Research Paper

Perplexity of utterances in untreated first-episode psychosis: an ultra–high field MRI dynamic causal modelling study of the semantic network

Maria Francisca Alonso-Sánchez*, MSc, PhD; Wolfram Hinzen*, PhD; Rui He, MSc; Joseph Gati, PhD; Lena Palaniyappan, MD, PhD

Background: Psychosis involves a distortion of thought content, which is partly reflected in anomalous ways in which words are semantically connected into utterances in speech. We sought to explore how these linguistic anomalies are realized through putative circuit-level abnormalities in the brain's semantic network. **Methods:** Using a computational large-language model, Bidirectional Encoder Representations from Transformers (BERT), we quantified the contextual expectedness of a given word sequence (perplexity) across 180 samples obtained from descriptions of 3 pictures by patients with first-episode schizophrenia (FES) and controls matched for age, parental social status, and sex, scanned with 7 T ultra–high field functional magnetic resonance imaging (fMRI). Subsequently, perplexity was used to parametrize a spectral dynamic causal model (DCM) of the effective connectivity within (intrinsic) and between (extrinsic) 4 key regions of the semantic network at rest, namely the anterior temporal lobe, the inferior frontal gyrus (IFG), the posterior middle temporal gyrus (MTG), and the angular gyrus. **Results:** We included 60 participants, including 30 patients with FES and 30 controls. We observed higher perplexity in the FES group, indicating that speech was less predictable by the preceding context among patients. Results of Bayesian model comparisons showed that a DCM including the group by perplexity interaction best explained the underlying patterns of neural activity. We observed an increase of self-inhibitory effective connectivity within the IFG, as well as reduced self-inhibitory tone within the pMTG, in the FES group. An increase in self-inhibitory tone in the IFG correlated strongly and positively with interregional excitation between the IFG and posterior MTG, while self-inhibition of the posterior MTG was negatively correlated with this interregional excitation. **Limitation:** Our design did not address connectivity in the semantic network during tasks that selectively activated the semantic network, which could corroborate findings from this resting-state fMRI study. Furthermore, we do not present a replication study, which would ideally use speech in a different language. **Conclusion:** As an explanation for peculiar speech in psychosis, these results index a shift in the excitatory–inhibitory balance regulating information flow across the semantic network, confined to 2 regions that were previously linked specifically to the executive control of meaning. Based on our approach of combining a large language model with causal connectivity estimates, we propose loss in semantic control as a potential neurocognitive mechanism contributing to disorganization in psychosis.

Introduction

Psychosis can clinically manifest in forms of speech involving distorted or delusional contents (e.g., "I was my husband," "I am Jesus") and problems of incoherence across the content of several sentences. Building content requires, first and foremost, the retrieval of concepts (e.g., husband, Jesus), which form constituents of the thoughts one is thinking and expressing. These concepts are retrieved internally from semantic memory, without a requirement of a perceptual stimulus functioning as their trigger. As such, they can be retrieved in the absence of sensorimotor stimulation, during internally driven thought. This intrinsic activity has been linked neuroanatomically to regions of the default mode network (DMN),[1] and independent evidence suggests that lexical–conceptual semantic processing involves a widely distributed neural network that strongly overlaps with the DMN.[2–5] This network broadly supports self-referential processing, which consists of retrieving and replaying episodes experienced by the self, constructing scenes, imagining the future,

Correspondence to: L. Palaniyappan, Room T-0104 T Pavilion, Douglas Mental Health University Institute, 6875 Boulevard LaSalle, Verdun, Que., H4H 1R3; lena.palaniyappan@mcgill.ca

*Co–first authorship.

Submitted Apr. 9, 2024; Revised May 31, 2024; Accepted June 5, 2024

Cite as: *J Psychiatry Neurosci* 2024 August 9;49(4). doi: 10.1503/jpn.240031

and generating internal ongoing narratives.[1,4,6] Retrieving concepts is as inherent to such mental activities as it is to any speech one externally produces, and specific patterns of conceptual associations are linked to activity in the DMN.[7] Anomalies in DMN activity are also among the most replicated neuroimaging findings across major disorders of thought, including psychosis and autism.[6,8] This suggests that semantic processing could shed light on such disorders,[9] providing a link that connects impoverished or disorganized forms of speech found in psychosis[10] to neurocognitive mechanisms. Few studies in psychosis, however, have specifically targeted the semantic network in this respect. An exception is a study from Matsumoto and colleagues[11] that analyzed semantic connections between brain representations of words characterizing movie scenes shown to participants during fMRI and found graph-theoretical properties of these connections to change in psychosis. A potential link between connectivity within this semantic network and natural language production in psychosis has not been investigated; thus, our aim was to provide empirical evidence for this link.

Lexical–semantic concepts (e.g., guitar) can be both multimodal, in the sense of integrating multiple perceptual modalities (e.g., visual, auditory, and tactile aspects of guitars), and amodal, representing even more abstract, modality-invariant conceptual information, which becomes more pertinent in concepts with a depleted conceptual content, as in the case of "come" or "during."[12] Recent work has identified both multimodal convergence zones, which retain some modality-specific information (while still not overlapping with sensorimotor cortices), and amodal cortical areas. The former includes the left posterior inferior parietal lobe — especially the angular gyrus, the posterior middle temporal gyrus (MTG), the anterior inferior frontal gyrus (IFG), and parts of the medial prefrontal cortex[13] — showing strong overlaps with the DMN.[5] A longstanding candidate for an amodal area representing conceptual information at the greatest abstraction from sensorimotor information has been the anterior temporal lobe.[14] Neither multimodal nor amodal lexical concepts, however, are sufficient to constitute meaningful mental activity. Just as humans do not tend to produce words like "guitar," "felt," or "during" in isolation, it seems just as unlikely that the corresponding concepts, occurring in isolation, would structure mental life during the default mode of thinking. Meaningful units of thoughts, words, or concepts need not only to co-occur with others, but do so in a specific mode of combination, not as mere lists or connected associatively, but rather as meaningful phrasal units like, "I really want a guitar," "How did I feel back then?" or "Come during lunch." In such units, words are connected grammatically.

Although the neural topography of such grammar-level (sentential) semantic processing is still unclear, recent evidence supports that this is likely to be as widely distributed as semantic processing at the level of single concepts, but also diverges from the latter, specifically in the temporal cortex, the IFG, and the inferior parietal lobe, including the angular gyrus.[15] Processing lexical concepts flexibly and adaptively in contextually appropriate ways also involves a specific system long identified under the label of semantic control, a subnetwork of the semantic network that specifically involves the posterior MTG and IFG and their functional connectivity.[16–20] Difficulties of semantic control in the form of a failure to inhibit retrieval of lexically related but contextually inappropriate words is attested among patients with psychosis.[21]

Against this background, our aim was to study the effective connectivity (i.e., the effect of 1 region on another, as well as self-connection) of 4 key regions of interest in the semantic cognition network (the IFG, posterior MTG, anterior temporal lobe, and angular gyrus). We also sought to relate effective connectivity to a computational metric of semantic coherence at the utterance level, called perplexity, as derived from large computational language models. Perplexity refers to an unexpected deviation of a spoken word from the semantic context of the utterance in which it is embedded; this can be quantified using language models that employ contextual word embeddings, such as Bidirectional Encoder Representations from Transformers (BERT).[22] We have recently used this natural language processing approach to demonstrate that patients with untreated first-episode schizophrenia (FES) showed an increase in perplexity, which related to overall symptom burden; medicated patients in a stable phase of illness did not differ from healthy participants in perplexity.[23]

