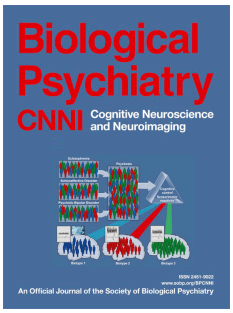
Journal Pre-proof



Speech as a graph: developmental perspectives on the organization of spoken language

Natália Bezerra Mota, Janaína Weissheimer, Ingrid Finger, Marina Ribeiro, Bárbara Malcorra, Lílian Hübner

PII: S2451-9022(23)00098-8

DOI: https://doi.org/10.1016/j.bpsc.2023.04.004

Reference: BPSC 1080

To appear in: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*

Received Date: 15 August 2022

Revised Date: 2 April 2023

Accepted Date: 10 April 2023

Please cite this article as: Mota N.B., Weissheimer J., Finger I., Ribeiro M., Malcorra B. & Hübner L., Speech as a graph: developmental perspectives on the organization of spoken language, *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2023), doi: https://doi.org/10.1016/j.bpsc.2023.04.004.

Title: Speech as a graph: developmental perspectives on the organization of spoken language

Running title: Speech as a graph in developing perspectives

Natália Bezerra Mota* [1,2], Janaína Weissheimer[3,4,7], Ingrid Finger[5,7], Marina Ribeiro[2,8], Bárbara Malcorra[2], Lílian Hübner[6,7]

1. Department of Psychiatry and Legal Medicine – Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil
2. Research department at Motrix Lab – Motrix, Rio de Janeiro, Brazil
3. Department of Modern Foreign Languages – Federal University of Rio Grande do Norte (UFRN), Natal, Brazil
4. Brain Institute - Federal University of Rio Grande do Norte (UFRN), Natal, Brazil
5. Department of Modern Languages – Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil
6. Department of Linguistics - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, Brazil
7. The National Council for Scientific and Technologica Development (CNPq), Brasília, Brazil
8. Bioinformatics Multidisciplinary Environment - Federal University of Rio Grande do Norte (BioME - UFRN), Natal, Brazil

* Corresponding author: Natália Bezerra Mota, Instituto de Psiquiatria UFRJ – Av. Venceslau Brás, 71 – Botafogo, Rio de Janeiro – RJ, Brazil; Zip Code: 22290-140. E-mail address: nataliamota@neuro.ufrn.br Phone: +55 84 994182269

Abstract

Language has been taken as a privileged window to investigate mental processes. More recently, descriptions of psychopathological symptoms have been analyzed with the help of natural language processing tools. An example is the study of speech organization using graph theoretical approaches that began around ten years ago. After its application in different areas, there is a need to characterize better what aspects can be associated with typical and atypical behavior throughout the lifespan, given variables related to aging, as well as biological and social contexts. The precise quantification of mental processes assessed through language may allow us to disentangle bio/social markers by looking at naturalistic protocols in different contexts. In the current review, we discuss ten years of studies in which word recurrence graphs were adopted to characterize the chain of thoughts expressed by individuals while producing discourse. Initially developed to understand formal thought disorder in the context of psychotic syndromes, this line of research has been expanded to understand atypical development in different stages of psychosis, differential diagnosis (such as dementia), as well as typical development of thought organization in school-age children/teenagers in naturalistic and school-based protocols. We comment on the effects of environmental factors, such as education and reading habits (in monolingual and bilingual contexts), in clinical and non-clinical populations at different developmental stages (from childhood to aging). Looking towards the future, there is an opportunity to employ word recurrence graphs to address complex questions that consider bio/social factors within a developmental perspective in typical and atypical contexts.

Introduction

What does it mean to have an organized mind? This is a hard concept to define, but in short, we can assume that the ability to express ideas and feelings through language and adopting a well-structured chain of thoughts is a sign of an organized mind. From early descriptions of major psychiatric syndromes, signs of a disorganized mind have been observed by trained professionals by listening carefully not only to the meaning but also to the flow of elements in a patient's narrative. A narrative is expected to introduce single elements of meaning through an organized sequence, eventually conveying a single unified story. When there is an unexpected sequence of words, we can recognize fragmented pieces of information or an irregular flow of words that hinder this unified meaning. Psychopathology presents descriptions for aberrant word sequences (such as derailment, tangentiality, fragmented speech, or even word salad) (1,2). Although such an approach is still useful, considering the narrative structure as a source of psychopathological information poses an important confounding factor, which is access to formal education (3), since people improve narrative structure by learning grammar and syntax rules and mastering specific linguistic aspects that support a complex narrative.

We can describe word sequence and its irregularities using psychopathological terms. However, with computational tools, we are now able to understand a narrative as a set of words organized in sequence from which specific meanings emerge. The original idea was to formally study the phenomena of planning a narrative using math tools to characterize these word sequences (Figure 1A). Looking at this process (planning a narrative), we described its final product (the narrative) as a system in which each element is a word, and the sequential information determines the recurrence pattern (Figure 1A). It is an indirect way to characterize thought, which is assessed through language. In addition, because the tool does not consider any linguistic rules, theoretically, it is not restricted to any particular language. Moreover, it could be used to study the biological and social factors that impact these word sequences. Thus, the sequence of words that individuals adopt to narrate a story is represented as a graph, where each element (word) can be represented as a node, and the temporal sequences can be represented as direct edges (4,5) (Figure 1A). By adopting this strategy, we characterize the relationship between nodes (words) in the phenomenon (narrative) that determines this sequence. Recurrences of words determine topological metrics, from short-range recurrences (such as repetitions of word associations) to long-range recurrences or connectedness (such as the number of words (nodes) connected in a single component) (Figure 1A). As the word sequence defines the relationship between nodes rather than the semantic content or linguistic relationships, the word recurrence pattern defines topological metrics. Therefore, these graphs are not based on meaning but on recurrence patterns. These topological signs translate the recurrence relationship between words into a narrative. In this sense, if any cognitive or mental process impacts the flow of narrative production, this might produce quantifiable markers that go beyond the typical and expected relationship between words given by grammar or syntactic rules.

