HHS Public Access

Author manuscript

Biol Psychiatry Cogn Neurosci Neuroimaging. Author manuscript; available in PMC 2021 August 01.

Published in final edited form as:

Biol Psychiatry Cogn Neurosci Neuroimaging. 2020 August ; 5(8): 770–779. doi:10.1016/j.bpsc.2020.06.004.

Using language processing and speech analysis for the identification of psychosis and other disorders

Cheryl Mary Corcoran, M.D.,
Icahn School of Medicine at Mount Sinai, James J. Peters Veterans Administration Medical Center

Guillermo Cecchi
Thomas J. Watson Research Center, IBM

Abstract

Increasingly, data-driven methods have been implemented to understand psychopathology. Language is the main source of information in psychiatry and represents “big data” at the level of the individual. Language and behavior are amenable to computational “natural language processing” (NLP) analytics, which may help operationalize the mental status exam. In this review, we highlight the application of NLP to schizophrenia and its risk states as an exemplar of its use, operationalizing tangential and concrete speech as reductions in semantic coherence and syntactic complexity, respectively. Other clinical applications are reviewed, including forecasting of suicide risk and detection of intoxication. Challenges and future directions are discussed, including biomarker development, harmonization and application of NLP more broadly to behavior, including intonation/prosody, facial expression and gesture, and the integration of these in dyads and during discourse. Similar NLP analytics can also be applied beyond humans to behavioral motifs across species, important for modeling psychopathology in animal models. Finally, clinical neuroscience can inform the development of artificial intelligence.

Keywords

language; semantics; syntax; speech graphs; schizophrenia; suicidal

INTRODUCTION

In psychiatry, behavior is the main source of data for diagnosis and treatment. There are no objective laboratory tests as in other fields of medicine. All DSM-5 criteria and diagnoses rest on *signs*, which are observed behaviors, and *symptoms*, which are described in narrative form through language by patients(1). Collateral information comes from families and

Corresponding Author: Cheryl Mary Corcoran, MD, Icahn School of Medicine at Mount Sinai, 1399 Park Avenue, 3-305F, Box 1230, Cheryl.corcoran@mssm.edu.

electronic health records, which are additional interpretations of behavior and narrative, through the lens of individuals other than the patient.

Current artificial intelligence (AI) relies on fast, probabilistic algorithms that perform tasks that normally require human intelligence; these algorithms are trained on large datasets or "big data". Language provides big data at the level of the individual. Automated "natural language processing" (NLP) methods could help operationalize the mental status exam in respect to language and behavior. If validated, such methods could transform the practice of psychiatry such that every patient might provide a behavioral sample as part of the mental status exam, to aid in diagnosis, characterize risk (e.g. psychosis, suicide, violence) and monitor treatment responses. Of note, speech is easy and inexpensive to capture and transcribe and requires no special equipment except a microphone and a recording device.

In this review, we describe data-driven methods to understand the structure of disturbances in language and behavior that underlie psychopathology, with a focus on "natural language processing" (NLP) analytics of transcribed speech. We use schizophrenia as an exemplar, as the structure of language is abnormal in schizophrenia, albeit subtly in early pre-psychosis stages. We focus on emerging literature on data-driven NLP studies in schizophrenia, identifying challenges, especially in respect to harmonization and best practices. We end by reviewing use of NLP for other clinical questions, such as risk assessment for suicidal ideation and behavior, and discuss future directions, including biomarker development, harmonization and the use of NLP to study behavior more broadly, to include intonation/prosody, facial expression and gesture, and integration of these during discourse. Finally, clinical neuroscience can inform the development of artificial intelligence.

What is "Natural Language Processing"?

Speech and language provide a rich source of data on human thought, including semantic and emotional content, semantic coherence (i.e. flow of meaning), and syntactic structure and complexity (i.e. usage of parts of speech). For the RDoC construct of "language production", the main paradigm is "linguistic corpus-based analyses of language output." "Corpus-based" means that the analyses are grounded in large-scale analyses of language made possible by machine-learning algorithms and the availability of large collections from the Internet. "Natural language processing" (NLP) is a field of study at the crossroads of computer science and linguistics, that uses automated analyses to understand natural human language; these analyses are typically corpus-based and probabilistic.

In medicine, NLP has been used primarily for analysis of electronic medical records (EMR), and can assist doctors with diagnosis, clinical trial screens, detection of drug-drug interactions and detection and prediction of adverse events (2). NLP for EMR requires definitions of concepts (e.g. SNOMED-CT or Systematized Nomenclature of Medicine – Clinical Terms), extensive data annotation, and expert opinion, and typically entails simple word counts(2). Examples of this use include lung cancer staging, heart failure prediction and diagnosis of diabetes mellitus (reviewed in (3)), prediction of suicide and accidental deaths(4), and dimensional characterization of psychopathology(5) and its genome-wide associations(6). NLP approaches to EMR are also used to identify symptom clusters and diagnoses from clinical notes (2).

In this review, we focus instead on direct NLP analyses of language of patients themselves, beyond the filter of clinicians' notes in the EMR. Most of this NLP research has focused specifically on psychotic disorders, in which the structure of language at the level of discourse coherence and complexity is disturbed, while negative content and use of first person-singulars is more transdiagnostic. In psychotic disorders, there are reductions in semantic coherence, or the flow of meaning in speech, such that individuals seem tangential. In psychotic disorders, especially schizophrenia, there are also reductions in complexity of speech, which mirror clinical terms of poverty of speech or concreteness.

NLP analyses of semantic coherence

Latent Semantic Analysis (LSA) is "a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text"(7). LSA provides a construction of meaning in language that resembles what the human mind does, i.e. contextualize the meaning of words and phrases in terms of prior experience with those words in different contexts(8). LSA does not require prior knowledge of grammar or vocabulary, and rests on the premise that word meaning is a function of the relationship of each word to every other word in the lexicon. As semantically similar words co-occur in texts more frequently than do unrelated words, the semantic similarity of two words is indexed by the frequency of their co-occurrence in a large corpus of text. This is commonly expressed as word embeddings, in which spatial proximity represents similarity (Figure 1). LSA captures words' meaning through vector representations in high (300–400) dimensional semantic space. Semantic coherence can be evaluated at the level of the word, or any level of word aggregate, including sentences. At the sentence level, semantic vectors for each sentence are calculated as the sum of vectors for each word within the phrase. The cosine of semantic vectors for successive phrases or word aggregates indexes semantic coherence (−1.0 for incoherence to 1.0 for coherence). LSA provides a consistent and sensitive computational approach for quantifying the disjointed flow of ideas that characterizes schizophrenia; it can overcome the subjectivity of clinical ratings and limitations of labor-intensive manual methods. Meaning representation is an active area of research, with recent developments such as Word2Vec and GloVe similarly used to analyze semantic coherence in psychiatry (9),(10),(11).