We hypothesized that thought disturbance in psychosis could be reflected both in connected speech, at the level of the contextual probability of 1 word to follow another (perplexity), and at the level of the semantic network and its semantic control subnetwork (effective connectivity). Specifically, we hypothesized a distortion in the normal causal flow of information across this network, which depends on an orchestrated balance between excitatory and inhibitory neuronal and regional interactions. Following previous work,[24] we chose spectral dynamic causal modelling (DCM) as a generative modelling framework for hypothesis testing about the directed, effective connectivity between regions.[25,26] Although the choice of functional magnetic resonance imaging (fMRI) and DCM constrains the number and size of brain regions that can be modelled in a network perspective, DCM applied to resting-state fMRI enables estimation of the parameters of a neural model, specifying the directed influence of 1 region on another based on their cross-spectral density (i.e., covariance of fMRI signals in the frequency domain). The cross-spectral density summarizes time-varying fluctuations and, crucially, retains phase or lag information inherent in complex cross-spectra.[27] This allows the study of self-connections (i.e., the influence of a given brain region on itself over time), enabling the estimation of the excitation–inhibition balance at the individual level and the putative changes or synaptic gain of a region in relation to variations extrinsic to it. In the context of our assessment of semantic perplexity in speech, we sought to test whether a DCM parametrized by perplexity scores — and, thereby, informing intrinsic neural connectivity through an automatically extractable external speech metric — would improve model performance. Furthermore, given the pervasive difficulties in construct–circuit mapping in psychiatric neuroscience,[28] we included another second-level model selection analysis to assess the extent to

which the computationally derived, language model–based measure could serve as an intermediary metric for studying brain connectivity, over and above clinically derived symptom scores. We sought to conduct this analysis using data from patients with FES with very minimal lifetime antipsychotic exposure to minimize the potential confound of illness chronicity and treatment on connectivity patterns, as well as symptom burden, preserving the variance necessary to observe the hypothesized relationship.

Methods

Participants

We recruited untreated patients with FES, matched with neurotypical control volunteers. We recruited patients from the Prevention and Early Intervention Program for Psychosis, a catchment area–based, high-fidelity early intervention program in London, Ontario, that received all patients with first-episode psychosis in the city in 2017–2019; 40% were referred to the program during an acute hospital stay. For this study, patients were assessed within the first week of referral to the first-episode psychosis team, with a requirement that patients have less than 2 weeks of lifetime antipsychotic exposure. As such, 1 defined daily dose worth of antipsychotic exposure was, on average, less than 3 days (calculated by converting various prescribed antipsychotic medication doses to a common equivalent and multiplying by the days of exposure).[29]

The recruitment procedures are described in our previous study.[24] Briefly, clinical assessments and speech sampling were completed on the same day of scanning, with the diagnosis of schizophrenia made using a consensus best-estimate method around 6 months after the initial assessment (and confirmed before fMRI analysis).[30] The diagnostic criteria were based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5). Informed consent was obtained from participants, and resting-state fMRI scans and picture descriptions were collected for both groups in the same session.

Clinical assessment

In the patient group, we confirmed the severity of symptoms with the 8-item Positive and Negative Syndrome Scale (PANSS). We assessed all participants with the Social and Occupational Functioning Assessment Scale; parental socioeconomic status (SES) was determined using the National Statistics Socioeconomic Classification.[31] We used the Edinburgh Handedness Inventory to define the handedness of the participants, the scale is scored from –12 (totally left-handed) to +12 (totally right-handed).[12]

Language assessment

In the language task, we asked the participants to describe 3 pictures from the Thematic Apperception test for 1 minute per picture. The interviewer encouraged participants to extend their responses with minimal prompts if they concluded before the designated time. The interview was recorded and later transcribed by research assistants blind to the diagnostic status. The evaluation of formal thought disorder was conducted using the Thought Language Index (TLI).[32]

Computational linguistic analysis

We segmented the transcribed picture descriptions into utterances, defined as syntactically independent units that provided new information to the discourse. We assessed the perplexity of an utterance as a metric of the unexpectedness of its comprising units. As we compared the observed word sequence (i.e., the original utterance) to the predicted word sequence (i.e., the prediction from a large language model), the cross-entropy between the 2 served as an indicator for how well the word sequence fit the language model. This indicator was first defined for causal language models and has been found to be a reliable marker of speech coherence that is sensitive to cognitive decline, capturing meaning at the discourse level.[33]

In causal modelling, autoregressive methods predict the next token in a sequence based on previous tokens.[34] Masked modelling involves predicting selected masked tokens within a sequence based on both the preceding and following context, enabling the use of bidirectional tokens to do the prediction.[34] Conceptually, this is akin to the well-known Cloze procedure in psycholinguistics. The perplexity measure was extended to masked language models like BERT in the form of pseudo-perplexity, using similar mathematical computations.[35] Specifically, using BERT, in a tokenized utterance (U):

$$U := (t_1, t_2, \ldots, t_n)$$

We formally defined the probability of token (t_i) as the log conditional probability of the utterance without this token:

$$U_{\setminus i} := (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)$$

The perplexity of the utterance was then defined as the exponential of the negative mean of the pseudo-log-likelihood scores provided by summing the conditional log probabilities of all tokens in the utterance:

$$perplexity(U_t) := Exp\left(-\frac{1}{n}\sum_{i=0}^{n} \log P_{LM}\left(t_i \middle| U_{\setminus i}\right)\right)$$

We computed the pseudo-perplexity scores for every utterance in the transcript, averaged to obtain a single score per participant, hereby referred to as perplexity.

Data analysis

We examined clinical, demographic, and linguistic data using both descriptive and statistical analyses. To compare the perplexity between groups we used generalized estimating equations after log-transformation and Tweedie distribution to make population-level (marginal) inferences, as reported in our previous work.[23,36] We analyzed categorical variables with χ^2 statistics. We report p values ($p < 0.05$ considered statistically significant), along with effect sizes, for continuous and categorical variables.

Resting-state fMRI acquisition and processing

We collected fMRI data at the Centre for Functional and Metabolic Mapping at the University of Western Ontario on a Siemens 7 T Plus scanner. We collected a total of 360 whole-brain functional images using a multi-band echo-planar imaging acquisition sequence with 20 ms of echo time, 1000 ms of repetition time, and a flip angle of 30°, in 63 slices with a multi-band factor of 3, an integrated parallel imaging technique factor of 3 and an isotropic resolution of 2 mm. We asked participants to lie still inside the scanner with their eyes open for the duration of the scan and not to think of anything in particular.

We processed the fMRI data with SPM12 in MATLAB.[37] The preprocessing workflow involved manually reorienting the anterior and posterior commissures. As part of the standard pipeline, we computed voxel displacement maps for inhomogeneities by subtracting the phase and magnitude. Realignment was performed using the realign and unwarp option, followed by coregistration and estimation of these realigned images. Finally, segmentation was carried out using native and Dartel-imported tissue probability maps, culminating in normalization to the Montreal Neurological Institute (MNI) space, with a Dartel template built from the T_1 images of all participants.

