Schizophrenia Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Towards a temporospatial framework for measurements of disorganization in speech using semantic vectors

Terje B. Holmlund [a,*], Chelsea Chandler [b], Peter W. Foltz [b], Catherine Diaz-Asper [c], Alex S. Cohen [d,e], Zachary Rodriguez [d,e], Brita Elvevåg [a,f]

[a] Department of Clinical Medicine, University of Tromsø - the Arctic University of Norway, Tromsø, Norway
[b] Institute of Cognitive Science, University of Colorado Boulder, United States of America
[c] Department of Psychology, Marymount University, United States of America
[d] Department of Psychology, Louisiana State University, United States of America
[e] Center for Computation and Technology, Louisiana State University, United States of America
[f] Norwegian Center for eHealth Research, University Hospital of North Norway, Tromsø, Norway

ARTICLE INFO

Keywords:
Language analysis
Speech
Coherence
Schizophrenia
Word vectors
Visualizations

ABSTRACT

Incoherent speech in schizophrenia has long been described as the mind making "leaps" of large distances between thoughts and ideas. Such a view seems intuitive, and for almost two decades, attempts to operationalize these conceptual "leaps" in spoken word meanings have used language-based embedding spaces. An embedding space represents meaning of words as numerical vectors where a greater proximity between word vectors represents more shared meaning. However, there are limitations with word vector-based operationalizations of coherence which can limit their appeal and utility in clinical practice. First, the use of esoteric word embeddings can be conceptually hard to grasp, and this is complicated by several different operationalizations of incoherent speech. This problem can be overcome by a better visualization of methods. Second, temporal information from the act of speaking has been largely neglected since models have been built using written text, yet speech is spoken in real time. This issue can be resolved by leveraging time stamped transcripts of speech. Third, contextual information - namely the situation of where something is spoken - has often only been inferred and never explicitly modeled. Addressing this situational issue opens up new possibilities for models with increased temporal resolution and contextual relevance. In this paper, direct visualizations of semantic distances are used to enable the inspection of examples of incoherent speech. Some common operationalizations of incoherence are illustrated, and suggestions are made for how temporal and spatial contextual information can be integrated in future implementations of measures of incoherence.

1. Introduction

> "'*The train of thought' can initially remain ordered, but later in many cases displays leaps, becomes disjointed, sometimes reaches complete incoherence*".
>
> (Kraepelin, 1921, p59)

'Coherence' - and thus incoherence - is a multidimensional concept of discourse. Coherence can be considered at multiple levels of linguistic complexity (e.g., semantic, syntactic) and operationalizations of the concept makes assumptions about prior contexts. The methodology by which speech coherence is assessed has important clinical and research implications, and indeed the Diagnostic and Statistical Manual of Mental Disorders notes that the key symptom of disorganized thinking is "inferred from the individual's speech" (American Psychiatric Association, 2022). However, this evaluation is a complex process. The framework and terminology a field employs to capture the phenomenon of interest are often a legacy of the metaphors that are prevalent in society at the time (Lakoff and Johnson, 1980), and this is indeed the case with (modern) psychiatry which had its origins at around the birth of the steam train and method of communication by telegraph. The lingering clues of this remain to date in our use of terms such as *derailment* and *pressured* or *telegraphic* speech. Not surprisingly then, the conceptualization by Kraepelin a century ago of incoherence in speech can appear to us today more as a figurative literary description than as an objective

* Corresponding author.
E-mail address: terje.holmlund@uit.no (T.B. Holmlund).

https://doi.org/10.1016/j.schres.2022.09.020
Received 31 March 2022; Received in revised form 5 September 2022; Accepted 6 September 2022

