lessons when they were young, along the same course of questions associated this topic, the interviewer would have adapted the subsequent requests for details in further asking *"Ti piaceva fare lezioni di tennis? Perché? Cosa trovavi più difficile? Cosa preferivi di questo sport?"*. This course of action allowed the interview to become more natural and provide something as close as possible to the spontaneity of spoken conversations.

Following de Boer's protocol (2020), the nature of each session as to analyze the interviewees' speech was disclosed after the interview took place.

The recordings took place at the AOU Federico II in a silent room, where the interviewer and interviewee sit opposite each other separated by a desk on which we positioned the recorder[1].

[1] We employed recorder Zoom H4n Pro which features high-quality stereo microphones that capture clear and natural audio with a resolution of 24-bit/96kHz.

4.4. Annotation criteria

The starting point for the choices of the current tag set were provided by the experience of the CLIPS and CIPPS Corpus as well as the more recent CIPP-ma (Corpus di Italiano Parlato Patologico della Malattia di Alzheimer[2]). Following analogous Italian corpora of neurotypical and non-neurotypical spontaneous speech allows the comparison of data and can spread some light on pathological speech uses in Italian.

The aim of our investigation was to be able to easily isolate *peri-* as well as *paralinguistic phenomena*. Hence, we opted for a simplification of the tag-names as well as their xml counterparts when applying regular expressions. According to De Mauro's distinction (2008), *peri-linguistic* phenomena involve those elements feebly framed or not-at-all framed within the oral or written language-system. Nonetheless, they are carriers of prosodic cues such as interjections, filled pauses and other disfluency phenomena.

Paralinguistic phenomena are elements that do not belong to language but to other semiotics areas and accompany linguistic production, such as gestures, vocal tone (e.g., screaming, whispering etc.), proxemics, and speech rate. In our study, due to the limitations of voice recordings, we relied on notes from doctors and firsthand observations of the interviews to identify these phenomena. Descriptions of these events were recorded in the "<note> … </note>" section at the end of each turn where the phenomenon occurred.

The phenomena division is reported below in Table 2:

Table 2: *Peri- and paralinguistic phenomena*

Phenomena	Type	Exempla
Peri-linguistic phenomena	False starts (new planning)	colora+ / a colori (CIPP-TRS, TRSD010)
	Fragmented words	il latte con i cerea+ (CIPPS-TRS, non-TRSF01)
	Repetition of same words	{<repetition> po<oo>r+ portare </repetition>} (CPP-TRS, TRSD01)
	Primary interjections	<ah> <beh>
	Filled pauses	<eeh>, <ehm>
	Empty pauses	<pause dur="0,504s"/>
	Elongations	pa<aa>rla<aa>ndo (CIPP-TRS, non-TRSE)
	Vocal phenomena	i.e., generic phenomena, <vocal>
	Interruption within the lexical word	tem_po<oo> (CIPP-TRS, non-TRSE01)
Paralinguistic phenomena	Laugh, giggle, tongue click, lip smacking, whispering, screaming, inspiration and exhale	Non verbal phenomena that may occur with the enunciative act or interrupt it.

Excerpts from two transcriptions:

> F#4: **<inspiration>** Dunque mi piacerebbe parlare se potessi **<inspiration>** con**<nn>** la signora **<ehm>** **<inspiration>** **<eeh>** Giorgia**<aa>** Meloni perché io tendo a destra **<inspiration>** **<lip smacking>** e**<ee>**

[2] The CIPP-ma Corpus gathers recorded speech of 20 patients affected with Alzheimer's disease and 18 healthy controls (HC), see https://www.lupt.unina.it/lisa/ and Dovetto (2025).

sono<**oo**> <**uu**>un<**nn**> / mi do {<**repetition**> del+ del+ del+ </**repetition**>} dell' *inte_ttuale perché ho scritto pure per<**rr**> questo<**oo**> <**inspiration**> giornale il secolo d' Italia <**pause** dur="0,509s"/> <**inspiration**> <**eeh**> mi piacerebbe <**dd**>d+ dialogare con lei<**ii**> rispondere<**ee**> alle sue domande<**ee**> e<**ee**> essere dei suoi insomma <**pause** dur="0,278s"/> <**inspiration**> nel mio poco

[Well, I would like to talk, if I could, with miss ehm eeh Giorgia Meloni, because I sway to the right and I am a / I think of myself as an inte_ctual because I wrote for this newspaper Il Secolo d' Italia eeh I would like to converse with her, answer her questions and be one of hers, basically, in my small way]

(CIPP-TRS, non-TRSB01)

G#39: Come ha imparato a nuotare ?
[How did you learn to swim?]

