Published in final edited form as:
J Neurolinguistics. 2010 May 1; 23(3): 270–284. doi:10.1016/j.jneuroling.2009.05.002.

An automated method to analyze language use in patients with schizophrenia and their first-degree relatives

Brita Elvevåg(1),*, **Peter W. Foltz**(2), **Mark Rosenstein**(3), and **Lynn E. DeLisi**(4)

(1) Clinical Brain Disorders Branch, National Institute of Mental Health, National Institutes of Health, Building 10, Room 4S235, MSC 1379, Bethesda MD 20892.

(2) Pearson Knowledge Technologies and University of Colorado, Institute for Cognitive Science, 4940 Pearl East Circle, Suite 200, Boulder, CO 80301

(3) Pearson Knowledge Technologies, 4940 Pearl East Circle, Suite 200, Boulder, CO 80301

(4) Department of Psychiatry, New York University Langone Medical Center and The Nathan S. Kline Institute for Psychiatric Research, 650 First Avenue, New York, NY 10016

Abstract

Communication disturbances are prevalent in schizophrenia, and since it is a heritable illness these are likely present - albeit in a muted form - in the relatives of patients. Given the time-consuming, and often subjective nature of discourse analysis, these deviances are frequently not assayed in large scale studies. Recent work in computational linguistics and statistical-based semantic analysis has shown the potential and power of automated analysis of communication. We present an automated and objective approach to modeling discourse that detects very subtle deviations between probands, their first-degree relatives and unrelated healthy controls. Although these findings should be regarded as preliminary due to the limitations of the data at our disposal, we present a brief analysis of the models that best differentiate these groups in order to illustrate the utility of the method for future explorations of how language components are differentially affected by familial and illness related issues.

Introduction

Schizophrenia is widely regarded as a neurodevelopmental disorder in which damage to the brain occurs many years before the illness expresses itself in a florid fashion (Weinberger, 1987; Murray & Lewis, 1987). Therefore it is assumed that even though the actual illness emerges in adulthood, evidence of deficits in brain function is present early in life, albeit in a less dramatic form. Indeed, findings of cognitive weakness being present before illness onset provide strong evidence for abnormal cortical development (David, Malmberg, Brandt, Allebeck & Lewis, 1997; for a review, see Elvevåg & Weinberger, 2001). Schizophrenia is also considered to be heritable via a polygenic mechanism, such that multiple genes exert relatively small effects that exceed a liability threshold (for a review, see Cannon, 2005). Therefore some similar deficits should be evident in family members, specifically first-degree relatives, although in a muted form.

*Corresponding author: Email: brita@elvevaag.net.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The bulk of this quest for deficits, in both probands as well as their unaffected relatives, has generally focused on cognitive domains (e.g., working memory, episodic memory, attention) that are considered to be at the very core of the pathology (Bilder et al., 2000; Egan et al., 2001; for reviews, see Elvevåg & Goldberg, 2000; Kuperberg & Heckers, 2000). Since deficits may index genetic liability, they are considered to be candidate intermediate phenotypes for schizophrenia and may be predictive of who develops the actual illness (e.g., see Aukes et al., 2008). Thus, even though schizophrenia is associated with a wide range of symptoms and cognitive deficits (all of which vary in terms of their frequency, predictive validity, specificity, course and amelioration by neuroleptic medication), it is deficits in cognition that have been regarded as the enduring feature of the illness, and has recently become the target for medication and treatment intervention (Marder & Fenton, 2004; Kern et al., 2008; Nuechterlein et al., 2008).

Within this approach, language variables have generally been represented by measures of vocabulary knowledge, reading pronunciation, and counts of the ability to generate as many words beginning with a specific letter or belonging to a specific category in a fixed period of time (e.g., 1 minute; for a meta-analysis, see Bokat & Goldberg, 2003). These measures provide very limited windows into language ignoring most aspects of communication, and category fluency for example is more likely tapping into verbal memory than language per se. Despite these rather narrow views of language, there have been some interesting findings. A recent meta-analysis of the cognitive deficits in unaffected first-degree relatives of schizophrenia patients found that of all the cognitive measures examined the largest effect size was with category fluency (d=.68; although this effect disappeared with more rigorous inclusion criteria, see Snitz, MacDonald & Carter, 2006). It is possible that examining the structure within the output of this fluency (i.e., the actual semantic search process itself) may provide useful clues concerning the underlying mechanisms. There is also much literature that adopts a wider approach to examine communicative (rather than linguistics) variables - such as features- in schizophrenia (e.g. see Gernsbacher, 1999 for an overview). Communication analysis is therefore likely to be of enormous value in elucidating the underlying vulnerabilities in this cognitive structure, since communication is a high-level cognitive function that provides a rich and extemporaneous data set reflecting the state of numerous underlying cognitive processes. The pattern and content of communication provides large amounts of information that can be traced back to individuals' cognitive abilities, knowledge and consequently overall mental state.