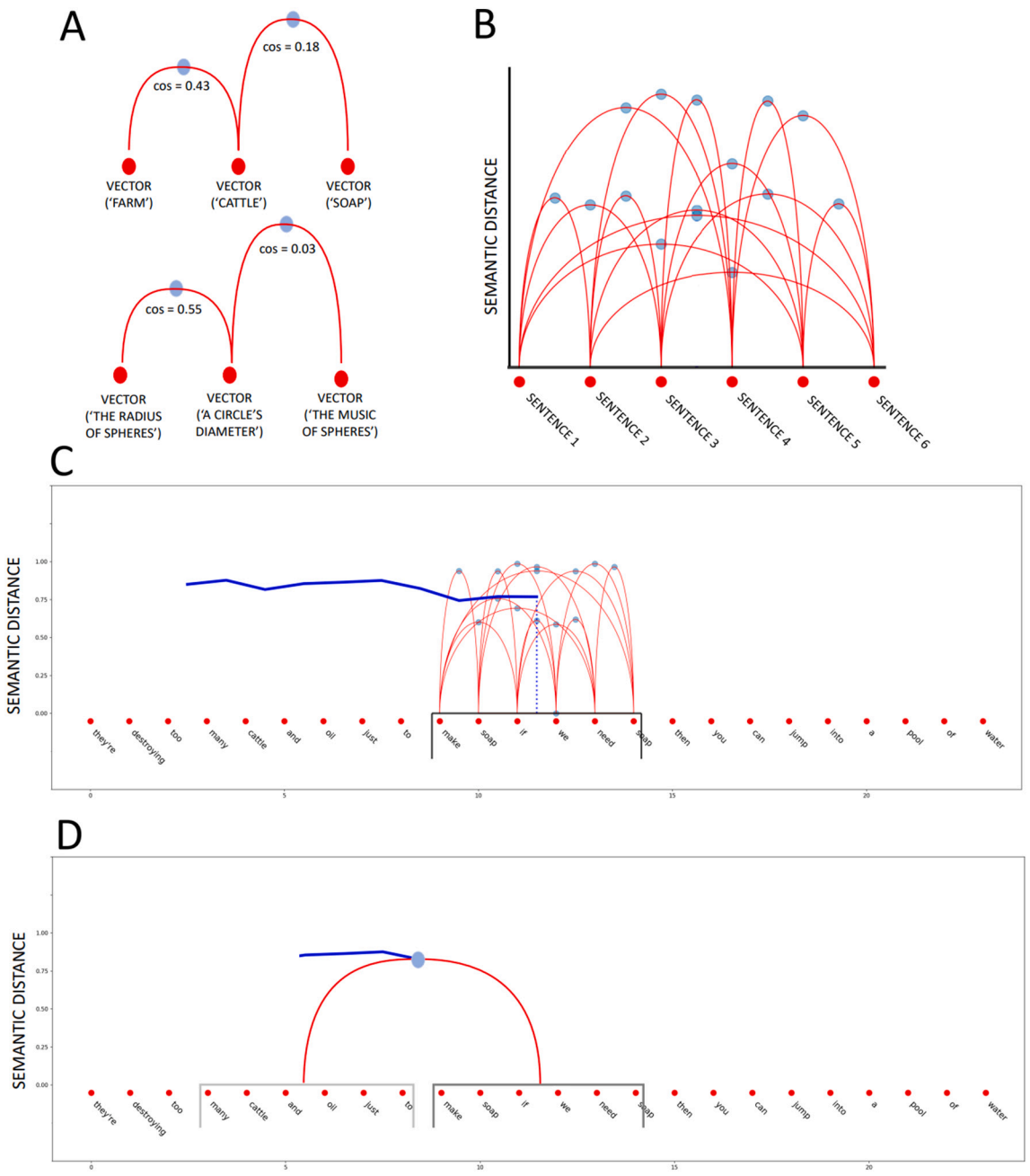

Fig. 1. Semantic distances can be used to quantify the conceptual connectedness in utterances. Panel A: At the smallest scale, the similarity, or relatedness, between single words can be quantified as the distance between them in a semantic space. By using a publicly available word2vec language model building on about 100 billion words from the Google News dataset, it is possible to show that the words "farm" and "cattle" have a (cosine) distance of 0.43 (red arc with blue dot), while the distance between "cattle" and "soap" is larger at 0.18. Conceptual relatedness can also be compared between sequences of words or sentences. A paraphrase of the sentence "Several doctors operated on the patient" has a low cosine distance to the original sentence. Two similar but not identical sentences are compared to the sentence "A circle's diameter", and it is possible to see how the conceptual distance, the leap, is bigger when a word like "music" is introduced. Panel B: When speech from patients is transcribed with punctuation, it is possible to measure the semantic distances between sentences. In this plot, each red dot indicates a vector representation of a sentence. If there are six sentences in a response from a patient, there are many different combinations of sentences to be compared for semantic distance, illustrated by the red arcs. The many comparisons can be summed up as average distance, or coherence, or other metrics such as standard deviation, minimum and maximum values. Panel C: A famous example of incoherent speech is one from the classic Thought, Language and Communication Scale of Nancy Andreasen (1986), here illustrated as a sequence with each word vector plotted as a red dot. Another common procedure is to examine semantic distances between word vectors within a moving "frame" or "window" of a given size (like six words in this example) and sum up the overall distance within the frame with the average distance (blue line). The frame is moved in sequential steps (see Supplementary material - Fig. S1 - for an animated version of this process), and values for each step can be summed in various ways to derive a score of overall connectedness. Incoherence, or conceptual "leaps", can therefore be expected to demonstrate high arcs representing low semantic similarity, but there may also be fluctuations, such as the lower mean of the illustrated frame where the word "soap" is repeated (cosine distance = 0). This type of single moving window procedure can examine incoherence on a short timescale, the connectedness of words uttered within a few seconds of each other. Panel D: The same incoherent utterance can be examined using a dual-window procedure, where the semantic distance between two sequential blocks of vectors is measured (in this case, window size = 6). Depending on the size of the window, this method can be used to examine conceptual "leaps" over a slightly longer timescale, from just a few seconds to as long as possible, namely comparing the first and second half of an utterance (if the window size is half the length of the utterance). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

quantification of a symptom. It is the premise of this paper that to reliably operationalize coherence of speech it will be crucial to move beyond the use of metaphors to the point where we have tools that are comparable and replicable (see Holmlund et al., 2021). The promise of this approach is replicability of measurements that are unambiguous of these putative *in vivo* complex thought processes, and generalizability of measures to diverse populations and communities.

The "leaps" in thoughts described by Kraepelin reveals an intuition about how language and the underlying cognitive processes can be conceptualized within a framework of time and space. A premise here is that thoughts are represented by the meaning of words, expressed via the medium of speech or writing at a specific time and generated within a particular physical location. There is a long tradition of conceptualizing the *distance* between thoughts from notions of a "*cognitive map*" (e.g., Tolman, 1948) as well as a "*psychological space*" (Shepard, 1987), which may well have a neurobiological basis (see e.g., Bellmund et al., 2018; Viganò and Piazza, 2020). If one accepts the premise that words represent the thoughts of the speaker, then the intuition of a "leap" in a psychological space becomes quantifiable as distances in these semantic spaces. Indeed, time and space can be quantified for scientific purposes with seconds and meters, but what about the meaning of a word? The language philosopher Ludwig Wittgenstein famously stated that "the meaning of a word is its use in the language" (Wittgenstein, 1953, section 43), hinting that systematic observation of communicative behavior provides a framework to examine meaning. Around the same time, the linguist John Rupert Firth stated that "you shall know a word by the company it keeps" (Firth, 1957). Together, these two position statements provide a philosophical foundation to the scientific study of meaning: by examining how words co-occur in language it is possible to quantify aspects of meaning, or semantics (and therefore the content of thoughts). Put differently, words that tend to occur in similar contexts are semantically related and thus should be *close* to one another in a derived word vector space. This has become known as the distributional hypothesis of semantics and leverages the distributions of words across large amounts of text (millions of written words) to derive semantic vector spaces. Distances in semantic vector spaces provide a useful analog to coherence in discourse (Foltz, 2007) in that leaps from one part of the space to another can provide an indication of how much change there is in the overall semantic content from one part of the discourse to the next. There are of course alternatives to the philosophical tradition of Wittgenstein (e.g., Sellars, 1963), and may be compelling arguments to be made that thought and language are separate systems (e.g., see Jackendoff, 1996). Nonetheless, viewing incoherence of speech through the lens of the above-mentioned philosophical framework has led to ideas and testable hypotheses on how word co-occurrence statistics can be different in speech from patients with schizophrenia, compared to healthy individuals.

Traditionally, such co-occurrence statistics have been derived from text corpora, and by employing mathematical techniques it is possible to obtain numerical (vector) representations of words, where the meanings of words are expressed as locations in high-dimensional "semantic spaces". Many mathematical techniques for the creation of word representations exist, and these options are increasing rapidly with the fast paced tempo of progress in the field of natural language processing. The first methods were presented in the late 1990s, notably with the "Hyperspace Analog to Language" (HAL; Burgess et al., 1998) and Latent Semantic Analysis (LSA; Foltz, 1996; Landauer and Dumais, 1997). More recent methods use deep neural networks and include word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), Embeddings from Language Models (ELMo; Peters et al., 2018), and Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019). LSA performs a singular value decomposition on the word type by document matrix to obtain lower dimensional vectors of each of the types. Word2vec is a neural network-based word embedding model trained on a large corpus of text to predict either a word given its context (continuous bag of words; CBOW) or the context surrounding a given word (skip gram). ELMo and BERT are deep neural language models that are built on Long Short Term Memory (LSTM) and transformer architectures, respectively. Generally, coherence metrics are computed as the cosine distance between consecutive vector representations of words, windows of words, phrases, or sentences. From a practical perspective, the goal is to have something that is both useful (e.g., provides predictions that are informative to a clinician or researcher) and explainable (e.g., that the operationalizations use constructs tied to underlying neurocognitive functions (see Foltz et al., 2022)).

There is no 'one size fits all' approach to choosing the right operationalization of disorganization in speech. Some operationalizations are suited for short timescales, some for long, and others yet for cases where connectedness to contextual cues are needed. What these aforementioned methods have in common is that words that share related meanings are 'close' to one another in these spaces, such as "cattle" and "farm", whereas the words "cattle" and "soap" will have longer distances between them. These words come from the now classic example of incoherence in Andreasen's (1986; p.477) *Thought, Language and Communication Scale* where when asked what they thought about the current political issues such as the energy crisis the patient is quoted as responding: "They're destroying too many cattle and oil just to make soap. If we need soap when you can jump into a pool of water" [quote abbreviated by us]. A language model (LSA) illustrates the distances between these three words ("farm", "cattle" and "soap") in Fig. 1 Panel A. To borrow from Kraepelin's quote cited in our introduction, the "leap" between some words or concepts are quantifiably larger, here illustrated by the cascades of such arcs, with the higher values indicating less coherent speech. It can also be useful to compute the conceptual relatedness between sequences of words or sentences as illustrated in Fig. 1 Panel B. These semantic spaces therefore provide a domain where the "leaps" of words and thoughts can be variously quantified for scientific investigations, enabling us to compute distance measurements in time, physical space and semantic space. All of these three domains *must* be properly operationalized and understood for a scientific analysis framework of language disorganization to have true clinical translation value.

