Title: Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of NLP automated measures of coherence

Parola Alberto[a,b*], Lin Jessica Mary[a,b*], Simonsen Arndis[b,c], Bliksted Vibeke[b,c], Zhou Yuan[d], Wang Huiling[e], Inoue Lana[f, g], Koelkebeck Katja[f, g], Fusaroli Riccardo[a,b,h]

[a] Department of Linguistics, Semiotics and Cognitive Science, Aarhus University, Aarhus, Denmark

[b] The Interacting Minds Center - Institute of Culture and Society, Aarhus University, Aarhus, Denmark

[c] Psychosis Research Unit - Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

[d] Institute of Psychology, Chinese Academy of Sciences, Beijing, China

[e] Department of Psychiatry, Renmin Hospital of Wuhan University, Wuhan, China

[f] LVR-Hospital Essen, Department of Psychiatry and Psychotherapy, Hospital and Institute of the University of Duisburg-Essen, Essen, Germany

[g] Center for Translational Neuro- & Behavioral Sciences (C-TNBS), University Duisburg Essen, Germany

[h] Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA

[*] These authors contributed equally to this work as first authors.

Correspondence concerning this article should be addressed to:

Alberto Parola, Jens Chr. Skous Vej 2, 8000 Aarhus C, Denmark

Phone number: +390116703065

E-mail: alberto.parola@cc.au.dk

Abstract (250 words)

Introduction

Language disorders – disorganized and incoherent speech in particular - are distinctive features of schizophrenia. Natural language processing (NLP) offers automated measures of incoherent speech as promising markers for schizophrenia. However, the scientific and clinical impact of NLP markers depends on their generalizability across contexts, samples, and languages, which we systematically assessed in the present study relying on a large, novel, cross-linguistic corpus.

Methods

We collected a Danish (DK), German (GE), and Chinese (CH) cross-linguistic dataset involving transcripts from 187 participants with schizophrenia (111DK, 25GE, 51CH) and 200 matched controls (129DK, 29GE, 42CH) performing the Animated Triangle task. Fourteen previously published NLP coherence measures were calculated, and between-groups differences and association with symptoms were tested for cross-linguistic generalizability.

Results

One coherence measure robustly generalized across samples and languages. We found several language-specific effects, some of which partially replicated previous findings (lower coherence in German and Chinese patients), while others did not (higher coherence in Danish patients). We found several associations between symptoms and measures of coherence, but the effects were generally inconsistent across languages and rating scales.

Conclusions

Using a cumulative approach, we have shown that NLP findings of reduced semantic coherence in schizophrenia have limited generalizability across different languages, samples, and measures. We argue that several factors such as sociodemographic and clinical heterogeneity, cross-linguistic variation, and the different NLP measures reflecting different clinical aspects may be responsible for this variability. Future studies should take this variability into account in order to develop effective clinical applications targeting different patient populations.

Keywords:

Natural language processing, digital phenotyping, thought disorder, schizophrenia spectrum disorder, semantic coherence, biomarker

Introduction

Language disturbances have been a hallmark of schizophrenia since the first definitions of the disorder (Bleuler, 1911; Kraepelin, 1919). They are particularly evident at the discourse level - ranging from reduced syntactic complexity to loss of semantic coherence and cohesion -, and are often associated with specific symptoms (e.g., formal thought disorders). Language disorders can seriously impair the patients' social functioning and communicative ability (e.g., Bliksted et al., 2014; Green et al., 2015; Gallagher & Varga; 2015; Parola et al., 2018; 2021) and are pervasive: they are an early distinctive feature of schizophrenia, preceding the onset of initial psychosis, and occur in individuals at high clinical risk as well as in patients' relatives (Bedi et al., 2015; Corcoran et al., 2018; Rezaii et al., 2019). Therefore, language disorders - and disorganized and incoherent speech in particular - could play a critical role for developing digital phenotyping of schizophrenia (Corcoran et al., 2020; de Boer et al., 2018; De Boer et al., 2020; Hitczenko et al., 2021).

Recent advances in natural language processing (NLP) techniques - e.g., topic modeling (Rezaii et al., 2019), word embeddings (Bedi et al., 2015; Corcoran et al., 2018; Elvevåg et al., 2007; Holshausen et al., 2014; Just et al., 2020; Tang et al., 2021; Voppel et al., 2021), speech graph analysis (Mota et al., 2014, 2017), and semantic density quantification (Rezaii et al., 2019) - could provide quantitative, cost-effective, and automated measures of incoherent speech. Indeed, automated analyses of linguistic content and coherence have variously found lower coherence and semantic density in schizophrenia, and individuals at high clinical risk.

These findings suggest that NLP techniques may complement clinical observations and constitute a window into the social and emotional features of the disorder (Cohen et al., 2021; Corcoran & Cecchi, 2020). However, a critical obstacle to any concrete use of these findings is that it is not clear whether the findings would replicate and generalize to new samples and populations, an overarching problem for clinical and social sciences (Hitczenko et al., 2021; Parola et al., 2020; Rocca & Yarkoni, 2021; Rybner et al., 2021). Indeed, a closer look reveals clearly contradictory results: linguistic measures are inconsistently associated with symptoms, and findings vary across different rating scales and samples (Bedi et al., 2015; Corcoran et al., 2018; Haas et al., 2020; Morgan et al., 2021; Pauselli et al., 2018; Sarzynska-Wawer et al., 2021; Tang et al., 2021).

Such inconsistencies may have several causes. Sample sizes are usually small (median size for participants with schizophrenia = 34.5), and given the heterogeneity of schizophrenia,

differences in demographic and clinical characteristics between studies can be quite large, and findings be overfit to the specific sample. Moreover, linguistic and/or cultural specificity may seriously affect coherence patterns, thus leading to differences in language impairments between samples collected in different countries, as well as differences in their specific association with symptomatology (Palaniyappan, 2021; Sumiyoshi et al., 2004, 2014; Wydell & Butterworth, 1999). For example, native speakers of different languages display differences in word retrieval and word processing strategies and in the use of pauses, hesitations and false starts (Sumiyoshi et al., 2004; Ishkhanyan et al., 2020; Palaniyappan, 2020). Thus it is not clear how well the findings of previous studies can generalize to different linguistic and/or cultural groups. Finally, automated measures of linguistic content have been operationalized in very different ways and differ substantially across studies- potentially reflecting different psychopathological dimensions - and these differences may account for the differences in findings.

To move beyond this situation, this study showcases a cumulative scientific approach capable of systematically assessing the impact of previous findings on current data and integrating the new findings into a global framework. Such a framework promotes the systematic assessment of previous findings and different automated measures of coherence across contexts and samples with different clinical, demographic, cultural and linguistic profiles (Corcoran et al., 2018; Fusaroli et al., 2021; Rybner et al., 2021). This will provide a more robust predictive performance assessment, but it may also provide more reliable foundations for theory development (e.g. generative modeling of incoherence) and accordingly improved understanding of language disturbances (Press et al., 2022; Rocca & Yarkoni, 2021).

First, we systematically reviewed the literature to identify replicable automated NLP coherence measures that characterize the language of patients with schizophrenia, and effect sizes of previous findings for critical comparison. Second, we assembled a large cross-linguistic (Danish, German, Chinese) corpus consisting of multiple speech transcriptions from patients with schizophrenia and matched controls. Third, we critically and systematically assessed how well previous findings generalized to the new corpus and were robust to language and sample variations. By comparing literature-informed (Brand et al., 2019; Fusaroli et al., 2021) to more traditional analyses (relying on regularizing priors), we can more directly assess how the data supports or strays from previously found patterns. Fourth, we provided a more systematic evaluation of the heterogeneities involved in the study. We explicitly model heterogeneity by

assessing the associations between measures of coherence and clinical ratings of psychopathology. Crucially, we were also able to estimate the robustness of coherence measures, within and between subjects, samples and languages, as well as their variability. Finally, we adopt an approach where we rely on open source software, extract features in a reproducible manner using openly available scripts, carefully describe the methodology used, and test the robustness of the results to variations in methodology.

The aim of the study is to directly address the problem of generalizability of previous NLP results and to develop a critical and systematic approach that can further the understanding of language disorders and their relationship with symptoms in schizophrenia.

Methods

Participants

We collected a Danish (DK), German (GE), and Chinese (CH) cross-linguistic dataset involving 187 participants with schizophrenia(111 DK, 25 GE, 51 CH) and 200 matched controls (HC) (129 DK, 29 GE, 42 CH). The samples for the present study were collected in separate studies assessing mentalizing ability in patients with schizophrenia and healthy controls. Information on demographics, IQ, psychopathology, and social functioning is summarized in **Table** 1. Detailed information on each study is reported in the **Supplementary Material 1** (**SM1**).

Table 1. Demographic and clinical characteristics of patients with schizophrenia (SCZ) and healthy controls (HC).