NLP analyses of complexity

Syntactic complexity is challenging to define and operationalize: approaches include measuring the length of "production units" such as sentences or clause, and usage of embedded or dependent clauses (12). While not capturing the full range of syntactic complexity, a basic NLP approach to assessing complexity is to use Part-of-Speech (POS) tagging(13), another probabilistic linguistic corpus-based algorithm which tags words according to grammatical function, such that sentences can be delineated and characterized in respect to syntax and complexity (longer sentences, use of dependent clauses, etc.) (Figure 2A); POS tagging is integral to the demarcation of sentences(14). Of note, POS tagging of grammatical function depends on the context in which words appear. For example, a word can serve as a noun or a verb (e.g. "dog", "leverage"), or a noun or an adjective (e.g. "fair") (13). Thus, tagging automation must be done in reference to a large text corpus, which acts as the basis for training in parsing words in a speech sample. Often,

the open access software of the Natural Language Tool Kit (nltk.org) is used to parse text and identify the grammatical function of words using the University of Pennsylvania (Penn) Treebank Tag-set, which has thirty-six "part-of-speech" tags, encompassing nouns, verbs, prepositions, etc.(13) POS tagging has been used to "fingerprint" authorship, as individual authors tend to use parts of speech in consistent and identifiable ways; its use in psychiatry and medicine thus far has been limited(14),(15).

Another approach is the use of speech graphs, where discourse is represented as a network with words as nodes and text proximity as edges, indexing co-occurrence patterns among words spoken or written in succession(16). Self-loops occur when edges connect a node (word) to itself, and multiple edges occur when two nodes (words) are connected by more than one edge(16). Network properties are assessed at the local level, describing the neighborhoods of individual nodes and occurrences of sub-graphs or components, and at the global level, in respect to statistical properties of the entire network. Speech graphs capture both semantic coherence (recurrence and deviation) and syntactic complexity(16). Speech graphs have the unique advantage of not requiring a corpus/semantic space for analysis, such that they can be used and generalized across cultures and languages, as large corpuses do not exist for all languages and also have culture-specific biases.

Language production in schizophrenia: a history of analyses

What aspects of language are normal in schizophrenia?

Language has a hierarchical structure that ranges from basic phonology to pragmatics. In schizophrenia, basic features of language appear largely intact, from the perspective of psycholinguistics(17), including phonology (pronunciation of phonemes), morphosyntax (grammatical rules, e.g. tense, subject-verb agreement, article use), and semantics (e.g. naming). Language impairment in schizophrenia commonly occurs at the level of semantic/discourse coherence and cohesion, which are the maintenance of flow of meaning and consistency of references across clauses and sentences.(17),(18) These are addressed below in clinical, manual and automated studies of language in schizophrenia and other psychotic disorders, including their risk states.

Andreasen's conceptualization of positive and negative thought disorder

Andreasen described schizophrenia as a disorder of language and communication in which the speaker "violates the syntactical and semantic conventions which govern language usage"(19). She developed the Thought and Language Communication (TLC) scale, which prioritizes inference of disorganization in thought simply from observing a "patient's speech and language behavior", "without complicated experimental procedures" and "without any attempt to characterize the underlying cognitive processes"(19), an approach consistent with automated "natural language processing" methods and approach described here, which takes natural language/speech itself as the substrate for linguistic analysis. Andreasen conceptualized "thought disorder" as having two domains: positive and negative. Positive thought disorder is expressed as the disruption in the normal flow of discourse (e.g. tangentiality, derailment, and circumstantiality) and can be captured by the notion of coherence. Negative thought disorder is expressed as the structural impoverishment of

discourse and can be described as a decrease in complexity (e.g. concreteness, poverty of speech). While “poverty of speech” can be indexed simply as the amount of speech spoken, more broadly, the heuristic of positive and negative thought disorder may correspond to NLP measures of coherence and complexity, respectively(20), as well as features derived from speech graph analysis(16). In support of this, TLC clinical ratings of thought disorder are associated with measures of LSA semantic coherence in schizophrenia patients (20).

Hoffman’s mathematical modeling of language and the “computational patient”

Beginning in the 1980’s, Hoffman and colleagues used theory and modeling to create a body of work on language in schizophrenia that informs the current emergence of automated NLP approaches. First, drawing on the hierarchical nature of language, he operationalized clinician measures of coherence in speech, with tests of classification based on random sampling of speech from a psychiatric population, finding high accuracy for discrimination of speech in schizophrenia from that of other disorders(21). Then, using discourse analysis, he showed that in schizophrenia, basic discourse structure was disrupted, while in mania, decreased coherence comes from increased shifts from one discourse structure to another(22), anticipating findings by Mota et al 2012, who similarly discriminated language in mania from schizophrenia, using speech graph methods.(14) Additionally, Hoffman found patients with schizophrenia generated more errors in meaning than the norm when required to build multi-sentence texts based on sets of input propositions; these errors were increased when syntax was either passive or complex, and these errors were not correlated with symptom severity, medications, or verbal skills.(23) In 1997, Hoffman modeled schizophrenia using neural network simulations of parallel, distributed processing systems, with reduction in connectivity leading to functional fragmentation of the attractor network, such that schizophrenia symptoms could be modeled, including decreased coherence in speech(24). This followed a special issue of the Journal of Nervous and Mental Disorders in 1994 that focused on the application of artificial intelligence to the question of psychotic speech, comprising papers by Hoffman and others. One example was the creation by Garfield and Rapp of semantic networks with “case frames” and “object taxonomies”, in which node-based and pathway-based reasoning rules could be violated to mimic speech in schizophrenia(25). In 2011, Hoffman built a “computational patient” or artificial neural network, whereby competing illness mechanisms were assessed for “goodness of fit” for breakdown of narrative coherence in real schizophrenia patients, finding exaggerated prediction error signaling during consolidation of episodic memories (e.g. “hyperlearning”) to offer the best fit(26), a model also supported by physiological and psycholinguistic studies by Kuperberg(18).

Elvevåg’s application of latent semantic analysis to language in schizophrenia

The first direct application of automated NLP to open-ended narratives by patients was done by Elvevåg in 2007, specifically applying LSA to speech produced by schizophrenia patients(20). Elvevåg and colleagues found that LSA semantic coherence differentiated speech in schizophrenia from the norm with 82% accuracy(20) and from unaffected adult siblings(27) with 86% accuracy. Decreased LSA semantic coherence also characterized older patients with schizophrenia(28), in whom it was related to poor adaptive functioning, independent of demographics and other symptoms. In these studies, LSA semantic

coherence was correlated with clinical ratings of thought disorder, as captured by Andreasen's scale for appraising the coherence of narrative language(20),(27). Elvevåg's finding of decreased coherence in schizophrenia extended beyond natural speech to also include single word associations in verbal fluency tasks(20), which was replicated more recently using GloVe(29). Further, Elvevåg and colleagues have used LSA to develop automated ratings of verbal recall on the Wechsler Memory Scale-Revised, that were correlated with human ratings(30).

