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

approximations to human cognitive semantic relations (see Foltz, Kintsch, and Landauer (1998), Landauer and Dumais (1997), and Landauer et al. (1998) for modeling language acquisition, semantic priming, semantic categorization and the effects of text coherence on comprehension). LSA has also been used as the critical component in successful commercial applications including automated essay scoring and information retrieval systems.

In the following experiments, the speech samples that we collected were analyzed within a single LSA-based semantic space that models general English language. The space was derived from a corpus of text that corresponded to the approximate amount and type of general reading that an average healthy person in the US would be exposed to by the first year of university. The corpus was comprised of 37,651 text samples with 92,408 unique words, totaling 69 MB of text (http://lsa.colorado.edu/). The analyses were done using 300 dimensions. This corpus at 300 dimensions has been used and tested in a wide range of analyses using LSA (see Landauer et al. (1998), and Landauer, Laham, Rehder, and Schreiner (1997) for examples). The speech samples were transcribed into machine-readable text. Then, using the derived semantic space, the semantic similarity of utterances both within and between the speech samples were computed: (i) how well one word relates to another; (ii) how well one sentence relates to the next sentence within a discourse; (iii) how well a person's answer relates to the question asked; (iv) or how well a person's answer to a question relates to another person's answer. It is crucial to appreciate that measures derived from LSA are automatically computed based on its model of semantic associations. Nevertheless, one measure of coherence will not be able to address all questions. Indeed, the different measures capture a range of aspects of what humans would characterize as discourse coherence (see Lorch and O'Brien, 1995).[1] In addition to LSA measures, in Experiment 4 other computational linguistically-derived statistical measures of the expected pattern of usage and order of words in the discourse are used in combination with the LSA coherence measures to assess more open-ended interview responses (see Jurafsky and Martin, 2000). Used individually and in combination, they provide an approach to assessing language coherence in schizophrenia and in other disorders in which there are language deviances.

Our goal here was to establish the validity of this automated, objective and reliable assay with which to assess coherence in discourse in schizophrenia directly, as compared to more established and conventional clinical measures of disordered thinking (*i.e.*, to evaluate methods with which to assay ThD, but not to suggest a candidate cognitive mechanism for ThD). In a series of studies we illustrate and evaluate a computational framework for analyzing coherence in discourse, and present methods with which to assay disordered thinking that both complement human clinical ratings or surpass them (*i.e.*, are more sensitive to subtle deviations in discourse). The validity and sensitivity of this computational approach may make it a valuable tool with which to explore how semantic sub-processing dysfunction impinges on communication as well as in the search for an understanding of various cognitive phenotypes in schizophrenia.

2. Methods

2.1. Participants

Participants were screened for and cleared of neurological, developmental learning, and substance-abuse problems. Patients fulfilled DSM-IV criteria for schizophrenia or schizoaffective disorder, as determined by the Structured Clinical Interview for DSM-IV (SCID), with three psychiatrists reaching a consensus diagnosis (for characteristics of patient and control samples, see Table 1 for Experiments 1, 3 and 4 and Table 2 for Experiment 2). Healthy control volunteers were recruited through the National Institutes of Health's volunteer panel. All control participants and outpatients were paid for their participation, and inpatients completed the study as part of their protocol for entering the hospital. All studies below were conducted according to the guidelines of the internal review board at the National Institute of Mental Health.

2.2. Thought disorder ratings

Speech generated in a 45 min semi-structured clinical interview with open-ended questions (including questions regarding symptoms, current events as well as why some people believe in God and what free-will is) was

[1] At first glance it may appear that our definition of coherence of thought as the semantic similarity or relationship of ideas to other ideas might run into trouble given the issue of perseveration in patients. However, word–word perseverations or phrase–phrase perseverations were very rare in our samples, although it is of course true that certain patients can become pre-occupied with one theme. Interestingly, the presence of the aforementioned types of perseverations would have in fact resulted in higher coherence scores in patients, but in fact they were low (i.e., evidence of minimal perseverations in our sample). Moreover, individual ratings of coherence (as determined by a human) were also low.

Table 1
Characteristics of samples in Experiments 1, 3 and 4

	Patients $n=26$ (19M, 7F)		Controls $n=25$ (10M, 15F)	
	M	SD	*M*	SD
Age (years)	33.77	7.63	35.44	12.94
WAIS-R IQ *	94.46	13.44	107.96	11.90
WRAT-R IQ **	103.58	11.50	109.32	8.68
Age at 1st hospitalization (years)	21.54	4.29	N/A	
Neuroleptic medication	24		0	
–Clozapine/ olanzapine/quetiapine	17		–	
–Risperidone	6		–	
–High potency drug [a]	1		–	
–Anticholinergics	2		–	
–Adjunctives [b]	14		–	
TLC scores			N/A	
–Global [c]	1.81 (range 0–3.875)	1.12		
–Poverty of speech	0.79 (range 0–3)	1.14		

Intellectual function was assessed with the Wide Range Achievement Test-Revised Reading (WRAT-R; Jastak and Wilkinson, 1984) and a short form of the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981).

[a] Haloperidol.

[b] Lithium, depakote, sertraline, lorazepam, venlafaxine, clonazepam, buspirone.

[c] Based on a single interviewer's clinical impression (TG) using a scale of 0–4; there was a wide range of scores in our patient sample (from absent to most severe).

* $p<.01$ (independent samples *t*-test).