Corpus	**Danish**		**German**		**Chinese**	
Diagnosis	SCZ N = 111	HC N =129	SCZ N= 25	HC N=29	SCZ N =51	HC N =42
N. of transcripts	N = 944	N = 1102	N=298	N= 346	N = 406	N = 321
Age	26.9 (9.57)	26.4 (8.78)	29.2 (8.48)	30.7 (7.56)	27.2 (7.22)	28.5 (7.72)
Education	12.8 (2.76)	14.8 (2.54)	12.1 (1.48)	12.3 (1.11)	12.7 (2.69)	14.4 (2.12)
Gender (n. of females and %)	46 (41%)	59 (46%)	11 (44%)	12 (41%)	23 (45%)	18 (43%)
Verbal IQ	89.12 (18.67)	102.02 (15.94)	NA	NA	96.03 (16.53)	101.72 (14.17)

SANS total	9.70 (4.39)	NA	NA	NA	7.61 (3.01)	NA
SAPS total	10.38 (4.90)	NA	NA	NA	7.16 (4.79)	NA
PANSS total	NA	NA	52.87	10.09	75.72	10.46
PANSS negative	NA	NA	14.12	3.90	20.01	5.25
PANSS positive	NA	NA	10.64	2.72	18.54	4.36
Illness duration (months)	8.89 (6.70)	NA	28.96 (47.42)	NA	62.97 (68.20)	NA

Speech samples

Speech samples were collected using the Animated Triangles task(Abell et al., 2000; Castelli et al., 2000). The task is used to measure theory of mind (ToM), and it consists of twelve video clips representing an interaction between animated triangles. In the four random clips the two triangles are moving randomly and unintentionally (e.g., bouncing about), in the four ToM clips the triangles are interacting intentionally (e.g., the large triangle trying to convince the small triangle to come outside), and in the other four clips they are merely performing an activity alone or together. The duration of each animation is approximately 40 seconds. The participants were asked to provide an interpretation of what was going on in each animation and their answers were audio-recorded and then transcribed by research assistants (see **SM2** for more details on the task). After the transcription, fillers such as 'uhm' and "ehm" (see **SM3** for more details) were removed. No other preprocessing was performed; specifically, interjections were not removed.

Speech pre-processing

We prepared the data for NLP-based analysis by using the UDPipe Natural Language Processing - Text Annotation in R (Straka et al., 2016). First, words were tokenized (identified as parts of speech), and then each transcript was parsed into phrases, using rules of grammar for each specific language. Words were then converted to the roots from which they are inflected, or lemmatized. The resulting pre-processed speech data yielded for each transcript a series of lemmatized words, maintaining the original order in which they were spoken. We then used fastText pre-trained models for the different languages (Bojanowski et al., 2017) to vectorize the tokenized speech samples, yielding a 300-dimensional vector for each word.

After that, we computed semantic coherence between words, i.e., word-to-word similarity, by calculating the cosine similarity between the corresponding vectors associated with each word. The cosine similarity values range between − 1 and 1, with − 1 representing the lowest similarity and 1 the highest similarity between two words (see **SM3** for more details).

Literature search and NLP measures of coherence

We systematically screened the current literature - following the indications of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Guidelines (PRISMA) statement (Rethlefsen et al., 2021) - to identify previous studies in the literature which quantified semantic coherence in schizophrenia using NLP automated methods (full details on the search on **SM4**). Among the final set of selected studies, we selected scalable measures of coherence, i.e. easier to apply to a larger set of languages (without ad hoc wordlists, etc), and to corpora with limited size (e.g. excluding training deep learning methods on the corpus). We found 14 studies using 14 different NLP measures to quantify semantic coherence in schizophrenia: we thus derived those coherence measures and tried to replicate their results on our corpus (**Table 2** and **SM4**). Median and interquartile range (IQR) were calculated for each of these measures (see **SM4**). We report in the main manuscript only results for median, while IQR results are reported in **SM5**. We opted to use median and IQR of coherence measures, even if previous studies used a wide variety of descriptors (e.g. standard deviation, maximum, minimum, etc..) because they are more robust to measurement errors. Note that any difference (e.g. in data preprocessing, or word embeddings employed) between the original studies and the current one is motivated by the goal of building a data analysis framework able to extract different measures of coherence and compare them in a scalable way across different samples and languages, and fully detailed in **SM4**.

Table 2. NLP coherence measures identified in the previous studies and derived in the present research.

NLP - Coherence Measures	Description	References
Similarity Mean	Average semantic similarity of each word to the immediately preceding word	Pauselli et al. (2018) Ryazanskaya & Khudyakova, 2020 Bar et al., 2019
Coherence 5	Average semantic similarity of each word in 5-words window	Pauselli et al. (2018)
Coherence 10	Average semantic similarity of each word in 10-words window	Pauselli et al. (2018)

Coherence-K2		
Coherence-K3		Bar et al. (2019)
Coherence-K4		
Coherence-K5		
Coherence-K6	Word-to-word variability at k inter-word distances	Corcoran et al. (2018)
Coherence-K7		
Coherence-K8		
Coherence-k9		Voppel et al., 2021
Coherence-k10		
First-order Coherence	Similarity of consecutive phrase vectors[1]	Bedi et al., (2015) Just et al., 2019 Morgan et al. (2021) Haas et al. (2020) Iter et al. (2018) Just et al. (2020) Sarzynska-Wawer et al., (2021)
Second-order Coherence	Similarity between phrases separated by another intervening phrase	Bedi et al. (2015) Sarzynska-Wawer et al., (2021)

Analysis of effect of diagnosis on (differences in) coherence measures

To estimate the differences between individuals with schizophrenia and HC in the different coherence measures, we used Bayesian multilevel regression models on the current data with each coherence measure as outcome, and diagnosis (schizophrenia vs. HC) and language (DK, GE, CH) as predictors. Within the same model, we separately assessed the effect of diagnosis for each language, and modeled varying effects of participants, i.e., intercepts and slopes, separately for each group and language. For each coherence measure, we built a model with weakly informative priors, i.e., expectations of no effects of diagnosis, thus conservatively regularizing the model parameters, reducing overfitting and leading to improved predictions (Gelman et al., 2020). We then built a second model with informed priors (when available), that is summary effect sizes (ES, see **SM5**), and compared results across the two models. We aimed to assess whether the effects of diagnosis are robust across changes of priors, and whether the skeptical or informed priors led to more robust inference, that is, in lower

[1]

estimated out-of-sample error - measured in terms of Leave-One-Out based stacking weights. To evaluate the potential role of gender (male vs. female), age and level of intelligence we built additional models, one per each moderator interacting with group separately in the three languages. We then reported the model estimates for the interaction, including credible (i.e., Bayesian confidence) intervals (CIs) and evidence ratios (ERs), i.e. evidence in favor of the effect observed against alternative hypotheses. When ER was weak (below 10, that is, less than ten times as much evidence for the effect as for alternative hypotheses), we also calculated the Evidence Ratio in favor of the null hypothesis. Further details are presented in the **SM5**. Note also that we report additional analyses in the **SM6** to assess the robustness of the findings: we repeated all analyses by: 1) including fillers and 2) including fillers and punctuation 3) explicitly assessing the association between transcript length (total number of words) and the different coherence measures. The results support our main findings and we report in the manuscript only qualitative divergences.

Analysis of the relationship between coherence measures and clinical ratings

To assess the relationship between the coherence measures and clinical ratings, we built Bayesian multilevel regression models with each coherence measure as outcome, and clinical features (one at a time) as ordinal predictors. We separately assessed the relationship between the different coherence measures and clinical ratings for each language, and modeled varying effects of participants, i.e. intercepts and slopes, separately for each language. This analysis was performed on the schizophrenia group only (see **SM5** for more details). All the code used for the analysis and the extracted features are openly available (see **SM7**).

Results

Effect of Diagnosis

The detailed results are reported in **Table 2** and **Figure 1**. We only partially replicated previous findings: reduced Similarity mean, Coherence 10, and Coherence-K in schizophrenia were found only in the Chinese and German corpus, while increased Coherence-K and increased Coherence 10 measures were found in the Danish corpus. Reduced first order coherence in schizophrenia was found in the Danish and Chinese corpus, while reduced second-order coherence was found in all the corpus. We found a lower total number of words in schizophrenia in the Danish and Chinese corpus. Globally, we found important differences within (i.e. between the diverse coherence measures) and between languages. In agreement

with the inconsistent replications, the informed models were more robust and generalizable to new data (LOO weights for informed models above .75) in less than half of the models (3 on 8, in 1 model there was no difference), indicating that prior findings were not fully representative of the current samples. Gender, age, education and level of intelligence of the participants also affected the group differences, although inconsistently across languages. Our results are robust to the inclusion of fillers and punctuations in the analysis, even if some differences are present and discussed in **SM6**. We also found a positive association between transcript length and the various coherence measures, that is longer transcripts tend to have higher coherence values.

Table 2. Estimated standardized mean difference (controls – patients with schizophrenia) for the fourteen coherence measures, as estimated separately by the informed and the skeptical models

Coherence measure	Group (HC – SCZ)	Group * Gender (M-F)	Age	Education	IQ
Similarity Mean	-0.83 (-1.11 -0.54)				
Skeptical DK	0.04 (-0.06 0.15) ER = 2.9 ER01 = 6.44	-0.06 (-0.27 0.16) ER = 2.03 ER01 = 4.81	**-0.01 (-0.02 0) ER = 21.52**	0.01 (-0.02 0.03) ER = 1.93 ER01 = 28.35	0 (-0.01 0) ER = 2.65 ER01 = 110.8
Skeptical GE	-0.03 (-0.19 0.12) ER = 1.76 ER01 = 5.3	-0.1 (-0.42 0.22) ER = 2.36 ER01 = 3.07	0 (-0.01 0.02) ER = 1.64 ER01 = 45.29	-0.01 (-0.07 0.06) ER = 1.3 ER01 = 11.66	0 (-0.83 0.82) ER = 1 ER01 = 0.97
Skeptical CH	**-0.27 (-0.47 -0.07) ER = 77.12**	-0.27 (-0.67 0.12) ER = 6.89	0 (-0.02 0.02) ER = 1.69 ER01 = 40.63	-0.01 (-0.06 0.03) ER = 2.23 ER01 = 17.69	0 (-0.01 0) ER = 4.3
Informed DK	**-0.1 (-0.2 0) ER = 16.7**	NA	NA	NA	NA
Informed GE	**-0.29 (-0.45 -0.14) ER = 2499**	NA	NA	NA	NA
Informed CH	**-0.52 (-0.68 -0.36) ER = Inf**	NA	NA	NA	NA
Stacking weight	**Skeptic Model 0.93**				
Coherence 5	-0.26 (-0.58 0.06)				
Skeptical DK	0.09 (-0.03 0.21) ER = 9.45	0.03 (-0.21 0.26) ER = 1.4 ER01 = 4.93	**-0.02 (-0.03 0) ER = 61.11**	0 (-0.03 0.03) ER = 1.48 ER01 = 26.68	0 (-0.01 0.01) ER = 1.49 ER01 = 128.8
Skeptical GE	-0.13 (-0.3 0.05) ER = 7.84	-0.23 (-0.57 0.1) ER = 6.7	0 (-0.02 0.02) ER = 1.03 ER01 = 48.47	0 (-0.07 0.07) ER = 1.1 ER01 = 11.37	0 (-0.82 0.83) ER = 1.01 ER01 = 1
Skeptical CH	**-0.41 (-0.62 -0.19) ER = 665.67**	-0.21 (-0.62 0.21) ER = 4.01	0 (-0.02 0.02) ER = 1.31 ER01 = 39.56	-0.02 (-0.07 0.03) ER = 2.28 ER01 = 14.54	-0.01 (-0.01 0) ER = 7.41
Informed DK	0.03 (-0.08 0.14) ER = 2.2 ER01 = 7.44	NA	NA	NA	NA
Informed GE	**-0.17 (-0.32 -0.03) ER = 41.19**	NA	NA	NA	NA