F#40: <**inspiration**> Mio padre <**pause** dur="1,127s"/> <**inspiration**> lui m' ha insegnato a nuotare <**pause** dur="0,345s"/> ma nuotavo bene <**pause** dur="0,911s"/> <**inspiration**> vasche su vasche <**pause** dur="1,296s"/> <**inspiration**> pure cinquanta sessanta vasche alla volta da cinquanta metri <**pause** dur="0,679s"/> <**inspiration**> <**pause** dur="1,030s"/> <**inspiration**> <**pause** dur="0,713s"/> <**eh**> <**pause** dur="0,798s"/> ce ne vuole per farlo mo<**oo**> se n' è scesa tutta la muscolatura non farei manco un chilome+ <**inspiration**> <**vocal**> andavo alla piscina di <**surname**> <**ehm**> <**pause** dur="0,334s"/> <**inspiration**> anni fa <**inspiration**> che poi conobbi anche in un' occasione <**pause** dur="2,049s"/> non è male come persona devo dire la verità , molte volte queste persone<**ee**> <**pause** dur="0,742s"/> <**inspiration**> dello spettacolo sembrano diverse invece<**ee**> fu molto disponibile forse faceva pure i suoi interessi {<**whispering**> inso+ </**whispering**>} <**inspiration**> però <**pause** dur="0,430s"/> <**inspiration**> <**pause** dur="0,521s"/> è una persona educata <**exhale**>

[My father. He taught me to swim but I swam well, tanks upon tanks, even fifty sixty fifty meter tanks at a time, eh it takes a lot to do it my muscle tone has decreased, I wouldn't even do a kilome+ <vocal> I went to the swimming pool of <surname> ehm years ago. I later met him on one occasion too, he's not a bad person, I have to tell the truth, a lot of times these people in the entertainment industry seem different, he was very amicable instead, maybe he was pursuing his own interests basic+]

5. Results

The following section presents the analysis on peri- and paralinguistic phenomena as well as a detailed analysis on interjections and pauses in past questions.

5.1. Peri- and paralinguistic phenomena

The first data that were drawn from our corpus concerned peri- and paralinguistic phenomena as well as pauses in order to evaluate eventual difficulties in planning and discourse management (de Boer et al. 2020). The data extracted from our sample size is shown in Figure 3, where the TRS group shows the percentage of peri-linguistic phenomena on the total token production.

These results may be linked to the concepts of fluency and disfluency (Lickley 2015), as patients' texts contain a higher occurrence of elements that are either loosely integrated or only partially framed within the language system (De Mauro 2008), including disfluencies. While effective communication is often associated with a perfectly structured and smoothly

delivered message, real-life oral interactions are naturally filled with disfluent phenomena, which may—but do not necessarily—lead to conversational breakdowns (Lickley 2015).

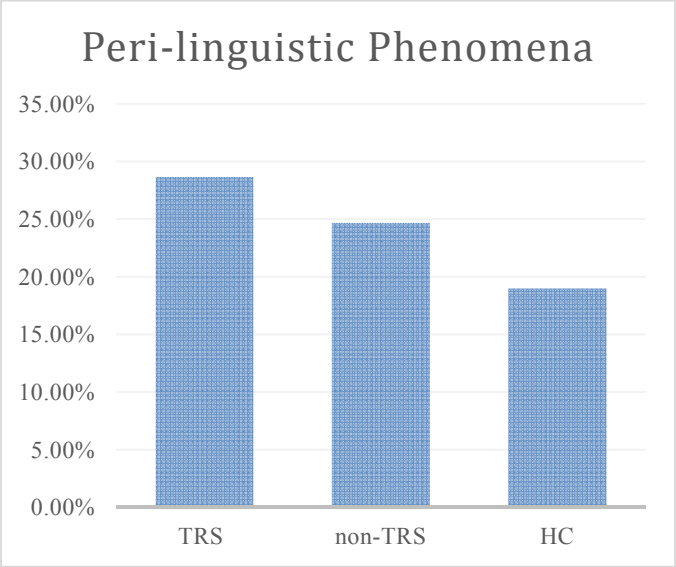


Figure 3: *The data were normalized based on a factor of 10, in accordance with the standard technique used in the literature. This normalization allows for a more accurate comparison across different datasets and ensures that the results are not skewed by variations in token production rates.*

Note that despite the small sample size due to the difficulty and high cost of time and resources for the acquisition, transcription, and annotation phase, results are promising, considering this is preliminary work. Extracted measures need to be validated on a large scale, but it exceeds the purpose of this work, aiming to validate the methodology.