In this paper, three critical problems - and potential solutions - are discussed: First, the variability of operationalizations of the concept of coherence and of word meanings can create a confusing landscape of methods. Put differently, it is not easy to know how different measurements of distance in the semantic domain translate to the concept of "incoherence". Indeed, the field risks a "black-box" situation, where methods seem to work as intended (e.g., for classification of research participants as patients or non-patients), but that may be for unintended or frankly wrong reasons. Luckily, the extension of our measurements into "conceptual space" lends itself to precise visualizations. The "leaps" are measured as distances, and here we present illustrations of how these distances are captured in current computational methods (see Fig. 1). Second, current computational tools are based on methods from natural language processing of text, and lack crucial temporal information about when words are spoken. Unlike language in text format, spoken language varies in its temporal delivery. This temporal pattern contains critical information about how words are connected. Hence, language models based on text may miss this critical information. Third, if information about the situation in which speech behavior happens is missing, crucial contextual information will be lost. By accounting for the context where speech is happening in physical space it is possible to create more nuanced models of whether *what* is said is coherent or not, similar to how clinicians intuitively account for contextual factors. We initiate this account by discussing the most basic notion of contextual information, namely spatial location, but ultimately contextual clues such as speaker demographics, speaker motivation or purpose, and previous conversational topics can and should be utilized. Addressing these three problems regarding measurements of incoherence and disorganization in speech has the objective of carving a path towards more robust and universal methods of operationalizing incoherence. We

will conclude that (1) interpretability of methods can be improved by explicit visualizations, (2) timestamping uttered words enables new and more nuanced information regarding the temporal aspect of coherence and (3) contextually anchored language models that incorporate situational information will allow more fine-grained information about whether or not speech is coherent within the limits of the local context (e.g., speech at a family gathering versus in an academic lecture hall).

2. Problem 1: incoherence in speech is conceptualized and computationally quantified in many different ways

Since the word "incoherence" can mean so many different things, this creates a problem for the specificity of our analyses. For example, to a clinician interviewing a patient with schizophrenia – informed by Andreasen's definition in her *Thought, Language and Communication Scale* (Andreasen, 1986; p.477 - 'A pattern of speech which is essentially incomprehensible at times') – incoherence refers to the comprehensibility of what is spoken. To a computational linguist examining incoherence at a discourse level, the term lexical cohesion (i.e., the opposite phenomenon) is often used, traditionally taken to mean that there is a sharing of semantically related or identical words in neighboring sentences (Halliday and Hasan, 1976) as well as syntactic markers indicating causal connections. Coherence, thus defined, is then examined with lexical and syntactic constraints, logical relations between concepts and events, and overall agreement with "world knowledge". To a neuroscientist, coherence in the brain can mean several different things depending upon whether their focus is cellular, circuitry or systems. Further, *why* coherence emerges and *what* causes it is a function of a variety of brain systems (e.g., Dapretto et al., 2005). Coherence measured by discourse using word embedding spaces - as is the focus of this paper - can also mean a variety of things.

Previously we suggested four approaches to compute semantic coherence using word embedding methods, namely using the semantic distance between one word and another; using the distances between larger units of language within a discourse; estimating how a person's answer relates to a question asked; and estimating how answers relate to another person's answer on the same question (Elvevåg et al., 2007). These approaches remain relevant today and can - with some generalizations - serve as the overarching categories within which the plethora of possible methods could fall. The first approach, word-to-word similarity (Fig. 1, panel A), has been used in several different ways to quantify connectedness between adjacent word responses in verbal semantic fluency tests (e.g., Holmlund et al., 2019; Kim et al., 2019; Pauselli et al., 2018). A notable variant of single word-to-word coherence measurements was used by Corcoran et al. (2018), where the semantic distances between words with inter-word distances of 5 to 8 were used to predict psychosis onset in clinical high-risk youths. The second general approach involves examining distances within and between larger units of speech, such as sentence-to-sentence similarity, phrase-to-phrase similarity or variants of using "windows" of text of various lengths (e.g., 6 words) with single- (Fig. 1, panel C) or dual-window variants (Fig. 1, panel D). Measuring coherence in longer units of connected discourse such as story recalls, free speech, and answering process questions has also uncovered differences between patients with schizophrenia and healthy volunteers (e.g., Bedi et al., 2015; Tang et al., 2021). The third general approach (combining the last two approaches mentioned earlier) involves comparing the semantic content of speech to some outside contextual information, such as the content of a preceding question (in discourse) or common speech in the same situation (e.g., other answers to the same question). This approach measures the topical coherence as to whether a response is related to a posed topic, as well as how much the discourse may deviate tangentially from the topic.

In addition to these main approaches, new variations of methods have been developed. The type of word vector spaces used have been updated over time, with the use of word2vec, GloVe (Iter et al., 2018), ELMo (Sarzynska-Wawer et al., 2021) and BERT (Tang et al., 2021). Also, the methods for computing semantic distance have seen innovation with explainability investigations into optimal moving-window sizes (Voppel et al., 2021), and the use of vector centroids (Xu et al., 2021) (for further studies see e.g., Iter et al., 2018; Just et al., 2019, 2020).

In essence, while varied, all approaches assess aspects of coherence. However, to operationalize a measure, it is necessary to understand the link between the output of the computational method with the neuropsychological phenomena being investigated (e.g., Foltz et al., 2022). A critical tool for clarifying the definition of incoherence being assessed is to make current verbal definitions visual. By improved visual representations of the methodology, it is possible to understand how coherence is computed, and therefore provide the user with a guide to decide if the specific way of operationalization is what they intended (i.e., if one is really measuring what one conceptualizes as incoherence). There are other studies that have illustrations to explain the methods beyond using just words and numbers (e.g., Hoffman et al., 2018). However, they often present the abstract principles behind the methods without providing illustrations of practical examples with real analyzed data (but notable exceptions include single-word analyses in verbal fluency studies, see e.g., Kim et al., 2019). The field will certainly benefit from coherence visualization software that can reliably and effectively demonstrate the resulting metrics "in situ" on transcripts or recordings. Such software efforts can generate a coherence plot for each and every datapoint in a study, ensuring complete transparency of the methods. Such transparency will nurture trust from clinicians provided the metrics are valid representations of incoherence. In a similar manner to the way a radiograph of a traumatized bone can reveal a fracture, a true "*coherograph*" can demonstrate where coherence breaks down in an utterance or a discourse. Whether or not it is possible to reach such a level of specificity and sensitivity remains to be seen. The methods have proven effective at enabling detection of group differences (i.e., patients versus nonpatients) in speech coherence, but it is possible that pinpointing incoherent parts of speech will be much more challenging.