After ten years of word recurrence graph research, we aim to review the findings generated so far to characterize both the atypical and typical development. Other computational approaches have been developed and tested for similar purposes, such as semantic coherence based on word embedding techniques (6–8). In this paper, we focused on evidence using the word recurrence graph to discuss biological and social factors that affect oral and

written discourse and discuss the limitations that could also impact novel language metrics, such as the speech elicitation protocol adopted (Figure 1B, Supplemental Information 1). Initially, we discuss word graphs within psychosis, analyzing markers at chronic and earlier stages of the pathology, their symptomatologic correlates, and differential diagnosis. Afterward, we review naturalistic school-based protocols employed to understand the typical development of markers and their cognitive and social correlates (such as formal education, reading/writing habits, and bilingualism). Finally, we present a developmental perspective of such an approach during typical and atypical aging, focusing on disorders such as dementia and their cognitive and social correlates.

Word graphs in psychosis

Chronic psychosis and connectedness

Word graphs were first created by representing a dream narrative as a graph (4). In the original study, (4) collected dream narratives from people with schizophrenia and bipolar disorder diagnosis and matched controls. A dream report was considered a spontaneous narrative, distinct from other speech elicitation protocols based on cued narratives (for example, based on picture descriptions, storyboards, or reading and re-telling stories) (Figure 1B). Syntactic analysis was performed to identify word classes such as subjects, objects, and predicates in a sentence and represented each as a node and the temporal sequence by directed edges. As those elements flow spontaneously in the narrative, a closed loop returning to the original node was created. Major topological differences were found between the two clinical groups. A Bayesian classifier (Naive Bayes classifier using cross-validation in 10 folds) performed the differential diagnosis with more than 90% sensitivity and specificity, even after controlling for verbosity. Despite its limitations of sample size (N=24), this was a breakthrough study, being the first to consider the recurrence of words as a quantifiable strategy to assess language organization and the use of machine learning techniques to identify groups.

At this point, it became clear that the technology needed to be automated, so the next step was to consider every single word as a node (5) (detailed discussion in Supplemental Information 1). We adopted the most naive point of view to define edges: the relationship between words did not consider any linguistic relationship or rules and was based directly on the sequence of nodes (Figure 1A). This decision was taken to avoid subjectivity in the analysis and to prevent adjusting the narrative to any normative data or linguistic rule. A second methodological step was to control for verbosity differences. In (5), the authors opted for splitting the narrative into windows of a fixed number of words (windows of 10, 20, and 30 words), jumping a fixed number of words to define the next graph (for example, one word), and performing consecutive graphs with some overlap between them (90% of overlap in the example of 10-word graphs with one word as a step) (Figure 1C). After performing all the fixed-word graphs, the mean values of each graph attribute are extracted, and analyzed the difference between groups.

In the second study on psychosis (N=60) (5), the authors analyzed participants' dreams and daily reports (another spontaneous narrative, Figure 1B). Only dream reports revealed the differences between groups (Naive Bayes classifier using cross-validation in 10 folds identified the groups with an area under the ROC curve (AUC) ranging from 0.72 to 0.94). The most informative attributes were related to connectedness: the number of edges in the graph (E),

the number of nodes inside the largest connected component (LCC - edges link all pairs of nodes, not considering edges directionality), and inside the largest strongly connected component (LSC - all pairs of nodes are linked and mutually reachable, considering the direction of the edges) (Figure 1A). The authors observed that the less connected the dream narrative, the more socially impacted the participant was, expressing higher levels of blunted affect, poor rapport, difficulty in abstract thinking, and keeping the flow of a conversation.

Since these studies had been conducted with Brazilian-Portuguese speakers (4,5), there was a need to replicate the results in other languages. Given the definition of a node as a single word not attached to any linguistic corpus or pre-processing rules, the method could be easily replicated in English in different cultural contexts, revealing similar topological patterns in discourse production from native English speakers (9–14). In (9), through the analysis of three image reports from participants diagnosed with schizophrenia and bipolar disorder exposed to the Thematic Apperception Test - TAT (15), it was possible to characterize clinical correlates using canonical correlations without collinearity of connectedness attributes with formal thought disorder measured by the subscales of the TLI (Thought Language Index - subscales Disorganization and Impoverishment) (16). Also, the canonical correlation between connectedness and cognitive assessments (performances in working memory and speed of cognitive processing) was significant, as well as the canonical correlation between connectedness and social functioning measured by the GAF (Global Assessment of Function - (17)) and SOFAS (Social and Occupational Functioning Assessment Scale - (18)). Symptomatology, cognitive performance, and social autonomy - all the behavioral measures associated with speech organization - were also mutually associated. Still, neither symptoms nor cognition or social autonomy was associated with neural-functional signals from functional magnetic resonance image - fMRI (combining gyrification index and the variance degree centrality of core hubs acquired at 10 minutes rest with eyes open in a 3 tesla MRI). On the other hand, the canonical correlation between connectedness and neuro-functional measures from fMRI was significant (9). The authors interpreted that only speech organization measured by word recurrence graphs was precise enough to establish the association with biological data assessed by fMRI.

Connectedness decline in different stages of psychosis: pathological development variance

The possibility of quantitatively characterizing negative symptomatology or even social and cognitive risk at earlier stages of psychosis became an important issue in mitigating possible cognitive and social impacts associated with such a diagnosis. In a further study (19) to characterize early detection of connectedness reduction, the authors added the comparison with a random graph distribution to the paradigm by performing 1000 random graphs with the same set of words and the same number of edges, shuffling the word sequence 1000 times (Figure 1D). If the narrative recurrence determines a topological pattern (such as increased LSC, LCC), shuffling the word sequence should disrupt this pattern. Indeed, narrative production from the control group revealed larger LSC and LCC than random graphs. Also, to better control for verbosity differences, the allowed time for the narrative production was limited to 30 seconds.