Effective connectivity

We modelled the resting-state data using a general linear model with a discrete cosine basis set, 6 head movement parameters, and the white matter and cerebrospinal fluid time series as regressors. We defined 4 volumes of interest a priori, based on the literature on the neural basis of the semantic network,[16,17,38] to construct a DCM architecture that could best explain the observed complex cross-spectra.[39,40] The volumes of interest were the anterior temporal lobe (–41, –15, –31), the inferior frontal gyrus (–48, 22, 20), the posterior medial temporal gyrus (–54, –42, 4), and the angular gyrus (–48, –64, 34), as shown in Figure 1. The volumes of interest were spherically defined with a radius of 8 mm. We specified a fully connected bilinear model without exogenous inputs. Having estimated the connectivity parameters of a standard bilinear DCM, we then tested for various group effects using parametric empirical Bayes. Effectively, this explains first-level (subject-specific) connectivity parameters in terms of main effects and interactions at the second (group) level using a general linear model. Because parametric empirical Bayes is a Bayesian approach, we were able to explicitly evaluate the evidence for models that contained a main effect of group, main effects of group and perplexity, and a full model with group effects, perplexity, and their interaction. We compared the free energy of the 3 models and calculated the posterior probability of the winning model. We opted for free energy over alternative methods such as the Akaike or Bayesian information criterion because of its robustness for model selection.[41]

Besides testing the relationship between computationally derived perplexity and the effective connectivity within the semantic network, we tested if this computational measure had better explanatory power than observer-rated clinical symptom scores. To this end, we conducted an additional second-level parametric empirical Bayes analysis incorporating PANSS positive symptom components (hallucinations, delusions, and conceptual disorganization scores) and severity of formal thought disorder (TLI score) into the model selection

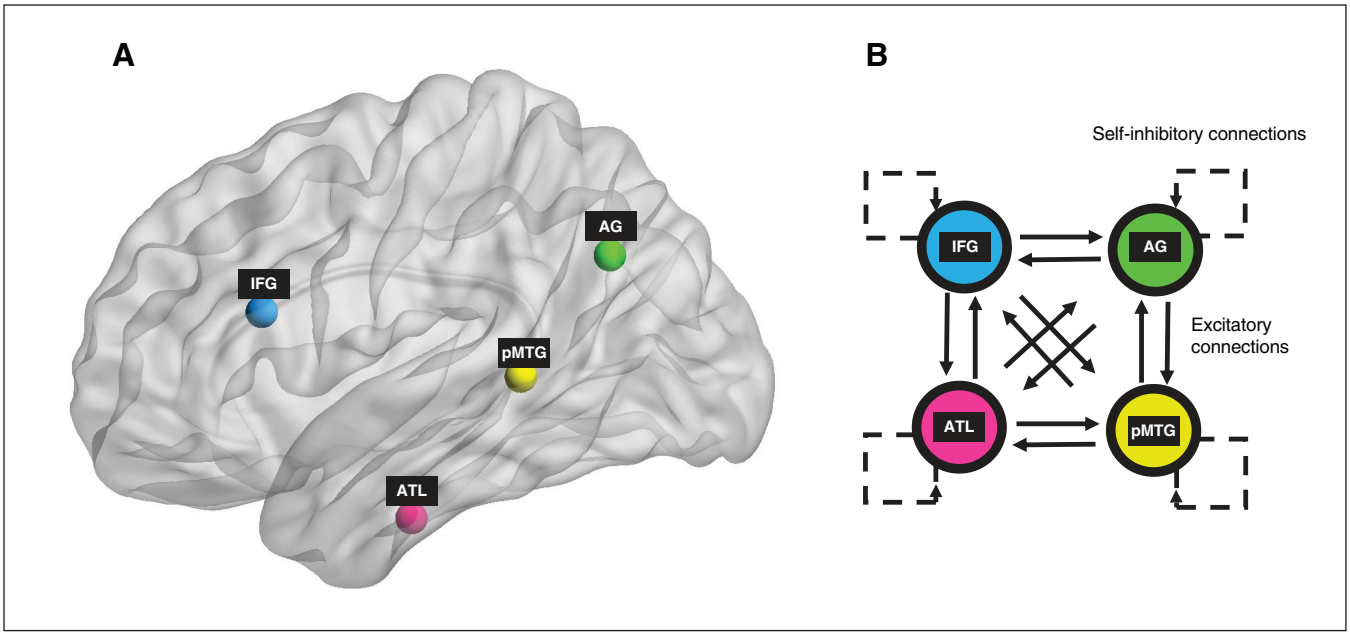

Figure 1: (A) Anatomic map of voxel positioning for dynamic causal modelling. (B) Self-inhibitory connections (dashed arrows) within each region and the bidirectional excitatory connections (solid arrows) between regions dynamic causal model. The plot was made with BrainNet Viewer. Note: AG = angular gyrus, ATL = anterior temporal lobe, IFG = inferior frontal gyrus, pMTG = posterior middle temporal gyrus.

process. This enabled us to determine the extent to which the perplexity metric specifically reflected the connectivity within the semantic network in a symptomatic patient with psychosis.

Finally, we analyzed the relationship between the within-node and between-nodes connectivity parameters of the winning model in the patient group to determine the degree to which changes within the network were interdependent. In dynamic causal models of this sort, there are 2 kinds of connectivity parameters. The first (intrinsic) parameters pertain to (recurrent) self-inhibition and are expressed in terms of log-scaling. The second (extrinsic) connectivity parameters concern directed effects between regions and are expressed in terms of rate constants (i.e., Hz). Both kinds of parameters were included in the parametric empirical Bayes analysis, looking for main effects of group and perplexity (and their interaction). Note that because self-connections are log-scale parameters, small values can be interpreted as proportional changes. For example, a self-inhibition estimate of 0.2 can be read as a 20% increase in self-inhibition, indicating a substantial decrease in excitability.

Results

Clinical, linguistic, and demographic data

We included 30 patients with FES and 30 controls. Clinical and demographic data are shown in Table 1. As expected from the group-matching procedure employed during recruitment, the 2 groups did not differ in age, sex, parental socioeconomic status, or education. Among patients with FES, 85% had English as their first language, compared with 90% among controls ($\chi^2 = 0.580$, $p = 0.446$, odds ratio 0.591); all the participants had English as their transactional language. The FES group had a higher mean perplexity score at 8.04 (standard deviation [SD] 3.31) than the control group (mean 6.82, SD 1.55). The group effect estimated by GEE was in line with our previous work,[23] showing a significant effect when analyzed by modelling random effects of the 3 pictures ($z = 2.38$, $p = 0.02$) and when averaging across the 3 pictures ($z = 2.08$, $p = 0.04$).

We performed a multiple regression analysis (restricted to the patient group) to evaluate the extent to which any of the 8 items of the PANSS related to perplexity. We first confirmed that multicollinearity was not a concern (all 8 variance inflation factors < 5). The regression model was not significant ($F_{8,21} = 1.16$, $R^2 = 0.31$, $p = 0.4$). None of the individual symptom predictors other than conceptual disorganization ($t = 2.49$, $p = 0.02$) were significant (all other $t < 1.71$, $p > 0.1$). When a similar regression was applied to the 6 items of the TLI (4 defining disorganized thinking [peculiar sentences, logic, words, and looseness] and 2 defining impoverished thinking [poverty of speech and weakening of goal]), we did not find any significant predictors (all $p > 0.2$, including for the model), indicating that perplexity measured from a picture description

Table 1. Participant characteristics

Characteristic	No. (%) of participants or mean ± SD		*p* value	Effect size (Cohen *d*)
	Patients with FES *n* = 30	Controls *n* = 30		
Age, yr, mean ± SD	21.4 ± 2.2	21.4 ± 3.5	1.00	< 0.00
Sex				
Female	8 (27)	8 (27)		
Male	22 (73)	22 (73)	0.764	0.180
Educational level				
< 12 yr	8 (30)	14 (47)		
> 12 yr	21 (70)	16 (53)	0.075	–0.985
PANSS				
Positive, mean ± SD	12.3 ± 3.5	–	–	–
Negative, mean ± SD	7.6 ± 4.3	–	–	–
Total, mean ± SD	26.4 ± 7.4	–	–	–
SOFAS, mean ± SD	39.0 ± 14.7	81.9 ± 5.0	< 0.001	3.829
Parental socioeconomic status			0.139	0.810
< 3	9 (30)	13 (44)		
≥ 3	21 (70)	17 (56)		
TLI, mean ± SD				
Total TLI	1.51 ± 1.3	0.3 ± 0.3	< 0.00	1.282
Impoverishment in thinking	0.49 ± 0.6	0.13 ± 0.2	0.002	0.804
Disorganization in thinking	1.0 ± 1	0.16 ± 0.2	< 0.00	1.164
Antipsychotic exposure × days at time of scan, mean ± SD	2.8 ± 3.7	–	–	–
Duration of untreated psychosis, mean ± SD	9.4 ± 13.5	–	–	–
Edinburgh Handedness Inventory, mean ± SD	9.8 ± 4.2	10.4 ± 3.1	0.511	0.259

FES = first-episode schizophrenia; PANSS = Positive and Negative syndrome scale; SD = standard deviation, SOFAS = Social and Occupational Functioning Assessment Scale, TLI = Thought Lanuage Index.

task captured the variations in raters' scoring of conceptual disorganization in the PANSS without these speech data, while human clinical raters were not able to detect speech-based deviations relating to perplexity from the speech data, assessed as part of rating the TLI.