In summary, proper visualizations can enable researchers and clinicians to understand where the source of (in)coherence is in a given segment of language. This understanding is critical in the choice of which computational approach to harness for a given language analysis as *the varied operationalizations measure different facets of potential incoherence*. In the future, visualizations can assist researchers and clinicians in understanding the methods they employ for specific segments of speech and aligning various implementations with the construct they are seeking to analyze. For clinicians, a dashboard-like output that pinpoints sections of incoherent speech, with metrics on how the patient's output relates to clinical reference materials can then aid in diagnosis and be a useful component of monitoring clinical states (see Fig. 28.2, p. 676, in Holmlund et al., 2020). The user interface of future clinical tools based on coherence metrics should therefore be created using established design principles and validated through interaction experiments (see; Rundo et al., 2020). Clearly, alignment of methods for coherence metrics would be useful, although it might limit innovation. While there is an increasing amount of consensus on the need to create standardized practices in this relatively new field, nonetheless, a simple agreement on methodological nomenclature and principles (e.g., word-level, sentence-level, dual/single windows, window sizes) is presently missing and an imperative first step towards a useful framework.

3. Problem 2: the temporal dimension - operationalizations of language coherence have been developed on transcriptions of speech and therefore are missing crucial temporal information

Wittgenstein (1953, section 108) noted broadly that "*The 'use' of words is extended in time*", and this extension in the temporal domain can provide important clues for how to improve computational metrics of incoherence. It has consequences for modeling speech, but importantly it means that transcripts of speech from a clinical setting are dissimilar

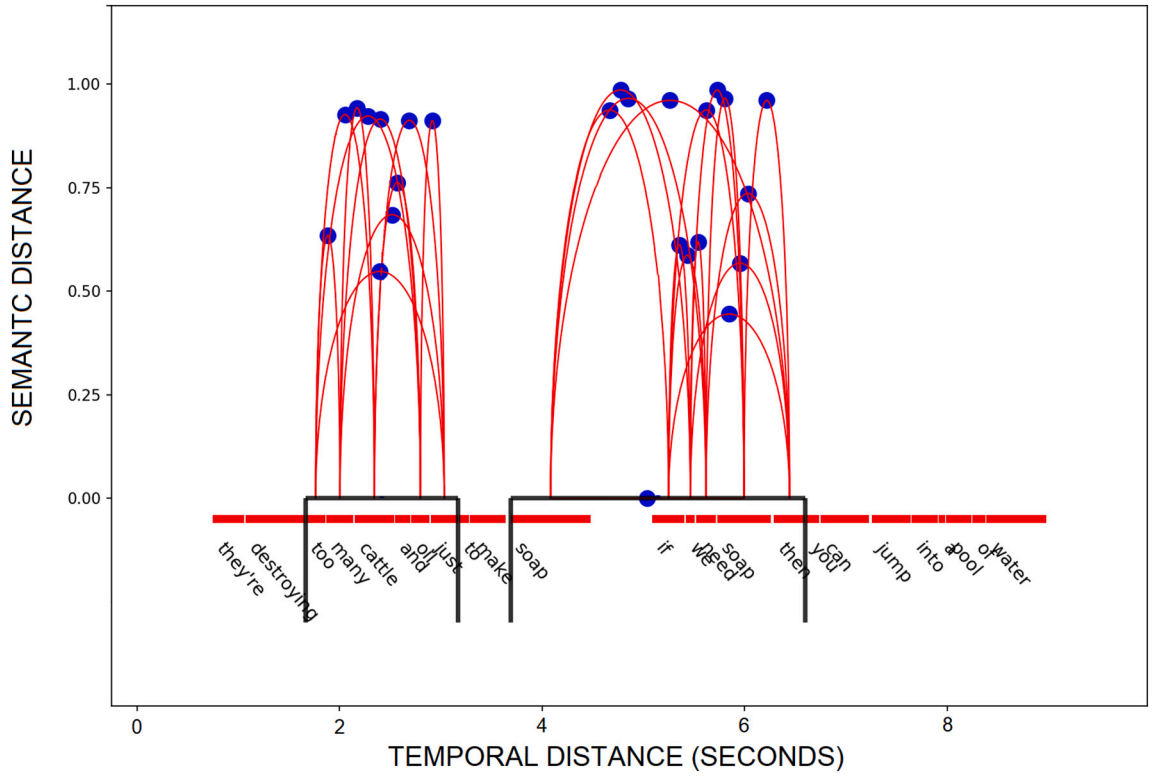

Fig. 2. The exact time that a word is uttered can be detected using automatic speech recognition tools, and when the same sentence is time stamped it is clear that words do not come equally spaced as illustrated in Fig. 1. This means that windows of a size of six words can have completely different temporal extents in a moving window procedure (see Supplementary material - Fig. S2 - for an animated version of this process). The difference in sizes is well illustrated in the moving-window procedure, where both windows are 6 words, but the first window spans 1.5 s and another window spans 3 s. This is a challenge if procedures are to be connected to putative underlying physiological processes where neural activity is integrated over certain timespans.

to the reference material (i.e., training data for language models) since speech is not an identical process to writing. Indeed, spontaneous speech is typically quite different from written language in many critical respects. From a conceptual level, writing is typically the result of some advanced planning; writers are afforded the opportunity to formulate and revise their thoughts in a structured and coherent manner. Additionally, sentences and paragraphs represent easily delimited units of thought in writing. In contrast, speech occurs in "real time" where the timing of utterances reflects an unfolding thought process. Units of thoughts may flow from one to the next without obvious delimiters. While it is unclear exactly how this impacts coherence metrics, it is notable that "sentences" as defined by utterance length in spontaneous face-to-face conversations have been found to be much shorter (median of 5 words) compared to sentences in news broadcasts and political debates (median 12 and 16 respectively; Wiggers and Rothkrantz, 2007). Moreover, attempts to translate spontaneous speech into text format are generally quite difficult, with challenges from an excess of filler sounds such as "uh" and utterances abruptly terminated before full sentences are formed. For example, text relies heavily on standardized punctuations for segmentation. In free speech, on the other hand, defining sentences can be difficult, and our own experience with dependency parsers (e.g., OpenIE; Angeli et al., 2015) is that they can fail in spectacular ways on real clinical language data. Models are known to have a hard time getting nested dependencies correct, particularly when they are long (Lakretz et al., 2021). The "gold standard" parsing, namely in human transcription, is not without challenges. This is because punctuation is a tool for increasing readability and "sentences" are, as such, a product of a subjective evaluation by the writer or transcriptionist. Indeed, with an informal browsing of transcription instructions one can get the impression that punctuation is often left up to the discretion of the transcriber's sense of style, although methods for automated punctuation do exist (e.g., Tilk and Alumäe, 2016).