** $p<.05$ (independent samples *t*-test).

rated using a standard clinical measure of ThD (the Scale for the Assessment of Thought, Language and Communication; TLC (Andreasen, 1986)). In this assessment the interviewer rated 18 defined abnormalities in speech (poverty of speech, illogicality, incoherence, clanging, neologisms, word approximations, poverty of content of speech, pressure of speech, distractible speech, tangentiality, derailment, stilted speech, echolalia, self-reference, circumstantiality, loss of goal, perseveration and blocking) (see Tables 1 and 2). A global score was derived that was based on a single interviewer's overall clinical impression (TG) of how impaired a person's communication was with a range of 0–4. This scoring is defined in Andreasen (1986) in which an overall impression of 0 indicates an absence of a TLC disorder, a 1 indicates a mild TLC disorder "but clinically significant", a 2 indicates a moderate TLC disorder "which leads to a moderate disturbance in communication at least from time to time", a 3 indicates a severe TLC disorder "significant enough to impair communication for a substantial part of the interview", and a global score of 4 indicates an extreme TLC disorder "so severe that communication is difficult or impossible most of the time" (p. 481). The global TLC scores were used to median split the patient group into those who displayed substantial clinical ThD *versus* those who displayed little ThD. In both patient cohorts this split resulted in a score of 2 or higher being characterized as high ThD and 1.75 or less being characterized as little ThD. Additionally, we employed a measure of verbal productivity (poverty of speech; see Bowie et al., 2005). In Experiment 2, patients were also rated on the Brief Psychiatric Rating Scale that measures the presence and severity of 24 symptom constructs (BPRS; here we employed the total score exclusive of 'conceptual disorganization') (Overall and Gorham, 1962).

Table 2
Characteristics of sample in Experiment 2

	Low ThD patients $n=11$		High ThD patients $n=10$	
	M	SD	*M*	SD
Age (years)	33.64	6.78	31.60	8.33
WAIS-R IQ	97.91	15.63	91.20	8.99
WRAT-R IQ	105.64	11.74	106.30	10.24
BPRS [a]	32 (range 20–46)	7.67	40 (range 19–62)	15.02
Neuroleptic medication	9		5	
–Clozapine/ olanzapine	4		5	
–Risperidone	5		0	
–High potency drug [b]	0		1	
–Anticholinergics	4		0	
–Adjunctives [c]	5		4	
TLC scores			N/A	
–Global [d]	1.30 (range 0–1.75)	0.53	2.75 (range 2–3.75)	0.59
–Poverty of speech	0.64 (range 0–2)	0.81	1.00 (range 0–3)	1.07

Patient groups were matched on age ($p>.5$), WRAT-R ($p>.5$) and WAIS-R ($p>.1$). There were twenty-four healthy control participants (11 females, 13 males) mean age 31.17 years (SD=9.34). Their mean score on the WRAT-R was 109.36 (SD=11.47) and on the WAIS-R was 111.65 (SD=13.21).

[a] The Brief Psychiatric Rating Scale (Overall and Gorham, 1962) — total score exclusive of conceptual disorganization.

[b] Haloperidol.

[c] Lithium, depakote, sertraline, lorazepam, venlafaxine, clonazepam, buspirone.

[d] No differences between the Cohorts #1 and #2 patient groups (Tables 1 and 2 respectively) on global TLC scores (using a scale of 0–4), $p>.05$.

3. Experiment 1: single word associations

3.1. Task

Participants verbalized the first word that came to mind when a series of 10 words was read one at a time to them ('God', 'food', 'boy', 'dark', 'hard', 'high', 'king', 'table', 'slow' and 'man').

3.2. LSA application

Assaying the underlying semantic associative network in patients is the first step towards understanding complex language production deviances. Given that "loose associations" contribute to the very definition of formal ThD we would expect word associations in patients with ThD to be less "usual" than in patients without ThD or in healthy controls. Thus, we examined whether our LSA derived coherence measure is sensitive enough to detect such subtle deviations in a simple word association task. The series of cue words and responses were electronically transcribed. For each person, the semantic similarity between each cue word and its response was computed (using the document to document comparison available at http://lsa.colorado.edu/), and the average coherence (cosine) scores were calculated (see Landauer et al. (1998) for details on the mathematics for using this approach).

3.3. Statistics

Group means (based on high and low ThD scores in the patient group, and the control group) for LSA derived cosines were compared by analysis of variance (ANOVA). LSA scores were also correlated with clinical global ThD scores and verbal productivity.[2]

3.4. Results

Patients' coherence scores were lower (0.32) than controls' (0.43) ($F(1,49)=8.66$, $p<0.01$). Patients were split (median) into a high and low ThD group based on their global TLC scores, resulting in a cut-off of 2 or greater comprising the high ThD group and a score of 1.75 or less comprising the low ThD group. In line with predictions, coherence was lower in patients with high levels of clinically rated ThD (global score ≥ 2; $n=11$) (0.25) as compared to patients with little ThD ($n=13$) (0.38) ($t(22)=2.35$, $p<0.05$) and healthy controls ($t(35)=-3.88$, $p<0.0001$). There was no difference between the latter two groups ($t(37)=-1.16$, $p>0.1$). Thus, patients with high levels of clinically rated ThD generated less usual word associations. Moreover, patients' coherence (as indexed by LSA) correlated significantly with their clinically rated global ThD ($r=-0.41$, $p<0.05$), but not with verbal productivity (poverty of speech; Bowie et al., 2005) ($r=0.26$, $p>0.1$).

LSA derived measures of association were able to detect subtle differences within the patient group, and the TLC clinical ThD ratings of the differences between the patients corroborated our automated experimental finding. Importantly LSA's measure was not an artifact of verbal productivity.

4. Experiment 2: verbal fluency

4.1. Task

Participants were asked to generate verbally as many 'animals' as they could in a period of 1 min. The series of responses per participant were transcribed electronically.