Mota's application of graph theory to quantify language disturbance in psychosis

Another approach to evaluate coherence and complexity in narrative language in schizophrenia and related psychotic disorders is the use of graph theory(31). In a speech graph, there are self-loops(32), which are the series of edges that ultimately connect a node (word) back to itself (Figure 2B), and there are multiple edges, in which two words are connected by more than one edge(16). Local measures describe the neighborhood of nodes and the occurrence of sub-graphs or components, while global measures reflect the statistical properties of the network of the entire text as a whole(16). Mota and colleagues found that local measures, in particular larger subgraphs or components, discriminated speech in schizophrenia from that of mania, consistent with earlier studies by Hoffman and Andreasen. (22) Later, Mota and colleagues found a machine learning classifier, comprised of local measures, accounted for 88% of the variance in negative symptoms in patients with first episode psychosis, and predicted schizophrenia diagnosis six months later with 92% accuracy; the classifier was validated in a second cohort, discriminating psychosis from the norm with 85% accuracy.(33) Normative developmental data exist for speech graph features, and patients with psychosis show early deviation from this normal trajectory (34). Cognitive, clinical and neural correlates exist for these speech graph features in psychosis, including slowed processing speed, higher clinical ratings of thought disorder, cortical gyrification and degree centrality in resting state functional connectivity.(35)

Could automated NLP approaches be used to predict schizophrenia?

In the preceding paragraph, we highlight a manuscript by Mota and colleagues that shows that speech graph features can predict a schizophrenia diagnosis six months later among individuals with first episode psychosis(33). But could NLP features predict the onset of schizophrenia-related psychosis among individuals at risk?

Clinical ratings of language in at-risk youths

Prospective risk cohort studies show that disorganization in language may predict the onset of psychosis in schizophrenia. In the New York High Risk Project, a longitudinal study of children of patients with schizophrenia or affective disorders, baseline clinical ratings at age nine of "positive thought disorder" (reduced coherence) and "negative thought disorder" (reduced complexity), as measured with Andreasen's TLC, and applied to videotape transcripts, predicted schizophrenia onset a decade later, with classification accuracy of 94% (36). In teens and young adults at clinical high risk (CHR) for psychosis, baseline clinical ratings of subtle language disturbance, weighted toward reduction in semantic coherence, has been consistently associated with increased transition rates to psychosis, including in

consortia(37) and large cohorts (38)(39). Mild to moderate “disorganized communication” was associated at one US consortium site with an eightfold increase in hazard for psychosis, carrying the greatest weight in the predictor model(40), and at another site, a doubling of hazard, both at baseline and as a stable trajectory over time(41).

Bearden’s manual linguistic analyses in at-risk youths

Beyond clinical ratings, manual linguistic analysis of baseline narratives by CHR teens identified linguistic features significantly predictive of psychosis, including counts of illogical thought content and poverty of content, and errors in use of pronouns or comparatives to refer to individuals or objects previously mentioned (e.g. “referential cohesion”)(42). The overall model using these manual linguistic features had prediction accuracy of 71%, better than the 35% for clinical ratings, and not accounted for by IQ(42). Illogical thought, poverty of content and errors in referential cohesion are all features previously described in schizophrenia by Andreasen, Hoffman and others.

Automated NLP analyses in at-risk youths by Cecchi, Wolff and Mittal

Each metric identified in Bearden’s manual linguistic study are amenable to automated NLP processing. Referential cohesion was assessed in CHR youths using the Coh-Metrix tool, as applied to written narrative descriptions elicited by a visual prompt. What Coh-Metrix does is apply part-of-speech (POS) tagging through a syntactic parser and identify roots and morphological forms (e.g. past tense, plurals), used to then identify relational connections across different parts of the text. These include overlap in use of words with themselves and with those that share a morphological stem, as well as with pronouns(43). Mittal and colleagues identified abnormal referential cohesion in CHR patients, specifically less overlap with morphological stems in writings by CHR patients than in healthy controls, associated with severity of subthreshold positive and disorganization symptoms, and lower verbal learning scores(44). Coh-Metrix has also been used to show decreased cohesion more broadly in first-episode psychosis, correlated with clinical ratings of disorganization(45).

The formal analysis of logical entailment is a challenging problem and an area of active research in NLP(46); however, illogical thought is suggested by reduction in semantic coherence, which in turn can be analyzed using word embeddings such as LSA. Poverty of content may be comprised in part by reduction in syntactic complexity, which can be analyzed using part-of-speech tagging. In a small proof-of-principle study, Bedi et al (14) applied LSA and POS tagging to transcripts of open-ended interviews conducted with CHR youths as part of a qualitative research study(47) for whom later psychosis outcome was known. Twenty-two LSA semantic and POS syntactic features were used to train and test the machine learning classification algorithm. Semantic features included minimum, mean, median and standard deviation for phrase-level semantic coherence; syntactic features included phrase length and rates of usage of different parts of speech. The best ML classifier included three parameters: minimum *semantic* coherence from one phrase to the next, and *syntactic* measures of complexity, including phrase length and usage of “complementizers” such as “which” and “that”, which introduce dependent clauses. This classifier correlated with prodromal symptom ratings but outperformed them in predicting psychosis. In canonical correlation, semantic indices correlated with positive symptoms whereas syntactic

indices correlated with negative symptoms (r's > 0.4). This same ML classifier also discriminated speech in schizophrenia from the norm with greater than 70% accuracy, both among English-speaking as well as Portuguese-speaking cohorts (data unpublished).

The predictive power of an NLP machine learning classifier focused on semantic coherence was cross-validated in speech samples from Bearden's manual linguistic study (42). As speech was elicited using a structured paradigm and responses were briefer (< 20 mean words per response vs. >150 words per response in the Bedi et al study(14)), a skip-gram approach was used to assess semantic coherence, generating nine semantic features significantly different among CHR converters, CHR non-converters, and healthy controls. POS tagging identified five syntactic variables that differed by group. Factor analysis yielded an ML classifier characterized by semantic coherence features and possessive pronoun usage, which in CHR datasets, had an intra-protocol accuracy of 83%, cross-protocol accuracy of 79%, and further 72% accuracy in discriminating speech of recent-onset psychosis patients from that of healthy individuals(15). Further, canonical correlation between automated and manual linguistic variables was highly significant ($r=0.71$, $p<10^{-6}$).