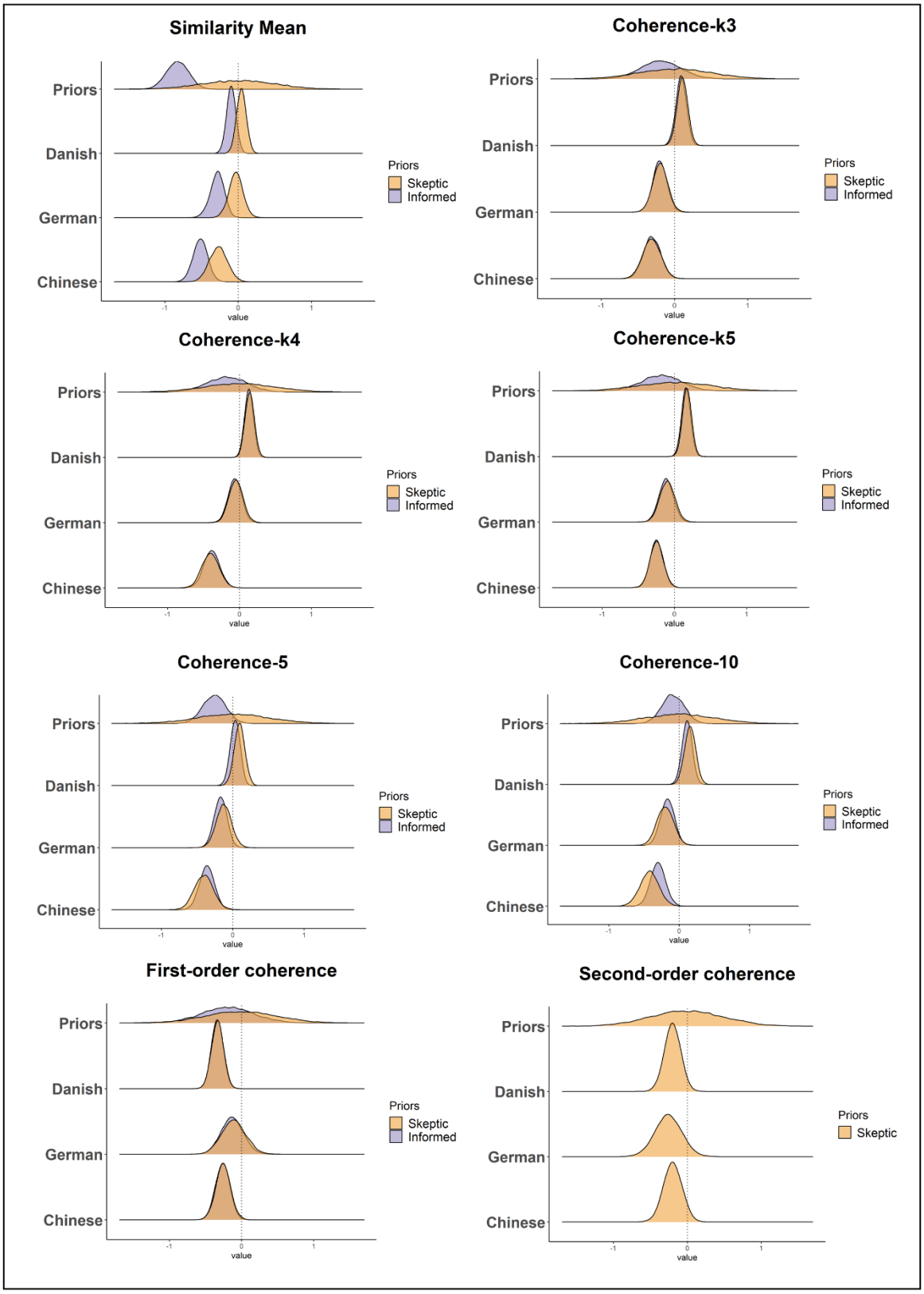
Informed CH	**-0.36 (-0.53 -0.2) ER = 3332.33**	NA	NA	NA	NA
Stacking weight	**Skeptic Model 0.77**				
Coherence 10	-0.09 (-0.41 0.23)				
Skeptical DK	**0.15 (0.03 0.28) ER = 41.02**	0.06 (-0.19 0.3) ER = 1.82 ER01 = 4.38	**-0.02 (-0.03 -0.01) ER = 499**	**-0.03 (-0.06 0.01) ER = 10.4**	0 (-0.01 0) ER = 2.47 ER01 = 103.79
Skeptical GE	**-0.2 (-0.39 -0.01) ER = 21.57**	**-0.33 (-0.71 0.07) ER = 11.18**	-0.01 (-0.02 0.01) ER = 2.16 ER01 = 41.21	-0.01 (-0.09 0.06) ER = 1.58 ER01 = 10.96	0.01 (-0.79 0.81) ER = 1.02 ER01 = 1
Skeptical CH	**-0.42 (-0.65 -0.2) ER = 1665.67**	-0.12 (-0.54 0.31) ER = 2.08 ER01 = 2.5	0 (-0.02 0.02) ER = 1.17 ER01 = 40.75	-0.03 (-0.08 0.02) ER = 4.27	-0.01 (-0.01 0) ER = 6.34
Informed DK	**0.11 (0 0.22) ER = 17.02**	NA	NA	NA	NA
Informed GE	**-0.17 (-0.33 -0.01) ER = 23.45**	NA	NA	NA	NA
Informed CH	**-0.3 (-0.47 -0.13) ER = 453.55**	NA	NA	NA	NA
Stacking weight	**Skeptic Model 0.5**				
Coherence K2	-0.33 (-0.88 0.23)				
Skeptical DK	**0.16 (0.04 0.27) ER = 87.5**	0.1 (-0.12 0.33) ER = 3.6	**-0.02 (-0.03 0) ER = 85.21**	-0.01 (-0.04 0.02) ER = 3.15	0 (-0.01 0) ER = 2.22 ER01 = 119.48
Skeptical GE	-0.12 (-0.27 0.04) ER = 8.09	**-0.29 (-0.59 0.01) ER = 16.89**	-0.01 (-0.02 0.01) ER = 4.66	-0.01 (-0.07 0.06) ER = 1.36 ER01 = 13.35	-0.01 (-0.84 0.79) ER = 1.02 ER01 = 1.04
Skeptical CH	-0.16 (-0.4 0.09) ER = 5.61	0.13 (-0.35 0.61) ER = 2.08 ER01 = 2.27	0 (-0.02 0.03) ER = 1.99 ER01 = 36.78	-0.02 (-0.07 0.02) ER = 3.54	0 (-0.01 0.01) ER = 2.44 ER01 = 86.35
Informed DK	**0.13 (0.02 0.24) ER = 41.74**	NA	NA	NA	NA
Informed GE	**-0.15 (-0.3 0) ER = 17.45**	NA	NA	NA	NA
Informed CH	**-0.2 (-0.42 0.02) ER = 14.58**	NA	NA	NA	NA
	Skeptic Model 1.0				
Coherence K3	-0.19 (-0.74 0.36)				
Skeptical DK	**0.11 (-0.01 0.22) ER = 14.77**	-0.05 (-0.27 0.18) ER = 1.73 ER01 = 4.89	**-0.01 (-0.03 0) ER = 48.75**	-0.01 (-0.04 0.02) ER = 3.75	0 (-0.01 0.01) ER = 1.62 ER01 = 125.78
Skeptical GE	**-0.2 (-0.36 -0.03) ER = 37.91**	**-0.3 (-0.62 0.03) ER = 14.11**	0 (-0.02 0.02) ER = 0.99 ER01 = 53.63	-0.02 (-0.09 0.05) ER = 2.09 ER01 = 10.79	0 (-0.83 0.81) ER = 1.03 ER01 = 1.03
Skeptical CH	**-0.31 (-0.51 -0.11) ER = 195.08**	0.09 (-0.31 0.49) ER = 1.84 ER01 = 2.83	0.01 (-0.01 0.03) ER = 2.04 ER01 = 37.18	-0.01 (-0.06 0.03) ER = 2.11 ER01 = 15.58	0 (-0.01 0) ER = 4.59
Informed DK	0.09 (-0.02 0.2) ER = 9.4	NA	NA	NA	NA