4.2. LSA application

The verbal fluency task is a widely used clinical test, and typically patients with schizophrenia generate fewer exemplars, but the reason for this is not clear (Bokat and Goldberg, 2003). For each person the coherence between word 1 to word 2, word 2 to word 3 and so on were computed (using the document to document matrix comparison available at http://lsa.colorado.edu/). Thus, if a person generated the sequence: dog, cat, fox, raccoon, bear, we calculated the similarity of each word to the next word (in LSA space) as follows: dog → cat (cosine = 0.36); cat → fox (cosine = 0.39); fox → raccoon (cosine = 0.38); raccoon → bear (cosine = 0.41), and thus derived the average coherence scores for each person.

4.3. Statistics

Group means (based on high and low ThD scores in the patient group (a median split of global TLC scores; ≤ 1.75 *versus* ≥ 2.0) and the control group) for LSA derived cosines were compared by ANOVA. LSA scores were also correlated with clinical global ThD scores, verbal productivity and BPRS scores.

[2] Because cosines (our coherence measure) are closely related to correlations (only the normalization is different), it is appropriate to apply Fisher's *r*-to-*z* transforms on the cosines before the analysis of variance. However, for all analyses, the *r*-to-*z* transforms did not change the results in any meaningful way, and so the raw cosines were used for all analyses reported in this paper.

4.4. Results

As in Experiment 1, patients were split (median) into a high and low ThD group based on their global TLC scores, resulting in a cut-off of 2 or greater comprising the high ThD group and a score of 1.75 or less comprising the low ThD group. The coherence scores of patients with high levels of clinically rated ThD ($n=10$) were lower than patients with low levels of clinically rated ThD ($n=11$) (0.24 *versus*. 0.29; $t(19)=2.80$, $p<0.01$), and to that of healthy control participants (0.32; $t(32)=-3.85$, $p<0.001$). There was no difference in terms of coherence between patients with low levels of clinically rated ThD and controls ($t(33)=-1.24$, $p>0.1$). Thus, in line with findings from Experiment 1, patients with high levels of clinically rated ThD generated less usual word associations (which may contribute somewhat to the slowness of word generation) (see Fig. 1). Importantly, our coherence measure was able to detect a deviance in coherence *within* the patient group, as captured in a significant correlation of our coherence measure with patients' global ThD ratings; $r=-0.53$, $p<0.01$. Interestingly, the typically used measure of word count was not useful in making this discrimination within the patient group (correlation between word count and global TLC ratings: $r=-0.35$, $p>0.1$) (mean of 16.9 *versus* 14.7 words for low and high ThD patients respectively; $t(19)=0.96$, $p>0.1$). Crucially, our coherence measure was not an artifact of verbal productivity as indicated by the absence of a significant correlation with poverty of speech ($r=-0.16$, $p>0.5$). Also, there was not a significant correlation between the total BPRS and LSA ($r=-0.36$, $p>0.1$).

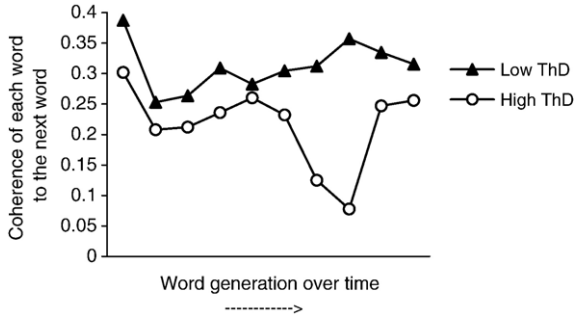

Fig. 1. A graphical illustration of the average word to word coherence from a word generation task (first 10 animal words) comparing a group of patients with clinically rated low levels of ThD (low ThD) to a group with high levels of ThD (high ThD).

We have thus validated our coherence measure and have shown that it is *at least* as sensitive to differences between patients and controls as simple word counts, and furthermore that our coherence measure is *more* sensitive than the usual measure of word count. These data also indicate that results were not task-specific or cohort-specific.

5. Experiment 3: interviews

5.1. Task

Participants were administered a structured interview that we had designed in order to elicit uninterrupted speech. Questions included queries such as "What would someone have to do if they wanted to smoke a cigarette/wash their hair/do their laundry", as well as "tell me the story of Cinderella/Romeo and Juliet". Responses to these types of questions tend to follow standard script-like themes. Questions also included more abstract themes such as "What is free-will/what is democracy/why do some people believe in God?" (A complete interview is available from the authors (BE) upon request). Participants generally spoke for a minute or two in response to each question, and no response was less than eight words.

Transcripts of these interviews were rated blindly by two staff members (including one of the authors, TG) for tangentiality, content and organizational structure (using a scale that we devised with anchors at 1 and 7; a score of 1 representing a response that was incisively related to the question, a conventional response or a coherent response, and a score of 7 representing a response that was unrelated to the question, a bizarre and impossible response or the whole response was incoherent—for tangentiality, content and organizational structure respectively).[3]

[3] We chose to create a short rating scale because participant responses were made to a narrow set of questions and were often brief, thus restricting the range of possible types of disorganization. Thus, we stressed three basic parameters that were consistent with the overall aims of the study: (i) tangentiality (included in part because of its ease of measurement), (ii) disorganization (important because we have stressed semantically relevant, logical connections among words and phrases), and (iii) content. Each of the three items were rated on a 1–7 scale, with clearly described anchor points. The consistency among clinical raters was good; the intra-class coefficients were 0.94, 0.97 and 0.85 for tangentiality, content and organizational structure respectively. Face validity of our scoring system was apparent in that patients scored higher (i.e., more thought-disordered) than control participants.