Another approach has focused on semantic content (the meaning of words themselves) and poverty (vs. richness) of content. A traditional way to measure semantic content is to use Linguistic Inquiry Word Count (LIWC)(48), a text analysis software that uses a pre-existing dictionary to assign words to categories; LIWC shows "hope" to be related to anger and sadness in schizophrenia, instead of to social and future words, or optimism(49). A traditional way to assess lexical diversity is to use Coh-Metrix to calculate type-token ratios (TTR), the ratio of unique words ("tokens") to total number of words; TTR is decreased in schizophrenia, and correlates with clinical ratings of thought disorder.(50) In a recent study, Rezaii, Walker and Wolff take newer approaches to measure both semantic content, using latent content analysis, and poverty of content, conceptualized as reduction in "semantic density"(51). Semantic density is assessed through "vector unpacking", which breaks down the meaning of a sentence into its core ideas. Following usual preprocessing (e.g. lemmatization, part-of-speech tagging) and use of Word2Vec word embeddings, vector unpacking was used to determine how many distinct meaning vectors are needed to recreate the meaning of a sentence, as an index of semantic density. Based on analyses of transcripts of clinical interviews, lower semantic density, plus the greater use of words related to voices and sounds, was predictive of psychosis transition with accuracy of ~90%(51). In another study, analysis of open-ended clinical interviews showed CHR and first episode patients produce more metaphoric content than healthy controls, using an NLP semantic method trained on a dataset human-labeled for metaphors(52).

NLP studies across psychiatry

A recent systematic review and meta-analysis found eighteen studies that used NLP methods to assess semantic features, primarily for psychosis, but also autism spectrum, dementia and Parkinson's disease. Overall cross-diagnostic effect size was of medium to large effect, with Hedges' g of .84 for autism spectrum and .96 for psychosis(53). In autism, LSA shows deficits in coherence of narrative recall (54), narrative response to visual prompts(55), and narrated descriptions of social relations(56); hence, disturbance in discourse coherence may

extend beyond psychosis to include autism spectrum, and potentially other disorders. In Parkinson's disease, LSA, POS tagging, and graph embedding applied to brief monologues, with machine learning, generated a classifier with 75% accuracy, based primarily on differential grammar patterns (57).

Some NLP linguistic features are transdiagnostic, specifically increased use of both first-person pronouns and of words with negative emotion content. LIWC has been applied to show that these features characterized separate online peer support groups for generalized anxiety, borderline personality, major depressive and obsessive-compulsive disorders, and schizophrenia (58), and may extend beyond mental illness forums (psychosis, depression, "Asperger's") to include spinal cord injury, cancer and physical symptoms (skin lesions, fatigue, and overall poor health), though not evangelical, anti-religion or conspiracy blogs(59). Therefore, a focus on self, reflected in first-person pronoun use, and semantic content evocative of negative emotions, may be features common to any mental and emotional distress, which is supportive by their predictive power for psychosis relapse(60), suicide attempt(61), and even death by suicide among poets (62). Individuals may construct narrative emplotments of their suffering in a sociocultural context, in order to manage and make sense of their suffering(63).

By contrast, there is a dearth of studies that have applied NLP analytics to spoken language across disorders, such that the extent to which discourse coherence and syntactic complexity uniquely characterize or distinguish psychosis and schizophrenia from other disorders is not yet known.

NLP also has promise for detecting states of intoxication. The same individuals, exposed in the laboratory on different days to MDMA, methamphetamine or placebo, exhibit disparate language patterns (64),(65). This raises the possibility that language analysis might be an alternative to breathalyzers, or possibly quantify intoxication.

Challenges and future directions

Biomarker development

Overall, application of automated NLP analytics to predict psychosis onset is relatively new, consisting of a few small studies, with limited cross-validation of classifiers. The focus thus far has been primarily on semantic content, coherence and density, and to some extent syntax. However, automated NLP analytics is rapidly expanding across industry and research such that there are several other approaches that can be used, including NLP analyses of metaphor, bizarreness, sentiment, and others(52).

Of note, "language production" is a construct within the Research Domain Criteria (RDoC), for which the main paradigm is "linguistic corpus-based analyses of language output", which are NLP analyses. NLP linguistic features such as coherence and complexity are putative biomarkers for psychosis for which there is a reasonable development path, which includes validation in multisite studies, tests of reliability and reproducibility, identification of sources of variability, standardization of protocols, and ultimately, assessments of acceptability, cost, utility and regulatory "context of use." Automated linguistic features

have been correlated with clinical ratings, including LSA semantic coherence, POS syntactic complexity, speech graph connectedness, and Coh-Matrix cohesion. Linguistic biomarkers may be informative about mechanisms when examined across circuit-based and physiological levels of analysis. In psychotic disorders, LSA coherence is associated with superior temporal activation(66), speech graph connectedness with gyrification and functional connectivity(35), and syntactic complexity (mean length of utterance) with integrity of white matter in language tracts(67). NLP analytics may align with neuroimaging studies of natural language production, implicating disruption of specific temporal windows(and hence hierarchical levels) of information processing in schizophrenia(68), consistent with Kuperberg's hierarchical generative framework of language processing, and its disruption in schizophrenia(18). NLP features that index coherence could be studied at the physiological level, in respect to abnormal "priming" event-related potentials in schizophrenia prior to language production(69).

Standardization and harmonization

Robust reduction in LSA semantic coherence in psychosis is reliably identified across several methods for eliciting speech, including single word associations, verbal fluency tasks and requests for narratives (e.g. how to do laundry, a description of free will)(20). However, standardization in speech elicitation is necessary as measures of coherence and complexity are context-dependent, and more evident in paradigms that provide less structure(23),(70). Across diagnoses of psychosis, autism, and dementia, analyses of full sentences discriminates pathology more than that of single words generated during tasks(53). While the use of visual prompts (e.g. pictures, cartoons) enables standardization in eliciting narrative, it may be less naturalistic than free speech(45). Certainly, if semantic content is a focus, then standardization is needed to minimize bias, and clinical interview provides the opportunity to include symptom content(51). Overall, the field must weigh the pros and cons of different strategies, and then standardize and harmonize methods across studies.

Issues with transcription

NLP algorithms are inherently statistical; in industry, such as banking or machine operation, accuracy of transcription is important as small error rates have serious consequences. However, applications in psychiatry do not require exacting standards, as they are based on statistical descriptions of features representing entire samples. Linguistic analyses can be robust to error rates as high as 25% in automated transcription programs(71).

Beyond language: intonation/prosody/facial expression/gesture

Beyond words themselves, human communication entails a rich repertoire of expression, including prosody, face emotion expression, gesture, eye contact. In schizophrenia, meta-analysis shows large effect size for abnormal duration and frequency of pauses(72), evident also in CHR individuals and related to negative symptoms(73). Baker and colleagues have developed a computational approach to voice and face expression analysis, finding increased brow flashes and greater smile variability when sitting alone were associated with clinical ratings of unusual thought content among schizophrenia patients (74). It may be that all behavior, more broadly, has coherence and syntax, which can be modeled in rodents by applying machine learning to sequences of behavioral "syllables" or "motifs" or "movemes"

(75). Temporal dynamics of natural language and behavior can be evaluated in respect to patterns of neural activation across temporal windows, both in humans(76) and in animals(77), to understand neural mechanisms. In both humans and animals, these can be observed across different states (e.g. exposure to drugs(64) and medications(77)) and also within dyads, in terms of discourse/dialogue in humans(78).