Effective connectivity

The interaction model including group, perplexity, and the interaction between the 2 variables, outperformed the other models (Figure 2A) without overfitting the data, indicating that the directed functional interactions within the semantic network that influenced perplexity differed between patients with FES and healthy controls.

The parametric empirical Bayes results further showed that, across participants, higher perplexity was associated with higher self-inhibition (i.e., reduced regional excitability) within all 4 nodes of interest (Figure 3A). Higher perplexity was also associated with lower excitatory influence from the anterior temporal lobe to the IFG and posterior MTG, from the IFG to the posterior MTG, and from the posterior MTG to the anterior temporal lobe (Figure 3A). On parsing the interaction between diagnostic group and perplexity, we noted higher perplexity among patients with FES in association with higher self-inhibition in the IFG, and among controls, higher perplexity related to higher self-inhibition in the posterior MTG (Figure 3B).

In addition, when comparing free energy across models that integrated PANSS components, TLI, and perplexity, this last measure outperformed its counterparts, establishing it as a more reliable proxy for mapping brain connectivity than the clinical rating scores.

Finally, to understand how the altered self-inhibitory tone in the IFG and posterior MTG related to extrinsic connections within the semantic cognition network in FES, we computed the correlation coefficients between the within-node and between-nodes parameters of the DCM in the patient group. Higher self-inhibition within the IFG was seen in the presence of more pronounced excitatory influence from the IFG to the posterior MTG ($r = 0.718$, $p < 0.001$) (Figure 4A). In addition, lower self-inhibition of the posterior MTG was seen with higher excitatory influence from the IFG to the posterior MTG ($r = –0.607$, $p < 0.001$) (Figure 4B). Taken together, these relationships indicate a pattern of reduced excitability (or synaptic gain) in the IFG and a release effect from the predicted inhibitory (brake-like) IFG control over the posterior MTG within the network, likely facilitating excitability (i.e., disinhibition) in the posterior MTG among patients with higher perplexity. Note that this classical correlation between posterior estimates is confounded by posterior correlations that may be caused by conditional dependencies during model fitting. We urge caution in interpreting the magnitude and causal structure of the ensuing correlations as they may reflect either real correlations or else posterior correlations, given that changing one or the other produces the same effect in the measured cross-spectra.

Discussion

To develop mechanistic insight into anomalous semantic associations (contextually unexpected words in utterances) in the descriptive narratives produced by people with psychosis, we charted the effective connectivity among key regions of the semantic network using a spectral DCM. The DCM was parametrized using an objectively quantified metric of unexpectedness in speech utterances (perplexity) based on a masked language model. We reasoned that semantic processing and the organization of concepts into meaningful units is a feature common to

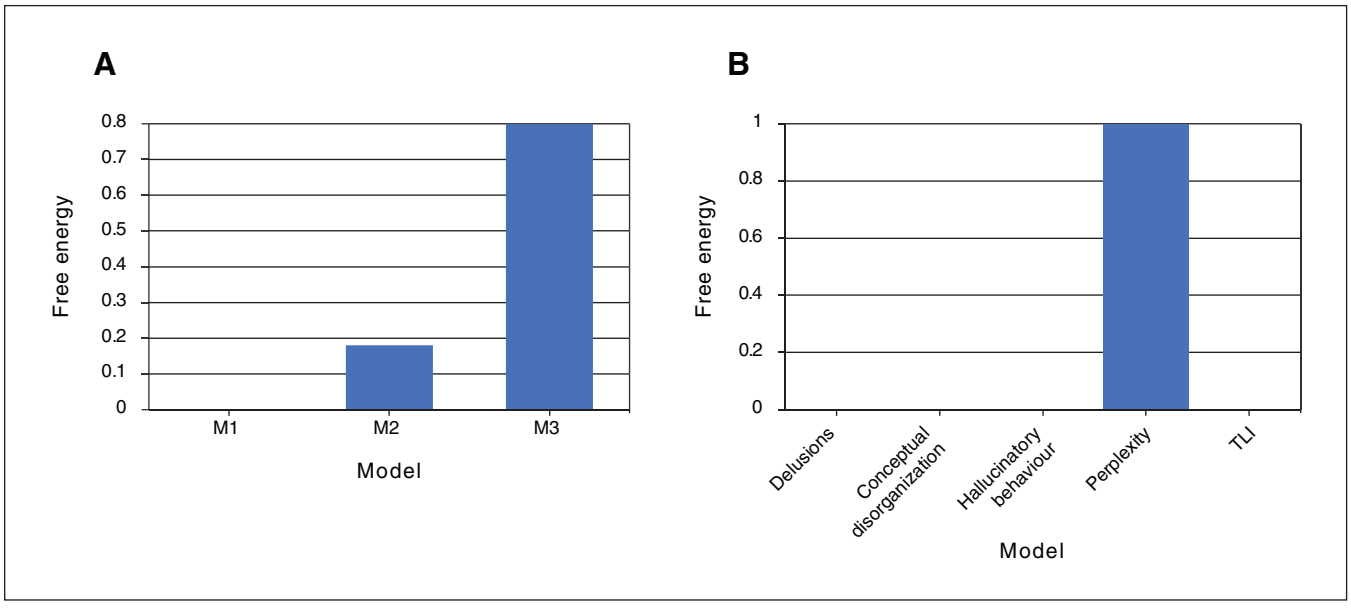

Figure 2: Model selection. (A) Free-energy comparison of the 3 models (M1: group; M2: group and perplexity; and M3: group, perplexity, and the interaction). The winning model is M3. (B) Model selection considering other symptoms rated on the Positive and Negative Symptom Scale and Thought and Language Index (TLI) scores. See Related Content tab for accessible version.

both spontaneous speech and internally driven thoughts during rest. Results confirmed our prediction that the semantic structure of spontaneous speech is not merely different in psychosis but informs understanding of the intrinsic functional organization of the semantic network, reinforcing a previous result that employed a more limited connectivity model.[24]

Across groups, perplexity was associated with a generalized increase in self-inhibition (i.e., reduced excitability) across all 4 brain regions of interest, together with a decrease in effective connectivity values (excitatory influences) between the IFG, the posterior MTG, and the anterior temporal lobe, but not the angular gyrus. The magnitude of association with self-inhibition (log-scales of 0.22–0.27) indicated that, across the semantic network nodes, perplexity explained 22%–27% of proportional variation in the ratio of output-to-input signal in pyramidal cells (i.e., synaptic gain). When accounting for group-specific interactions with perplexity, patients with FES displayed a heightened self-inhibitory tone (or reduced excitability) in the