Evaluating the temporal relationships between words can provide critical information about how thought processes are occurring. To illustrate this point with reference to Fig. 1, consider Panel A. Despite the high arcs illustrating less coherence between "cattle" and "soap" it is not clear when these words were uttered and how close they were to each other within the conversation. Fig. 2, on the other hand, features these two words visualized as part of its original spoken (in this case spoken by the first author) and recorded utterance. As a first step, the visualizations using timestamped words make explicit the temporal spans over which the semantic distance measurements (i.e., a 6-word window coherence metric) are computed. This elucidates the time-scale within which the comparison method is relevant (e.g., are the items compared 1 s apart, or 20 s?), and how words and semantic concepts are inter-related. Temporal information can allow for segmentation approaches that improve upon limitations by forcing punctuation solutions to spoken language. The timestamping of words further allows for new types of time series data, which can be analyzed with established signal processing methods (for a notable recent example of time-series analysis, see: Xu et al., 2022). An important limitation of traditional moving window techniques defined by *n* number of words is demonstrated in Fig. 2, namely that a window size of for example 6 words can span dramatically different distances in the temporal domain. Such issues can be detected, visualized and understood only by adding in temporal information. Future methods might consider using temporally defined windows (e.g., words within a 2 s window, a relevant time-scale to capture delta-band electrophysiological activity related to language understanding, see e.g., Lo et al., 2022), and thus ground the "use of words" within a specific temporal framework.

While most language models process language as a sequence of words or parts of words, some recent developments in modeling hold the

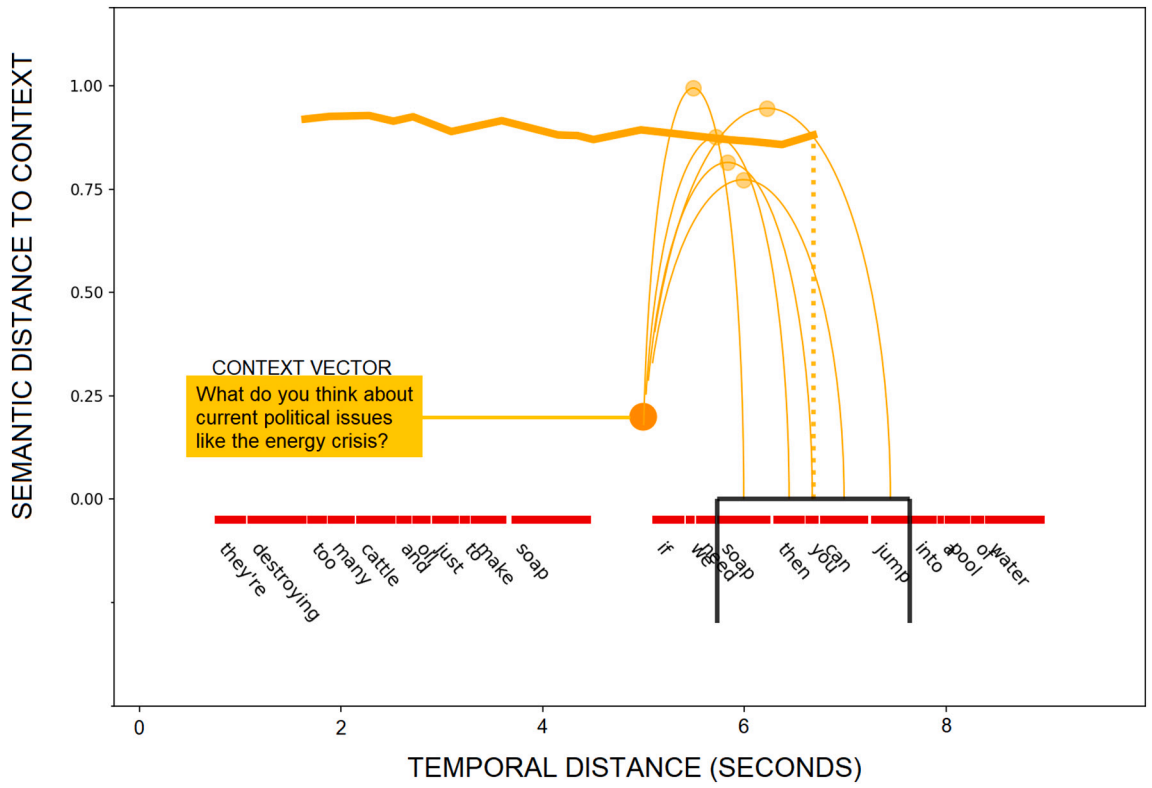

Fig. 3. Semantic distance can also be measured between an utterance and some external context. In this example, words within a window (size = 5) are examined for their distance to the question posed before the incoherent speech examined in Figs. 1 and 2. Here, instead of direct comparisons between individual words, each word's distance (orange arcs and line) to a vectorized "context" (orange box and ball), namely the question, is visualized. Higher arcs mean words are less connected to the conceptual content of the question, indicating incoherence in the form of tangentiality [Note: these two concepts are not fully overlapping and are separated by Andreasen, 1986]. Choosing a small window size increases the resolution with which one can assess tangentiality: A small window can, in theory, pinpoint the sections of speech that are not connected to the question. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

promise of true integration of a more full range of temporal aspects of speech. Based on the architecture of the previously mentioned BERT model (Devlin et al., 2019), the HuBERT model directly processes audio waveform information rather than lexical information (Hsu et al., 2021). Such methods may in the future be able to capture the unique patterns and signals found in speech and massively improve the way quantitative tools can "listen" to speech in clinical settings. The new models may also aid in improving our ability to define when incoherent speech occurs, with output that alerts to temporally defined sections of recordings (e.g., in a 3 s window) that are incoherent with the previous utterances in a conversation. Interestingly, the approach taken with HuBERT has also been expanded to include both audio and video data in the same model (Shi et al., 2022), demonstrating that these powerful approaches can integrate information about various aspects of human behavior through time. As with all data-driven models, it will be crucial to have suitable training material. While the current HuBERT model is based on audiobooks (Hsu et al., 2021), it is plausible that models trained on spontaneous speech recordings will have more relevance for detecting patterns of incoherent speech in clinical settings. This, of course, is a matter to be evaluated by future experimental designs.

4. Problem 3: situational or contextual information is necessary to improve the sensitivity of coherence measurements in clinical settings

In conversation, a clinician has a clear sense of where the conversation is taking place (e.g., in a hospital ward versus an encounter in the street) and what the situation is (e.g., a serious admission interview versus a casual chat about the weather), and based on this information is able to form clear expectations of the language that will be produced in the current context. Recalling the philosophical standpoint that is foundational to the methods for quantification of word meaning, namely that "the meaning of a word is its use in the language" (Wittgenstein, 1953, section 43), there are important implications that contextual information has with respect to language usage and coherence. In short, people use words differently in different physical and sociocultural contexts. Indeed, by using a range of different coherence measures to examine cohorts of schizophrenia patients from three different countries, Parola et al. (2022) found generalizability to be limited across the languages, samples and measures. Obviously this has consequences for how language models should be built for clinical purposes, and for how those models should be utilized for increased sensitivity to signs of pathology in the speech of patients. If clinicians are to rely on and trust computational measurements of speech incoherence, then the ability to account for context will be crucial (see Fig. 3).