In order to ensure robust results to validate our methodology, we have taken into account two different statistical tests well-known in the literature, namely Newman-Keuls and Dunnett's tests (Masumi et al. 2024; Dybowski et al. 2025), for they have proven very reliable in comparing small-scale datasets with heterogeneous data (Jiang et al. 2023).

First, we used the Newman-Keuls test as an exploratory follow-up which allows us to examine pairwise comparisons more closely (Table 3). Since Newman-Keuls is less strict than other post hoc tests like Tukey's HSD, it is more sensitive to potential differences that other scales might have missed. This helps us identify trends between groups (TRS, non-TRS, and HC) that could be useful for future research, even if they are not statistically confirmed in this dataset.

Table 3: *Neuman-Keuls test, peri-linguistic phenomena*

Groups (peri-linguistic phenomena)	Difference	Test Statistic	p-value	Significant
HC vs non-TRS (df = 2)	-267.3333	0.2475	0.8668	*No*
HC vs TRS (df = 3)	-442.0000	0.4092	0.9553	*No*
non-TRS vs TRS (df = 2)	-174.6667	0.1617	0.9128	*No*

Furthermore, we employed Dunnett's test because of the small dataset when comparing multiple groups such as (TRS and non-TRS) to a single control group HC (Table 4). In this

regard, compared to stricter tests like Scheffé's or broader tests like Tukey's HSD, Dunnett's test reduces the chance of false positives while still being sensitive to meaningful differences.

Table 4: *Dunnett's Test, peri-linguistic phenomena*

Group vs. Group (Contrast) (peri-linguistic phenomena)	**Difference**	**Test Statistic**	**p-value**	**Significant**
non-TRS vs HC	267.3333	0.1750	0.9778	*No*
TRS vs HC	442.0000	0.2893	0.9411	*No*

Neither test yielded statistically significant differences, indicating that the observed variations in our data are not strong enough to be considered meaningful. Interjections are a natural component of normative speech, frequently appearing in everyday communication and serving important pragmatic functions. However, their excessive presence can lead to fragmented discourse, potentially impacting fluency. This phenomenon has been observed in the non-TRS group compared to the HC group and, to an even greater extent, in the TRS group. The increased frequency of interjections in these populations may reflect underlying cognitive or linguistic difficulty, influencing overall speech planning.

5.1.2. Interjections

We conducted a sampling study to determine whether the occurrence of *eh* functions as a discourse-related filler indicating programming difficulty or as an emotionally driven marker reflecting the speaker's heightened involvement in the conversation.

We analyzed all the occurrences of the interjections *eh* and *beh* in a sample made of 50 turns for each group (see Figure 4).

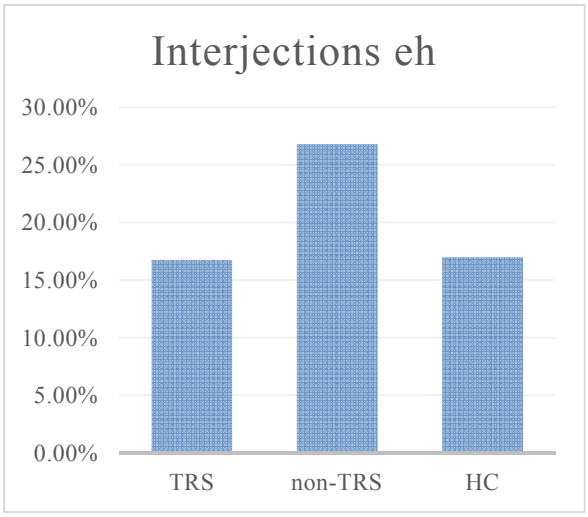


Figure 4: *Interjections eh in 10 TRS, 10 non-TRS and 10 HC percentage on the total interjections produced*

While the three groups tend to produce more *eh* than *beh*, non-TRS group also uses the interjection *eh* more frequently, compared to the other groups. *Eh* is an expression typically linked to a sense of discomfort; when speakers are unsure it's common to hear phrases like

"*eh, non lo so*" [Eng. *eh, I don't know*]. An increased use of *eh* may suggest that patients have difficulty finding words, maintaining a natural flow in their speech, or explaining their ideas. This result is consistent with de Boer et al. (2020) and it may depends on the type of medications which impact the patients' linguistic production.