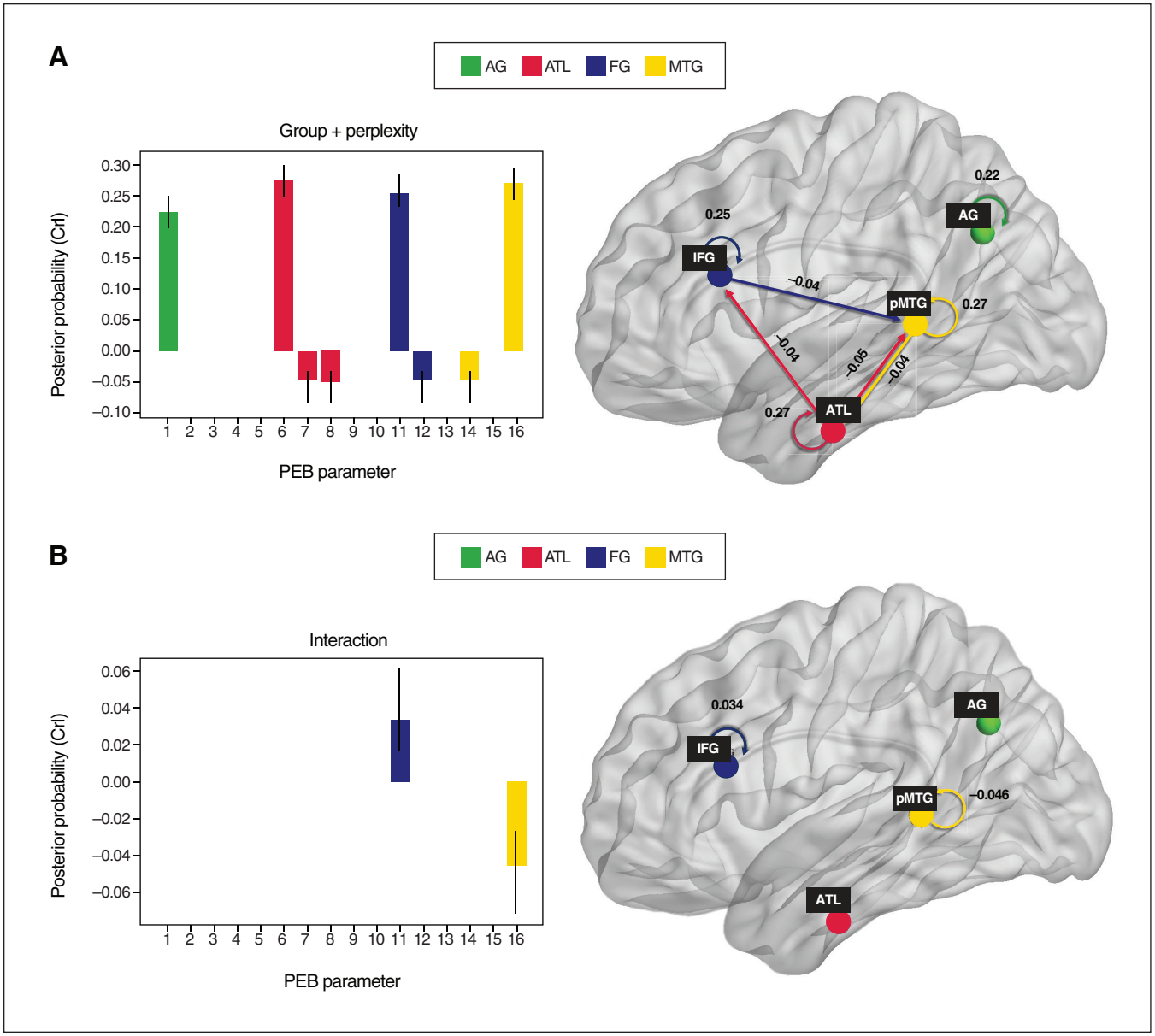

Figure 3: Second-level effect (A) on connections modulated by perplexity and (B) of group on connections modulated by the interaction between the covariates in brain areas of interest (1 = angular gyrus [AG] to AG, 2 = AG to anterior temporal lobe [ATL], 3 = AG to inferior frontal gyrus [IFG], 4 = AG to posterior middle temporal gyrus [pMTG], 5 = ATL to AG, 6 = ATL to ATL, 7 = ATL to IFG, 8 = ATL to pMTG, 9 = IFG to AG, 10 = IGF to ATL, 11 = IFG to IFG, 12 = IFG to pMTG, 13 = pMTG to AG, 14 = pMTG to ATL, 15 = pMTG to IFG, 16 = pMTG to pMTG). Black lines correspond to the 90% credible interval (CrI). All the posterior probabilities (colour bars) shown are higher than 95%. Brain images show the posterior probabilities of the inhibitory and excitatory brain connectivity of the group, perplexity values, and their interaction. See Related Content tab for accessible version.

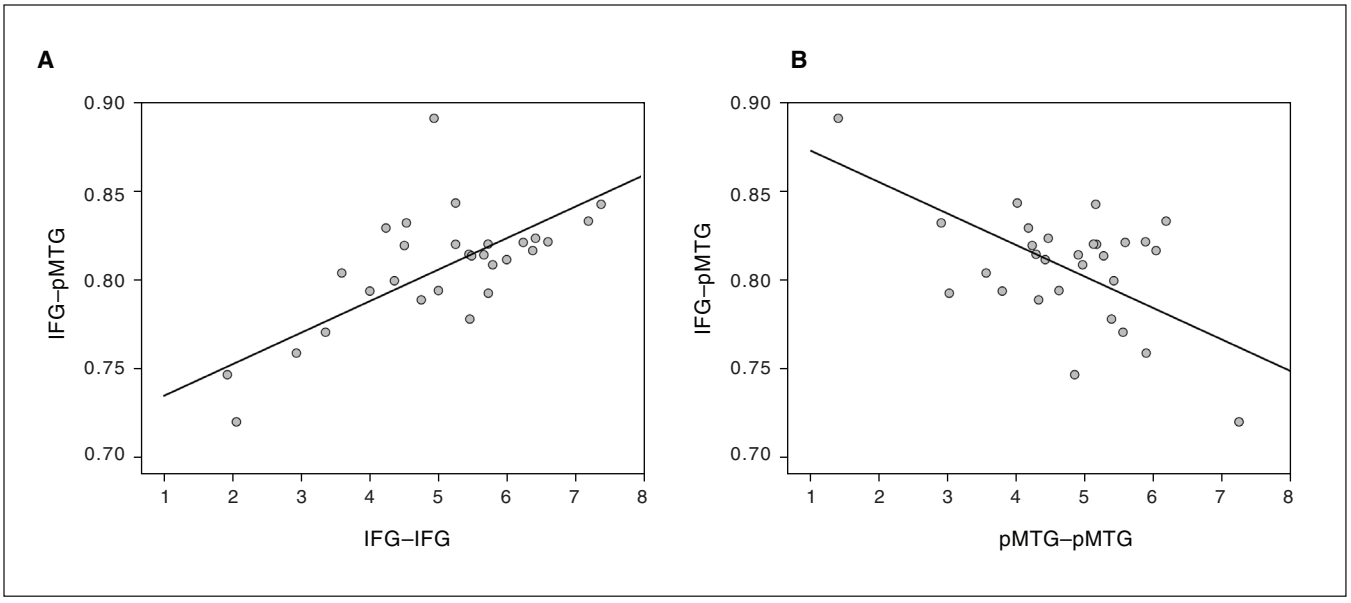

Figure 4: (A) Posterior correlation between the self-inhibitory connection strength of the inferior frontal gyrus (IFG) and the excitatory connectivity from the IFG to the posterior middle temporal gyrus (pMTG). (B) Posterior correlation between the self-inhibitory connection strength of the pMTG and the excitatory connectivity from the IFG to pMTG.