Currently, language models used for analyzing speech from psychiatric patients in the published literature are built on text from various sources, which may or may not represent the diversity of how language is used in varied situations, and this is a problem for the generalizability to clinical applications. The semantic space that is generated in a popular "off-the-shelf" pre-trained implementation of word2vec (available here: https://code.google.com/archive/p/word2vec/) is based on the "Google News dataset", which can serve as a powerful example. Words from news articles are typically used to convey information about issues of public interest, often with a broad geographic coverage, such as international politics. In this context, the word "oil" would most commonly be used in proximity to words such as "economy" and "pipeline". In contrast, in a psychiatric clinical setting, the words used

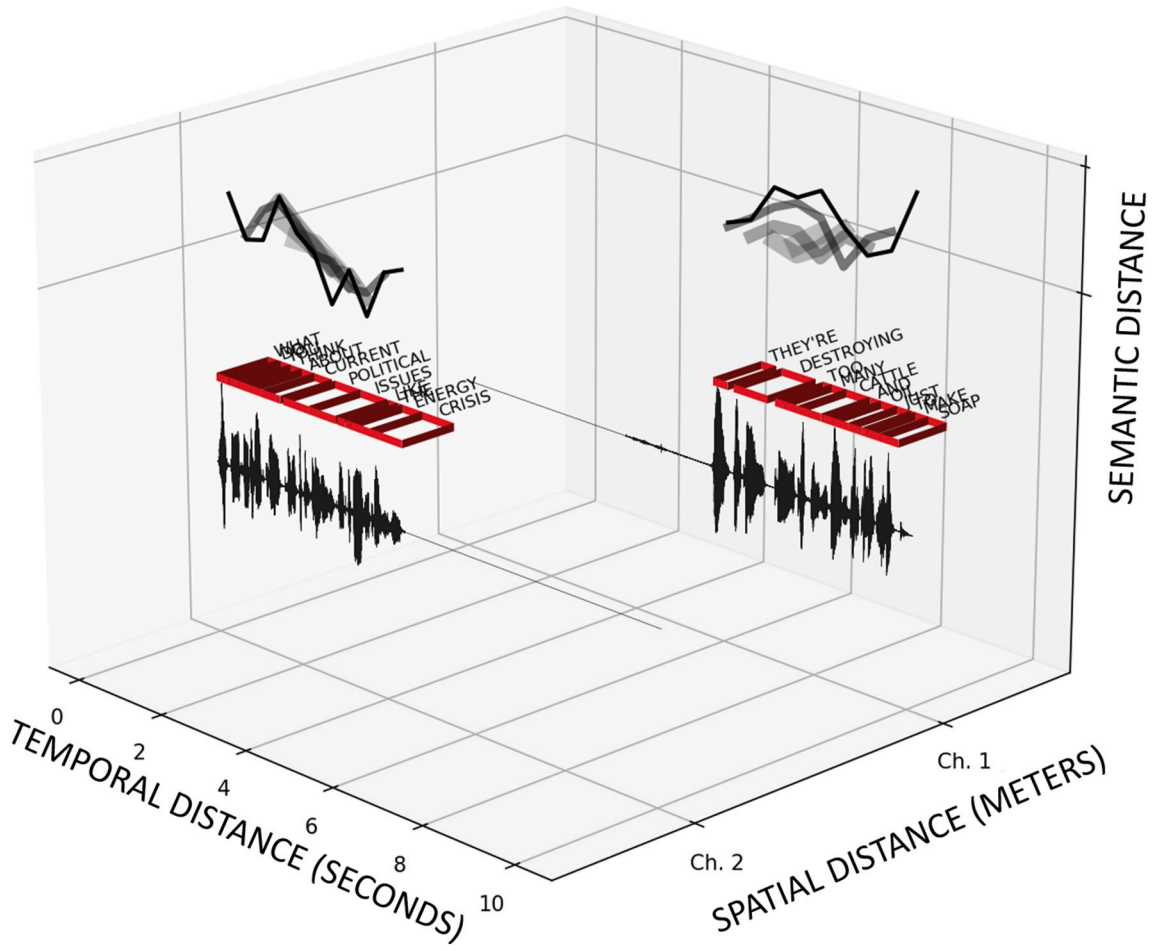

Fig. 4. Merging information about temporal, spatial and semantic information is possible and opens up new opportunities for both research tools and clinical applications. This figure represents sound recordings (black audio waveforms) and illustrates how speech unfolds over a temporal axis, with temporal boundaries of individual words (derived from automatic speech recognition) marked as red boxes. On the vertical axis, information about the degree of semantic incoherence is exemplified with data from within-channel dual-window distances (black lines). Notable values in this domain can inform clinical decisions if properly operationalized. On the left-facing horizontal axis temporal information about word vectors can be expressed, and if such information is combined with physiological data (e.g., electrophysiological- or magnetic resonance imaging data), it may increase our understanding of what unfolds in the brain at the time incoherent utterances are made. Real-time processing of speech can also allow for biofeedback approaches that alert for pending breakdown of communication. On the right-facing horizontal axis the different speaker channels are placed on a spatial axis, indicating that the location (and ultimately a situation) of an utterance can be quantitatively determined in future systems. Combined, the temporospatial context can inform measurements of semantic coherence by making sure that the evaluation is relevant to the actual situation, not based on language from other contexts (e.g., what is common language in written news reports). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

are most often describing personal matters, such as symptoms or the history of an individual. In such a setting, the word "oil" would be less likely to be used, but if it was, it would be more likely to occur in a conversation about nutrition and co-occur with words like "soy" or "pasta". The result of such differences can serve to reduce the relevance of measurements of semantic distances, potentially leading to both over- and underestimation of speech coherence. Further, in the construction of semantic spaces from language material from a broad cross-section of society, cultural biases unfortunately are also included in the models. These data-driven approaches can ideally be considered "neutral" representations of how language is used, but it turns out that they may include undesirable biases when it comes to certain groups (e.g., increasing barriers for people with disabilities; Hutchinson et al., 2020). Indeed, it has been demonstrated that coherence measurements are sensitive to cultural biases in the datasets, and can end up perpetuating such biases (Hitczenko et al., 2021).

Of course, models built on language with generic or wide-ranging topics do have their advantages. Importantly, they can cover a vast number of different words and topics. More context-specific models would be vulnerable to the occurrence of out-of-lexicon words, triggering results that are uninterpretable (or at least *should not* be interpreted). Such a problem can be counteracted by building hybrid models, where the "backbone" is built on a broad corpus, but the model for clinical application is fine-tuned to the specific place and situation in which it is to be applied. Such a fine-tuned approach holds promise of increased sensitivity in detecting incoherent speech, but it does come at the cost of decreased generalizability of methods. For example, a method developed and used at a rural clinic in one country will not be directly transferable and applicable in a more populous area even within that same language speakers within the same country.