An additional advantage of a focus on robust measurement tools of cognition is that they can be used to more specifically define and explore the underlying psychopathology of the disorder and also focus specifically on aspects peculiar to schizophrenia, such as disorganized thinking, as evidenced by disorganized speech. Although it may be argued that 'unconventional' use of language is simply a characteristic of the acute psychotic state and subsides when the psychosis does, studies show that even in the stable state, several characteristics of language processing are not 'conventional' in people with schizophrenia (e.g., Li et al., 2007a,b; Sommer et al., 2001, 2003). However, if 'unconventional' use of language are trait abnormalities the assumption is that communication, as a complex combination of cognitive processes, may account for some of the genetic burden if it can be usefully assayed. Indeed, there is a strong theoretical rationale for analyzing language samples from individuals at high genetic risk for schizophrenia, as the neural pathways for language processing are likely related to the underlying pathophysiology of the disorder (DeLisi, 2001; Li et al., 2007a,b).

Since brain pathology is already detectable by the time of first episode (and probably progressing in the prodromal phase before symptoms appear), one goal would be to detect a range of subtle discourse deviations in individuals prior to the emergence of overt symptomatology. The clinical importance of this link is illustrated by recent studies reporting that early detection and treatment of some signs of illness in adolescents may prevent poor outcome and the progression of symptoms to frank psychosis (McGlashan, Miller & Woods, 2001; McGlashan, Miller, Woods, Rosen, Hoffman & Davidson, 2003a; McGlashan et al., 2003b; Woods et al., 2003). Similarly, although nonschizophrenic relatives of patients are not thought disordered, there are significant deviations as compared to healthy unrelated comparison participants on subtle measures of communicating meaning. This includes peculiar verbalizations (Shenton et al., 1989) and reference failures in speech (Docherty et al., 1999; 2004). "These communication variables, and especially referential disturbances, involve the use of vague, unstable, or idiosyncratic concepts. They also have been characterized as demonstrative of a lack of awareness on the part of the speaker of the perspective of the listener, of what the listener needs to understand the speaker's meaning" (p.399; Docherty et al., 2004). Moreover, it has been shown that patients with chronic schizophrenia tend to produce fewer words and less complex sentence structure than either their well siblings or controls and that those in early stages of their illness display less 'conventional' language use (DeLisi, 2001; Shedlack et al., 1997). However, very early on they likely process language in a 'different' manner than controls from the general population as can be visualized by a reduction in the 'usual' left-sided language lateralization in people with chronic schizophrenia and in their well siblings at high-risk for developing the disorder (Li et al., 2007a,b). One possible explanation of this difference is that a progressive brain structural process is occurring beginning early in the development of the disorder, detectable as a functional, but not obvious structural deficit, but then later in the illness a progressive frontal and temporal lobe deterioration is detected by enlargement of the lateral ventricles and loss of brain volume (DeLisi et al., 1995, 1997b), although this hypothesis remains controversial to date.

Automated Communication Analysis

Recent advances in computational linguistics and statistical-based semantic analysis have shown the power of automated communication analysis. For example, computational language analysis techniques have become the basis for improving search engines (e.g., Manning, Raghaven & Schütze, 2008; Berry & Browne, 2005) and spam filtering (e.g., Zang, Zhu & Yao, 2004), for performing automated analysis of team communication, as well as for improving computational models of cognitive processing. Computational language analysis combines language features with machine learning techniques to associate these features with aspects of human cognition or performance. A wide variety of predictive language features have emerged across diverse domains such as education (e.g., Graesser et al., 2004, text analysis in psychology - reviewed in Pennebaker, Mehl & Niederhoffer, 2003) and tasks in natural language processing (surveyed in Jurafsky & Martin, 2008). For analysis, it is often useful to divide the features into three classes that have frequently been found valuable in modeling cognition and performance with language data. The first class of features is often described as surface features and provides metrics on the surface level of the language, such as word counts, mean syllables per word, mean speech rate, and mean sentence length. The second class derives from information theoretic statistical features such as n-gram likelihoods (first formalized by Shannon, 1948; for a historical perspective see Pierce, 1980; and for a more modern treatment see Chapter 6 of Manning & Schütze, 1999). This class of features measures how likely word patterns in a passage would occur in large samples of English text, as well as other regularities in the patterns of language use, giving a measure of how "English like" the patterns are. For instance, n-gram likelihood-based measures use the correlation structure of language estimated from large corpora of text to detect the probability that a word (or set of words) may follow one another. These features capture such aspects of language as syntactic complexity, flow and word choice. The third class of features involves statistics-based semantics measures, based on Latent Semantic Analysis (LSA). LSA is a cognitive modeling tool and computational technique for matching discourse content (for technical details see Landauer, Foltz & Laham, 1998). Its special capabilities for communications analysis are: (i) that it can represent the whole conceptual content of verbal communication rather than surrogates such as keywords, titles, abstracts, or overlap counts of literal words; (ii) that its representation of the similarity of two words, sentences, passages, or documents closely simulates human judgments of the overall similarity of their meanings; (iii) and that it assesses two passages on the same topic but phrased in different vocabulary as being semantically similar. The similarities detected by LSA can be quite subtle, permitting quantitative comparisons of different participants' semantic coherence as well as their choices and combinations of words. In the present approach we measure the degree to which participants from the same experimental groups cluster in terms of their use of semantically related communication. This is accomplished by using a nearest-neighbor technique in which we compare one participant's response to other participants' responses in which we know which experimental group the other participants are from. For example, this approach provides measures of the degree to which a participant's discourse is more like that of a patient or a control (see Eastman & Weiss, 1978). The nearest neighbor technique used in the modeling described here is called the k-near technique (see Landauer, Foltz & Laham, 1998 for more details on the application of this approach).