5.2. LSA application

Although word–level analyses are useful experimental measures for assessing associational processes (*e.g.*, Experiments 1 and 2), in everyday-life assessments of coherence are made through conversations, and the ecological utility of our automated measure needs to be validated by its assessment of speech samples generated with minimal artificial experimental constraints. In semi-structured interviews and subsequent two-way dialogue we can examine the coherence between the utterance (*i.e.*, unit of speech) from the interviewer to the resulting response (from the patient), and then the coherence between this response of the patient to the next unit of utterance by the clinician, and so on in order to assess the general flow of the conversation. In coherent dialogue, speakers continually assess the others' understanding of the conversation, and thus adjust speech accordingly (*i.e.*, a common knowledge-base is built up during the course of the dialogue (Bamberg and Moissinac, 2003; Clark and Haviland, 1977; Zwaan and Singer, 2003). In incoherent dialogues, the listener may have to fill in the missing gaps (either overtly or covertly) and thus much of the coherence may be driven by the interviewer, therefore such dialogue may not be ideally suited to fully automated discourse analysis for deviances in coherence, hence our use of structured questions.

Since our approach to coherence is based on a theoretical model of knowledge and discourse representation there are numerous questions that can be addressed using a variety of measures. In two-way dialogue, we can examine the coherence *between* all that an interviewer said and all that a patient said. This approach is akin to more traditional measures indexing tangentiality. We can examine *within* a patient the coherence of what they say initially to what they say subsequently. Thus, we could examine the coherence of one word to the next word, or one set of words to the next set of words, or one set of sentences to the next set of sentences, and so on. We can also examine the coherence of patients *to* other patients who have responded to similar tasks (see Experiment 4 below). Of course, a crucial question concerns whether our coherence measures capture something that is similar to that detected by psychiatric raters. Here we first examine coherence between question and response in order to illustrate how the time-course of a response can be examined (*e.g.*, to compute tangentiality over the course of the response). Second, we sought to localize where in sentence production speech becomes incoherent by analyzing the coherence between question and response using a variety of "moving" window sizes.

We used a "moving windows" method to compute the similarity of one part of the discourse to the next. A moving window has the same basic effect as a moving average in that data fluctuations are smoothed thus allowing patterns to be seen more clearly. In this method, a cosine is generated for each clause or window of words, indicating the similarity of those words in response to the question asked. As the window moves across the patient's response and farther away in time from the interviewer's question, one would expect that the text in the response would be less related to the question asked. Thus, the cosines should decrease as the window moves. Therefore, such an approach detects how quickly the discourse moves away from the original topic or question asked. The coherence was computed between the question and the response using window sizes two, three, four, five, six, seven and eight words. Then this "window" was moved to the right and the coherence between the original question and this "window" of the answer was re-computed (hence the term "moving window" or more accurately moving clause). This comparison of each window to question continued until the end of the text.

5.3. Statistics

The coherence values of the similarity of the response to the question can be plotted to look at the time-course of a response. In cases where there is a sufficient length of response, the regression line for the cosine as a function of distance from the original question can be computed and the slope found (see Foltz et al. (1998) for a similar approach to written discourse), and the greater the slope the more tangential the response.

To address where it is in sentence production that speech becomes incoherent, as well as validating such an approach with relatively short speech samples, speech in response to four questions was analyzed ("What would someone have to do if they wanted to smoke a cigarette/wash their hair/get a can of soda and tell the story of Romeo and Juliet"). These responses were selected because there was a sufficient amount of speech (*i.e.*, at least two sentences at the very minimum) generated from all participants. We used repeated measures ANOVA to assess differences between the groups and amongst the window sizes.

5.4. Results

Since patients and controls talked at considerable length in response to the question "Tell me the story of Cinderella", the slope was computed for these responses. We found a significant correlation between the slope of the coherence values and the blind human ratings of tangentiality for these same responses ($r=0.44$, $p<0.01$).

Examining deviations within patients we found greater divergence of coherence as the window size increased for patients rated clinically as having high levels of ThD than for patients rated clinically as having low levels of ThD, or controls (see Fig. 2 Panel A). Importantly, this greater divergence of coherence as the window size increased was also the case when the entire sample (patients and controls collapsed) simply was grouped as a function of a median split of blind psychiatric ratings of coherence and tangentiality of these transcripts (see Fig. 2 Panel B). In both Panels A and B, there was not a significant effect of group ($p>0.1$), although there was of window size ($p<0.0001$), since the coherence (cosine) typically increases with a bigger window size due to greater contextual overlap (*i.e.*, more information). Importantly, in both cases there was a significant group x window size interaction (ANOVA, $p<0.01$) due to a difference at a window size of 8 words. These results were not due to a difference in the number of words generated (when number of words was used as a covariate). Specifically, in Panel A the results of the ANOVA were as follows: group, $F(2,41)=1.32$, $p>0.1$; window size, $F(6,246)=187.06$, $p<0.0001$; group x window size, $F(12,246)=4.66$, $p<0.0001$. In Panel B the results of the ANOVA were: group, $F(1,43)=2.47$, $p>0.1$; window size, $F(6,258)=232.33$, $p<0.0001$; group x window size, $F(6,258)=3.59$, $p<0.01$. With reference to both Panels A and B, post-hoc t-tests comparing the groups at each window size (corrected for multiple tests of comparison using a modified Bonferroni method (Holm, 1979)) did not reveal any significant group differences in window sizes 2 to 7, but a significant difference in window size 8 between low and high ThD patients and between high ThD patients and controls in Panel A and between low ThD and high ThD in Panel B.