Language, objectivity and ground truth

While it is possible to assign objectivity or a truth value to some linguistic expressions (e.g. “The red apple is on the counter”), this does not generalize to the broader context of language use (e.g. “Give me the keys!”). NLP provides the distinct advantage of consistency: however imperfect the algorithms discussed in this review may be at capturing our intuitions of speech coherence and complexity, they are designed to be consistently reproducible in different experimental designs by different researchers. In this sense, NLP can contribute to a common understanding of language in the context of mental health and align psychiatry with modern epistemological notions of objectivity as consensual agreement(79)(80)

What the future may bring: language and the convergence of psychiatry and artificial intelligence

The analytic features used in the studies of psychosis reviewed here as an exemplar are, to a large extent, relatively simple in comparison with the seemingly inexhaustible richness of language to shape human communication, behavior and mental life, ranging from simple transactional utterances to poetry read across millennia. Climbing to higher analytic echelons will require a closer collaboration between psychiatry and artificial intelligence (AI), which we claim is possible precisely because language theory has influenced AI since its very inception. In 1939, Alan Turing, the founding father of AI, attended and actively participated in a seminar at Oxford presented by Ludwig Wittengenstein(81), arguably the most important philosopher of the 20th century, during which Wittgenstein developed and debated the ideas he would later publish in his Philosophical Investigations(81), in which he proposes to understand language not as “representation” but as a process that acquires meaning by the way participants use it in “language games”. Turing’s seminal paper, Computer Machinery and Intelligence(82) can be understood as a direct application of the game-theoretic perspective on language: if you can play the game like a human, I cannot claim you are not human, because this is how I know others are human. While performance in various games, from chess to arcade to Go, have become gold standards for measuring advances in AI, conversational language still lags noticeably behind. In this regard, psychiatry can not only benefit from tools to quantify the complexity of human dialogue and communication, but also provide significant and consequential insights and a challenging arena to test and sharpen AI developments.

ACKNOWLEDGMENTS

R01MH107558, R01MH115332

FINANCIAL DISCLOSURE: The authors report no biomedical financial interests or potential conflicts of interest.

REFERENCES:

1. American Psychiatric Association (2013): DSM-5 Diagnostic Classification. Diagnostic and Statistical Manual of Mental Disorders. 10.1176/appi.books.9780890425596.x00diagnosticclassification
2. Corcoran CM, Benavides C, Cecchi G (2019): Natural language processing: Opportunities and challenges for patients, providers, and hospital systems. Psychiatr Ann 49 10.3928/00485713-20190411-01
3. Zeng Z, Deng Y, Li X, Naumann T, Luo Y (2019): Natural Language Processing for HER-Based Computational Phenotyping. IEEE/ACM Trans Comput Biol Bioinforma. 10.1109/TCBB.2018.2849968
4. McCoy TH, Castro VM, Roberson AM, Snapper LA, Perlis RH (2016): Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry. 10.1001/jamapsychiatry.2016.2172
5. McCoy TH, Yu S, Hart KL, Castro VM, Brown HE, Rosenquist JN, et al. (2018): High Throughput Phenotyping for Dimensional Psychopathology in Electronic Health Records. Biol Psychiatry. 10.1016/j.biopsych.2018.01.011
6. McCoy TH, Castro VM, Hart KL, Pellegrini AM, Yu S, Cai T, Perlis RH (2018): Genome-wide Association Study of Dimensional Psychopathology Using Electronic Health Records. Biol Psychiatry. 10.1016/j.biopsych.2017.12.004
7. Landauer TK, Foltz PW, Laham D (1998): An introduction to latent semantic analysis. Discourse Process. 10.1080/01638539809545028
8. Landauer TK, Dumais ST (1997): A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychol Rev. 10.1037/0033-295X.104.2.211
9. Mikolov T, Corrado G, Chen K, Dean J (2013): word2vec-v1. Proc Int Conf Learn Represent (ICLR 2013).
10. Pennington J, Socher R, Manning CD (2014): GloVe: Global vectors for word representation. EMNLP 2014 – 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference 10.3115/v1/d14-1162
11. Altszyler E, Ribeiro S, Sigman M, Fernández Slezak D (2017): The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text. Conscious Cogn. 10.1016/j.concog.2017.09.004
12. Szmrecsanyi B (2004): On operationalizing syntactic complextity. JADT 2004 7es Journées Internationales d'Analyse Statistique Des Données Textuelles.
13. Santorini B (1990): Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). Univ Pennsylvania 3rd Revis 2nd Print 10.1017/CBO9781107415324.004
14. 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 10.1038/npjschz.2015.30
15. 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 10.1002/wps.20491
16. 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. 10.1371/journal.pone.0034928
17. Covington MA, He C, Brown C, Naçi L, McClain JT, Fjordbak BS, et al. (2005): Schizophrenia and the structure of language: The linguist's view. Schizophr Res. 10.1016/j.schres.2005.01.016
18. Brown M, Kuperberg GR (2015): A hierarchical generative framework of language processing: Linking language perception, interpretation, and production abnormalities in schizophrenia. Front Hum Neurosci. 10.3389/fnhum.2015.00643
19. Andreasen NC, Grove WM (1986): Thought, language, and communication in schizophrenia: diagnosis and prognosis. Schizophr Bull. 10.1093/schbul/12.3.348

20. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE (2007): Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. Schizophr Res 10.1016/j.schres.2007.03.001
21. Hoffman RE, Kirstein L, Stopek S, Cicchetti D V. (1982): Apprehending schizophrenic discourse: A structural analysis of the Listener's task. Brain Lang 10.1016/0093-934X(82)90057-8
22. Hoffman RE, Stopek S, Andreasen NC (1986): A Comparative Study of Manic vs Schizophrenic Speech Disorganization. Arch Gen Psychiatry 10.1001/archpsyc.1986.01800090017003
23. Hoffman RE, Hogben GL, Smith H, Calhoun WF (1985): Message disruptions during syntactic processing in schizophrenia. J Commun Disord 10.1016/0021-9924(85)90020-6
24. Hoffman RE (1997): Neural Network Simulations, Cortical Connectivity, and Schizophrenic Psychosis. MD Comput. 10.1142/9789812819819_0020
25. Garfield DAS, Rapp C (1994): Application of artificial intelligence principles to the analysis of "crazy" speech. J Nerv Ment Dis. 10.1097/00005053-199404000-00002
26. Hoffman RE, Grasemann U, Gueorguieva R, Quinlan D, Lane D, Miikkulainen R (2011): Using computational patients to evaluate Illness mechanisms in schizophrenia. Biol Psychiatry. 10.1016/j.biopsych.2010.12.036
27. Elvevåg B, Foltz PW, Rosenstein M, DeLisi LE (2010): An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. J Neurolinguistics. 10.1016/j.jneuroling.2009.05.002
28. Holshausen K, Harvey PD, Elvevåg B, Foltz PW, Bowie CR (2014): Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. Cortex. 10.1016/j.cortex.2013.02.006
29. Pauselli L, Halpern B, Cleary SD, Ku B, Covington MA, Compton MT (2018): Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. Psychiatry Res. 10.1016/j.psychres.2018.02.037
30. Rosenstein M, Diaz-Asper C, Foltz PW, Elvevåg B (2014): A computational language approach to modeling prose recall in schizophrenia. Cortex 10.1016/j.cortex.2014.01.021
31. Sigman M, Cecchi GA (2002): Global organization of the Wordnet lexicon. Proc Natl Acad Sci U S A. 10.1073/pnas.022341799
32. Ma'Ayan A, Cecchi GA, Wagner J, Rao AR, Iyengar R, Stolovitzky G (2008): Ordered cyclic motifs contribute to dynamic stability in biological and engineered networks. Proc Natl Acad Sci U S A. 10.1073/pnas.0805344105
33. 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. 10.1038/s41537-017-0019-3
34. 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 10.1038/s41537-018-0067-3
35. 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 Neuro-Psychopharmacology Biol Psychiatry. 10.1016/j.pnpbp.2018.07.007
36. Gooding DC, Ott SL, Roberts SA, Erlenmeyer-Kimling L (2013): Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: Findings from the New York High-Risk Project. Psychol Med. 10.1017/S0033291712001791
37. Addington J, Liu L, Buchy L, Cadenhead KS, Cannon TD, Cornblatt BA, et al. (2015): North American Prodrome Longitudinal Study (NAPLS 2): The prodromal symptoms. J Nerv Ment Dis. 10.1097/NMD.0000000000000290
38. Nelson B, Yuen HP, Wood SJ, Lin A, Spiliotacopoulos D, Bruxner A, et al. (2013): Long-term follow-up of a group at ultra high risk ("Prodromal") for psychosis the PACE 400 study. JAMA Psychiatry 10.1001/jamapsychiatry.2013.1270
39. Demjaha A, Valmaggia L, Stahl D, Byrne M, McGuire P (2012): Disorganization/cognitive and negative symptom dimensions in the at-risk mental state predict subsequent transition to psychosis. Schizophr Bull. 10.1093/schbul/sbq088

40. Cornblatt BA, Carrión RE, Auther A, McLaughlin D, Olsen RH, John M, Correll CU (2015): Psychosis prevention: A modified clinical high risk perspective from the recognition and prevention (RAP) Program. Am J Psychiatry. 10.1176/appi.ajp.2015.13121686
41. DeVylder JE, Muchomba FM, Gill KE, Ben-David S, Walder DJ, Malaspina D, Corcoran CM (2014): Symptom trajectories and psychosis onset in a clinical high-risk cohort: The relevance of subthreshold thought disorder. Schizophr Res 159 10.1016/j.schres.2014.08.008
42. Bearden CE, Wu KN, Caplan R, Cannon TD (2011): Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. J Am Acad Child Adolesc Psychiatry. 10.1016/j.jaac.2011.03.021
43. McNamara DS, Graesser AC, McCarthy PM, Cai Z (2012): Automated evaluation of text and discourse with Coh-Metrix. Automated Evaluation of Text and Discourse with Coh-Metrix 10.1017/CBO9780511894664
44. Gupta T, Hespos SJ, Horton WS, Mittal VA (2018): Automated analysis of written narratives reveals abnormalities in referential cohesion in youth at ultra high risk for psychosis. Schizophr Res. 10.1016/j.schres.2017.04.025
45. Mackinley M, Chan J, Ke H, Dempster K, Palaniyappan L (2020): Linguistic determinants of formal thought disorder in first episode psychosis. Early Interv Psychiatry 10.1111/eip.12948
46. Chen Q, Ling Z, Jiang H, Zhu X, Wei S, Inkpen D (2017): Enhanced LSTM for natural language inference. ACL 2017 – 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 10.18653/v1/P17-1152
47. Ben-David S, Birnbaum ML, Eilenberg ME, DeVylder JE, Gill KE, Schienle J, et al. (2014): The subjective experience of youths at clinically high risk of psychosis: A qualitative study. Psychiatr Serv 65 10.1176/appi.ps.201300527
48. Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015): The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin 10.15781/T29G6Z
49. Bonfils KA, Luther L, Firmin RL, Lysaker PH, Minor KS, Salyers MP (2016): Language and hope in schizophrenia-spectrum disorders. Psychiatry Res 10.1016/j.psychres.2016.08.013
50. Minor KS, Willits JA, Marggraf MP, Jones MN, Lysaker PH (2019): Measuring disorganized speech in schizophrenia: Automated analysis explains variance in cognitive deficits beyond clinician-rated scales. Psychol Med. 10.1017/S0033291718001046
51. Rezaii N, Walker E, Wolff P (2019): A machine learning approach to predicting psychosis using semantic density and latent content analysis. npj Schizophr. 10.1038/s41537-019-0077-9
52. Gutiérrez ED, Corlett PR, Corcoran CM, Cecchi GA (2017): Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings 10.18653/v1/d17-1316
53. de Boer JN, Voppel AE, Begemann MJH, Schnack HG, Wijnen F, Sommer IEC (2018): Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. Neuroscience and Biobehavioral Reviews 10.1016/j.neubiorev.2018.06.008
54. Losh M, Gordon PC (2014): Quantifying Narrative Ability in Autism Spectrum Disorder: A Computational Linguistic Analysis of Narrative Coherence. J Autism Dev Disord. 10.1007/s10803-014-2158-y
55. Lee M, Martin GE, Hogan A, Hano D, Gordon PC, Losh M (2018): What's the story? A computational analysis of narrative competence in autism. Autism. 10.1177/1362361316677957
56. Luo SX, Shinall JA, Peterson BS, Gerber AJ (2016): Semantic mapping reveals distinct patterns in descriptions of social relations in adults with autism spectrum disorder. Autism Res 10.1002/aur.1581
57. García AM, Carrillo F, Orozco-Arroyave JR, Trujillo N, Vargas Bonilla JF, Fittipaldi S, et al. (2016): How language flows when movements don't: An automated analysis of spontaneous discourse in Parkinson's disease. Brain Lang 10.1016/j.bandl.2016.07.008
58. Lyons M, Aksayli ND, Brewer G (2018): Mental distress and language use: Linguistic analysis of discussion forum posts. Comput Human Behav. 10.1016/j.chb.2018.05.035