In the (19) study, 21 first-episode psychosis patients (FEP) and 21 matched controls were interviewed, following the patients for at least six months until their diagnosis was complete according to the DSM IV criteria. At their first consultation, they were experiencing psychotic

symptoms such as delusions and hallucinations. Still, according to DSM IV, the patients had to be followed for at least six months for the diagnosis to be established. Eleven participants received the diagnosis of schizophrenia, and 10 received the diagnosis of bipolar disorder with psychosis. Dream and affective image reports (Figure 1B) (collected at the patient's first consultation) presented lower connectedness for those later diagnosed with schizophrenia disorder, and a Bayesian classifier (Naive Bayes classifier using cross-validation in 10 folds) was able to identify this group with more than 90% accuracy. Moreover, 64% of the participants in the schizophrenia group presented narratives with connectedness measures similar to their random graph distribution (against only 5% of the control and 30% of the bipolar group). Combining connectedness attributes generated from dreams and negative image narratives, a multilinear correlation was performed with negative symptoms after extracting collinear measures. Connectedness attributes were negatively associated with negative symptoms and explained 88% of negative symptomatology variance, allowing for the definition of the "Disorganization Index." The results expressed how those subtle signs can be identified early, detecting symptoms hard to capture by attentive listening alone.

Recently, considering the difficulty of recalling a dream, or the side effects from exposure to negative affective images, a study investigated the validity of a protocol based exclusively on three positive images on FEP (N=24) and matched controls (N=33) (20). Connectedness explained 53% of negative symptomatology variance. Interestingly, the proportion of positive emotional words in the narrative was associated with connectedness only in the FEP group, indicating a possible interaction of emotional recognition and expression and the ability to produce a well-connected narrative under negative symptomatology, with an important contribution of formal education. While negative symptomatology contributed to reduced connectedness, formal education helped to keep connectedness with similar strength, an association mediated by emotional expression. This association between emotional processing and language pattern was independently replicated by (13): a combination of disorganized or underproductive speech markers (that include recurrence graph attributes like LSC) was associated with social cognitive performance (that includes emotional processing).

These differences during the early stages of psychosis have also been replicated in other language and cultural contexts, even in clinical high-risk (CHR) patients (10,11). Administering the same TAT (15) to participants (a group on the FEP, on CHR, followed over seven years to verify transition for psychosis and controls), it was possible to replicate lower connectedness in the FEP group, and in CHR participants that presented a transition to psychosis (11). Also, there were negative correlations between connectedness and formal thought disorder symptomatology measured by TLI (16) (mostly with negative symptoms) and also positive correlations with IQ performance (11).

The word recurrence graph analysis revealed that it could be an exciting strategy to further our understanding of structural aspects of narrative planning and production. Before using language as a pathological marker, however, it is crucial to understand how other confounding factors, such as education, also affect language. The studies conducted so far had a limited sample size, with sparse cultural and linguistic representativeness. It is important to understand its associations and complementarities with other NLP-based markers as presented in (10) and restrictions on speech elicitation protocols (detailed discussion in Supplemental Information 1). As grammar/syntax rules shape the sequence of words, they also impact recurrences, which are important confounding factors within this approach. This

could be especially important if there is a large population variability regarding levels of formal education (3). Before jumping to application purposes, it is important to understand how acquiring new language rules (by learning written language structure or another language) contributes to topological variations in word recurrence graphs. Also, from the clinical point of view, it is worth considering that there is a dominant developmental role in the etiology of many psychiatric disorders.

Typical developmental perspective across the lifetime and differential diagnosis

Speech organization through the lifespan - the role of education and social context

From a developmental perspective, given the dynamic nature of language, it is imperative to understand possible patterns in the development of speech organization and the social/cognitive correlates that support it. Focusing on the changes in social and linguistic behavior when a child begins formal education, (22) collected narratives based on affective pictures limited to 30 seconds in a naturalistic and school-based protocol. Seventy-six second-grade children from low socioeconomic status backgrounds in Brazil were investigated (ages from 6 to 8 years old). The narratives were collected in the middle of the school year, together with the intelligence quotient (IQ) and theory of mind (ToM). Four months later, reading and math scores were acquired from a national assessment. Larger long-range and fewer short-range recurrences were associated with better IQ, ToM, and reading performance. Notably, the association with reading performance was independent of IQ and ToM performances (the correlation between connectedness and reading adjusted by IQ or ToM kept significant levels) (22). In the following year, when participants were in 3rd grade, the authors investigated whether short-term and working memory performance (visual-spatial and verbal) were related to graph markers (23). Exclusively verbal short-term memory was associated with connectedness measures (LSC). The correlation with reading fluency was significant only in the second grade, pointing to possible dynamics associated with the reading acquisition process (23).

As learning new language rules, second language acquisition can also impact word recurrence graphs. The context of biliteracy (simultaneous reading acquisition in two different languages) follows the same direction, increasing long-range recurrences in both languages in association with syntactic complexity (24). Moreover, when a second language acquisition occurs in literate individuals, the increase in connectedness occurs exclusively in the second language, in association with proficiency improvement (25) (detailed discussion in Supplemental Information 2).

The previous evidence shows a change in the word recurrence graph developmental pattern: from a short-range recurrent narrative to a long-range recurrent and well-connected narrative in parallel with general intelligence, theory of mind, verbal memory, and academic performance (22,23). In line with this developmental trajectory, (26) have also observed the impact of attention deficit and hyperactive disorder (ADHD) symptoms in narratives produced by adults, which affected short-range recurrence (L1 or self-loop, the repetition of the same word in sequence) and long-range recurrence markers (LSC). The higher the hyperactivity-impulsivity symptomatology, the more L1 and lower LSC were found in the narratives.