Informed GE	**-0.2 (-0.36 -0.04) ER = 58.88**	NA	NA	NA	NA
Informed CH	**-0.31 (-0.5 -0.12) ER = 242.9**	NA	NA	NA	NA
Stacking weight	**Skeptic Model 1.0**				
Coherence K4	-0.18 (-0.73 0.37)				
Skeptical DK	**0.14 (0.04 0.24) ER = 112.64**	0.08 (-0.12 0.27) ER = 2.88 ER01 = 4.93	**-0.01 (-0.02 0) ER = 26.32**	**-0.02 (-0.05 0) ER = 10.92**	0 (-0.01 0) ER = 2.22 ER01 = 126.98
Skeptical GE	-0.05 (-0.21 0.11) ER = 2.35 ER01 = 4.52	**-0.27 (-0.59 0.05) ER = 11.22**	0 (-0.02 0.01) ER = 1.33 ER01 = 51.33	0.01 (-0.05 0.07) ER = 1.52 ER01 = 12.59	0 (-0.85 0.83) ER = 0.99 ER01 = 0.99
Skeptical CH	**-0.4 (-0.6 -0.21) ER = 1249**	0.05 (-0.34 0.43) ER = 1.39 ER01 = 3.12	0 (-0.02 0.02) ER = 1.46 ER01 = 41.52	-0.01 (-0.06 0.03) ER = 1.91 ER01 = 15.95	0 (-0.01 0) ER = 4.04
Informed DK	**0.13 (0.04 0.22) ER = 91.59**	NA	NA	NA	NA
Informed GE	-0.07 (-0.22 0.09) ER = 3.11	NA	NA	NA	NA
Informed CH	**-0.39 (-0.57 -0.2) ER = 1249**	NA	NA	NA	NA
Stacking weight	**Informed Model 1.0**				
Coherence K5	-0.18 (-0.73 0.37)				
Skeptical DK	**0.17 (0.08 0.27) ER = 356.14**	0.08 (-0.12 0.27) ER = 2.86 ER01 = 4.67	**-0.02 (-0.02 -0.01) ER = 262.16**	**-0.03 (-0.06 -0.01) ER = 77.12**	0 (-0.01 0) ER = 2.54 ER01 = 121.5
Skeptical GE	-0.1 (-0.27 0.07) ER = 4.98	**-0.3 (-0.63 0.03) ER = 14.48**	0 (-0.02 0.02) ER = 1.13 ER01 = 49.29	-0.02 (-0.08 0.05) ER = 2.12 ER01 = 11.92	0 (-0.83 0.83) ER = 1 ER01 = 0.99
Skeptical CH	**-0.25 (-0.4 -0.1) ER = 262.16**	0.08 (-0.22 0.39) ER = 1.99 ER01 = 3.36	0 (-0.01 0.02) ER = 1.33 ER01 = 49.15	0 (-0.04 0.04) ER = 1.04 ER01 = 22.36	0 (-0.01 0.01) ER = 1.72 ER01 = 112.57
Informed DK	**0.16 (0.06 0.26) ER = 262.16**	NA	NA	NA	NA
Informed GE	-0.12 (-0.28 0.05) ER = 7.5	NA	NA	NA	NA
Informed CH	**-0.25 (-0.39 -0.1) ER = 453.55**	NA	NA	NA	NA
Stacking weight	**Informed Model 0.88**				
Coherence K6					
Skeptical DK	0.07 (-0.03 0.16) ER = 7.16	0.03 (-0.16 0.21) ER = 1.51 ER01 = 6.25	**-0.01 (-0.02 0) ER = 32.56**	**-0.02 (-0.05 0) ER = 17.94**	0 (-0.01 0) ER = 5.04
Skeptical GE	**-0.18 (-0.35 0) ER = 18.08**	**-0.49 (-0.84 -0.12) ER = 58.17**	0 (-0.02 0.02) ER = 1.23 ER01 = 50.42	-0.03 (-0.1 0.03) ER = 3.58	0.01 (-0.81 0.82) ER = 1.03 ER01 = 0.96
Skeptical CH	**-0.27 (-0.45 -0.09) ER = 165.67**	0.04 (-0.31 0.39) ER = 1.4 ER01 = 3.41	0.01 (-0.01 0.02) ER = 2.53 ER01 = 40.3	-0.01 (-0.06 0.03) ER = 2.07 ER01 = 17.41	0 (-0.01 0) ER = 3.45
Coherence K7					

Skeptical DK	**0.11 (0.01 0.22) ER = 26.25**	0.06 (-0.16 0.28) ER = 2.13 ER01 = 4.68	**-0.01 (-0.02 0) ER = 10.48**	-0.02 (-0.05 0.01) ER = 4.26	0 (-0.01 0.01) ER = 1.81 ER01 = 118.59
Skeptical GE	-0.06 (-0.22 0.11) ER = 2.55 ER01 = 4.42	-0.13 (-0.48 0.22) ER = 2.85 ER01 = 2.72	-0.01 (-0.02 0.01) ER = 3.04	-0.03 (-0.1 0.04) ER = 3.5	0 (-0.84 0.85) ER = 0.99 ER01 = 0.96
Skeptical CH	**-0.32 (-0.5 -0.13) ER = 262.16**	-0.07 (-0.43 0.29) ER = 1.68 ER01 = 3.11	0.01 (-0.01 0.02) ER = 2.4 ER01 = 39.96	-0.01 (-0.05 0.03) ER = 2.16 ER01 = 18.76	0 (-0.01 0) ER = 7.12
Coherence K8					
Skeptical DK	**0.13 (0.02 0.23) ER = 46.39**	-0.01 (-0.22 0.2) ER = 1.21 ER01 = 5.55	**-0.02 (-0.03 -0.01) ER = 157.73**	-0.02 (-0.05 0.01) ER = 7.58	0 (-0.01 0) ER = 5.06
Skeptical GE	-0.11 (-0.28 0.06) ER = 5.86	**-0.31 (-0.62 0.02) ER = 15.53**	0 (-0.02 0.01) ER = 1.23 ER01 = 54.38	-0.02 (-0.08 0.05) ER = 1.94 ER01 = 10.78	0 (-0.81 0.82) ER = 1.01 ER01 = 1.03
Skeptical CH	**-0.22 (-0.4 -0.03) ER = 35.5**	-0.06 (-0.43 0.3) ER = 1.53 ER01 = 3.11	0.01 (-0.01 0.03) ER = 3.28	-0.03 (-0.07 0.02) ER = 5.89	0 (-0.01 0) ER = 2.8 ER01 = 99.73
Coherence K9					
Skeptical DK	**0.12 (-0.02 0.27) ER = 11.92**	-0.07 (-0.38 0.23) ER = 1.83 ER01 = 3.49	0 (-0.02 0.01) ER = 2.06 ER01 = 56.5	**-0.04 (-0.07 0) ER = 20.01**	**-0.01 (-0.01 0) ER = 10.6**
Skeptical GE	-0.09 (-0.27 0.09) ER = 4.12	-0.23 (-0.6 0.15) ER = 5.33	0 (-0.02 0.01) ER = 2.2 ER01 = 45.88	-0.01 (-0.08 0.07) ER = 1.19 ER01 = 10.72	0 (-0.82 0.82) ER = 1.01 ER01 = 1
Skeptical CH	**-0.36 (-0.56 -0.17) ER = 768.23**	-0.13 (-0.5 0.25) ER = 2.52 ER01 = 2.67	-0.01 (-0.02 0.01) ER = 2.01 ER01 = 38.42	0 (-0.05 0.04) ER = 1.07 ER01 = 18.68	0 (-0.01 0) ER = 3.84
Coherence K10					
Skeptical DK	**0.12 (0.02 0.21) ER = 33.6**	-0.05 (-0.24 0.14) ER = 2.05 ER01 = 5.48	**-0.01 (-0.02 0) ER = 127.21**	-0.02 (-0.04 0.01) ER = 7.8	0 (-0.01 0) ER = 5.76
Skeptical GE	-0.08 (-0.25 0.08) ER = 3.87	-0.13 (-0.47 0.22) ER = 2.65 ER01 = 2.85	-0.01 (-0.02 0.01) ER = 2.3 ER01 = 44.11	-0.01 (-0.08 0.05) ER = 1.7 ER01 = 11.88	0 (-0.82 0.83) ER = 1.01 ER01 = 0.98
Skeptical CH	**-0.37 (-0.56 -0.17) ER = 3332.33**	-0.14 (-0.53 0.25) ER = 2.68 ER01 = 2.43	0 (-0.03 0.01) ER = 1.93 ER01 = 38.98	0 (-0.05 0.04) ER = 1.11 ER01 = 18.66	0 (-0.01 0) ER = 3.82
First-order Coherence					
Skeptical DK	**-0.33 (-0.47 -0.2) ER = Inf**	0.1 (-0.16 0.36) ER = 2.87 ER01 = 3.88	0 (-0.02 0.01) ER = 1.91 ER01 = 58.2	0.02 (-0.02 0.05) ER = 3.64	0 (-0.01 0.01) ER = 1.22 ER01 = 117.51
Skeptical GE	-0.11 (-0.38 0.17) ER = 3.01	0.11 (-0.42 0.66) ER = 1.76 ER01 = 1.96	0 (-0.02 0.02) ER = 1.24 ER01 = 41.35	**0.08 (0.01 0.16) ER = 25.04**	0 (-0.81 0.82) ER = 1.03 ER01 = 0.98
Skeptical CH	**-0.26 (-0.42 -0.09) ER = 157.73**	0.17 (-0.14 0.49) ER = 4.55	0.01 (-0.01 0.03) ER = 3.36	-0.02 (-0.06 0.02) ER = 4.51	0 (-0.01 0.01) ER = 1.6 ER01 = 116.25
Informed DK	**-0.34 (-0.46 -0.21) ER = Inf**	NA	NA	NA	NA
Informed GE	-0.13 (-0.38 0.12) ER = 4.14	NA	NA	NA	NA
Informed CH	**-0.27 (-0.42 -0.11) ER = 311.5**	NA	NA	NA	NA
Stacking weight	**Skeptic Model 0.64**				