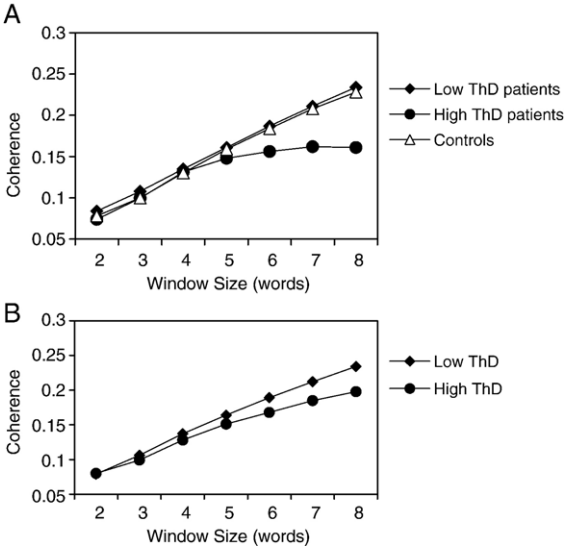

Fig. 2. Examination of the size of clause at which speech becomes incoherent, with a moving window of 2 to 8 words. In Panel A, control participants were compared to patients who were split as a function of a median split of standard clinical ratings of ThD. In Panel B, participants (patients and controls collapsed) are split (median) simply as a function of the blind ratings of the actual scripts.

Our results indicate that we are able to capture some aspects of organizational structure, tangentiality and content that psychiatric raters also detect. Moreover, it is very clear that problems in coherence in language are amplified over larger units. Importantly, our measure seems to be able to detect subtle deviations at larger speech units (as well as at the word level; see Experiments 1 and 2). We note that this analysis is difficult to do at a sentence level because unlike written language humans rarely speak distinct sentences, thus it is a very subjective matter of determining the end or beginning of sentences. Theoretically, within conversations, a window of eight words would capture much coherence found between sentence level discourse. We chose not to go beyond a window size of eight words, because of the practical reason that some responses were short and we would have to exclude them. Thus, our findings suggest that the problems underlying incoherence may involve higher level discourse planning (Barch and Berenbaum, 1996; Hoffman, 1986) as well as simpler associationist processes.

6. Experiment 4: story-telling

6.1. Task

A sub-set of the speech samples from the structured interview in Experiment 3 was employed. These speech samples were generated in response to questions about what someone would do in order to do their laundry and descriptions of the story of Cinderella. Two human raters scored each response for tangentiality, coherence and content (consistency among raters was good and so their ratings were averaged).

6.2. LSA application

We sought to examine the semantic similarity of patients *to* other patients who had responded to similar tasks. This is because if we have already judged the appropriateness (*e.g.*, organizational flow, tangentiality, content, and so on) of other people's responses, then by comparing the similarity of a persons' response to pre-judged responses we can determine the extent to which two people are discussing the same topic in a similar manner (*i.e.*, ranking of formal thought disorder of that person), as well as how well a particular patient's response matches a particular diagnostic category (*e.g.*, patient *versus* control). For example, if asked the question, "Tell me the story of Cinderella", one would expect people to provide a range of answers, yet still the responses would all be largely similar. Any response that goes off-topic or becomes unusual in terms of content would tend to deviate semantically to a greater extent than other responses. The degree of this deviation can be measured with LSA by measuring the distance of a response to any other individual response or to the centroid of all other individuals' responses, which can be taken to represent the equivalent of the semantic content of the "average" answer. These measures of distances can then be used for a range of types of predictions. We wished to evaluate whether our automated approach could predict the scores of organizational structure, tangentiality and content that have been given by blind human raters of the same discourse. In essence, this approach predicts the psychiatric ratings of a participant's response based on the response's similarity to other participant responses for which the clinical ratings are known. Thus, given N responses that have been rated by clinicians, how well can the system predict the rating or diagnostic category of the $N+1$th response?

6.3. Statistics

Regressions were used to predict the average human ratings. A series of LSA-based measures were derived in order to determine which measures, individually and in combination, most successfully modeled the human ratings. The primary LSA measures that were derived were based on measuring the semantic similarity of a participant's response to the 10 most similar responses generated by other participants to that question that had already been coded by the human raters. A predicted rating was then assigned to that participant's response based on the ratings given to the 10 similar responses and weighted by their similarity. For example, if a participant's response was similar to some responses that received ratings of 2 and some that received ratings of 3, it would tend to receive a predicted rating between 2 and 3 (see Landauer et al. (1998) for additional details and for use of this approach for scoring student essays and related applications). Several additional measures were computed on each response that computed statistical measures of word flow and word frequencies of the kind used in speech recognition and natural language processing (see Jurafsky and Martin, 2000). The rationale behind this approach is that responses that are semantically and structurally similar to other responses should receive similar scores by clinician raters. In using the 300 separate dimensions of LSA along with a variety of other measures and modeling how to put them together to mimic clinical judgments, we are doing a type of high dimensional triangulation to the human expert. Thus, although LSA is scoring each response blindly, it is making its judgments based on how clinicians have scored semantically similar responses in the past. Stepwise regressions were used to determine the combination of measures that best modeled the human raters, resulting in a single predicted score for each participant's response. Resulting models typically had 3–5 variables. The predicted scores from the stepwise regressions were compared against the actual blind human ratings for the ratings. In order to verify generalizability of the models, a jack-knife procedure was used in which the prediction for each participant's response was predicted by the data from all other participant's responses but not itself. Thus, all the correlations reported provide a more conservative measure that can show how well the results may generalize to new datasets.

Additionally, using the same predictor variables described above we used a stepwise discriminant function analysis on the Cinderella and laundry responses to classify patients and controls. We sought to address the question of whether we can use LSA to predict group membership (patient *versus* control), how LSA compares to blind human raters (three expert psychiatrists) on this task, and whether LSA could detect ThD, using same classification of ThD as in Experiments 1 and 2. We report the cross-validation prediction results, which like the jack-knife procedure described above, provides measures of the generalizability of the results to new datasets using the same prediction model.