59. Fineberg SK, Leavitt J, Deutsch-Link S, Dealy S, Landry CD, Pirruccio K, et al. (2016): Self-reference in psychosis and depression: A language marker of illness. Psychol Med. 10.1017/S0033291716001215
60. Birnbaum ML, Ernala SK, Rizvi AF, Arenare E, Van Meter A R, De Choudhury M, Kane JM (2019): Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. npj Schizophr. 10.1038/s41537-019-0085-9
61. Coppersmith G, Leary R, Crutchley P, Fine A (2018): Natural Language Processing of Social Media as Screening for Suicide Risk. Biomed Inform Insights 10.1177/1178222618792860
62. Agurto C, Pataranutaporn P, Eyigoz EK, Stolovitzky G, Cecchi G (2018): Predictive Linguistic Markers of Suicidality in Poets. Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018 10.1109/ICSC.2018.00051
63. Barker S (2017): Subject to pain: Ricoeur, Foucault, and emplotting discourses in an illness narrative. Subjectivity. 10.1057/s41286-017-0035-9
64. Bedi G, Cecchi GA, Slezak DF, Carrillo F, Sigman M, De Wit H (2014): A window into the intoxicated mind? Speech as an index of psychoactive drug effects. Neuropsychopharmacology. 10.1038/npp.2014.80
65. Agurto C, Cecchi GA, Norel R, Ostrand R, Kirkpatrick M, Baggott MJ, et al. (2020): Detection of acute 3,4-methylenedioxymethamphetamine (MDMA) effects across protocols using automated natural language processing. Neuropsychopharmacology 10.1038/s41386-020-0620-4
66. Tagamets MA, Cortes CR, Griego JA, Elvevåg B (2014): Neural correlates of the relationship between discourse coherence and sensory monitoring in schizophrenia. Cortex 10.1016/j.cortex.2013.06.011
67. de Boer JN; van Hoogdalem M, Mandl RCW; Brummelman J; Voppel AE, Begemann MJH van DEWFSI (2020): Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. npj Schizophr 6.
68. Silbert LJ, Honey CJ, Simony E, Poeppel D, Hasson U (2014): Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc Natl Acad Sci U S A. 10.1073/pnas.1323812111
69. Kuperberg GR, Delaney-Busch N, Fanucci K, Blackford T (2018): Priming production: Neural evidence for enhanced automatic semantic activity preceding language production in schizophrenia. NeuroImage Clin. 10.1016/j.nicl.2017.12.026
70. Barch DM, Berenbaum H (1997): The effect of language production manipulations on negative thought disorder and discourse coherence disturbances in schizophrenia. Psychiatry Res. 10.1016/S0165-1781(97)00045-0
71. Holmlund TB, Chandler C, Foltz PW, Cohen AS, Cheng J, Bernstein JC, et al. (2020): Applying speech technologies to assess verbal memory in patients with serious mental illness. npj Digit Med. 10.1038/s41746-020-0241-7
72. Cohen AS, Mitchell KR, Elvevåg B (2014): What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments. Schizophrenia Research. 10.1016/j.schres.2014.09.013
73. Stanislawski E, Bilgrami Z, Sarac C, Cecchi G, Corcoran C (2019): S19. ANALYZING NEGATIVE SYMPTOMS AND LANGUAGE IN YOUTHS AT RISK FOR PSYCHOSIS USING AUTOMATED LANGUAGE ANALYSIS. Schizophr Bull. 10.1093/schbul/sbz020.564
74. Baker JT, Pennant L, Baltrušaitis T, Vijay S, Liebson ES, Ongur D, Morency L-P (2016): Toward Expert Systems in Mental Health Assessment: A Computational Approach to the Face and Voice in Dyadic Patient-Doctor Interactions. Iproceedings. 10.2196/iproc.6136
75. Datta SR (2019): Q&A: Understanding the composition of behavior. BMC Biology. 10.1186/s12915-019-0663-3
76. Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011): Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci. 10.1523/JNEUROSCI.3684-10.2011
77. Datta SR, Anderson DJ, Branson K, Perona P, Leifer A (2019): Computational Neuroethology: A Call to Action. Neuron 10.1016/j.neuron.2019.09.038

78. Sichlinger L, Cibelli E, Goldrick M, Mittal VA (2019): Clinical correlates of aberrant conversational turn-taking in youth at clinical high-risk for psychosis. Schizophrenia Research 10.1016/j.schres.2018.08.009

79. Habermas J (2001): Remarques sur vérité et justification. Raisons Polit. 10.3917/rai.002.0217

80. Rovelli C (1996): Relational quantum mechanics. Int J Theor Phys. 10.1007/BF02302261

81. Wittgenstein L, Diamond C (2019): Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge, 1939. Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge, 1939 10.7208/chicago/9780226308609.001.0001

82. Turing AM (2009): Computing machinery and intelligence. Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer 10.1007/978-1-4020-6710-5_3

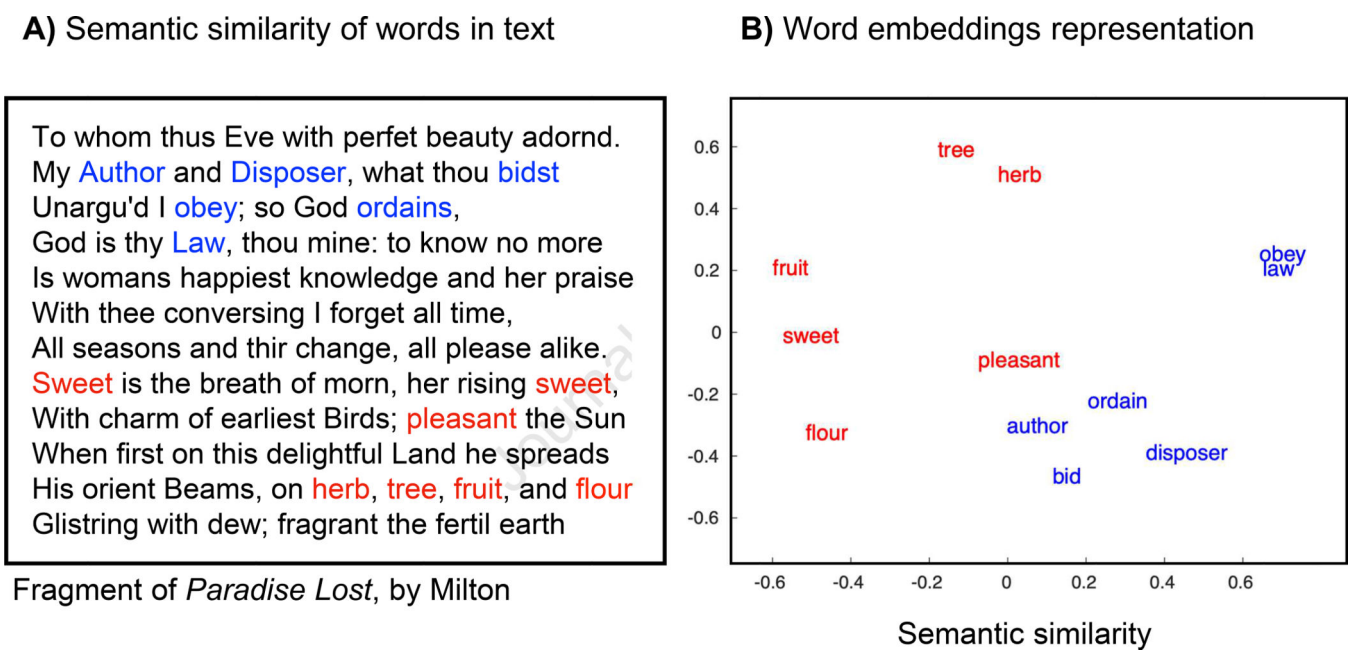

Figure 1.