LSA-based methods have been evaluated favorably within the cognitive, commercial training and assessment, and clinical domains. In terms of cognition in general, LSA produces good approximations to human cognitive semantic relations (for modeling language acquisition, semantic priming, semantic categorization and the effects of text coherence on comprehension, see Foltz, Kintsch & Landauer, 1998; Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998). Computational language analysis has been used as the critical component in successful commercial applications including automated essay scoring and information retrieval systems. However, for current purposes we emphasize that this technique has proven to be a valid assay of disordered language production in schizophrenia, used to both complement human clinical ratings as well as to experimentally parse this incoherence in a theory-driven manner.

A clinical "proof of concept" was established in our previous work (Elvevåg, Foltz, Weinberger & Goldberg, 2007). This initial study of the technology's applicability to clinical contexts - specifically operationalizing and indexing thought disorder in patients with schizophrenia - demonstrated that LSA can be used effectively to evaluate psychiatric patients based on open-ended verbalizations. We used speech samples from various sources, including word associations, verbal fluency, and different types of discourse: narrative speech (i.e., storytelling) and expository speech (i.e., descriptions of abstract concepts). The analysis methods included comparisons of the groupings of words either within or across speech samples, including assessment of overall coherence, interactions of the grouping parameters and coherence, comparisons with human-rated diagnoses, relationship to diagnostic category, and relationship to cognitive test scores. We also have preliminary positive results from unstructured interviews across various topics. Although not as sensitive as the structured interviews, the results demonstrate the impressive and diverse potential of these methods.

In addition, a number of ongoing studies suggest that computational language analysis is a robust and objective measure of 'unconventional' language use in schizophrenia, and has the potential to be sensitive to different levels of language disturbance when specific probes that target different levels of conceptual organization are used. Specifically, these studies and simulations have shown that computational language analysis can effectively evaluate patients with schizophrenia based on open-ended verbalizations. Overall, these automatically derived language scores have distinguished patients from controls surprisingly accurately (and patients from other patients) as well as predicted clinical ratings, using both large discourse samples as well as using responses that constitute only a few words.

Modeling approach

In the present study, we adopt the modeling approach used in Elvevåg et al., (2007). In this approach, the goal is to identify communication features which reliably distinguish participants from different experimental groups. The communications from the participants are first analyzed into sets of language features. These features comprise variables representing surface features, statistical features and semantic-based features. The variables from these three classes of features are then used as predictor variables and discriminating category (e.g., experimental group) is used as a response variable within a linear discriminant analysis in order to measure how well the language features can distinguish between groups.

While all three classes of features are important for modeling communication, we have found the statistical-based semantic features often provide the most sensitive distinctions (see Foltz, Martin, Abdelali, Rosenstein & Oberbreckling, 2006). It is noteworthy however, that within each of the three categories of communication features, many of the individual features are highly correlated. This is due to the fact that the level of syntactic and semantic information in communication can be highly related. This does imply that it is best to interpret a model at the higher level of surface, statistical and semantic features since attributing strong significance to the exact features chosen for the model is less informative than understanding the nature of the participation of the larger classes of language features that are predictive of the underlying cognitive deficits.