6.4. Results

Using just the LSA-based measures to predict scores was quite successful with a significant correlation with blind human ratings of these responses. Adding the additional statistical language based measures (which could account for some aspects of grammar and syntax) accounted for additional variance. For the Cinderella responses, the correlation between predicted and actual ratings for organizational structure, tangentiality and content were 0.70, 0.73, and 0.79 respectively (all $p < 0.001$). The most common variables chosen by the regressions were those that measured the semantic similarity of each response to the other responses. Responses less like the others tended to have higher human ratings of tangentiality and less content and coherence. Thus, LSA is able to accurately predict the raters' scores for the responses. However, it should be noted that not all question types resulted in an accurate predictions. For responses to the laundry question, correlations were still significant, but not as strong as those found in the Cinderella responses, $r = 0.48$ ($p < 0.01$), 0.75 ($p < 0.001$), and 0.46 ($p < 0.01$) (for organizational structure, tangentiality and content respectively).

In the discriminant function analysis for classification of patients and controls, we obtained correct classification 82.4% of the time (78.4% using cross-validation) for the Cinderella question. For the laundry question, we obtained correct classification of patients and controls 80.4% of the time (78.4% using cross-validation), but with the caveat that the measures that were most discriminating assessed aspects of language not central to this paper (*e.g.*, syntax). For the discriminant function classification of patients with high *versus* low ThD levels, we obtained 87.5% correct classification for the Cinderella questions and 87.5% correct classification for the laundry question (both using cross-validation). Thus, the methods used provided high rates of accurate classification for both questions.

As a comparison, three psychiatrists (with an average of 14 years of clinical experience) blindly categorized group membership (patient or control) from transcripts of responses to the Cinderella and laundry questions. There were good intra-class correlations between the three psychiatrists' binary classifications of each transcript (0.84 and 0.64 for Cinderella and laundry responses respectively). The human raters performed at 71.9% for the Cinderella stories and 65.4% for the laundry stories. Importantly, the pattern of correct/incorrect categorization made by LSA was comparable to that of the human raters. Interestingly, on average when LSA was "correct", in 48% of cases at least one of the psychiatrists was incorrect.

The results described above in Experiment 4 indicate that LSA detects the amount and quality of relevant content information in a response and can use that to predict measures of organizational structure, content and tangentiality, as well as to predict whether the response was from a patient or control. Thus, through testing of different question types, it is possible to determine the questions that will elicit responses that LSA will be maximally sensitive to the detection of both between and within group differences. With the appropriate choice of question types,[4] LSA can predict the presence of schizophrenia from the responses about as well as trained clinicians. Nevertheless, LSA may be more or less sensitive to different question types. Indeed, the length of the responses to the laundry question were shorter than those to the Cinderella question and there was likely not sufficient information contained in the responses to permit LSA to provide as accurate measures of the semantic content. We recognize that there have been studies showing that speech in patients with schizophrenia can be distinguished from speech from healthy controls, and it is likely that ThD does contribute to the discriminability since ThD is known to affect language as indexed by measures such as TLC scores. The final discriminant function analysis shows that LSA is able to accurately discriminate ThD level, even though the ThD measure for each patient was based on a separate sample of discourse. Nevertheless, the measures were able to discriminate patients with both low and high ThD from controls.

7. General discussion

The goal of this research was to apply an automated, theoretically-based model of semantic content to the analysis of discourse. In testing LSA over a range of different types of patient discourse, the approach to quantifying incoherence was successful: LSA was able to detect differences between patient and healthy control groups and within the patient group itself in simple word generation tasks. LSA was also sensitive to very subtle deviations in standard

[4] Interestingly, we have found comparable accuracy of predicting group membership in responses to questions that we *apriori* expected to result in especially varied and unusual content (unpublished observations).

(lengthy) clinical interviews that were detectable by clinicians. Moreover, it was demonstrated that LSA could detect incoherent speech in patients with high levels of formal thought disorder when analyzing larger units of discourse (*e.g.*, five or greater words). Given the importance of this value in working memory capacity limits, it would be interesting to examine the relationship of this to working memory in future studies. Finally, LSA measures were successfully used to predict whether the discourse belonged to a patient or control as well as the actual clinical ratings of tangentiality, content and organizational structure.

We recognize and appreciate that there are several other well validated rating scales of thought disorder. For instance the Communication Disturbances Index (Docherty, 2005) emphasizes ambiguous referents in speech samples, while the Thought Disorder Index (TDI) (Solovay et al., 1987; Niznikiewicz et al., 2002) rates unusual word combinations and content (as obtained from Rorschach ink blot test responses and to explanations of proverbs). We believe that it would be useful to assess the relationship of LSA ratings of speech samples to these rating scales. Based on our understanding of how LSA works and empirical evidence from Experiment 4 demonstrating a relationship between LSA and content, we would predict that LSA to TDI correlations might be especially robust. We also note that the LSA studies performed here do not specify or favor any one candidate cognitive mechanism. In a sense, LSA may be largely agnostic to those cognitive mechanisms implicated in thought disorder.