This developmental path during elementary school mirrors an important association with social development through learning. Is growing up immersed in a rich social environment enough

to produce these changes in language patterns, or do formal education and exposure to literacy instruction matter the most? To disentangle this issue, (27) investigated the recurrence pattern change in a diverse population. The sample ranged from individuals aged 2 to 60 years old (N = 214) with distinct educational paths (from zero to 20 years of formal education), including typical adults with low levels of education and individuals with psychosis. The trajectories were studied by adopting an asymptotic model that predicted an accelerated development from the beginning, reaching maturation at a certain point (when changes seem to be more stable), which allowed the authors to estimate the time of maturation counted as years of age or as years of education[1]. First, following the asymptotic model[1], there was an increase in lexical diversity (measured by the number of nodes in the word graph), long-range recurrence (LSC), and the size of the graph (estimated by the average shortest path - ASP), and an abrupt decrease of short-range recurrence (RE), better explained by years of education (R^2 from 0.52 to 0.95) than years of age in multilinear and adjust correlations. In addition, in an independent sample of illiterate adults, long-range recurrence (connectedness - LSC) scores were similar to those obtained by preschoolers. Both scores were distinct from those of literate adults, showing that formal education contributed to this increase more than age. Moreover, the decrease of short-range recurrence matured during the first year of elementary school, while connectedness (LSC) needed 13 years of education (high school level) to reach this maturation point. The group with psychotic symptoms failed to show the same associations with age or education (27), probably expressing another factor that could be more relevant to this context: symptomatology. Given the complex interaction between both developmental trajectories (typical and atypical paths combined in a single mind), cultural, language, and socioeconomic diversity should be considered when interpreting the results (3,21). Daily advance in communication patterns creates different languages and could impact oral narratives. The written language's impact on orality is one example of it. The same dynamics found in a child's oral narrative development were found in written language development through a historical timeline by analyzing more than 700 historical books spanning 5,000 years of literature history (28) (Figure 2A, detailed discussion in Supplemental Information 3). Looking at past advances in languages and their impacts could help us understand how novel communication patterns (like those mediated by social media) could also impact orality.

Typical and atypical aging considering the social context

Regarding aging, connected speech has been used to characterize language production in typical older adults taking into account sociodemographic aspects, and to support the diagnosis of Mild Cognitive Impairment (MCI) or dementia, mainly Alzheimer's Disease (AD).

For instance, (29) employed word recurrence graph analysis to analyze performance in a semantic verbal fluency task in three groups of older adults (AD, MCI, and healthy controls) (N=100). The results showed that graph attributes differed significantly among AD, MCI, and healthy controls, with denser and less direct graph networks in the cognitive impairment contexts. More recently, (30) have applied graph analysis in assessing semantic fluency in individuals with Parkinson's Disease (PD), AD, and healthy controls. The results revealed that

[1] Asymptotic model: $f(t) = f0+(f_\infty - f0)(1-\exp(-t/\tau))$ on which f is the graph attribute, $f0$ is the initial value, f_∞ is the asymptotic value, t is time and τ is the time when f reachs the asymptotic value).

the speech graphs of PD individuals showed higher density, shorter diameter, and shorter average shortest path (ASP) than those of healthy controls but lower density, longer diameter, and longer ASP length than those of AD individuals.

While the studies of (29) and (30) focused on the production of isolated words, (31) applied word recurrence graph analysis to assess language production in a more naturalist picture-narrative task (The dog story) (32). The authors studied differences in oral production between healthy older adults (N=48) and older adults diagnosed with AD (N=24), both groups with low educational levels and socioeconomic status. The AD group produced less connected narratives than the control group, with fewer edges and smaller LSC. The authors also investigated whether connectedness would correlate with episodic, semantic, and working memory scores. Semantic memory correlated moderately with LCC in the AD group, showing that the lower the cognitive performance in the semantic memory task (the ability to name objects), the lower the graph connectedness. On the other hand, episodic memory correlated with LSC in the control group (the ability to remember details was associated with connectedness). These results indicate that semantic memory can precede episodic memory to produce well-connected narratives in AD.

Considering the growing life expectancy worldwide and, consequently, dementia rates, (33) investigated the effect of education and reading and writing habits (RWH) on the oral narratives of typical older adults. One hundred and eighteen individuals, predominantly of low education levels and socioeconomic status, produced an oral narrative from the same sequence of pictures used in (31). Results showed that the advancing age led to increased RE and reduced connectedness, revealing the same pattern of narrative connectedness throughout life found in (27). In addition, canonical correlation analyses revealed associations between two sets of variables (age, education, and reading and writing habits as part of the first set, while RE, LCC, and LSC as part of the second set). Therefore, short- and long-range recurrences were explained by the combination of age, education, and RWH and not by any of these variables in isolation. The strength of the reading and writing habits coefficient compensated for the aging effect, suggesting a protective effect of reading on cognition in a low educational level population.

Taken together, these results demonstrate the relevance of word recurrence graph analyses to discriminate between typical and atypical aging, as well as to characterize typical aging populations regarding sociodemographic and cultural profiles, such as educational level and reading and writing habits.

Conclusion

The current work aimed at discussing how we can gain insights into the complex cognitive process of planning a narrative by understanding its recurrence pattern, considering both biological and social variables. However, we need to address the cognitive basis that supports this behavior and the factors that impact it. The hypothesis is that, as language allows complex communication, it evolves according to its use in social life, supported by biological aspects and environmental input. Word recurrence graphs can reveal connectedness patterns beginning at the initial stages of language acquisition, as it does not rely on the meaning or linguistic rules. At the initial stages of language acquisition, biological factors drive cognitive/sensorial development in association with input from a variety of diverse

environments when young individuals are exposed to spoken language. Factors such as the richness of vocabulary, exposure to different languages, and opportunities for social interactions will contribute to enlarging and diversifying the repertoire of language structures. Social factors such as migration, socioeconomic status, violence, and trauma could also indirectly impact this early exposure to language structure, as well as the quality of social interactions that underlie language development.