IFG jointly with a reduced inhibitory tone (or higher excitability) in the posterior MTG, both of which correlated, though in opposite directions, with the effective frontotemporal connectivity from the IFG to posterior MTG. Concisely, the physiological correlate of perplexity within the IFG (i.e., reduced pyramidal excitability) is exaggerated in FES, and this occurs in conjunction with a likely disinhibitory effect on the posterior MTG via increased excitatory interregional influence among patients.

In psychosis, the key differential effect of a change in effective connectivity was restricted to the 2 subregions of the semantic network, the IFG and the posterior MTG, which are associated with semantic control (i.e., the adaptive fitting of retrieved concepts into contexts).[16] Connectivity in, and to or from, the anterior temporal lobe and angular gyrus, by contrast, was unaffected. This suggests that changes in the effective connectivity of the semantic network in psychosis are limited to the frontotemporal axis, wherein a functional connection between the inferior frontal and posterior temporal lobes has been established as the basis of integrating form and meaning.[42] This is consistent with our argument that building any kind of coherence requires more than retrieving and associating single concepts. Concepts need to enter meaningful structural configurations expressing thoughts, which, at the level of expressed speech, manifest as utterances. It is also precisely at this level of semantic coherence — that of grammatical structure-building — on which our perplexity metric operates. As semantic control is the interface between semantic and executive processing, the grammatical system, viewed as a system exerting executive (and inhibitory) control in the generation of meaning, may be key to the loss of semantic coherence in psychosis.

Our ability to secure evidence for an empirical Bayesian model that included the effects of perplexity (and its interaction with group) lends an important validity to the effective connectivity estimates. Perplexity ratings were based on data acquired independently from the resting-state fMRI data used for DCM. In short, if the DCM had returned inefficient estimators of effective connectivity within the semantic network, there would have been no evidence for an effect of perplexity (or its interaction with the group).

Among patients with psychosis and high perplexity, we observed lower self-inhibitory connection of the IFG but higher self-inhibitory connection of the posterior MTG. In spectral DCM, the inhibitory self-connections represent excitability or synaptic gain, the ratio between output and input signals in (pyramidal) neuronal populations (Figure 5). These are estimated through an autocovariance function (i.e., the relationship of a time series with a shifted version of itself, estimated in the frequency domain for cross-spectral density).[43] Assuming stationarity of the time series, higher negative values indicate a faster decay of the covariance and, thus, a lower effect of a recent input on the output (i.e., lower excitability or poor synaptic gain). At the neural implementation level, the observed pattern of lower excitability or synaptic gain in the IFG in relation to perplexity among patients may indicate glutamatergic (pyramidal) dysfunction.[44] The higher excitability or synaptic gain in the posterior MTG may represent a disinhibition effect that is facilitated through the excitatory influence from the IFG and posterior MTG or a compensatory interneuron down-regulation after a period of (currently unobserved) lower excitability in the posterior MTG per se, as suggested by Adams and colleagues.[44] Computationally, this can be interpreted as representing a shorter memory span for inputs to the IFG neuronal populations, with a longer span for inputs at the posterior MTG node of the semantic network among

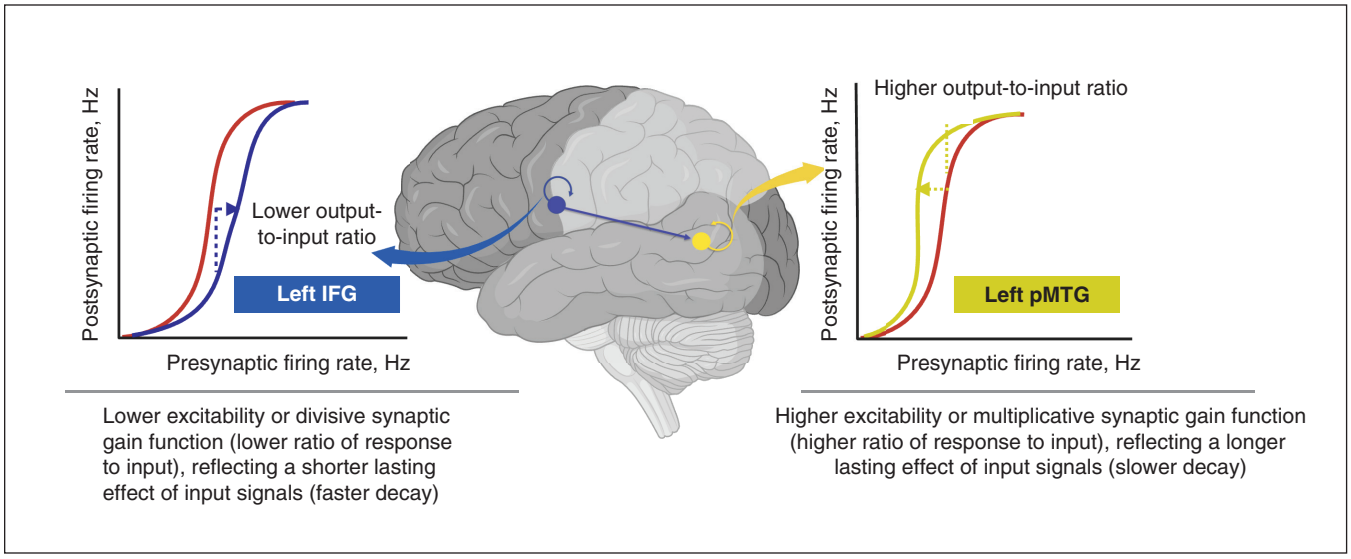

Figure 5: Interpretation of excitability or synaptic gain in the spectral dynamic causel model of patients with high perplexity. Neuronal response to input is shown as a hypothetical sinusoidal plot (red) between pre- and postsynaptic firing rates in Hertz units (Hz). In pyramidal neuronal populations of the left inferior frontal gyrus (IFG), this function shifts to the right among patients with high perplexity, with a lower ratio between output and input signals (reduced synaptic gain or excitability). In pyramidal neuronal populations of the left posterior middle temporal gyrus (pMTG), this function shifts to the left, with a higher ratio between output and input signals (higher synaptic gain or excitability). These patterns may indicate reduced glutamatergic tone in the IFG and disinhibition (or reduced modulation by inhibitory interneuronal population) in the pMTG.

patients. On a speculative note, this may reflect that semantic control through grammar is weakened in schizophrenia with higher perplexity, as the posterior MTG is increasingly recognized as a site for syntactic integration of meaning, which is serialized during word order generation via the IFG for language production.[45] At the same time, posterior MTG is also a key site with a direct interface with executive control systems, which may serve to provide syntax-based contextual control to reduce a more automatic spreading of activations across semantic representations associated with the angular gyrus and anterior temporal lobe regions.[46,47] Higher synaptic gain or excitability of the posterior MTG among patients with higher perplexity indicates that that inputs of the angular gyrus and anterior temporal lobe (i.e., concepts linked to words retrieved during production) may linger around for longer in discourse production. Lower synaptic gain or excitability of the IFG in turn indicates that MTG inputs (i.e., hierarchical syntax) may have less influence on serial word production than expected. Taken together, our findings suggest an overall disruption in the influence that grammar can have on the generation of coherent, meaningful messages (i.e., semantic control) during speech production in psychosis.