The non-TRS group produced about 10% more *eh* compared to HCs and TRS.

The totality of the occurrences of <eh> seem to underline a difficulty in discourse planning (Bazzanella 1995). Furthermore, while 67% of the occurrences extracted convey the speaker's emotional load which we associate with a sense of discomfort as reported in the examples below, 33% of cases are used as a filler in order to keep the floor.

(1) Dime+ / me ne dimentico quindi non sapre+, **eh**, boh credo che mi metterei a posto e poi (devoid of emotional load)
[Forg+ / I forget about it so I wouldn't kno+ eh, boh that I would settle down]

(2) Come? *Bi+ *bi+ ehm scusi beauti+ beaut+ beu+ *bi+, **eh**, beautifu+ beaut+ *bi+ *bil+ ehm mi aiuti eh per cortesia (emotional load)
[What? *Bi+ *bi+ ehm sorry beauti+ beaut+ beu+ *bi+, eh, beautifu+ beaut+ *bi+ *bil+ ehm, help me please]

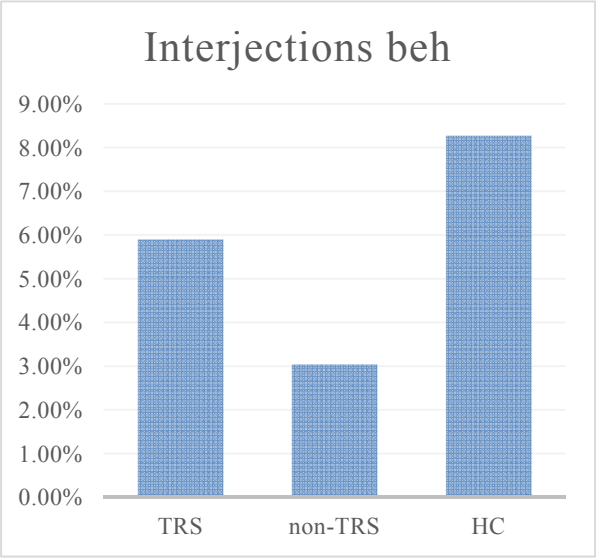


Figure 5: *Interjections beh in 10 TRS, 10 non-TRS and 10 HC percentage on the total interjections produced*

Poggi's (1995) taxonomy serves as a reference framework for interpreting *beh* as an amplified expression of doubt or hesitation. Bazzanella (1995) theorizes that *beh* reflects the speaker's sense of inadequacy, prompting them to soften their response in terms of relevance, pertinence, and exactness.

The occurrences of <beh> subsume an 'esercitive function', where it holds a clarificatory/explicative value. In the other 15% of cases, the interjection <beh> plays the role of a discourse marker, which signals a brief halt related to indecision or disagreement.

(3) Mi rallegra mi diverte m+ / *ai *fini mi diverte, **beh**, eeh dottore questa è una domanda da (disagreement related to the fact that the speaker does not know what to say).
[It lights me up, it cheers me up i+ / *at the *ends it cheers me up **beh**, eeh, doctor this is a question]

(4) La cosa migliore, la più divertente ? **Beh** ho visto un mio amico che / che sta fuori da tanti anni (explicative value).
[The best thing, the most fun ? Beh, I saw a friend of mine that/ that has been abroad for years]

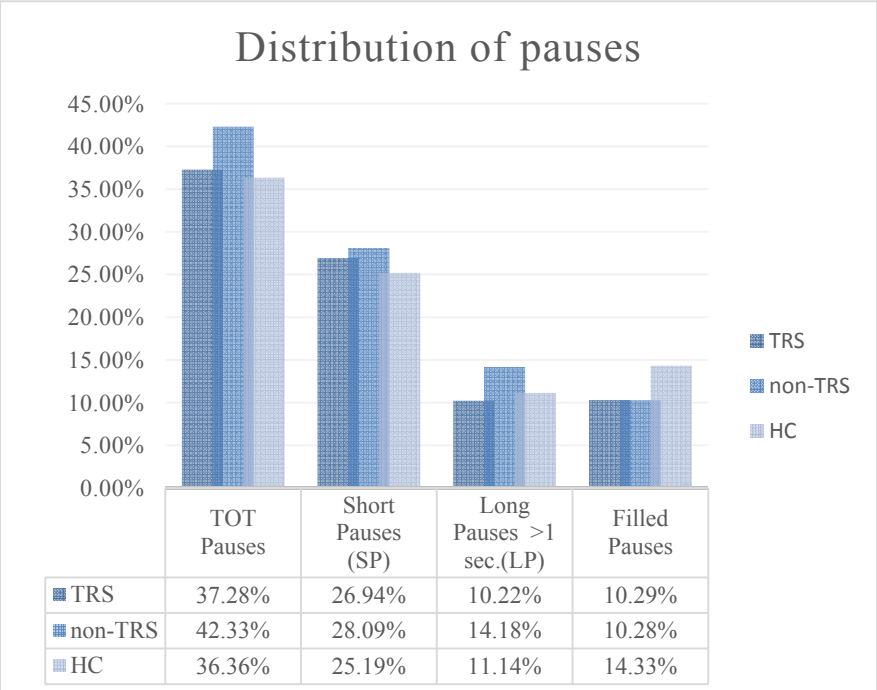
In contrast, the HC group used *beh* more extensively. As native Italian speakers, we associate *beh* with a moment of word-searching, often followed by clarification or reformulation.