Second-order Coherence					
Skeptical DK	**-0.2 (-0.38 -0.03) ER = 30.85**	0.16 (-0.2 0.51) ER = 3.53	-0.01 (-0.03 0) ER = 7.45	0.01 (-0.03 0.05) ER = 1.45 ER01 = 20.66	0 (-0.01 0.01) ER = 1.28 ER01 = 96.27
Skeptical GE	**-0.26 (-0.56 0.04) ER = 12.09**	**-0.39 (-1.02 0.24) ER = 5.88**	-0.01 (-0.03 0.01) ER = 3.92	0.02 (-0.07 0.1) ER = 1.69 ER01 = 9.19	0 (-0.83 0.82) ER = 1 ER01 = 1
Skeptical CH	**-0.2 (-0.4 0.01) ER = 17.02**	0.2 (-0.18 0.58) ER = 4.21	0.01 (-0.01 0.03) ER = 2.64 ER01 = 35.37	0.02 (-0.03 0.06) ER = 2.34 ER01 = 15.3	0 (-0.01 0.01) ER = 1.83 ER01 = 113.99
Total Words					
Skeptical DK	**-0.21 (-0.37 -0.05) ER = 80.3**	0.09 (-0.2 0.39) ER = 2.29 ER01 = 3.25	-0.01 (-0.02 0.01) ER = 4.75	NA	0 (0 0.01) ER = 3.37
Skeptical GE	-0.1 (-0.37 0.18) ER = 2.59 ER01 = 2.52	0.19 (-0.31 0.7) ER = 2.75 ER01 = 1.75	0.01 (-0.02 0.03) ER = 1.98 ER01 = 32.27	NA	0 (-0.81 0.81) ER = 1.01 ER01 = 1.01
Skeptical CH	**-0.25 (-0.49 -0.01) ER = 21.17**	**0.47 (0.04 0.93) ER = 25.95**	0 (-0.02 0.02) ER = 1.41 ER01 = 42.09	NA	0 (-0.01 0.01) ER = 1.29 ER01 = 108.88

Figure 1. Comparing informed and skeptical expectations and results. Each panel presents a separate coherence measure, with the x-axis corresponding to standardized mean differences (schizophrenia - HC) equivalent to Hedges' g, with estimates above 0 indicating higher scores for patients with schizophrenia.



Effect of Symptoms

Detailed results are reported in **Table 4, Table 5 and Table SM5_B**. Globally, clinical ratings of symptoms correlate with NLP measures of semantic coherence. We found several associations between Coherence-k, Coherence 5-10 and First-order coherence measures, and SAPS ratings (Global SAPS, SAPS FTD, Derailment, Tangentiality, Incoherence). However, these associations were not robust across languages, and the direction of the effects were often inconsistent. We found more consistent associations SANS ratings and several coherence measures, i.e. higher SANS ratings (global SANS, global alogia, poverty of content and poverty of speech) were generally associated with reduced coherence. We found a very weak association between PANSS symptoms and coherence measures (see **Table SM5)**. Most of the correlations were between small and moderate (5-20% of explained variance), and varied across languages and rating scales.

Table 4. Estimated standardized relation between coherence measures and clinical features (SAPS). ER indicates the evidence ratio for the difference, ER01 the evidence ratio for the null effect.

Rating scales	**SAPS – Global**	**SAPS – Formal Thought Disorder**	**SAPS _ Illogicality**	**SAPS - Incoherence**	**SAPS - Derailment**	**SAPS – Tangentiality**
Mean Similarity						
Skeptical DK	-0.01 (-0.05 0.02) ER = 2.46 ER01 = 26.83	-0.01 (-0.03 0.01) ER = 2.84 ER01 = 36.43	0.01 (-0.02 0.03) ER = 1.62 ER01 = 34.67	-0.02 (-0.06 0.02) ER = 3.27	0 (-0.02 0.03) ER = 1.21 ER01 = 38.41	0.02 (-0.01 0.06) ER = 8.26
Skeptical CH	0.02 (-0.05 0.09) ER = 2.04 ER01 = 12.45	-0.01 (-0.21 0.18) ER = 1.18 ER01 = 7.34	-0.02 (-0.34 0.27) ER = 1.25 ER01 = 5.79	0.37 (-0.21 1.04) ER = 7.56	-0.01 (-0.23 0.19) ER = 1.2 ER01 = 6.85	0.05 (-0.08 0.22) ER = 2.63 ER01 = 7.35
Coherence 5						
Skeptical DK	-0.03 (-0.09 0.03) ER = 3.03	0.01 (-0.03 0.04) ER = 1.57 ER01 = 25.78	-0.01 (-0.07 0.04) ER = 2.03 ER01 = 17.53	**-0.08 (-0.16 -0.01) ER = 29.3**	0 (-0.05 0.05) ER = 1.26 ER01 = 19.98	**0.06 (0 0.13) ER = 15.48**
Skeptical CH	0.04 (-0.08 0.16) ER = 2.25 ER01 = 7.57	0.02 (-0.26 0.34) ER = 1.14 ER01 = 4.44	0.02 (-0.35 0.43) ER = 1.18 ER01 = 3.8	0.49 (-0.15 1.17) ER = 9.55	0.09 (-0.21 0.45) ER = 2.33 ER01 = 3.57	-0.02 (-0.28 0.24) ER = 1.18 ER01 = 4.42
Coherence 10						

Skeptical DK	0.02 (-0.05 0.1) ER = 2.09 ER01 = 12.36	**0.06 (0 0.11) ER = 23.29**	0.01 (-0.06 0.08) ER = 1.52 ER01 = 15.05	-0.02 (-0.12 0.08) ER = 1.5 ER01 = 10.65	0.04 (-0.02 0.1) ER = 5.4	**0.08 (0.01 0.17) ER = 24.21**
Skeptical CH	0.06 (-0.08 0.21) ER = 3.39	0.1 (-0.2 0.45) ER = 2.63 ER01 = 3.28	-0.03 (-0.48 0.38) ER = 1.25 ER01 = 3.26	**0.57 (-0.08 1.24) ER = 14.23**	0.22 (-0.07 0.61) ER = 8.58	0.13 (-0.18 0.47) ER = 3.25
Coherence K5						
Skeptical DK	**0.03 (0 0.05) ER = 30.91**	**0.02 (0 0.03) ER = 51.17**	0.01 (-0.01 0.03) ER = 3.01	0.02 (-0.01 0.04) ER = 4.58	**0.03 (0.01 0.05) ER = 332.33**	**0.03 (0.01 0.05) ER = 114.38**
Skeptical CH	-0.01 (-0.06 0.04) ER = 1.79 ER01 = 19.98	-0.01 (-0.16 0.14) ER = 1.17 ER01 = 9.64	-0.05 (-0.33 0.15) ER = 2.03 ER01 = 8.11	**0.38 (-0.1 1.03) ER = 11.74**	0.02 (-0.12 0.18) ER = 1.4 ER01 = 9.27	-0.01 (-0.14 0.12) ER = 1.13 ER01 = 10.01
Coherence K6						
Skeptical DK	0.01 (-0.02 0.04) ER = 3.17	**0.02 (0.01 0.05) ER = 61.5**	0.02 (-0.01 0.04) ER = 8.79	**0.03 (0 0.06) ER = 11.61**	0.01 (-0.01 0.04) ER = 4.59	**0.03 (0 0.05) ER = 20.58**
Skeptical CH	0.03 (-0.03 0.1) ER = 4.77	0.1 (-0.05 0.32) ER = 5.78	0.02 (-0.25 0.32) ER = 1.33 ER01 = 6.27	**0.41 (-0.09 1.04) ER = 11.96**	**0.14 (-0.02 0.42) ER = 11.22**	0.06 (-0.09 0.23) ER = 2.72 ER01 = 6.79
Coherence K7						
Skeptical DK	0 (-0.03 0.02) ER = 1.34 ER01 = 40.01	0.01 (-0.01 0.02) ER = 2.59 ER01 = 55.07	**0.02 (0 0.04) ER = 12.1**	0 (-0.04 0.03) ER = 1.34 ER01 = 31.01	0.01 (-0.01 0.02) ER = 2.15 ER01 = 47.66	0 (-0.02 0.03) ER = 1.64 ER01 = 42.87
Skeptical CH	0.01 (-0.04 0.06) ER = 2.15 ER01 = 18.77	0.08 (-0.04 0.26) ER = 6.17	0.01 (-0.23 0.27) ER = 1.05 ER01 = 7.95	**0.37 (-0.11 1.02) ER = 11.1**	**0.11 (-0.02 0.31) ER = 10.52**	0.03 (-0.11 0.18) ER = 1.99 ER01 = 7.71
Coherence K8						
Skeptical DK	-0.01 (-0.05 0.03) ER = 1.81 ER01 = 25.29	0.02 (-0.01 0.04) ER = 5.42	0.01 (-0.02 0.04) ER = 2.12 ER01 = 28.31	0.03 (-0.02 0.07) ER = 5.98	**0.03 (0 0.06) ER = 19.48**	**0.03 (0 0.07) ER = 15.04**
Skeptical CH	0.01 (-0.04 0.06) ER = 2.12 ER01 = 17.84	0.07 (-0.07 0.26) ER = 3.97	-0.1 (-0.44 0.11) ER = 3.55	0.31 (-0.27 0.94) ER = 5.8	**0.14 (0 0.37) ER = 20.05**	0.11 (-0.04 0.3) ER = 9.05
First-order Coherence						
Skeptical DK	**-0.07 (-0.14 0) ER = 21.9**	-0.04 (-0.09 0.01) ER = 9.51	-0.01 (-0.09 0.07) ER = 1.27 ER01 = 12.88	0.01 (-0.09 0.1) ER = 1.28 ER01 = 10.19	**-0.06 (-0.13 0.01) ER = 11.66**	-0.05 (-0.12 0.02) ER = 8.92