In terms of the strength of connection correlations within the patient group, we discovered that the excitatory connectivity between the IFG and posterior MTG related to lower excitability within the IFG but higher excitability within the posterior MTG. Given that these correlations were sampled from posterior distributions in our study, experimental validation during online speech production or perturb-and-measure approaches such as brain stimulation or pharmacological challenges (e.g., ketamine) are needed to clarify the observed relationship. Although a frontotemporal functional axis between these 2 regions is critical to semantic control,[18] they nonetheless have been observed to exhibit a differential profile.[42] Traditionally, the IFG has been seen as the prime region related to formal aspects of syntactic complexity, while the posterior MTG has been associated with meaning at both the lexical and structural levels.[39,45] For example, the IFG is involved in the processing of more complex object-relative clauses than simpler subject-relative clauses,[48] and to cognitive control over recursively embedded hierarchical structures in mathematics, music, and language,[49] in addition to its role in semantic retrieval and control, verbal working memory, and verbal creativity.[50,51] The left IFG has also already been identified as a region with inhibitory control over distributed networks.[52] In this context, reduced inhibitory tone within the IFG is likely to have a release effect on the posterior MTG, consistent with our observations. A differential effect of perplexity in psychosis on these 2 regions is therefore not unexpected and triggers new and crucial neurocognitive questions about mechanisms behind clinically important phenomena identified under such labels as derailment, incoherence, or tangentiality in spontaneous speech. These mechanisms are more likely to involve an interaction between regions and neurocognitive systems, rather than being localizable in specific regions just as, in language, semantic coherence is the overall effect of multiple interacting systems. Semantic memory as a storage of lexical–conceptual semantic representations and grammatical organization are 2 of these systems, and it is only from their joint operation that any kind of coherence can arise.

The between-group differences in perplexity in this sample has already been reported by He and colleagues[23] in an overlapping sample with fMRI data. To our knowledge, the only other study that estimated a perplexity metric recruited a larger but diagnostically heterogeneous sample of patients with psychotic disorders and reported higher within-subject changes in perplexity over time in relation to positive symptom severity.[53] In our study, the BERT-derived metric of perplexity specifically related to the observer-rated scores of conceptual disorganization, rather than to other positive or negative symptoms. Crucially, the perplexity metric surpassed clinical ratings in explaining the interindividual variations in the connectivity of the brain's semantic network. Computational model-based behavioural metrics have been touted to provide a more reliable instantiation of latent constructs than rating scale measurements,[28] and our empirical demonstration shows that a computationally derived natural language processing metric could be more proximal to brain connectivity than symptom scores in psychosis. The recruitment of an early-stage sample with less than 3 days of lifetime antipsychotic exposure, in an untreated state with active psychotic symptoms, ensured minimal medication and chronicity-related related confounds. Further studies, preferably in different languages and clinical samples, are required to verify and validate this observation.

Limitations

We obtained the speech sample outside the scanner; we used resting-state fMRI data for inferences on connectivity. While this reduced motion-related confounds and enabled us to incorporate DMN nodes (angular gyrus) and draw inferences about spontaneous thought processes, the parameters of connectivity within the semantic network may not be the same during online speech. Furthermore, the temporal resolution of MRI limits the scope of connectivity models; this is especially relevant when considering the speed of word production in everyday speech. We caution the readers when generalizing our findings in the context of everyday discourse.

Conclusion

We used DCM of brain connectivity to understand the likely origins of deviation of meaning in psychosis-related speech. The large-language model–based quantification of perplexity in speech indicated a diminished use of top–down semantic control processes in SDD, which was, in turn, attributable to aberrant synaptic gain function of neural populations within the semantic network. These results expand understanding of the cortical mechanisms that give rise to atypicality in thought, language, and communication in psychosis.

Acknowledgements: The authors thank Drs. Kara Dempster, Priya Subramanian, Julie Richard, Hooman Ganjavi, Sabrina D. Ford, and Michael MacKinley for assistance in clinical assessments and Claudio Palominos Flores, Han Zhang, and Philipp Homan for discussions.

Affiliations: From CIDCL, Escuela de Fonoaudiología, Universidad de Valparaíso, Valparaíso, Chile (Alonso-Sánchez); the Department of Translation & Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain (Hinzen, He); the Intitut Català de Recerca i Estudis Avançats (ICREA), Barcelona, Spain (Hinzen); the Robarts Research Institute, Schulich School of Medicine and Dentistry, Western University, London, Ont. (Gati, Palaniyappan); the Department of Medical Biophysics, Schulich School of Medicine and Dentistry, Western University, London, Ont. (Gati, Palaniyappan); the Douglas Mental Health University Institute, Department of Psychiatry, McGill University, Montréal, Que (Palaniyappan).

Competing interests: Maria Francisca Alonso-Sánchez is a member of the steering committee of the Discourse in Psychosis consortium. Rui He reports support meeting or travel support from the European Research Council and the Department of Science and Technology of Guangdong Province. Lena Palaniyappan reports speaker or consultant fees from Janssen Canada and Otsuka Canada, SPMM Course Limited, and the Canadian Psychiatric Association, book royalties from Oxford University Press, and investigator-initiated educational grants from Janssen Canada, Sunovion, and Otsuka Canada. No other competing interests were declared.

Contributors: Maria Francisca Alonso-Sánchez, Wolfram Hinzen, Joseph Gati, and Lena Palaniyappan contributed to the conception and design of the work. Joseph Gati and Lena Palaniyappan contributed to data acquisition. Rui He contributed to data analysis and interpretation. Maria Francisca Alonso-Sánchez, Wolfram Hinzen, and Lena Palaniyappan drafted the manuscript. All of the authors revised it critically for important intellectual content, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work.

Funding: Lena Palaniyappan is supported by the Canada First Research Excellence Fund, awarded to the Healthy Brains, Healthy Lives initiative at McGill University (through a New Investigator Supplement to Lena Palaniyappan) and the Monique H. Bourgeois Chair in Developmental Disorders. This study was funded by a Canadian Institutes of Health Research Foundation Grant (FDN 154296) to Lena Palaniyappan. Data acquisition was supported by the Canada First Excellence Research Fund to BrainSCAN, Western University (Imaging Core); the Innovation Fund for Academic Medical Organization of Southwest Ontario and Digital Research Alliance Canada (Compute Canada) Resources (no. 1530) were used in the storage and analysis of imaging data. Lena Palaniyappan receives a salary award from the Fonds de recherche du Quebec – Santé. This work was also supported by FONDECYT Regular 1230532 from the Agencia Nacional de Investigación y Desarrollo (to Maria Francisca Alonso-Sánchez). Rui He was supported by the China Scholarship Council (no. 202108390062).

Disclaimer: Lena Palaniyappan is co-editor in chief of the *Journal of Psychiatry and Neuroscience* but was not involved in the editorial decision-making process for this article.

References

1. Smallwood J, Bernhardt BC, Leech R, et al. The default mode network in cognition: a topographical perspective. *Nat Rev Neurosci* 2021;22:503-13.
2. Binder JR, Desai RH, Graves WW, et al. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb Cortex* 2009;19:2767-96.
3. Andrews-Hanna JR, Smallwood J, Spreng N. The default network and self-generated thought: component processes, dynamic control, and clinical relevance. *Ann N Y Acad Sci* 2014;1316:29-52.
4. Yeshurun Y, Nguyen M, Hasson U. The default mode network: where the idiosyncratic self meets the shared social world. *Nat Rev Neurosci* 2021;22.