A possible limitation of LSA might be that it is most sensitive to the differences between patients with high levels of ThD *versus* those with little or no ThD. However, the issue of sensitivity to the differences among low ThD patients and healthy controls is similar to that seen by blind human raters in our own data (in that the blind human ratings do not always accurately predict group membership). Choice of question may increase sensitivity (both for LSA and clinical raters). Indeed, this issue requires additional explicit evaluations of what measures work well on what types of questions and specifically what questions are best for eliciting various language disturbances of clinical interest. A non-obvious difference between LSA and human raters is that for the current analyses, LSA was trained on a corpus of text that represents the material more typical of written text, while humans acquire some of their semantic knowledge through exposure to spoken discourse (see Landauer and Dumais, 1997). Nevertheless, this same corpus has been used for other experiments related to spoken discourse (see Landauer et al., 1998). This limitation notwithstanding, our results nonetheless demonstrate that a model based on written text is able to capture much of the variance between the groups in spoken discourse. We recognize that our LSA measure could be monitoring illness severity, but severity is manifested in the linguistic domain and thus can be reliably assessed with LSA. This is because an important aspect of the severity of illness in schizophrenia is displaying disorganized speech (see *e.g.*, Allen et al., 1993). We have not analyzed our data with regards to gender as have others (Solovay et al., 1987); we know of no evidence suggesting that LSA would rate male and female language differentially. However, medication is clearly an important issue (but not addressed in the current experiments) that merits a future study specifically designed to have sufficient statistical power to address the potential effects of medication on LSA measures.

A key question is how important word order is when assessing coherence in language. Clearly for humans this is an important criteria, but this is not the case in some LSA-based analyses in which word order is not considered nor typically needed to assess the meaning of a unit of text (Landauer et al., 1997). Nevertheless, if one were especially concerned about bizarre word orders (in some language disorder for example), then one could use LSA to compare the coherence of each sentence to each next sentence, use a moving window approach, or add statistical models of normal word order as was done in this research. Words "scrambled" across the moving windows would generally result in lower cosines than normal text, thereby capturing some aspects of word order, although not all aspects of syntax. In addition, the LSA-based measures can be used in combination with syntactic measures, such as *n*-gram models, as was done in our analyses that predicted human scores of disordered thinking and categorized patients and controls (Jurafsky and Martin, 2000).

Therefore, LSA alone can capture many of the aspects of word order without directly considering word order, but additional measures can be employed if they are shown to provide additional predictive power. Furthermore, this combined approach means that the measures do not confound organization and content. In principle, models of schizophrenia could examine independently the role that semantic, syntactic, and other organizational discourse features have in the disease.

Our aim was to examine a patient group in whom speech problems are a hallmark in order to establish the sensitivity of our measure. Now that we believe

we have successfully done this in schizophrenia, the next step of course is to examine the specificity (in a future study) of the measures in other patient groups (*e.g.*, bipolar disorder, semantic dementia, Alzheimer's etc). Indeed, disturbances in content and coherence of formal thought and abnormalities of verbal expression are cardinal symptoms in a large proportion of neuropsychological disorders, ranging from traumatic brain damage, focal stroke, degenerative dementia to psychoses. We would expect different illnesses to produce a different array of performance profiles on various assays of coherence (as for example in the four studies reported here). However, the usefulness of these putative differences will depend on the ease of obtaining such measures (*i.e.*, quick and cost-effective) as well as whether such measures are sufficiently sensitive so as to be useful in the treatment process (*e.g.*, monitoring subtle but important clinical fluctuations) as well as in early stages of the diagnostic process (*e.g.*, risk predictive). By capitalizing on a modern computational linguistic approach to knowledge representation, we have demonstrated that it is possible to analyze large quantities and varieties of discourse for deviance in language for clinical and empirical purposes. We have developed novel analyses and models that both complement and are more objective (and sensitive) and reliable than clinicians' ratings.

We have presented a framework for analyzing discourse, and tests of an automatic tool with which to perform such analyses. As a framework, we have adopted a novel, but theoretically established approach to modeling semantics of discourse, focusing on such factors as the choice of words, expression of meaning, relatedness of discourse and coherence. The semantic structure derived by LSA can be successfully applied to modeling and measuring semantic content as expressed through speech. We have evaluated our approach by analyzing discourse from patients with schizophrenia, which suggests that disordered thinking occurs at both association and higher level planning stages. As a tool, we suggest that development of this more objective and reliable tool with which to assess coherence in discourse will be useful in clinical research. We have shown that by measuring semantic coherence we can differentiate patients from controls with reasonable accuracy, as well as determine the severity of incoherence in language. Improving and increasing the manner by which communication from a patient can be analyzed may ultimately improve the level of psychiatric intervention that a patient receives (*e.g.*, such as by accurate monitoring of treatment effects). This tool may be useful also in characterizing memory and language abnormalities and their relationship to incoherent discourse, and thus help establish how incoherence in schizophrenia, for example differs from that in other clinical conditions in which incoherent discourse is of interest (*e.g.*, in distractible speech in mania, and in the degradation of content in language longitudinally in Alzheimer's disease). Moreover, development of a mechanistic account of incoherence at the word and sentence level will enhance understanding of how patients represent, retrieve and use meaning in discourse, and provide a tool with which to search for anomalies in discourse that may be psychopathological risk factors.

Acknowledgements

We are grateful for the assistance from numerous people: Christina Bokat, Georgia Bushell, Michael Egan, Joscelyn E Fisher, Sara Gilliam, Kate Goldhaber, David Goldsmith, Thomas K Landauer, Jennifer Iudicello, K. Megan Kerbs, Joel Kleinman, Stefano Marenco, Meredith Melinder, Jim Parker, Andrew Schwartz and Thomas Weickert. The authors are grateful for the use of the website http://lsa.colorado.edu/ for generating a number of the analyses in the paper.

Contributors

Brita Elvevåg participated in the study design, data collection and analysis and the writing of the manuscript. Peter Foltz participated in the experimental design, and he contributed to the data analysis and the writing of the manuscript. Daniel Weinberger participated in the data interpretation and writing. Terry Goldberg participated in the study design, data collection and contributed to the data analysis and writing of the manuscript. All authors read and approved this manuscript.

References

Allen, H.A., Liddle, P.F., Frith, C.D., 1993. Negative features, retrieval processes and verbal fluency in schizophrenia. British Journal of Psychiatry 163, 769–775.