Considering typical language development in a literate society, individuals are introduced to formal grammar and syntax rules more systematically during written language acquisition. The lack of schooling for an individual immersed in a literate society will impact the further development of language structure (Figure 2A, (27)). Avoiding short-range recurrences is an explicit strategy encouraged during written production, for example, and drops abruptly in oral production when a child begins reading and writing (Figure 2A, (27)). Also, the development of written skills allows us to have an external repository of language that does not rely on our limited verbal memory capacity, and this development enables us to build more complex connections and produce more connected constructions in written language compared to oral language (Figure 2A, (27)). This does not seem to happen in languages with oral traditions, as expressed in Amerindian leaders' oral tales (Figure 2A, (28)), even though those leaders needed to memorize their cultural traditions and teach them to the following leader. The lack of external repositories of memories is assumed to generate a distinct developmental path for language structure, keeping short-range recurrence to the detriment of global connectedness as a mnemonic strategy (Figure 2A, (28)).

The increase in thought complexity has slow dynamics, maturing during high school, probably relying on different grammar and syntax rules, but also on other factors. Social interactions are more complex, with various biological changes associated with cognitive development. This is the same developmental stage associated with psychosis risk. As psychosis impacts cognitive abilities (especially social cognition), the impact on language structure seems to return to its original pattern. As it occurs after the maturation of short-range recurrences, it does not impact this marker. However, as connectedness is still maturing, it is impacted. Therefore it drops during CHR (Figure 2A, (11)) and after the FEP (Figure 2A, (19,20)). Despite formal education exposure, it still presents a similar pattern at chronic stages (27).

Moreover, during typical development in a literate society, the oral language keeps its connectedness because exposure to a connected written language remains due to work or social life. Biological factors that could impact mental health or increase risks for a neurological event, such as a stroke, could change this stability. Still, even if no neurological incident occurs, a reduction in cognitive abilities is expected during typical aging. Biological factors related to the loss of cognitive and sensorial accuracy associated with social factors, such as retirement, and reduced social interactions, should contribute to a decrease in language connectedness, which is abrupt during dementia (Figure 2A, (31)) and seems to be mitigated by a cognitive reserve (Figure 2A,(33)).

Looking at word recurrence as a graph has also inspired other successful methods to define nodes and edges, including referential relationships, pointing to future directions (12,34,36,38). It is important to consider that the recurrence pattern studied here considers recurrence at the individual words level, different from considering it at the theme level (or the recurrence of concepts or ideas). There might be some overlap and divergences between

both, and even stronger effects at theme-level recurrences as it fits better with psychopathological concepts. Future studies could consider automated strategies to find themes/ideas to be represented as nodes instead of individual words and study their associations with cognitive processes. The association with other computational speech markers can also shed light on complex processes such as social cognition and its diversities (13). Language is at the roots of our social connectedness, supported by a biological basis that changes over time (21,35). Inclusive interventions that allow for social rehabilitation respecting socio-cultural specificities and considering cultural competencies (37) can be evaluated over time by acknowledging the impact of all factors (linked to typical and atypical development) on these language markers. There is a potential benefit in using these and other NLP markers to aid mental health evaluations or to track symptoms. Still, before these uses, we must understand its markers' developmental curve. Understanding this variance may help us characterize recovery beyond symptomatology, promoting quality of life improvement, more than simply mitigating suffering.

Figures:

Figure 01: Methodological illustration of word recurrence graphs. A. An example of a narrative generated after a request to report a story based on a picture previously seen. First, biological and social factors impact the interaction between individuals, such as hearing accuracy and social skills. To formulate the answer, there is an initial mental process that integrates picture information while planning the narrative. This planning includes the choice of words and the sequence. While producing the narrative, biological and social factors could impact language structure. Inspired by the psychopathological strategy to assess mental states through language, the spoken word sequence is represented as a graph. Each word is a node (red dots), representing their sequence by direct edges (arrows). There is no need for pre-processing, such as tokenization, or removing stopwords. The recurrence pattern determines markers of short-range recurrence (such as RE) and long-range recurrence (such as LCC and LSC). **B.** Speech elicitation strategies include spontaneous (such as dream reports) and cued narratives (based on pictures, storyboards, or reading and recalling a story). **C.** To control verbosity differences, a sliding window strategy is often adopted. In this example, consecutive windows of 10 words separated by five words were represented as a graph. After calculating graph attributes for all the 10-word windows, an average of all graph attributes is calculated. **D.** Random graph analysis compares the original narrative sequence with N random graphs created with the same nodes (words), with the same amount of edges, and shuffling word sequence N times. By calculating the zscore of the original graph compared with the random graph distribution, it is possible to estimate how similar this original narrative structure is to a random pattern.

Figure 2: Structural pattern across studies from a developmental perspective considering biological and social aspects. A. Summary of LSC (connectedness marker) from 30-word graphs as a function of age across studies. Blue squares represent a typical development in a literate society, red squares represent results from a pathological development, and purple squares represent results from non-literate populations. All the squares are linked to a study referenced: in black, studies conducted in Brazilian Portuguese; in blue, in English; and in green, original tales from Amerindian leaders (mostly on Kalapalo). The evidence from Brazilian Portuguese and English shows a smaller LSC associated with psychosis. Still, the English dataset showed higher LSC values than the Brazilian Portuguese dataset. **B.** Summary of biological and social factors that interact in each developmental stage and could impact language structure.

Acknowledgments:

We acknowledge the educators and mental health professionals that build the educational system and the communitarian assistance network in Brazil that enable preventive and rehabilitation interventions for those in mental suffering in our country. The current work did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Financial Disclosures:

Dr. Mota, Ribeiro, and Malcorra work at the Motrix, an EduTech startup. Dr. Mota has been a consultant to Boehringer Ingelheim. Dr. Weissheimer has a grant as a productivity researcher at CNPq (National Council of Technological and Scientific Development #306659/2019-0), Dr. Finger has a grant as a productivity researcher at CNPq (National Council of Technological and Scientific Development # 309715/2021-0), and Dr. Hübner has a grant as a productivity researcher at CNPq (National Council of Technological and Scientific Development #312123/2019-1) and by the Coordination of Superior Level Staff Improvement (CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, # 88887.584264/2020-00).