5.2. Silent intervals

While interjections like *eh* can signal a planning difficulty, another common marker of such difficulties is the use of pauses, which serve as a strategic tool for speakers to manage cognitive load and speech production. Silent intervals are reported below. Greater overall production of silent pauses was observed in both groups of patients that can be linked to a disruption in speech production and delayed processing (de Boer et al. 2020; Çokal et al. 2019).

In particular, the non-TRS group shows higher numbers in terms of production of both short silent pauses and long silent pauses, which could be related to a difficulty in lexical retrieval (Allen et al. 1993) as widely documented within the pathology, as well as an overall sense of uncertainty.

Notably, our findings indicate a higher occurrence of pauses within the non-TRS group, reinforcing their connection to speech planning disruptions.



	TOT Pauses	Short Pauses (SP)	Long Pauses >1 sec.(LP)	Filled Pauses
TRS	37.28%	26.94%	10.22%	10.29%
non-TRS	42.33%	28.09%	14.18%	10.28%
HC	36.36%	25.19%	11.14%	14.33%

Figure 6: *Distribution of Pauses, namely empty and filled pauses e.g., eeh and ehm. This number is expressed as a percentage of the total number of disfluencies produced.*

The higher percentage of filled pauses produced by HCs can be interpreted following Matsumoto and colleagues (2020), who obtained similar findings. He associates the greater

production of filled pauses with the speaker's intention to maintain his communicative turn while pausing to reprogram his speech.

Upon a more subtle verification, we report that filled pauses are evenly distributed across parts of speech in all three groups, indicating that they do not specifically occur before any particular word type.

Interestingly, in our data set, the TRS group produced half of filled pauses at the beginning of the turn, compared to both HCs and non-TRS patients[3], hence pointing to a possible difficulty in signaling their intent to speak (Rose 1998).

5.3. Paralinguistic features

Regarding paralinguistic features, our findings (Table 5) are consistent with those of Dovetto and Gemelli (2013), who noted that a patient with *Wahnstimmung* frequently interrupted their speech with pauses and deep breaths. This disrupted flow of discourse reflects the cognitive and emotional strain often experienced in such conditions, underscoring the significance of paralinguistic elements in communication.

Table 5: *Paralinguistic phenomena in 10 TRS, 10 non-TRS and 10 HC. These percentages are normalized on token number.*

	TRS	**non-TRS**	**HC**
Inspiration	52.80%	67.00%	64.84%
Exhale	13.14%	12.85%	5.49%
Tongue clicks	3.74%	2.31%	3.02%
Lip smacking	9.83%	8.43%	5.26%
Throat clearing	0.30%	0.26%	0.11%
Whispering	9.70%	4.22%	5.04%
Screaming	0.30%	0.40%	0.00%
Giggle	3.62%	3.56%	12.43%
Laugh	0.72%	0.40%	1.12%

One possible explanation for the high occurrence of breaths (both inhaling and exhaling) in the non-TRS group, which also produced more and longer pauses, is that these phenomena are linked to increased cognitive effort and speech planning demands. Breaths, like pauses, can serve as a mechanism to manage speech production, particularly when speakers face difficulties in formulating their utterances. The frequent and prolonged pauses in this group may indicate a need for more time to structure speech, while the accompanying breaths could reflect the physiological effort involved in sustaining communication under such conditions.