Skeptical CH	**0.08 (-0.01 0.18) ER = 13.56**	-0.06 (-0.34 0.16) ER = 1.97 ER01 = 5.28	**-0.25 (-0.71 - 0.01) ER = 20.98**	0.26 (-0.47 0.97) ER = 3.32	0.06 (-0.21 0.34) ER = 2.06 ER01 = 4.23	-0.09 (-0.36 0.13) ER = 3.06
Second-order Coherence						
Skeptical DK	-0.01 (-0.1 0.08) ER = 1.26 ER01 = 11.31	0 (-0.07 0.06) ER = 1.16 ER01 = 17.25	0.05 (-0.04 0.14) ER = 4.96	0.11 (-0.04 0.26) ER = 7.73	-0.02 (-0.12 0.06) ER = 1.94 ER01 = 11.04	-0.04 (-0.12 0.03) ER = 4.59
Skeptical CH	-0.01 (-0.13 0.11) ER = 1.3 ER01 = 8.53	-0.16 (-0.51 0.09) ER = 6.1	**-0.27 (-0.73 0.04) ER = 12.22**	0.19 (-0.61 0.99) ER = 2.14 ER01 = 1.24	-0.06 (-0.4 0.24) ER = 1.81 ER01 = 4.08	-0.1 (-0.4 0.15) ER = 3.1
Total Words						
Skeptical DK	**-0.09 (-0.16 - 0.02) ER = 52.57**	**-0.07 (-0.12 - 0.03) ER = 427.57**	-0.02 (-0.09 0.04) ER = 2.52 ER01 = 13.52	**-0.08 (-0.17 0) ER = 20.66**	**-0.07 (-0.13 - 0.01) ER = 47**	-0.02 (-0.09 0.04) ER = 2.63 ER01 = 11.82
Skeptical CH	0.05 (-0.03 0.13) ER = 5.7	0.07 (-0.14 0.33) ER = 2.47 ER01 = 5.34	0.05 (-0.28 0.43) ER = 1.66 ER01 = 4.47	**0.8 (0.4 1.39) ER = Inf**	0.17 (-0.04 0.48) ER = 8.95	0.01 (-0.17 0.19) ER = 1.37 ER01 = 6.58

Table 5. Estimated standardized relation between coherence measures and clinical features (SANS). ER indicates the evidence ratio for the difference, ER01 the evidence ratio for the null effect.

Rating scales	SANS - Global	SANS - Alogia	SANS - Poverty of content	SANS - Poverty of speech
Mean Similarity				
Skeptical DK	-0.02 (-0.06 0.02) ER = 3.31	-0.02 (-0.06 0.01) ER = 7.85	-0.02 (-0.04 0.01) ER = 5.04	**-0.04 (-0.09 -0.01) ER = 39.27**
Skeptical CH	-0.07 (-0.2 0.04) ER = 6.08	-0.06 (-0.25 0.1) ER = 2.91 ER01 = 6.88	0.11 (-0.54 0.78) ER = 1.87 ER01 = 2.28	-0.09 (-0.29 0.05) ER = 5.75
Coherence 5				
Skeptical DK	-0.02 (-0.09 0.05) ER = 1.92 ER01 = 13.29	-0.02 (-0.07 0.03) ER = 2.43 ER01 = 16.81	0.02 (-0.03 0.07) ER = 3.11	-0.04 (-0.11 0.02) ER = 6.33
Skeptical CH	-0.12 (-0.32 0.06) ER = 6.41	**-0.19 (-0.55 0.04) ER = 10.88**	0.23 (-0.45 0.92) ER = 3.13	**-0.26 (-0.58 -0.06) ER = 63.52**
Coherence 10				
Skeptical DK	0.06 (-0.03 0.16) ER = 7.08	0.05 (-0.02 0.12) ER = 6.78	**0.07 (0.01 0.13) ER = 27.04**	0.03 (-0.05 0.13) ER = 2.52 ER01 = 10.45
Skeptical CH	-0.08 (-0.31 0.15) ER = 2.51 ER01 = 4.06	-0.14 (-0.48 0.14) ER = 4.14	0.27 (-0.43 0.96) ER = 3.61	**-0.26 (-0.63 -0.01) ER = 24.32**
Coherence K5				
Skeptical DK	**0.04 (0.02 0.07) ER = 299**	**0.03 (0.01 0.05) ER = 87.24**	**0.03 (0.01 0.04) ER = 135.36**	0.01 (-0.01 0.05) ER = 3.61
Skeptical CH	**-0.08 (-0.18 0) ER = 21.22**	-0.09 (-0.27 0.04) ER = 6.98	**0.36 (0 0.95) ER = 19.55**	**-0.12 (-0.32 -0.01) ER = 31.79**
Coherence K6				

Skeptical DK	**0.04 (0 0.07) ER = 25.32**	0.02 (-0.01 0.05) ER = 8.45	0.01 (-0.01 0.04) ER = 6.24	0.02 (-0.01 0.07) ER = 7.23
Skeptical CH	-0.03 (-0.15 0.07) ER = 2.49 ER01 = 8.6	-0.05 (-0.25 0.1) ER = 2.57 ER01 = 6.95	0.29 (-0.23 0.9) ER = 6.72	**-0.13 (-0.36 0) ER = 17.46**
Coherence K7				
Skeptical DK	0.02 (-0.01 0.05) ER = 8.98	0.01 (-0.02 0.03) ER = 2.03 ER01 = 43.71	0 (-0.02 0.02) ER = 1.73 ER01 = 47.77	0 (-0.03 0.02) ER = 1.44 ER01 = 42.9
Skeptical CH	-0.03 (-0.11 0.05) ER = 2.44 ER01 = 11.22	-0.01 (-0.16 0.12) ER = 1.17 ER01 = 10.28	0.1 (-0.51 0.72) ER = 1.91 ER01 = 2.42	-0.04 (-0.21 0.09) ER = 2.55 ER01 = 8.76
Coherence K8				
Skeptical DK	0.03 (-0.02 0.08) ER = 6.04	**0.04 (0.01 0.09) ER = 38.47**	**0.03 (0 0.06) ER = 16**	0.03 (-0.02 0.09) ER = 4.67
Skeptical CH	-0.02 (-0.11 0.06) ER = 1.97 ER01 = 11.02	0.01 (-0.15 0.18) ER = 1.31 ER01 = 8.48	0.05 (-0.61 0.73) ER = 1.41 ER01 = 2.68	0 (-0.16 0.15) ER = 1.08 ER01 = 9.78
First-order Coherence				
Skeptical DK	**-0.15 (-0.26 -0.05) ER = 239**	**-0.09 (-0.17 -0.02) ER = 89.91**	**-0.07 (-0.13 -0.01) ER = 31.09**	**-0.07 (-0.18 0.01) ER = 11.15**
Skeptical CH	**-0.23 (-0.41 -0.09) ER = 170.43**	**-0.3 (-0.64 -0.09) ER = 161.16**	-0.16 (-0.85 0.56) ER = 2.27 ER01 = 1.71	**-0.28 (-0.63 -0.05) ER = 50.72**
Second-order Coherence				
Skeptical DK	**-0.1 (-0.21 0) ER = 16.49**	**-0.08 (-0.16 -0.01) ER = 43.44**	-0.05 (-0.13 0.02) ER = 5.76	-0.05 (-0.13 0.03) ER = 6.96
Skeptical CH	**-0.33 (-0.57 -0.13) ER = 314.79**	**-0.23 (-0.54 -0.02) ER = 27.71**	-0.01 (-0.72 0.73) ER = 1.04 ER01 = 1.72	**-0.27 (-0.58 -0.06) ER = 85.96**
Total Words				
Skeptical DK	**-0.1 (-0.18 -0.02) ER = 36.04**	**-0.1 (-0.18 -0.04) ER = 351.94**	-0.04 (-0.1 0.02) ER = 7.2	**-0.09 (-0.18 -0.02) ER = 49.42**
Skeptical CH	-0.1 (-0.24 0.02) ER = 9.85	**-0.17 (-0.47 0.02) ER = 12.86**	**0.45 (0.03 1.08) ER = 23.79**	**-0.24 (-0.54 -0.07) ER = 192.55**

Discussion

Current developments in clinical Natural Language Processing promise to revolutionize clinical practice. However, the scientific and clinical impact of these developments rely on the possibility to generalize NLP-based results across different samples and contexts. In this study, we assessed how the results of previous NLP studies measuring semantic coherence in schizophrenia generalize to a large novel cross-linguistic corpus. Globally, we found that only one previous result, i.e., reduced second-order coherence in schizophrenia, generalized across our entire corpus. Other results were replicated only for some specific languages: Chinese and German patients with schizophrenia showed lower coherence than controls across several,

but not completely overlapping, measures; while Danish patients showed a mixed pattern with higher semantic coherence than controls across multiple measures, and lower coherence in first- and second-order coherence.

A first possible explanation is *clinical heterogeneity*. Schizophrenia is a highly heterogeneous disorder (Gratton & Mittal, 2020; Hitczenko et al., 2021; Schnack, 2019), and participants from different studies are likely to have different clinical profiles, thus contributing to inconsistent results. In particular, reduced semantic coherence has traditionally been associated with formal thought disorders and positive symptoms more in general (e.g. Andreasen, 1979). However, this assumption has only inconsistently been supported by the literature, with highly varying statistical significance and size of the effects found across studies: while some studies found a strong correlation between semantic coherence and measures of formal thought disorder (Bilgrami et al., 2022; Elvevåg et al., 2007), most others reported uncertain results (Bedi et al., 2015; Haas et al., 2020; Just et al., 2020; Morgan et al., 2021; Pauselli et al., 2018; Sarzynska-Wawer et al., 2021; Tang et al., 2021). Our study emphasized the lack of a clear picture. We found indeed only inconsistent associations between semantic coherence and formal thought disorders. Surprisingly, we found more reliable associations of coherence with higher ratings of negative symptoms (alogia, poverty of content). These results, overall, may indicate that the relationship between NLP measures and clinical ratings of symptoms is affected by several factors, such as the rating scale used, the clinical characteristics of the specific sample, the NLP metrics employed, and the statistical design adopted (e.g., ordinal modeling vs. group splitting).