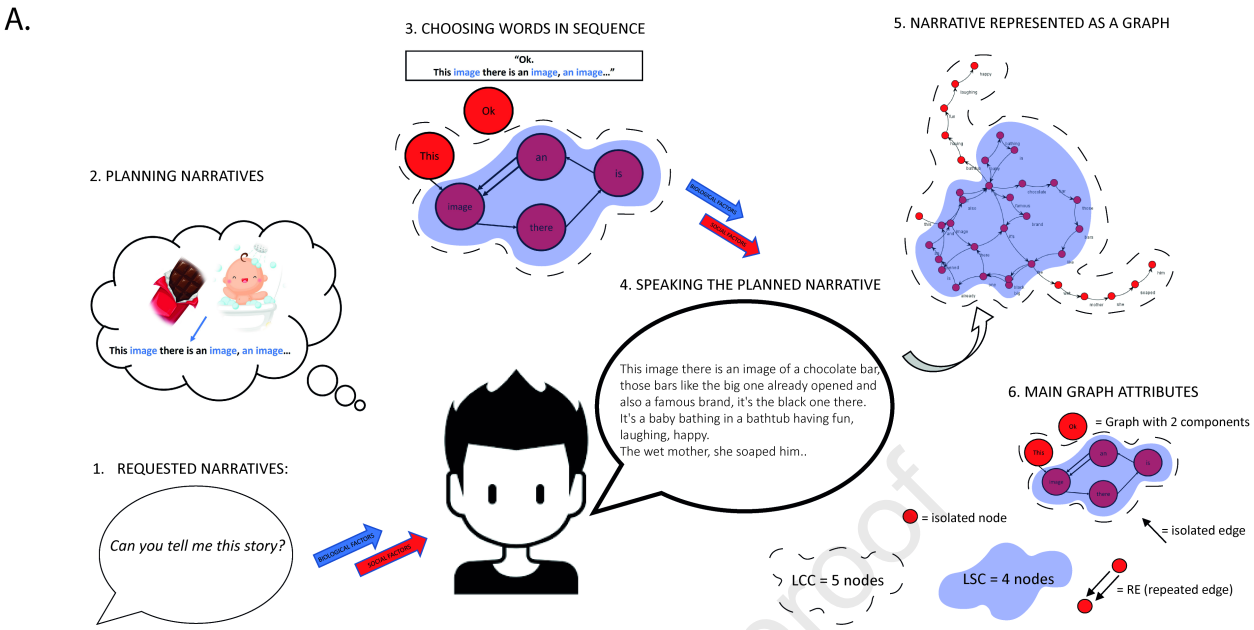
References:

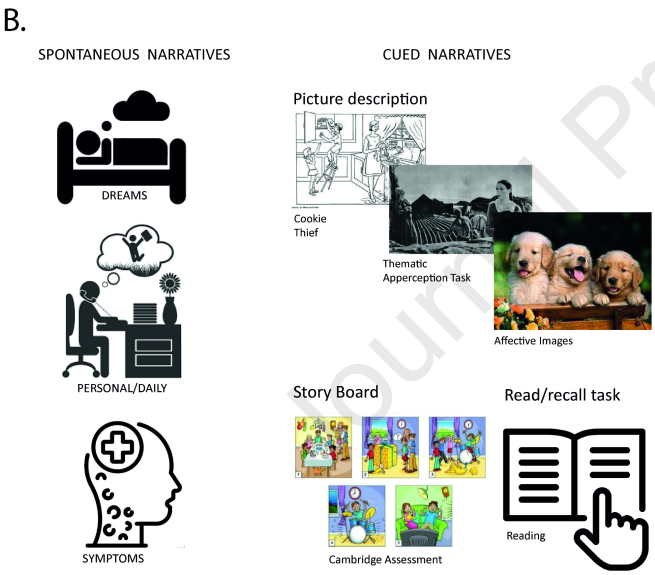
1. Association AP (2016): *DSM-5® Classification*.
2. Holzman PS, Shenton ME, Solovay MR (1986): Quality of Thought Disorder in Differential Diagnosis. *Schizophr Bull* 12: 360–372.
3. Mota NB (2023): How can computational tools help to understand language patterns in mental suffering considering social diversity. *Psychiatry Res* 319. https://doi.org/10.1016/j.psychres.2022.114995
4. Mota NB, Vasconcelos NAP, Lemos N, Pieretti AC, Kinouchi O, Cecchi GA, *et al.* (2012): Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One* 7. https://doi.org/10.1371/journal.pone.0034928
5. Mota NB, Furtado R, Maia PPC, Copelli M, Ribeiro S (2014): Graph analysis of dream reports is especially informative about psychosis. *Sci Rep* 4. https://doi.org/10.1038/srep03691
6. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE (2007): Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr Res* 93: 304–316.
7. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, *et al.* (2015): Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr* 1. https://doi.org/10.1038/npjschz.2015.30
8. Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, *et al.* (2018): Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* 17: 67–75.
9. Palaniyappan L, Mota NB, Oowise S, Balain V, Copelli M, Ribeiro S, Liddle PF (2019): Speech structure links the neural and socio-behavioural correlates of psychotic disorders. *Prog Neuropsychopharmacol Biol Psychiatry* 88. https://doi.org/10.1016/j.pnpbp.2018.07.007
10. Morgan SE, Diederen K, Vértes PE, Ip SHY, Wang B, Thompson B, *et al.* (2021): Natural Language Processing markers in first episode psychosis and people at clinical high-risk. *Transl Psychiatry* 11. https://doi.org/10.1038/s41398-021-01722-y
11. Spencer TJ, Thompson B, Oliver D, Diederen K, Demjaha A, Weinstein S, *et al.* (2021): Lower speech connectedness linked to incidence of psychosis in people at clinical high risk. *Schizophr Res* 228: 493–501.
12. Nikzad AH, Cong Y, Berretta S, Hänsel K, Cho S, Pradhan S, *et al.* (2022): Who does what to whom? graph representations of action-predication in speech relate to psychopathological dimensions of psychosis. *Schizophrenia* 8: 58.
13. Tang SX, Cong Y, Nikzad AH, Mehta A, Cho S, Hänsel K, *et al.* (2022): Clinical and computational speech measures are associated with social cognition in schizophrenia spectrum disorders. *Schizophr Res*. https://doi.org/10.1016/J.SCHRES.2022.06.012
14. Tang SX, Kriz R, Cho S, Park SJ, Harowitz J, Gur RE, *et al.* (2021): Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *NPJ Schizophr* 7. https://doi.org/10.1038/s41537-021-00154-3
15. Murray A (1943): *Thematic Apperception Test.* Cambridge, MA. US.: Harvard University Press.
16. Liddle PF, Ngan ETC, Caissie SL, Anderson CM, Bates AT, Quested DJ, *et al.* (2002): Thought and language index: An instrument for assessing thought and language in schizophrenia. *British Journal of Psychiatry* 181: 326–330.