5. Tong J, Binder JR, Fernandino L, et al. A distributed network for multimodal experiential representation of concepts. *J Neurosci* 2022;42:7121–30.
6. Menon V. ll Perspective 20 years of the default mode network: a review and synthesis. *Neuron* 2023;111:2469-87.
7. Benedek M, Jurisch J, Koschutnig K, et al. Elements of creative thought: Investigating the cognitive and neural correlates of association and bi-association processes. *NeuroImage* 2020;210:116586.
8. Padmanabhan A, Lynch CJ, Schaer M, et al. The default mode network in autism. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2017;2:476-86.
9. Hinzen W, Palaniyappan L. The 'L-factor': language as a transdiagnostic dimension in psychopathology. *Prog Neuropsychopharmacol Biol Psychiatry* 2024;131:110952.
10. Palaniyappan L. Dissecting the neurobiology of linguistic disorganisation and impoverishment in schizophrenia. *Semin Cell Dev Biol* 2022;129:47-60.
11. Matsumoto Y, Nishida S, Hayashi R, et al. Disorganization of semantic brain networks in schizophrenia revealed by fMRI. *Schizophr Bull* 2022;49:498-506.
12. Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 1971;9:97-113.
13. Kuhnke P, Kiefer M, Hartwigsen G. Task-dependent recruitment of modality-specific and multimodal regions during conceptual processing. *Cereb Cortex* 2020;30:3938–59.
14. Jackson RL, Hoffman P, Pobric XG, et al. The semantic network at work and rest: differential connectivity of anterior temporal lobe subregions. *J Neurosci* 2016;36:1490-501.
15. Anderson AJ, Kiela D, Binder JR, et al. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *J Neurosci* 2021;41:4100-19.
16. Jackson RL. The neural correlates of semantic control revisited. *NeuroImage* 2021;224:117444.
17. Noonan KA, Jefferies E, Visser M, et al. Going beyond inferior prefrontal involvement in semantic control : evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. *J Cogn Neurosci* 2013;25:1824–50.
18. Zhang Q, Wang H, Luo C, et al. The neural basis of semantic cognition in Mandarin Chinese: a combined fMRI and TMS study. *Hum Brain Mapp* 2019;5412–23.
19. Aboud KS, Nguyen TQ, Del Tufo SN, et al. Rapid interactions of widespread brain networks characterize semantic cognition. *J Neurosci* 2023;43:142-54.
20. Chiou R, Humphreys GF, Jung JY, et al. Controlled semantic cognition relies upon dynamic and flexible interactions between the executive 'semantic control' and hub-and-spoke 'semantic representation' systems. *Cortex* 2018;103:100-16.
21. Almeida VN, Radanovic M. Semantic priming and neurobiology in schizophrenia: a theoretical review. *Neuropsychologia* 2021;163:108058.
22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv;2023. Available: http://arxiv.org/abs/1706.03762 (accessed 2024 Feb. 28).
23. He R, Palominos C, Zhang H, et al. Navigating the semantic space: unraveling the structure of meaning in psychosis using different computational language models. *Psychiatry Res* 2024;333:115752.
24. Alonso-Sánchez MF, Limongi R, Gati J, et al. Language network self-inhibition and semantic similarity in first-episode schizophrenia: a computational-linguistic and effective connectivity approach. *Schizophr Res* 2023;259:97-103.
25. Zeidman P, Jafarian A, Corbin N, et al. A guide to group effective connectivity analysis, part 1: first level analysis with DCM for fMRI. *Neuroimage* 2019;200:174-90.
26. Zeidman P, Jafarian A, Seghier ML, et al. A guide to group effective connectivity analysis, part 2: second level analysis with PEB. *NeuroImage* 2019;200:12-25.
27. Li B, Daunizeau J, Stephan KE, et al. Generalised filtering and stochastic DCM for fMRI. *Neuroimage* 2011;58:442-57.
28. Palaniyappan L, Homan P, Alonso-Sanchez MF. Language network dysfunction and formal thought disorder in schizophrenia. *Schizophr Bull* 2023;49:486-97.
29. WHO Collaborating Centre for Drug Statistics Methodology. *Guidelines for ATC classification and DDD assignment*. Oslo; 2024.
30. Leckman JF, Sholomskas D, Thompson D, et al. Best estimate of lifetime psychiatric diagnosis: a methodological study. *Arch Gen Psychiatry* 1982;39:879-83.
31. Rose, David, Pevalin, David J. *A Researcher's Guide to the National Statistics Socio-economic Classification*. First London: Sage; 2003.
32. Liddle PF, Ngan ETC, Caissie SL, et al. Thought and Language Index: an instrument for assessing thought and language in schizophrenia. *Br J Psychiatry* 2002;181:326-30.
33. Colla D, Delsanto M, Agosto M, et al. Semantic coherence markers: the contribution of perplexity metrics. *Artif Intell Med* 2022; 134:102393.
34. Feder A, Oved N, Shalit U, et al. CausaLM: causal model explanation through counterfactual language models. *Comput Linguist* 2021:1-54.
35. Salazar J, Liang D, Nguyen TQ, et al. Masked language model scoring. *Proc 58th Annu Meet Assoc Comput Linguist* 2020;2699–712.
36. Pekár S, Brabec M. Generalized estimating equations: a pragmatic and flexible approach to the marginal GLM modelling of correlated data in the behavioural sciences. *Ethology* 2018;124:86–93.
37. FIL Methods Group. SPM12 Manual. 2017;15:1–508.
38. Jefferies E, Thompson H, Cornelissen P, et al. The neurocognitive basis of knowledge about object identity and events: dissociations reflect opposing effects of semantic coherence and control. *Philos Trans R Soc Lond B Biol Sci* 2020;375:20190300.
39. Friston KJ, Kahan J, Biswal B, et al. A DCM for resting state fMRI. *Neuroimage* 2014;94:396-407.
40. Marreiros AC, Kiebel SJ, Friston KJ. Dynamic causal modelling for fMRI: a two-state model. *Neuroimage* 2008;39:269-78.
41. Penny WD. Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage* 2012;59:319-30.
42. Friederici AD, Gierhan SME. The language network. *Curr Opin Neurobiol* 2013; 23:250-4.
43. Novelli L, Friston K, Razi A. Spectral dynamic causal modeling: a didactic introduction and its relationship with functional connectivity. *Netw Neurosci* 2024;8:178-202.
44. Adams RA, Pinotsis D, Tsirlis K, et al. Computational modeling of electroencephalography and functional magnetic resonance imaging paradigms indicates a consistent loss of pyramidal cell synaptic gain in schizophrenia. *Biol Psychiatry* 2022;91:202-15.
45. Matchin W, Hickok G. The cortical organization of syntax. *Cereb Cortex* 2020;30:1481-98.
46. Teige C, Cornelissen PL, Mollo G, et al. Dissociations in semantic cognition: Oscillatory evidence for opposing effects of semantic control and type of semantic relation in anterior and posterior temporal cortex. *Cortex* 2019;120:308-25.
47. Davey J, Thompson HE, Hallam G, et al. Exploring the role of the posterior middle temporal gyrus in semantic cognition: integration of anterior temporal lobe with executive processes. *Neuroimage* 2016;137:165-77.
48. Rogalsky C, Matchin W, Hickok G. Broca's area, sentence comprehension, and working memory: an fMRI study. *Front Hum Neurosci* 2008;2:1–13.
49. Zaccarella E, Papitto G, Friederici AD. Brain and cognition language and action in Broca's area: computational differentiation and cortical segregation. *Brain Cogn* 2021;147:105651.
50. Cogdell-Brooke LS, Sowden PT, Violante IR, et al. A meta-analysis of functional magnetic resonance imaging studies of divergent thinking using activation likelihood estimation. *Hum Brain Mapp* 2020;41:5057-77.
51. Gonen-Yaacovi G, Cruz de Souza L, Levy R, et al. Rostral and caudal prefrontal contribution to creativity: a meta-analysis of functional imaging data. *Front Hum Neurosci* 2013;7:465.
52. Swick D, Ashley V, Turken U. Left inferior frontal gyrus is critical for response inhibition. *BMC Neuroscience* 2008;11:1–11.
53. Girard JM, Vail AK, Liebenthal E, et al. Computational analysis of spoken language in acute psychosis and mania. *Schizophr Res* 2022; 245:97–115.