The TRS speech seems to be characterized by a greater presence of whispering phenomena, consisting of voicelessness, increased airflow, reduced pitch variations and soft volume (Tartter 1989). This result might display issues in maintaining prosodic stability,

[3] The normalized percentage of filled pauses at the beginning of the conversational turn amounts to 1.13% in TRS, 2.2% in non-TRS and 3.46% in HCs.

possibly due to increased cognitive load or uncertainty during speech production. Such variations in pitch could signal hesitation, reduced confidence, or a struggle to maintain fluency, further reinforcing the idea that paralinguistic features play a crucial role in managing communication, especially in challenging speaking conditions.

5.4. Pauses in past questions

ToM, i.e., the ability to recognize, predict, and interpret others' mental states, is essential for social interaction and goal-directed behavior. It enables individuals to integrate multiple sources of information to infer intentions and navigate communication effectively. ToM is closely linked to pragmatics, which governs understanding beyond literal meaning through language, gestures, and context.

Research indicates that ToM is impaired in schizophrenia (Doody et al. 1998), affecting the ability to relate intentions to actions and monitor social cues (Frith 1992). This impairment contributes to pragmatic difficulties, particularly in irony and metaphor comprehension (Abu-Akel 1999; Langdon et al. 2002).

Furthermore, schizophrenia disrupts temporal cognition, as patients struggle to conceptualize time, favoring spatial relations instead (Minkowski 2004). These deficits highlight ToM's role in both social competence and one's perception of lived time.

Following this line, we investigated the production of pauses in past questions which highlighted longer pauses and greater hesitation when discussing past situations, pointing to challenges in lexical retrieval and overall temporal processing.

Table 6: *Normalized silent intervals for past questions. Number of pauses/tokens (max. 1)*

	TRS	**non-TRS**	**HC**
Past	0.758	0.439	0.061

This pattern appears to stem from social withdrawal, as individuals become absorbed in themselves, their beliefs, and the immediacy of the present. Their difficulty in expressing experiences beyond the 'now' highlights a sense of temporal dislocation, where thoughts seem suspended and motionless, contributing to the idea that schizophrenia is marked by a distinct atemporality. This is consistent with studies that have shown that individuals with schizophrenia often struggle with recalling and narrating past events, exhibiting impoverished autobiographical narratives and disorganized temporal sequencing (Mediavilla et al. 2021). Although the difference between TRS and non-TRS does not seem numerically significant, this result lays the foundation for further analyses on a larger sample.

6. Conclusion and future work

This study introduced the CIPP-TRS Corpus, the first corpus explicitly designed to differentiate between TRS and non-TRS in spoken language. This study aimed to identify potential differences between the two patient groups by comparing their linguistic patterns to

a healthy control group (HC). By incorporating both qualitative and quantitative analyses, the research sought to provide a comprehensive understanding of how linguistic markers vary across these populations.

Our hypotheses were that (a) TRS showed more fragmented discourse and richer peri-linguistic and paralinguistic phenomena, and that (b) pauses were more present in non-TRS patients, but due to the severity of their condition, the TRS group displayed longer pauses for past questions. Moreover, literature widely documents a difficulty in moving across the line of time (Minkowski 2004), which can directly signify that questions regarding future and/or hypothetical events could be deemed as challenging for patients.

In the end, the research hypotheses were partially disconfirmed. Although differences in the spontaneous linguistic production of TRS patients and non-TRS patients have been highlighted, these do not always indicate more severe linguistic impairment in TRS patients. By analyzing peri- and paralinguistic features, we identified preliminary trends highlighting disruptions in speech planning, lexical retrieval difficulties, and temporal disorientation in TRS and non-TRS patients. While both patient groups exhibited increased disfluencies, interjections, and pauses, our findings suggest that the non-TRS group appeared more compromised than TRS for what concerns the amount of silent intervals and a higher production of interjection '*eh*' during the interview. These results, while aligned with de Boer and colleagues (2020), suggest that evaluating the linguistic production of patients requires not only considering the severity of the condition but also the type of medication used for treatment. This aspect will be further examined and clarified in the next phase of the study.

These observations align with theory of mind (ToM) impairments and the lived-time hypothesis, reinforcing the notion that individuals with schizophrenia, particularly those with TRS, experience temporal and cognitive rigidity that affects their ability to process past and future events. Although this is a pilot study, aimed at validating the methodology and tested at the moment on a small-scale dataset, results are promising, and this corpus represents an essential first step toward a more nuanced understanding of linguistic markers of schizophrenia.