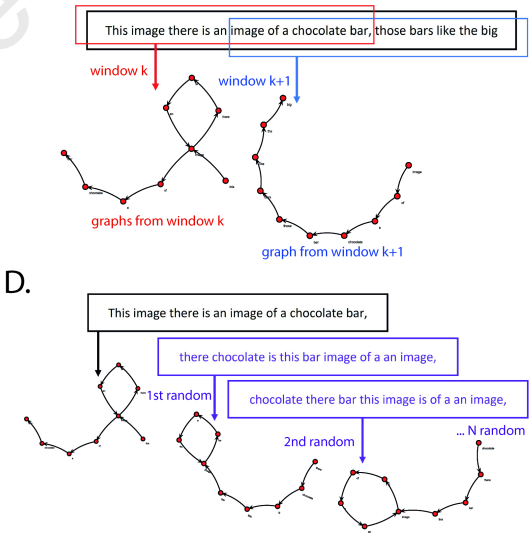
17. First MB, Spitzer RL, Gibbon M, Williams JB (1996): *Structured Clinical Interview for the DSM-IV Axis I Disorders (SCID PTSD Module)*. Retrieved from http://www.scid4.org/
18. Goldman HH, Skodol AE, Lave TR (1992): Revising Axis V for DSM-IV: A Review of Measures of Social Functioning. *Am J Psychiatry* 149: 1148–1156.
19. Mota NB, Copelli M, Ribeiro S (2017): Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *NPJ Schizophr* 3. https://doi.org/10.1038/s41537-017-0019-3
20. Mota NB, Ribeiro M, Malcorra BLC, Atídio JP, Haguiara B, Gadelha A (2022): Happy thoughts: What computational assessment of connectedness and emotional words can inform about early stages of psychosis. *Schizophr Res*. https://doi.org/10.1016/j.schres.2022.06.025
21. Palaniyappan L (2021): More than a biomarker: could language be a biosocial marker of psychosis? *NPJ Schizophr* 7: 42.
22. Mota NB, Weissheimer J, Madruga B, Adamy N, Bunge SA, Copelli M, Ribeiro S (2016): A Naturalistic Assessment of the Organization of Children's Memories Predicts Cognitive Functioning and Reading Ability. *Mind, Brain, and Education* 10. https://doi.org/10.1111/mbe.12122
23. Mota NB, Callipo R, Leite L, Torres AR, Weissheimer J, Bunge SA, *et al.* (2019): *Verbal Short-Term Memory Underlies Typical Development of "Thought Organization" Measured as Speech Connectedness*.
24. Lemke CE, Weissheimer J, Mota NB, de Souza Brentano L, Finger I (2021): The Effects of Early Biliteracy on Thought Organisation and Syntactic Complexity in Written Production by 11-Year-Old Children. *Language Teaching Research Quarterly* 26: 1–17.
25. Botezatu MR, Weissheimer J, Ribeiro M, Guo T, Finger I, Mota NB (2022): Graph structure analysis of speech production among second language learners of Spanish and Chinese. *Front Psychol* 13. https://doi.org/10.3389/fpsyg.2022.940269
26. Coelho RM, Drummond C, Mota NB, Erthal P, Bernardes G, Lima G, *et al.* (2021): Network analysis of narrative discourse and attention-deficit hyperactivity symptoms in adults. *PLoS One* 16. https://doi.org/10.1371/journal.pone.0245113
27. Mota NB, Sigman M, Cecchi G, Copelli M, Ribeiro S (2018): The maturation of speech structure in psychosis is resistant to formal education. *NPJ Schizophr* 4. https://doi.org/10.1038/s41537-018-0067-3
28. Pinheiro S, Mota NB, Sigman M, Fernández-Slezak D, Guerreiro A, Tófoli LF, *et al.* (2020): The History of Writing Reflects the Effects of Education on Discourse Structure: Implications for Literacy, Orality, Psychosis and the Axial Age. *Trends Neurosci Educ* 21. https://doi.org/10.1016/j.tine.2020.100142
29. Bertola L, Mota NB, Copelli M, Rivero T, Diniz BS, Ribeiro MAR, *et al.* (2014): Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, Mild Cognitive Impairment and normal elderly controls. *Front Aging Neurosci* 6. https://doi.org/10.3389/fnagi.2014.00185
30. Zhang G, Ma J, Chan P, Ye Z (2022): Graph Theoretical Analysis of Semantic Fluency in Patients with Parkinson's Disease. *Behavioural Neurology* 2022. https://doi.org/10.1155/2022/6935263
31. Malcorra BLC, Mota NB, Weissheimer J, Schilling LP, Wilson MA, Hübner LC (2021): Low speech connectedness in alzheimer's disease is associated with poorer semantic memory performance. *Journal of Alzheimer's Disease* 82: 905–912.
32. Hübner LC, Loureiro FS, Smidarle AD, Tessaro B, Siqueira ECG, Jerônimo GM, *et al.* (2019): Bateria de Avaliação da Linguagem no Envelhecimento (BALE). In:

Zimmermann N, Delaere FJ, Fonseca RP, editors. *Tarefas Para Avaliação Neuropsicológica 3: Avaliação de Memória Episódica, Percepção, Linguagem e Componentes Executivos Para Adultos*. São Paulo: Memnon, pp 188–218.