A second possible explanation is *socio-demographic heterogeneity*. Gender, age and socio-economic status are known to impact speech production and therefore likely semantic coherence, and relatedly to affect the expression of specific symptoms such as thought disorder (e.g., Palaniyappan, 2021). Indeed, recent studies seem to suggest that NLP algorithms, and coherence measures in particular, can be biased by socio-demographic variables such as racial identity. For instance, Hitczenko et al. (2022) found the NLP algorithms rated entirely asymptomatic Black American participants as having language patterns consistent with thought disorder, thus leading to biased predictions. Because not all previous studies systematically adjusted their estimates for all key socio-demographic variables, we could only speculate that they affected previous findings. However, our results show that gender, age and education do affect the difference by group in coherence patterns.

A third possible explanation is *cross-linguistic (and relatedly cultural) variation*. Different languages present different linguistic structures and usage patterns (Evans & Levinson, 2009), and indeed computational measures of different linguistic aspects, including semantic coherence, have been shown to vary across languages (Palaniyappan, 2021; Sumiyoshi et al., 2004, 2014; Wydell & Butterworth, 1999; Dideriksen, et al., 2020). Nevertheless, previous literature has implicitly assumed the existence of NLP markers of schizophrenia independent of language and cultural groups analyzed. Our findings disconfirm this preconception. For example, Danish patients were generally more coherent than controls, whereas Chinese and German patients were less coherent than controls. One could speculate that this is driven by the general opacity of the Danish sound structure, which has been shown to relate to higher redundancy in speech (Dideriksen et al., 2020; Trecca et al., 2021). However, we still lack an overview across diverse languages and even more a systematic theory-driven approach that would identify relevant cross-linguistic contrasts to assess the impact of linguistic constraints on semantic coherence and other linguistic measures (e.g.,Çokal et al., 2019; Deffner et al., 2021).

Finally, the *variety in measures employed* in assessing clinical features as well as semantic coherence could be contributing to inconsistent results. Different clinical scales do not completely overlap in terms of included symptoms and symptom definition, and likely imply different representations of the underlying psychopathological dimensions (e.g., Marder & Galderisi, 2017). For example, Bedi et al. (2015) have found a significant association only between a specific combination of symptoms and linguistic measures, including coherence: this may suggest as measures of coherence may be related to more complex psychopathological dimensions than to specific symptoms, and they can vary across scales. Not least, we found important differences between the different NLP measures of coherence, both within and between languages. For example, while we found that Danish patients showed lower coherence than controls on first- and second-order coherence, they instead showed higher coherence on the different coherence-k measures. Coherence-k represents semantic similarity between single words irrespective of the sentence structure, while first- and second- order coherence represents semantic similarity between different sentences. Our results show how measuring semantic coherence at different levels of granularity can yield very different results and highlight the importance of using different NLP measures when assessing patterns of coherence in schizophrenia. They also point to the need for validation

studies (e.g., Bilgrami et al., 2022) specifically aimed at assessing the psychometric properties of the different coherence measures, their relationship to clinical ratings, and their relationship to each other. Indeed, our robustness analysis have shown as different preprocessing options (e.g. transcript length and punctuation) can affect the various coherence measures. This is in line with previous previous studies (Elvevåg et al., 2007; Iter et al., 2018) and with recent evidence showing how some of these features (sentence length) can interact with socio-demographic characteristics and generate bias (Hitczenko et al., 2022). We thus advocate for larger theory-driven validation studies incorporating socio-demographic, linguistic and clinical variability, in order to consistently identify and statistically account for sources of variation in the decreased semantic coherence. A promising venue is the establishing of large normative datasets in order to determine whether deviations from normative values have clinical significance, or they only reflect the characteristics of a specific sample (Marquand et al., 2016, 2019).

Self-correcting approach

In this study we provided a concrete application of a cumulative yet self-critical scientific approach. We relied on a review of previous literature to design the current study and to set up priors for our analyses. At the same time, we critically attempt to replicate previous findings and compare statistical inferences relying on informed priors with inferences relying on skeptical priors. Our findings highlight some of the issues with previous literature. Indeed, we found only partial replications and informed models were often less robust and generalizable to new data than skeptical ones. Both pieces of evidence point towards a lack of generalizability of the previous literature, be it due to unreliable estimates, sparsity of previous data or lack of representativity of the previous samples compared to the current ones. Thus, even though the informed priors per se are not directly useful to generally improve our estimates, the comparison with skeptical ones provides valuable checks of the inferential robustness and of how the current study relates to the previous literature.

Limitation and future perspectives

One of the limitations of the present study is that we used pre-trained algorithms to quantify semantic coherence. This may have limited the sensitivity of the algorithms in detecting linguistic patterns specific to the characteristics of our corpus. On the other hand, using pre-trained algorithms allowed us to compare different coherence measures across different samples and languages in a scalable and easily replicable way. Another limitation is that we

focused on and compared single measures of semantic coherence: future studies should focus more on developing replicable (i.e., freely accessible and computationally feasible) machine learning pipelines to test the generalizability of large multidimensional patterns of features to account for shared variance across features, speech tasks, and languages. Finally, another limitation is the large variability in terms of clinical and demographic features of our multilingual corpus. While this variability may have contributed to the differences in the main results, we argue that the sample size of the samples of our corpora is larger than average (median for participants with schizpohrenia = 34.5), and that this setting provided us with a concrete basis for evaluating the heterogeneity of NLP measures across samples and languages.

Conclusions

In this study we showcased how a cumulative, self-correcting and replicable approach can be used to test the generalizability and robustness of NLP results in schizophrenia across different languages, samples and measures. Overall, we found large cross-linguistic variability in NLP-based assessments of semantic coherence in schizophrenia, with different sources of heterogeneity interacting at different levels. Future studies should take this variability into account in order to devise effective clinical applications able to target different ranges of patients and identify the presence of potential bias.

Role of the funding source

A.P is supported by a Marie Skłodowska-Curie Actions – H2020-MSCA-IF-2018 grant (ID: 832518, Project: MOVES). A.S is supported by the Carlsberg Foundation. The project was supported by seed funding from the Interacting Minds center. K. K has been supported by Japan Society for the promotion of Science (JSPS) to me (PE 07550).

Reference List

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cogn. Dev.* https://doi.org/10.1016/S0885-2014(00)00014-9

Andreasen, N. C. (1979). Thought, Language, and Communication Disorders: II. Diagnostic Significance. *Arch. Gen. Psychiatry*, *36*(12), 1325–1330. https://doi.org/10.1001/ARCHPSYC.1979.01780120055007

Bar, K., Zilberstein, V., Ziv, I., Baram, H., Dershowitz, N., Itzikowitz, S., & Vadim Harel, E. (2019). Semantic Characteristics of Schizophrenic Speech. *ArXiv Prepr. ArXiv1904.07953.*, 84–93. https://doi.org/10.18653/v1/w19-3010

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophr.*, *1*(1), 15030. https://doi.org/10.1038/npjschz.2015.30

Bilgrami, Z. R., Sarac, C., Srivastava, A., Herrera, S. N., Azis, M., Haas, S. S., Shaik, R. B., Parvaz, M. A., Mittal, V. A., Cecchi, G., & Corcoran, C. M. (2022). Construct validity for computational linguistic metrics in individuals at clinical risk for psychosis: Associations with clinical ratings. *Schizophr. Res.* https://doi.org/10.1016/J.SCHRES.2022.01.019

Bleuler, E. (1911). *Dementia Praecox or the Group of Schizophrenias*. International University Press.

Bliksted, V., Fagerlund, B., Weed, E., Frith, C., & Videbech, P. (2014). Social cognition and neurocognitive deficits in first-episode schizophrenia. *Schizophr. Res.*, *153*(1–3), 9–17. https://doi.org/10.1016/j.schres.2014.01.010

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.*, *5*, 135–146. https://doi.org/10.1162/TACL_A_00051/43387

Brand, C. O., Ounsley, J. P., Van der Post, D. J., & Morgan, T. J. H. (2019). Cumulative Science via Bayesian Posterior Passing. *Meta-Psychology*, *3*. https://doi.org/10.15626/mp.2017.840

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns. *Neuroimage*, *12*(3), 314–325. https://doi.org/10.1006/NIMG.2000.0612

Cohen, A. S., Schwartz, E., Le, T. P., Cowan, T., Kirkpatrick, B., Raugh, I. M., & Strauss, G. P. (2021). Digital phenotyping of negative symptoms: the relationship to clinician ratings. *Schizophr. Bull.*, *47*(1), 44–53. https://doi.org/10.1093/schbul/sbaa065

Çokal, D., Zimmerer, V., Turkington, D., Ferrier, N., Varley, R., Watson, S., & Hinzen, W. (2019). Disturbing the rhythm of thought: Speech pausing patterns in schizophrenia, with and without formal thought disorder. *PLoS One*, *14*(5), 1–14. https://doi.org/10.1371/journal.pone.0217404

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67–75. https://doi.org/10.1002/wps.20491

Corcoran, C. M., & Cecchi, G. A. (2020). Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, *5*(8), 770–779. https://doi.org/10.1016/j.bpsc.2020.06.004

Corcoran, C. M., Mittal, V. A., Bearden, C. E., E. Gur, R., Hitczenko, K., Bilgrami, Z., Savic, A., Cecchi, G. A., & Wolff, P. (2020). Language as a biomarker for psychosis: A natural language processing approach. *Schizophr. Res.*, *226*, 158–166. https://doi.org/10.1016/J.SCHRES.2020.04.032

De Boer, J. N., Brederoo, S. G., Voppel, A. E., & Sommer, I. E. C. (2020). Anomalies in language as a biomarker for schizophrenia. *Curr. Opin. Psychiatry*, *33*(3), 212–218. https://doi.org/10.1097/YCO.0000000000000595

De Boer, J. N., Voppel, A. E., Begemann, M. J. H., Schnack, H. G., Wijnen, F., & Sommer, I. E. C. (2018). Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. *Neurosci. Biobehav. Rev.*, *93*, 85–92. https://doi.org/10.1016/J.NEUBIOREV.2018.06.008