Beyond its theoretical contribution, the CIPP-TRS Corpus may have useful clinical implications. The linguistic markers identified in this research—such as disfluencies, hesitation patterns, and signs of temporal disorientation—show promise as non-invasive indicators that could complement existing clinical assessment methods. With further validation in larger samples, such markers might help improve the early detection of schizophrenia and contribute to distinguishing treatment-resistant from non-resistant profiles, thereby supporting timely intervention. Longitudinal tracking of these features could also provide additional information on treatment response and functional outcomes, alongside traditional clinical scales. Although preliminary, linguistic analysis may also hold potential for assisting differential diagnosis with other conditions that present overlapping symptoms, as well as for highlighting specific functional impairments without the need for invasive or costly procedures (e.g., fMRI, MRI, CT).

Moving forward, we plan on expanding the corpus size to refine our statistical analyses and incorporating prosodic, syntactic, and pragmatic features to deepen our understanding of speech disturbances in schizophrenia. Future studies will also investigate disfluency clustering,

speech rate variations, and the positioning of interjections to differentiate TRS and non-TRS speech profiles. Further research will delve into various types of interjections, not only based on their type but also on their position within speech, as recurring patterns or specific co-occurrences may open new research directions.

References

Abu-Akel, A. 1999. Impaired theory of mind in schizophrenia. *Pragmatics and Cognition* 7: 247–282.

Allen, P. A., D. J. Madden, T. A. Weber, and K. E. Groth. 1993. Influence of age and processing stage on visual word recognition. *Psychology and Aging* 8(2): 274.

American Psychiatric Association. 2013. *DSM-5 TM guidebook: The essential companion to the Diagnostic and statistical manual of mental disorders, fifth edition*. 5th ed. Washington, DC: American Psychiatric Publishing.

Bazzanella, C. 1995. I segnali discorsivi. In L. Renzi, G. Salvi, and A. Cardinaletti (eds.), *Grande grammatica italiana di consultazione*, 431–452. Bologna: il Mulino.

CIPPS Corpus. n.d. LISA / Lingua e salute (del corpo, della persona, della comunità). Centro Interdipartimentale di ricerca LUPT, Università degli Studi di Napoli Federico II. Retrieved August 26, 2025, from https://www.lupt.unina.it/lisa/.

Çokal, D., V. Zimmerer, D. Turkington, N. Ferrier, R. Varley, S. Watson, and W. Hinzen. 2019. Disturbing the rhythm of thought: Speech pausing patterns in schizophrenia, with and without formal thought disorder. *PloS One* 14(5): e0217404.

de Bartolomeis, A., L. Vellucci, A. Barone, M. Manchia, V. De Luca, F. Iasevoli, and C. U. Correll. 2022. Clozapine's multiple cellular mechanisms: What do we know after more than fifty years? A systematic review and critical assessment of translational mechanisms relevant for innovative strategies in treatment-resistant schizophrenia. *Pharmacology & Therapeutics* 236: 108236.

De Boer, J. N., A. E. Voppel, S. G. Brederoo, F. N. K. Wijnen, and I. E. C. Sommer. 2020. Language disturbances in schizophrenia: The relation with antipsychotic medication. *NPJ Schizophrenia* 6(1): 24.

De Mauro, T. 2008. *Lezioni di linguistica teorica*. Roma-Bari: Laterza.

Doody, G. A., M. Götz, E. C. Johnstone, C. D. Frith, and D. G. Owens. 1998. Theory of mind and psychoses. *Psychological Medicine* 28(2): 397–405.

Dovetto, F. M. 2023. *Speech in schizophrenia*. Roma: tab.

Dovetto, F. M. 2025. (ed.) *Il linguaggio tra invecchiamento e demenza: Parlato e malattia di Alzheimer nel corpus CIPP-ma*. Roma: tab.

Dovetto, F. M., and M. Gemelli. 2013. Il corpus CIPPS. In *Il parlar matto: Schizofrenia tra fenomenologia e linguistica. Il corpus CIPPS*. 2nd ed. with DVD-ROM. Roma: Aracne.

Dybowski, F. P., D. S. Scott, and C. A. Tamminga. 2025. Pharmacological reduction of reverse-translated hippocampal hyperactivity in mouse: Relevance for psychosis. *Neuropsychopharmacology*.

Ekman, P. 1972. Universals and cultural differences in facial expressions of emotion. In J. K. Cole (ed.), *Nebraska Symposium on Motivation*, 207–282. Lincoln, NE: University of Nebraska Press.

Frith, C. D. 1992. *The cognitive neuropsychology of schizophrenia*. Hove, UK: Lawrence Erlbaum Associates.

Frith, C. D., and R. Corcoran. 1996. Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine* 26(3): 521–530.