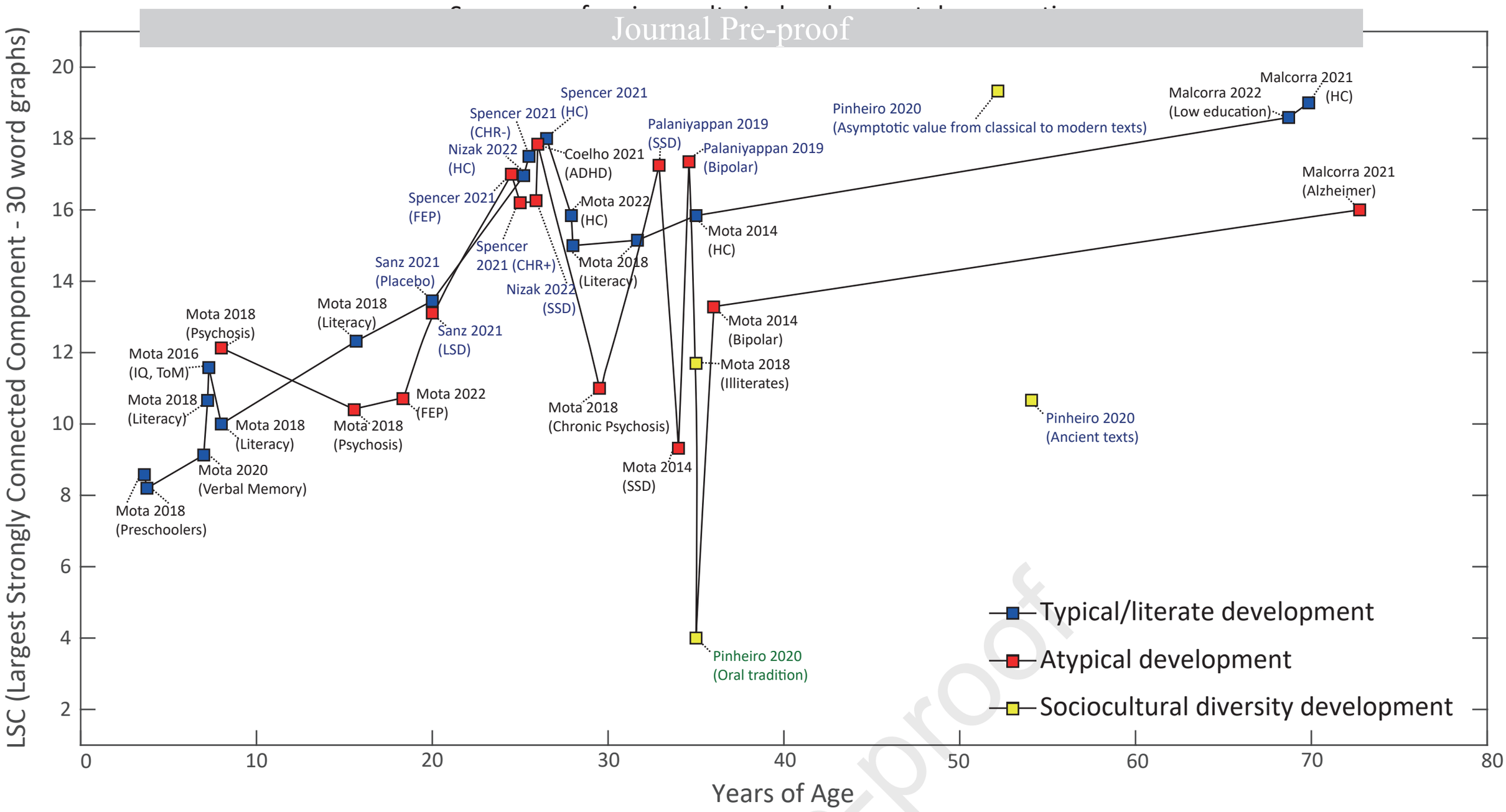
33. Malcorra BLC, Mota NB, Weissheimer J, Schilling LP, Wilson MA, Hübner LC (2022): Reading and writing habits compensate for aging effects in speech connectedness. *NPJ Sci Learn* 7: 13.
34. Nettekoven CR, Diederen K, Giles O, Duncan H, Stenson I, Olah J, *et al.* (2023): Semantic Speech Networks Linked to Formal Thought Disorder in Early Psychosis. *Schizophr Bull* 49: S142–S152.
35. Quesque F, Coutrot A, Cox S, de Souza LC, Baez S, Cardona JF, *et al.* (2022): Does culture shape our understanding of others' thoughts and emotions? An investigation across 12 countries. *Neuropsychology*. https://doi.org/10.1037/neu0000817
36. Palominos C, Figueroa-Barra A, Hinzen W (2023): Coreference Delays in Psychotic Discourse: Widening the Temporal Window. *Schizophr Bull* 49: S153–S162.
37. Mota NB, Pimenta J, Tavares M, Palmeira L, Loch AA, Hedin-Pereira C, Dias EC (2022): A Brazilian bottom-up strategy to address mental health in a diverse population over a large territorial area – an inspiration for the use of digital mental health. *Psychiatry Res* 311. https://doi.org/10.1016/j.psychres.2022.114477
38. Ciampelli S, de Boer JN, Voppel AE, Corona Hernandez H, Brederoo SG, van Dellen E, *et al.* (2023): Syntactic Network Analysis in Schizophrenia-Spectrum Disorders. *Schizophr Bull* 49: S172–S182.







A.



B.