Deffner, D., Rohrer, J. M., & McElreath, R. (2021). A Causal Framework for Cross-Cultural Generalizability. *PsyAirXiv.* https://doi.org/10.31234/OSF.IO/FQUKP

Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2020). Quantifying the interplay of conversational devices in building mutual understanding. *Pre-print*. https://doi.org/10.31234/OSF.IO/A5R74

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. https://doi.org/10.1016/J.SCHRES.2007.03.001

Fusaroli, R., Grossman, R., Bilenberg, N., Cantio, C., Jepsen, J. R. M., & Weed, E. (2021). Toward a cumulative science of vocal markers of autism: A cross-linguistic meta-analysis-based investigation of acoustic markers in American and Danish autistic children. *Autism Res.* https://doi.org/10.1002/AUR.2661

Gallagher, S., & Varga, S. (2015). Social cognition and psychopathology: a critical overview. *World Psychiatry*, 14(1), 5-14.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. *ArXiv Prepr. ArXiv2011.01808*. https://doi.org/10.48550/arxiv.2011.01808

Gratton, C., & Mittal, V. A. (2020). Embracing the Complexity of Heterogeneity in Schizophrenia: A New Perspective From Latent Clinical-Anatomical Dimensions. *Schizophr. Bull.*, *46*(6), 1337–1338. https://doi.org/10.1093/schbul/sbaa122

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. In *Nature Reviews Neuroscience*. https://doi.org/10.1038/nrn4005

Haas, S. S., Doucet, G. E., Garg, S., Herrera, S. N., Sarac, C., Bilgrami, Z. R., Shaik, R. B., & Corcoran, C. M. (2020). Linking language features to clinical symptoms and multimodal imaging in individuals at clinical high risk for psychosis. *Eur. Psychiatry*, *63*(1). https://doi.org/10.1192/J.EURPSY.2020.73

Hitczenko, K., Cowan, H. R., Goldrick, M., & Mittal, V. A. (2022). Racial and Ethnic Biases in Computational Approaches to Psychopathology. *Schizophr. Bull.*, *48*(2), 285–288. https://doi.org/10.1093/SCHBUL/SBAB131

Hitczenko, K., Mittal, V. A., & Goldrick, M. (2021). Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods. *Schizophr. Bull.*, *47*(2), 344–362. https://doi.org/10.1093/SCHBUL/SBAA141

Holshausen, K., Harvey, P. D., Elvevåg, B., Foltz, P. W., & Bowie, C. R. (2014). Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex*, *55*(1), 88–96. https://doi.org/10.1016/J.CORTEX.2013.02.006

Iter, D., Yoon, J. H., & Jurafsky, D. (2018). Automatic detection of incoherent speech for diagnosing schizophrenia. *Proc. 5th Work. Comput. Linguist. Clin. Psychol. From Keyboard to Clin. CLPsych 2018 2018 Conf. North Am. Chapter Assoc. Comput. Linguist.*

Hum. Lang. Technol., 136–146. https://doi.org/10.18653/v1/w18-0615

Just, S. A., Haegert, E., Kořánová, N., Bröcker, A. L., Nenchev, I., Funcke, J., Heinz, A., Bermpohl, F., Stede, M., & Montag, C. (2020). Modeling Incoherent Discourse in Non-Affective Psychosis. *Front. Psychiatry*, *11*, 846. https://doi.org/10.3389/FPSYT.2020.00846/BIBTEX

Just, S., Haegert, E., Kořánová, N., Bröcker, A.-L., Nenchev, I., Funcke, J., Montag, C., & Stede, M. (2019). Coherence models in schizophrenia. *Proc. Sixth Work. Comput. Linguist. Clin. Psychol. Assoc. Comput. Linguist.*, 126–136. https://doi.org/10.18653/v1/w19-3015

Kraepelin, E. (1919). *Dementia Precox and Paraphrenia*. University of Edinburgh.

Marder, S. R., & Galderisi, S. (2017). The current conceptualization of negative symptoms in schizophrenia. *World Psychiatry*, *16*(1), 14–24. https://doi.org/10.1002/WPS.20385

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., & Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Mol. Psychiatry 2019 2410*, *24*(10), 1415–1424. https://doi.org/10.1038/s41380-019-0441-1

Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biol. Psychiatry*, *80*(7), 552–561. https://doi.org/10.1016/J.BIOPSYCH.2015.12.023

Morgan, S. E., Diederen, K., Vértes, P. E., Ip, S. H. Y., Wang, B., Thompson, B., Demjaha, A., De Micheli, A., Oliver, D., Liakata, M., Fusar-Poli, P., Spencer, T. J., & McGuire, P. (2021). Natural Language Processing markers in first episode psychosis and people at clinical high-risk. *Transl. Psychiatry 2021 111*, *11*(1), 1–9. https://doi.org/10.1038/s41398-021-01722-y

Mota, N. B., Copelli, M., & Ribeiro, S. (2017). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *Npj Schizophr. 2017 31*, *3*(1), 1–10. https://doi.org/10.1038/s41537-017-0019-3

Mota, N. B., Furtado, R., Maia, P. P. C., Copelli, M., & Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis. *Sci. Reports 2014 41*, *4*(1), 1–7. https://doi.org/10.1038/srep03691

Palaniyappan, L. (2021). More than a biomarker: could language be a biosocial marker of psychosis? *Npj Schizophr. 2021 71*, *7*(1), 1–5. https://doi.org/10.1038/s41537-021-

00172-1

Parola, A., Berardinelli, L., & Bosco, F. M. (2018). Cognitive abilities and theory of mind in explaining communicative-pragmatic disorders in patients with schizophrenia. *Psychiatry Res.*, *260*, 144–151. https://doi.org/10.1016/j.psychres.2017.11.051

Parola, A., Simonsen, A., Bliksted, V., & Fusaroli, R. (2020). Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. In *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2019.11.031

Parola, A., Brasso, C., Morese, R., Rocca, P., & Bosco, F. M. (2021). Understanding communicative intentions in schizophrenia using an error analysis approach. *NPJ schizophrenia*, 7(1), 1-9.

Pauselli, L., Halpern, B., Cleary, S. D., Ku, B., Covington, M. A., & Compton, M. T. (2018). Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. *Psychiatry Res.* https://doi.org/10.1016/j.psychres.2018.02.037

Press, C. ., Yon, D. ., & Heyes, C. (2022). Building better theories. *Curr. Biol.*, *32*(1), R13–R17.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Syst. Rev.*, *10*(1), 39. https://doi.org/10.1186/S13643-020-01542-Z/TABLES/1

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *Npj Schizophr. 2019 51*, *5*(1), 1–12. https://doi.org/10.1038/s41537-019-0077-9

Rocca, R., & Yarkoni, T. (2021). Putting Psychology to the Test: Rethinking Model Evaluation Through Benchmarking and Prediction: *Https://Doi.Org/10.1177/25152459211026864*, *4*(3). https://doi.org/10.1177/25152459211026864

Ryazanskaya, G., & Khudyakova, M. (2020). Automated Analysis of Discourse Coherence in Schizophrenia: Approximation of Manual Measures. *Lr. 2020 Lang. Resour. Eval. Conf. 11-16 May 2020.*, 98.

Rybner, A., Trenckner Jessen, E., Damsgaard Mortensen, M., Nyhus Larsen, S., Grossman, R., Bilenberg, N., Cantio, C., Richardt, J., Jepsen, M., Weed, E., Simonsen, A., & Fusaroli, R. (2021). Vocal markers of Autism Spectrum Disorder: assessing the generalizability of machine learning models. *BioRxiv*, 2021.11.22.469538.

https://doi.org/10.1101/2021.11.22.469538

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.*, *304*, 114135. https://doi.org/10.1016/J.PSYCHRES.2021.114135

Schnack, H. G. (2019). Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophr. Res.*, *214*, 34–42. https://doi.org/10.1016/J.SCHRES.2017.10.023

Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proc. Tenth Int. Conf. Lang. Resour. Eval.*, 4290–4297.

Sumiyoshi, C., Ertugrul, A., Yağcioğlu, A. E. A., Roy, A., Jayathilake, K., Milby, A., Meltzer, H. Y., & Sumiyoshi, T. (2014). Language-dependent performance on the letter fluency task in patients with schizophrenia. *Schizophr. Res.*, *152*(2–3), 421–429. https://doi.org/10.1016/J.SCHRES.2013.12.009

Sumiyoshi, C., Sumiyoshi, T., Matsui, M., Nohara, S., Yamashita, I., Kurachi, M., & Niwa, S. (2004). Effect of orthography on the verbal fluency performance in schizophrenia: examination using Japanese patients. *Schizophr. Res.*, *69*(1), 15–22. https://doi.org/10.1016/S0920-9964(03)00174-9

Tang, S. X., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R. E., Bhati, M. T., Wolf, D. H., Sedoc, J., & Liberman, M. Y. (2021). Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *Npj Schizophr. 2021 71*, *7*(1), 1–8. https://doi.org/10.1038/s41537-021-00154-3

Trecca, F., Tylén, K., Højen, A., & Christiansen, M. H. (2021). Danish as a Window Onto Language Processing and Learning. Lang. Learn., 71(3), 799–833. https://doi.org/10.1111/LANG.12450

Voppel, A. E., de Boer, J. N., Brederoo, S. G., Schnack, H. G., & Sommer, I. E. C. (2021). Quantified language connectedness in schizophrenia-spectrum disorders. *Psychiatry Res.*, *304*, 114130. https://doi.org/10.1016/J.PSYCHRES.2021.114130

Wydell, T. N., & Butterworth, B. (1999). A case study of an English-Japanese bilingual with monolingual dyslexia. *Cognition*, *70*(3), 273–305. https://doi.org/10.1016/S0010-0277(99)00016-5