Kane, J. M., O. Agid, M. L. Baldwin, O. Howes, J. P. Lindenmayer, S. Marder, and C. U. Correll. 2019. Clinical guidance on the identification and management of treatment-resistant schizophrenia. *The Journal of Clinical Psychiatry* 80(2): 2783.

Khan, U., M. Habibur Rahman, Md. Salauddin Khan, M. D. Hossain, M. S. Morsaline Billah. 2022. Bioinformatics and network-based approaches for determining pathways, signature molecules, and drug substances connected to genetic basis of schizophrenia etiology. *Brain Research* 1785: 147889.

Kircher, T., A. Krug, M. Stratmann, S. Ghazi, C. Schales, M. Frauenheim, L. Turner, P. Fährmann, T. Hornig, M. Katzev, M. Grosvald, R. Müller-Isberner, and A. Nagels. 2014. A rating scale for the assessment of objective and subjective formal thought and language disorder (TALD). *Schizophrenia Research* 160(1–3): 216–221.

Jaspersen, O. 1922. *Language: Its nature, development, and origin*. London: Allen & Unwin.

Jiang, W., J. Zhou, and B. Liang. 2023. An improved Dunnett's procedure for comparing multiple treatments with a control in the presence of missing observations. *Mathematics* 11(14): 3233.

Langdon, R., M. Davies, and D. Coltheart. 2002. Understanding minds and understanding communicated meanings in schizophrenia. *Mind & Language* 17(1–2): 37–67.

Lavelle, M., P. G. Healey, and R. McCabe. 2013. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia Bulletin* 39(5): 1150–1158.

Lickley, R. J. 2015. Fluency and disfluency. In M. A. Redford (ed.), *The handbook of speech production*, 445–469. Hoboken, NJ: John Wiley & Sons.

Masoumi, S. M., M. R. Youssefi, and S. S. R. Shojaei. 2024. Exploring the interplay of chronic toxoplasmosis and NMDAR dysfunction: Insights into schizophrenia-like behaviors and therapeutic potential. *Open Veterinary Journal* 14(7): 1634–1643.

Mediavilla, R., M. López-Arroyo, J. Gómez-Arnau, C. Wiesepape, P. H. Lysaker, and G. Lahera. 2021. Autobiographical memory in schizophrenia: The role of metacognition. *Comprehensive Psychiatry* 109: 152254.

Minkowski, E. 2004. *Il tempo vissuto: Fenomenologia e psicopatologia*. Translated by G. Terzian. Torino: Einaudi. (Original work published 1968).

Pennisi, A. 1998. *Psicopatologia del linguaggio: Storia, analisi, filosofie della mente*. Rome: Carocci.

Pennisi, A. 2022. Psychopathology of language, DMN, and embodied neuroscience: A unifying perspective. *Il Mulino – Riviste Web, Reti, Saperi, Linguaggi* 1: 21.

Poggi, I. 1995. Le interiezioni. In L. Renzi, G. Salvi, and A. Cardinaletti (eds.), *Grande grammatica italiana di consultazione*, 453–470. Bologna: il Mulino.

Premack, D., and G. Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1(4): 515–526.

Raso, T., B. N. R. de Melo Rocha, J. V. Salgado, B. F. Cruz, L. M. Machado Mantovani, and H. Mello. 2023. The C-ORAL-ESQ project: A corpus for the study of spontaneous speech of individuals with schizophrenia. *Language Resources & Evaluation*.

Rose, R. L. 1998. The communicative value of filled pauses in spontaneous speech. MA diss., University of Birmingham.

Savy, R. 2006. Specifiche per la trascrizione ortografica annotata dei testi raccolti. In F. Albano Leoni and R. Giordano (eds.), *Italiano parlato: Analisi di un dialogo*. Napoli: Liguori. http://www.clips.unina.it/.

Sperber, D., and D. Wilson. 2002. Pragmatics, modularity and mind-reading. *Mind & Language* 17(1–2): 3–23.

Tartter, V. C. 1989. What's in a whisper? *The Journal of the Acoustical Society of America* 86(5): 1678–1683.

Cite this article as:

Tagliaferro, L., Fiorentino, L., Guarasci, R., Franzese, L., Saia, V. M., Spennato, G., Iasevoli, F., Vitelli, R., de Bartolomeis, A., & Dovetto, F. M. (2025). The CIPP-TRS Corpus: Corpus Construction and Preliminary Analyses. *LingBaW. Linguistics Beyond and Within, 11*, 253–269.