Semantic similarity of words in text and word embeddings representation

A passage from Paradise Lost analyzed with LSA to demonstrate how content and proximity are related. Left panel: words in blue and red were selected to highlight ideas of order and pleasure, respectively; moreover, they appear contiguous in the text. Right panel: 2D projection of LSA vectors showing how the words cluster according to their meaning.

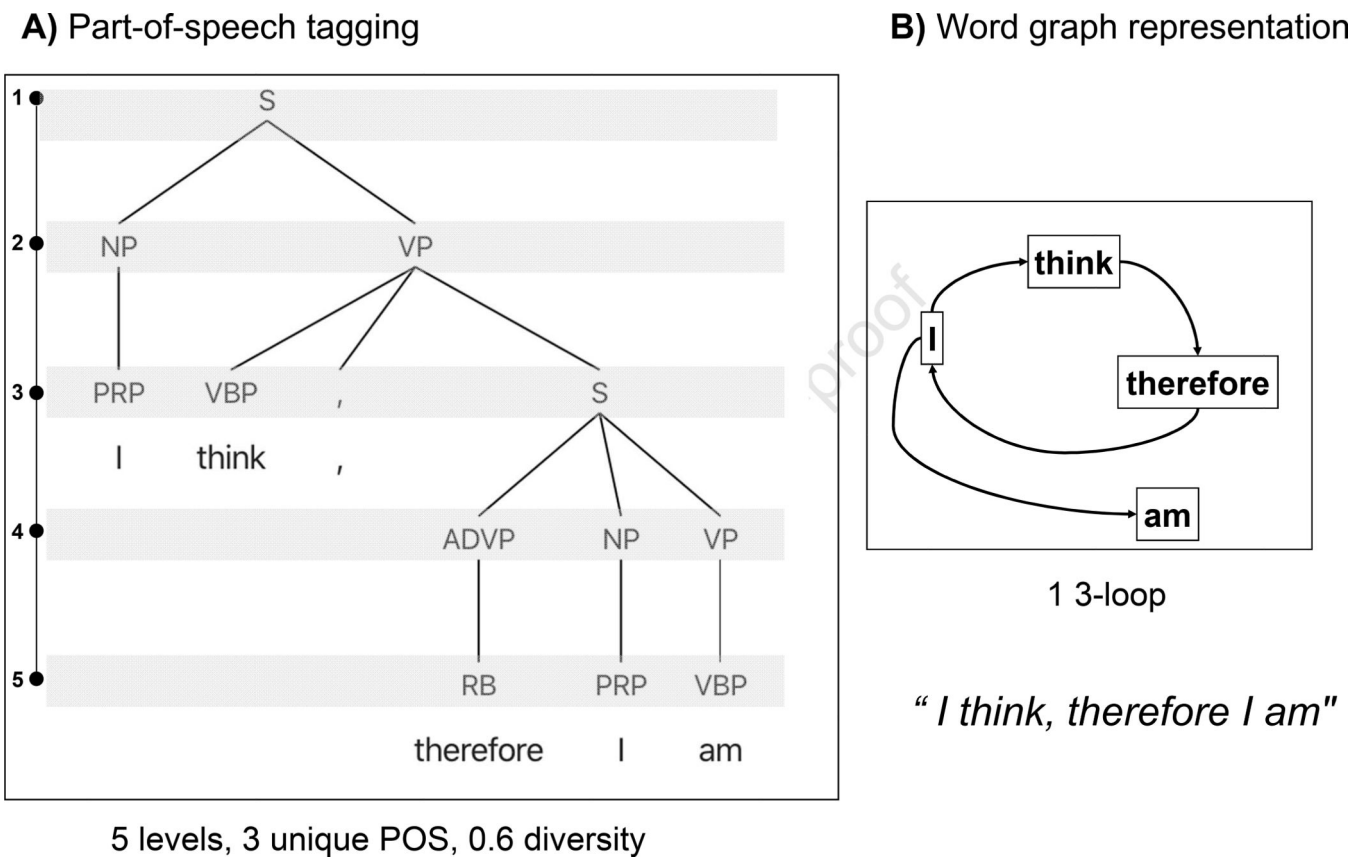

Figure 2.

A) Part-of-speech tagging **B)** Word graph representation

Figure 2A: part-of-speech (POS) tagging. The decomposition of the sentence "I think, therefore I am" into POS also generates a phrase tree structure (Noun Phrase, Verb Phrase, etc.) whose depth and diversity (POS/length) can be used as measures of complexity.

Figure 2B: Graph representation. The directed graph identifies a recurrence to the word "I".

Table 1

Highlights of language studies in psychotic disorders and their risk states

Topic	Authors
Conceptualization of positive and negative thought disorder in schizophrenia and mania.	Andreasen and Grove, 1986(19)
Disruption of basic discourse structure in schizophrenia but abrupt shifts of intact discourse in mania	Hoffman et al., 1986(22)
Increased errors in meaning in production of complex sentences in schizophrenia patients	Hoffman et al., 1985(23)
Modeling of decreased coherence of speech in schizophrenia by decreasing connectivity in neural network simulations of parallel distributed processing systems	Hoffman et al., 1997(24)
Abnormal predictive coding provides best fit for breakdown of narrative coherence in computational patient with schizophrenia	Hoffman et al., 2011(26)
Use of Latent Semantic Analysis (LSA) to identify decreased coherence of speech in schizophrenia, as compared with the norm and with unaffected siblings	Elvevåg et al., 2007(20); Elvevåg et al., 2010(27)
Use of speech graphs to distinguish speech in schizophrenia and mania from the norm, to characterize negative symptoms, to predict schizophrenia diagnosis, and to track early deviations in a developmental trajectory	Mota et al., 2012(16); Mota et al., 2017(33); Mota et al., 2018(34)
Use of manual linguistic analysis to prediction psychosis among clinical high risk (CHR) patients, identifying illogical thought, poverty of content and errors in referential cohesion,	Bearden et al., 2011(42)
Use of LSA and part-of-speech tagging to identify measures of coherence and complexity that predict psychosis in CHR	Bedi et al., 2015(14); Corcoran et al., 2018(15)
Use of latent semantic content analysis and semantic density to predict psychosis in CHR	Rezaii et al. 2019(51)
Use of Coh-Metrix to identify abnormal referential cohesion in CHR related to verbal learning deficits and symptom severity	Gupta et al., 2018(44)