Establishing methods that effectively and reliably incorporate contextual information for coherence measurements will need to fulfill several requirements. A first requirement would be extensive localized data collection, enabling language models tailored to the place and situation where they are to be used. A concrete example of clinical relevance could be recording speech from all clinical consultations conducted across psychiatric wards in a single city, and using the transcriptions to build novel models or improve existing models (e.g., by fine-tuning). Even if such models cannot be generalized to and used in other locations, the path towards robust clinical applications will

depend on a uniform and consensus-based manner with which to build the models using appropriate localized data. Establishing consensus around this will take substantial effort, but there are already consortium approaches to standardize speech data collection for psychosis research (e.g., Discourse in psychosis' consortium - https://discourseinpsychosis.org/; see also: Palaniyappan et al., 2022). A second requirement would be a carefully constructed procedure for how to define and quantify the most important characteristics of the temporospatial context within the situation. That is to say, it will be critical to ascertain which aspects of a situation (e.g., location, time of day, weather, current events, and so on) are necessary to consider as potential influences on a patient's language output. Including those necessary aspects into subsequent quantitative analyses will require new types of multimodal models. Important steps have recently been taken towards that end, for example with Google's Pathway Language Model (PaLM; Chowdhery et al., 2022), that has been trained on an all-encompassing dataset of language production in digital form across languages and contexts (e.g., from conversations, books or computer code). Even more inclusive in terms of domains included in the training material, DeepMind's recent large transformer model, Gato, includes data from both language and images, and even certain quantifications of actions of robotic arms or items in computer games, all represented and modeled in a combined fashion across domains (Reed et al., 2022). Such a multimodal approach has obvious appeal for modeling human behavior and clinical assessments, where visual appearance, movement patterns and sounds are all important (Holler and Levinson, 2019). In short, to be able to capture how speech is incoherent in a clinical context a "world model" is needed, not just a language model. For these new powerful models to have applications in psychiatry, the crux will be to find ways for the resulting word (or audio, visual) embeddings to quantify and express the conceptual leaps or abnormal behavioral signs that best capture a clinical disorder.

5. Concluding remarks

The problems with defining incoherence presented in this paper showcase how future study designs and clinical tools will necessarily need exact definitions of coherence that optimally capture the interests of an investigator, whether they are a clinician, a computational linguist, a neuroscientist or a computer scientist. Importantly, it needs to become clear why a coherence measurement is conducted and what neuropsychological constructs or physiological processes the coherence measurements are representing (Foltz et al., 2022). Beyond study design, it will also be necessary and pragmatic to build language models that are based on temporally and contextually appropriate methods. Using models, for example, based on text from news feeds to define coherence in conversational speech during a psychiatric interview will be problematic. Transparency of how methods are producing results will be key in this regard, and information should be explicit for all users and stakeholders. Clinical relevance will be improved when methods can localize *where* and *when* coherence breaks down in speech, and this will increase understanding of disease in brain processes and the nature of language production. Ultimately these coherence metrics can be a part of larger systems for monitoring mental states in patients with schizophrenia. So, by unifying temporal, spatial and semantic information into the same framework (Fig. 4), exploring and defining the "distances" that are most relevant to describe pathological conditions in humans, progress is indeed possible. Indeed, this is in line with another, less famous quote from the philosophical foundation of computational semantic analysis: "*We are talking about the spatial and temporal phenomenon of language, not about some non-spatial, non-temporal phantasm.*" (Wittgenstein, 1953, section 108).

Looking ahead, a possible futuristic scenario might be to leverage these superior temporal and contextually relevant measures of semantic incoherence for clinical intervention purposes. For example, it is conceivable that fast computation and real-time estimations of coherence could be a core component in biofeedback alerting speakers to increased disorganization in conversations, akin to more established audio-visual biofeedback solutions used in speech development and misarticulation (e.g., Byun and Hitchcock, 2012). Techniques for interventions aimed at short-timescale events such as phoneme utterances may not be directly transferable to longer and more complex events such as the entire discourse itself, but this technological approach could help pinpoint *when* and *where* communication is likely to break down. This may well prove to be useful for both patients and clinicians if developed carefully. Quite possibly there may be other permutations of this that are useful. However, whether or not it will prove useful, or even harmful, is a matter for rigorous examination in controlled clinical trials.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.schres.2022.09.020.

Role of the funding source

The funding source had no role in this publication.

Declaration of competing interest

None of the authors report conflicts of interest.

Acknowledgements

Terje B. Holmlund is supported by a Helse Nord grant (#PFP-1301-16).

References

American Psychiatric Association, 2022. Diagnostic And Statistical Manual of Mental Disorders, 5th ed. American Psychiatric Publishing, Arlington, VA.

Andreasen, N., 1986. Scale for the assessment of thought, language, and communication (TLC). Schizophr. Bull. 12 (3), 473–482. https://doi.org/10.1093/schbul/12.3.473.

Angeli, G., Premkumar, M.J.J., Manning, C.D., 2015. Leveraging linguistic structure for open domain information extraction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics And the 7th International Joint Conference on Natural Language Processing, Vol. 1. Association for Computational Linguistics, Beijing, China, pp. 344–354.

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

Bellmund, J., Gärdenfors, P., Moser, E.I., Doeller, C.F., 2018. Navigating cognition: spatial codes for human thinking. Science 362. https://doi.org/10.1126/science.aat6766.

Burgess, C., Livesay, K., Lund, K., 1998. Explorations in context space: words, sentences, discourse. Discourse Process. 25 (2–3), 211–257. https://doi.org/10.1080/01638539809545027.

Byun, T.M., Hitchcock, E.R., 2012. Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation. Am. J. Speech-Lang. Pathol. 21 (3), 207–221. https://doi.org/10.1044/1058-0360(2012/11-0083).

Chowdhery, A., Narang, S., Devlin, J., et al., 2022. PaLM: Scaling Language Modeling With Pathways. arXiv:2204.02311v3 [cs.CL]. https://doi.org/10.48550/arXiv.2204.02311.

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

Dapretto, M., Lee, S.S., Caplan, R., 2005. A functional magnetic resonance imaging study of discourse coherence in typically developing children. Neuroreport 16 (15), 1661–1665. https://doi.org/10.1097/01.wnr.0000183332.28865.11.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1. Association for Computational Linguistics, pp. 4171–4186.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr. Res. 93 (1), 304–316. https://doi.org/10.1016/j.schres.2007.03.001.

Firth, J., 1957. A synopsis of linguistic theory, 1930-55. In: Studies in Linguistic Analysis. Special Volume of the Philological Society. Blackwell, Oxford, pp. 1–31.

Foltz, P.W., 1996. Latent semantic analysis for text-based research. Behav. Res. Methods Instrum. Comput. 28 (2), 197–202. https://doi.org/10.3758/BF03204765.

Foltz, P.W., 2007. Discourse coherence and LSA. In: Landauer, T.K., Kintsch, W., McNamara, D., Dennis, S. (Eds.), Handbook of Latent Semantic Analysis. Lawrence Erlbaum Publishing, Mahwah, N.J.

Foltz, P.W., Chandler, C., Diaz-Asper, C., Cohen, A.S., Rodriguez, Z., Holmlund, T.B., Elvevåg, B., 2022. Reflections on the nature of measurement in language-based automated assessments of patients' mental state and cognitive function. Schizophr. Res. https://doi.org/10.1016/j.schres.2022.07.011.

Halliday, M.A.K., Hasan, R., 1976. Cohesion in English. In: English Language Series. Longman, London.

Hitczenko, K., Mittal, V.A., Goldrick, M., 2021. Understanding language abnormalities and associated clinical markers in psychosis: the promise of computational methods. Schizophr. Bull. 47 (2), 344–362. https://doi.org/10.1093/schbul/sbaa141.

Hoffman, P., Loginova, E., Russell, A., 2018. Poor coherence in older people's speech is explained by impaired semantic and executive processes. eLife 7. https://doi.org/10.7554/elife.38907.

Holler, J., Levinson, S.C., 2019. Multimodal language processing in human communication. Trends Cogn. Sci. 23 (8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006.

