Available online at www.sciencedirect.com
ScienceDirect
Schizophrenia Research 93 (2007) 304 – 316
SCHIZOPHRENIA RESEARCH
www.elsevier.com/locate/schres

Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia

Brita Elvevåg [a,*], Peter W. Foltz [b], Daniel R. Weinberger [a], Terry E. Goldberg [a]

[a] *Clinical Brain Disorders Branch, National Institute of Mental Health, Bethesda MD, United States*
[b] *Department of Psychology, New Mexico State University, Las Cruces NM, United States*

Received 6 November 2006; received in revised form 27 February 2007; accepted 2 March 2007
Available online 16 April 2007

Abstract

Incoherent discourse, with a disjointed flow of ideas, is a cardinal symptom in several psychiatric and neurological conditions. However, measuring incoherence has often been complex and subjective. We sought to validate an objective, intrinsically reliable, computational approach to quantifying speech incoherence. Patients with schizophrenia and healthy control volunteers were administered a variety of language tasks. The speech generated was transcribed and the coherence computed using Latent Semantic Analysis (LSA). The discourse was also analyzed with a standard clinical measure of thought disorder. In word association and generation tasks LSA derived coherence scores were sensitive to differences between patients and controls, and correlated with clinical measures of thought disorder. In speech samples LSA could be used to localize where in sentence production incoherence occurs, predict levels of incoherence as well as whether discourse "belonged" to a patient or control. In conclusion, LSA can be used to assay disordered language production so as to both complement human clinical ratings as well as experimentally parse this incoherence in a theory-driven manner.

Keywords: Psychosis; Language; Semantic; Thought disorder

Communicating ideas and thoughts through the medium of language is a fundamental aspect of human social behavior. Discourse is perceived as coherent when ideas relate to a global theme and follow a logical sequence determined by one's knowledge of the world. In contrast, discourse is perceived as incoherent when the flow of ideas seems disjointed or when loose associations between words are present, or tangential if there are digressions from the topic. Such incoherent discourse, often termed formal thought disorder (ThD), occurs in a variety of psychiatric and neurological conditions. In particular, patients with schizophrenia (whom we studied) display abnormalities in the use of language during spontaneous speech. Importantly, the neural substrates of these language deviances are likely related to the underlying pathophysiology of the disorder (DeLisi, 2001).

Coherence is a widely used concept in both discourse psychology and clinical diagnosis. The concept of coherence encompasses the idea of an orderly flow of information within a discourse, including how well the discourse is connected within and across words, sentences, utterances, documents and between people. We define "coherence" of speech as the semantic similarity or relationship of ideas to other ideas. Crucially, it is a patient's verbal self-presentation as elicited in a clinical interview and subjectively evaluated that remains an

* Corresponding author.
E-mail address: brita@elvevaag.net (B. Elvevåg).

0920-9964/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.schres.2007.03.001

essential diagnostic tool in psychiatry, and assessing the coherence of this discourse is fundamental. As a symptom, ThD forms a major component of the observed phenomenology (present in 20–50% of patients with schizophrenia (Andreasen and Black, 2005; Breier and Berg, 2003), is an important criterion in the diagnosis of schizophrenia (Bleuler, 1911; Kraepelin, 1919; McKenna and Oh, 2005) and may have prognostic significance (Andreasen and Grove, 1986; Harrow and Marengo, 1986). Neuroleptic medication generally improves all symptoms, including speech coherence (Spohn et al., 1986). Clearly disordered thinking is a fundamental aspect of the brain dysfunction associated with the schizophrenia illness. However, establishing a primary cognitive mechanism responsible for ThD has not been straightforward, both because the underlying pathology is multidimensional (Cuesta and Peralta, 1999; Harrow et al., 1982) and because reliable fine-grained ratings of ThD are difficult to make (for an overview see McKenna and Oh, 2005). Thus, a valid, reliable and objective measure of discourse coherence would be of potential value in indexing ThD and useful for prognosis, in assessing treatment responsiveness, and in research concerning the associated brain dysfunction.

Hitherto, attempts to examine deviance and incoherence in formal thought disordered patients have generally focused on analyzing word level deviancies or examining sensitivity to linguistic anomalies in sentences, and their relationship to clinical ratings of ThD. Previous textual analysis of discourse has examined speech predictability and the quantity of information conveyed, and has employed cloze procedures, type–token ratios or readability indices (Manschreck et al., 1981). However, these relatively simple linguistic measures do not fully capture the richness of human discourse, and are time-consuming and subjective in scoring. With the advent of powerful computing techniques, and recent developments in computational linguistics and cognitive modeling, automated methods capable of analyzing coherence of discourse have been developed. We have capitalized on this technology to develop and validate an objective and reliable tool with which to measure coherence in language in schizophrenia, which may also be applicable to a variety of disorders where language deviances occur.

1. An automated approach to coherence

Latent Semantic Analysis (LSA) is a computational model of human knowledge acquisition and a practical application for concept-based text analysis (for details see Landauer and Dumais, 1997; http://lsa.colorado.edu/). The underlying premise for deriving a model of meaning is that words used in similar contexts tend to be more semantically similar to each other than words in different contexts. LSA acquires a representation of semantic knowledge based on the automated analyses of millions of words of natural discourse, and by solving the relations between word and passage meanings using Singular Value Decomposition (SVD, a matrix algebra technique related to factor analysis). In LSA the discourse is first represented as a matrix, where each row represents a unique word in the text and each column represents a text passage or other unit of context (*e.g.*, a paragraph). The entries in this matrix are the frequency of the word in the context. An SVD of the matrix is then applied which results in a 100–500 dimensional "semantic space". The dimensions are automatically derived as part of the solution of the SVD analysis, and a possible interpretation of the dimensions is that they are analogous to the semantic features often postulated as the basis of word meaning. However, interpretation of those features is technically and conceptually quite complicated (see Landauer et al., 1998).

In the derived semantic space, words, sentences, paragraphs, or any other unit of text are represented as vectors by the sum of the vectors of the individual words contained in the text. The word and large unit of text vectors can be compared against each other in order to measure the amount of semantic similarity. In this paper, the cosine of the angle (range −1 to +1) between two vectors is the key measure of semantic similarity, with greater cosine values indicating greater degrees of similarity (for an introduction and more details, see Landauer et al., 1998).

In essence, LSA is inducing the semantic similarities of language based on the pattern of usage of words across a large corpus of text. The information about all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. This similarity can then be estimated through analyses of large text corpora. Thus, to LSA, the meaning of a word is defined by the contexts in which it appears, and the meaning of a context is defined by the words that appear in it. The result is that text vectors that share semantic content but have no terms in common can be highly similar. For example, consider the following phrases: "The radius of spheres" and "A circle's diameter" have a cosine similarity of 0.55, whereas "The radius of spheres" and "The music of spheres" have a cosine of only 0.01. In other words, LSA's technique captures a much deeper "latent" structure than simple word–word correlations and clusters, and this may be why LSA produces good