Andreasen, N.C., 1986. Scale for the assessment of thought, language and communication (TLC). Schizophrenia Bulletin 12, 474–482.

Andreasen, N.C., Black, D.W., 2005. Introductory Textbook of Psychiatry. American Psychiatric Association, Washington DC.

Andreasen, N.C., Grove, W.M., 1986. Thought, language, and communication in schizophrenia: diagnosis and prognosis. Schizophrenia Bulletin 12, 348–359.

Bamberg, M., Moissinac, L., 2003. Discourse development. In: Graesser, A.C., Gernsbacher, M.A., Goldman, S.R. (Eds.), Handbook of Discourse Processes. Lawrence Erlbaum Publishing, Mahwah, NJ.

Barch, D.M., Berenbaum, H., 1996. Language production and thought disorder in schizophrenia. Journal of Abnormal Psychology 105, 81–88.

Bleuler, E., 1911. Dementia Praecox or the Group of Schizophrenia. International Universities Press, New York. (English translation; 1950).

Bokat, C.E., Goldberg, T.E., 2003. Letter and category fluency in schizophrenic patients: a meta-analysis. Schizophrenia Research 64, 73–78.

Bowie, C.R., Tsapelas, I., Friedman, J., Parrella, M., White, L., Harvey, P.D., 2005. The longitudinal course of thought disorder in geriatric patients with chronic schizophrenia. American Journal of Psychiatry 162, 793–795.

Breier, A., Berg, P.H., 2003. The psychosis of schizophrenia. Biological Psychiatry 46, 361–364.

Clark, H., Haviland, S., 1977. Comprehension and the given-new contract. In: Freedle, R. (Ed.), Discourse Production and Comprehension. Erlbaum, Hillsdale, NJ, pp. 1–40. 24.

Cuesta, M.J., Peralta, V., 1999. Thought disorder in schizophrenia: testing models through confirmatory factor analysis. European Archives of Psychiatry and Clinical Neuroscience 249, 55–61.

DeLisi, L.E., 2001. Speech disorder in schizophrenia: review of the literature and exploration of its relations to the uniquely human capacity for language. Schizophrenia Bulletin 27, 481–496.

Docherty, N.M., 2005. Cognitive impairments and disordered speech in schizophrenia: thought disorder, disorganization, and communication failure perspectives. Journal of Abnormal Psychology 114, 269–278.

Foltz, P.W., Kintsch, W., Landauer, T.K., 1998. The measurement of textural coherence with Latent Semantic Analysis. Discourse Processes 25, 285–307.

Harrow, M., Marengo, J.T., 1986. Schizophrenic thought disorder at follow-up: its persistence and prognostic significance. Schizophrenia Bulletin 12, 373–393.

Harrow, M., Grossman, L.S., Silverstein, M.L., Meltzer, H.Y., 1982. Thought pathology in manic and schizophrenic patients: its occurrence at hospital admission and seven weeks later. Archives of General Psychiatry 39, 665–671.

Hoffman, R.E., 1986. Verbal hallucinations and language production processes in schizophrenia. Behavioral and Brain Sciences 9, 503–548.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

http://lsa.colorado.edu/.

Jastak, S., Wilkinson, G.S., 1984. The Wide Range Achievement Test-Revised administration manual, Rev. ed. Jastak, Wilmington, Del.

Jurafsky, D., Martin, J.H., 2000. Speech and Language Process: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, Upper Saddle River, NJ.

Kraepelin, E., 1919. Dementia Praecox and Paraphrenia (Barclay RM, Trans.). RE Krieger, New York. (Reprinted 1971).

Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240.

Landauer, T.K., Laham, D., Rehder, B., Schreiner, M.E., 1997. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In: Shafto, M.G., Langley, P. (Eds.), Proceedings of the 19th Annual Meeting of the Cognitive Science Society. Erlbaum, Mahwah, N.J, pp. 412–417.

Landauer, T.K., Foltz, P.W., Laham, D., 1998. Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284.

Lorch, R.F., O'Brien, E.J. (Eds.), 1995. Sources of Coherence in Reading. Erlbaum, Hillsdale, NJ.

Manschreck, T.C., Maher, M.A., Ader, D.N., 1981. Formal thought disorder, the type-token ratio, and disturbed voluntary motor movement in schizophrenia. British Journal of Psychiatry 139, 7–15.

McKenna, P.J., Oh, T.M., 2005. Schizophrenic Speech: Making Sense of Bathroots and Ponds that Fall in Doorways. Cambridge University Press, Cambridge.

Niznikiewicz, M.A., Shenton, M.E., Voglmaier, M., Nestor, P.G., Dickey, C.C., Frumin, M., Seidman, L.J., Allen, C.G., McCarley, R.W., 2002. Semantic dysfunction in women with schizotypal personality disorder. American Journal of Psychiatry 159, 1767–1774.

Overall, J.E., Gorham, D.R., 1962. The brief psychiatric rating scale. Psychological Reports 10, 799–812.

Solovay, M.R., Shenton, M.E., Holzman, P.S., 1987. Comparative studies of thought disorder. I. Mania and schizophrenia. Archives of General Psychiatry 44, 13–20.

Spohn, H.E., Coyne, L., Larson, J., Mittelman, F., Spray, J., Hayes, K., 1986. Episodic and residual thought disorder in chronic schizophrenics: effect of neuroleptics. Psychopharmacology 21, 582–587.

Wechsler, D., 1981. Wechsler Adult Intelligence Scale-Revised. Psychological Corporation, San Antonio, Texas.

Zwaan, R.A., Singer, M., 2003. Text comprehension. In: Graesser, A.C., Gernsbacher, M.A., Goldman, S.R. (Eds.), Handbook of Discourse Processes. Lawrence Erlbaum Publishing, Mahwah, NJ.