Holmlund, T.B., Cheng, J., Foltz, P.W., Cohen, A.S., Elvevåg, B., 2019. Updating verbal fluency analysis for the 21st century: applications for psychiatry. Psychiatry Res. 273, 767–769. https://doi.org/10.1016/j.psychres.2019.02.014.

Holmlund, T.B., Diaz-Asper, C., Elvevåg, B., 2021. The reality of doing things with (thousands of) words in applied research and clinical settings: a commentary on Clarke et al. (2020). Cortex 136, 150–156. https://doi.org/10.1016/j.cortex.2020.08.024.

Holmlund, T.B., Fedechko, T.L., Elvevåg, B., Cohen, A.S., 2020. Chapter 28: Tracking language in real time in psychosis. In: Badcock, J.C., Paulik-White, G. (Eds.), A Clinical Introduction to Psychosis: Foundations for Clinical And Neuropsychologists. Elsevier, pp. 663–685. https://doi.org/10.1016/b978-0-12-815012-2.00028-6.

Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. In: IEEE/ACM Transactions on Audio, Speech, And Language Processing, vol. 29, pp. 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., Denuyl, S., 2020. Social biases in NLP Models as barriers for persons with disabilities. arXiv: 2005.00813 [cs.CL]. https://doi.org/10.48550/arxiv.2005.00813.

Iter, D., Yoon, J., Jurafsky, D., 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In: Proceedings of the Fifth Workshop on Computational Linguistics And Clinical Psychology: From Keyboard to Clinic, Association for Computational Linguistics, New Orleans, LA, pp. 136–146. https://doi.org/10.18653/v1/W18-0615.

Just, S.A., Haegert, E., Kořánová, N., Bröcker, A.L., Nenchev, I., Funcke, J., Montag, C., Stede, M., 2019. Coherence models in schizophrenia. In: ACL Anthol. Proceeding, 126–136. https://doi.org/10.18653/V1/W19-3015.

Jackendoff, R., 1996. How language helps us think. Pragmat. Cogn. 4 (1), 1–34. https://doi.org/10.1075/pc.4.1.03jac.

Just, S.A., Haegert, E., Koranova, N., Broecker, A.-L., Nenchev, I., Funcke, J., Heinz, A., Bermpohl, F., Stede, M., Montag, C., 2020. Modeling incoherent discourse in non-affective psychosis. Front. Psychiatry 11, 1–11. https://doi.org/10.3389/fpsyt.2020.00846.

Kim, N., Kim, J.-H., Wolters, M.K., MacPherson, S.E., Park, J.C., 2019. Automatic scoring of semantic fluency. Front. Psychol. 10, 1–16. https://doi.org/10.3389/fpsyg.2019.01020.

Kraepelin, E., 1921. Einführung in Die Psychiatrische Klinik, , 4th ed.Vol. 1. Verlag von Johann Ambrosius Barth, Leipzig, Germany.

Lakoff, G., Johnson, M., 1980. Metaphors we live by. University of Chicago Press.

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., Dehaene, S., 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. Cognition 213,. https://doi.org/10.1016/j.cognition.2021.104699.

Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104 (2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211.

Lo, C.-W, Tung, T.-Y, Ke, A.H, Brennan, J.R, 2022. Hierarchy, not lexical regularity, modulates low-frequency neural synchrony during language comprehension. Neurobiology Lang. 3 (4), 538–555. https://doi.org/10.1162/nol_a_00077.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. ArXiv:1301.3781 [Cs] http://arxiv.org/abs/1301.3781.

Palaniyappan, L., Alonso-Sanchez, M.F., MacWhinney, B., 2022. Is collaborative open science possible with speech data in psychiatric disorders? Schizophr. Bull. 48 (5), 963–966. https://doi.org/10.1093/schbul/sbac058.

Parola, A., Lin, J.M., Simonsen, A., Bliksted, V., Zhou, Y., Wang, H., Inoue, L., Koelkebeck, K., Fusaroli, R., 2022. Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of NLP automated measures of coherence. Schizophr. Res. S0920-9964 (22) https://doi.org/10.1016/j.schres.2022.07.002, 00274-2.

Pauselli, L., Halpern, B., Cleary, S.D., Ku, B., Covington, M.A., Compton, M.T., 2018. Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. Psychiatry Res. 263, 74–79. https://doi.org/10.1016/j.psychres.2018.02.037.

Pennington, J., Socher, R., Manning, C., 2014. GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. In: Walker, M., Ji, H., Stent, A. (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1. Association for Computational Linguistics, pp. 2227–2237. https://doi.org/10.18653/v1/N18-1.

Reed, S, Zolna, K, Parisotto, E, et al., 2022. A Generalist Agent arXiv:2205.06175v2 [cs. AI]. https://doi.org/10.48550/arXiv.2205.06175.

Rundo, L., Pirrone, R., Vitabile, S., Sala, E., Gambino, O., 2020. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. J. Biomed. Inform. 108, 103479 https://doi.org/10.1016/j.jbi.2020.103479.

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

Sellars, W., 1963. Science, Perception And Reality. Humanities Press, New York.

Shepard, R.N., 1987. Toward a universal law of generalization for psychological science. Science 237 (4820), 1317–1323. https://doi.org/10.1126/science.3629243.

Shi, B., Hsu, W.-N., Lakhotia, K., Mohamed, A., 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. https://doi.org/10.48550/arXiv.2201.02184 arXiv:2201.02184v2 [eess.AS].

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

Tilk, O., Alumäe, T., 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Proceedings Interspeech 2016, pp. 3047–3051. https://doi.org/10.21437/Interspeech.2016-1517.

Tolman, E.C., 1948. Cognitive maps in rats and men. Psychol. Rev. 55 (4), 189–208. https://doi.org/10.1037/h0061626.

Viganò, S., Piazza, M., 2020. Distance and direction codes underlie navigation of a novel semantic space in the human brain. J. Neurosci. 40 (13), 2727–2736. https://doi.org/10.1523/JNEUROSCI.1849-19.2020.

Voppel, A., de Boer, J., Brederoo, S., Schnack, H., Sommer, I., 2021. Quantified language connectedness in schizophrenia-spectrum disorders. Psychiatry Res. 304, 1–8. https://doi.org/10.1016/j.psychres.2021.114130.

Wiggers, P., Rothkrantz, L.J.M., 2007. Exploratory analysis of word use and sentence length in the spoken Dutch corpus. In: Matoušek, V., Mautner, P. (Eds.), Text, Speech And Dialogue: 10th international conference, TSD 2007. Lecture Notes in Computer Science, vol. 4629. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74628-7_48.

Wittgenstein, L., 1953. Philosophical Investigations. Macmillan, New York.

Xu, W., Portanova, J., Chander, A., Ben-Zeev, D., Cohen, T., 2021. The centroid cannot hold: comparing sequential and global estimates of coherence as indicators of formal thought disorder. AMIA Annu. Symp. Proc. 1315–1324. PMID: 33936508; PMCID: PMC8075468.

Xu, W., Wang, W., Portanova, J., Chander, A., Campbell, A., Pakhomov, S., Ben-Zeev, D., Cohen, T., 2022. Fully automated detection of formal thought disorder with time-series augmented representations for detection of incoherent speech (TARDIS). J. Biomed. Inform. 126 https://doi.org/10.1016/j.jbi.2022.103998.