In this paper we re-analyze data collected by DeLisi and colleagues by adopting this automated modeling approach. As will be discussed, there are limits to the modeling possible with this data, but even within the limitations we can develop models that separate populations of interest. We opportunistically use this data to model an additional population set, to build a deeper understanding of the models, and to analyze the characteristics of the models that drive the separation. Our research involves combining automated communications analysis with the language deviations observed in schizophrenia to derive statistical models that may relate language features to underlying biologically relevant factors.

Methods

Participants

We analyzed speech samples from a series of studies of DeLisi, some of which have been published (DeLisi et al., 2001; Shedlack et al., 1997) and were transcribed into an electronic format (N=83; 53 patients and 30 controls). Participants were asked to talk about whatever came to mind, perhaps what they did yesterday or what they would like to be doing. Individuals were selected as part of two large cohort studies, one a family study of individuals from families with a high density of schizophrenia (see DeLisi et al., 2002 for cohort details) and another a longitudinal study of first episode patients with schizophrenia over time (see DeLisi et al., 1997b for cohort details). Three people were trained to conduct the interviews, however, different testers did not test different cohorts. The participants were randomly distributed among testers. The participants were classified as Family or non-Family, with the category Family being further broken down into Well or Patient. The Non-Family category was further broken down into Control or Patient (see Table 1).

Design (and Caveats)

The following series of analyses were conducted on an existing set of data that was collected over time and contains language differences resulting from the varying purposes and contexts driven by the original research goals. Different probe questions were used depending on the ability of the participant to spontaneously produce verbal material for the family and non-family groups, the latter where the bimodal distribution of word density suggests that the test was administered slightly differently within groups, specifically most non-family participants were asked to tell a story about what they did today and what they like, while most family participants produced quite heterogeneous answers. For instance the Family data appears more homogeneous than the non-Family data using just simple surface features such as word count. Of note are the large differences in the lengths of the speech samples by the family group and the non-family group, with the family speech samples being ~300 words shorter (see Table 1). These limitations impact the language-based analysis we can perform with this dataset, since analysis across the family, non-family divide risks confounding the different experimental conditions and the disease related signals we are attempting to detect. Specifically, building a model of language features that provides a three-part split between patients, well family members and controls is neither possible nor reasonable with this data. Instead, we turn to two-class classification models and their analysis to understand classes of language features that separate various paired categories of participants. This resulted in developing four separate models which use language features as predictor variables and discriminating category as the response variable to measure how well language features can distinguish between groups.

The comparisons least impacted by the differing experimental conditions are separations within each of the family set and the non-family set, so we report on a model that separates Patient family participants from Well family participants and a model that separates Control non-family participants from Patient non-family participants. Using the entire dataset, we report on a model that separates all patients from all non-patients, Patient family participants + Patient non-family participants from Control non-family participants + Well family participants. Note that this model includes the introduced noise of the different experimental conditions, but is not confounded with them, so any separation will be attributable to disease related features. Finally, we mention an analysis to separate Control non-family participants from Well family participants. While the language features that allow this separation are of great interest, using this dataset we cannot tease out the disease factors from the experimental design factors, so more definite results must wait for new data.

Preparation

The data consists of 83 speech samples, one per participant. Each speech sample was transcribed by a single transcriber which provided a consistent basis for the transcriptions. Questions or prompts from the interviewer, were eliminated, but the meta-comments of the participants were left in (e.g., “Keep talking about what I like?”). Presented below are samples of approximately the initial 100 words of the transcripts from two participants, the first a patient and the second a control to illustrate the type of transcript data used in modeling.

> Patient:
>
> “I like to play basketball do you want me to talk about it I like doing lay-ups uh I like the three point shot I like watching basketball I just watched the NCAA championship but it could have been a closer game I I I was glad that the two teams that were in it I thought it was going to be a good game Michigan Bob Buschman I thought they had a good team and Duke. Duke’s a great team Christian Laetner Bobby Hurley I like the NBA too I like to watch that I like the knicks the knicks are my favorite team”
>
> Control:
>
> “I like to go water skiing uh um uh I enjoy the ocean uh salt water uh and the sunlight particularly in the summertime I like going over jumps while skiing uh I like sharp cuts and the spray of the water and the sun glistening through the water the light glistening through the water Seeing fish in the ocean, and the smell of the outboard motor um um I I don’t know sometimes the am a little I am a little afraid of sharks

> when you see the fish in the ocean I am a little concerned about the sharks um so I stay close to shore as I can uh sometimes scaring the swimmers”

The transcriber retained some of the speech disfluencies, such as “um” and repeated words, and provided some punctuation as well. No attempt was made to use the disfluencies as features in the models reported here though that is a path for further investigation. Values for surface level features, such as word count, and information theoretic features, such as n-gram likelihood, were derived directly from the raw transcript text, while for semantic features the text was first normalized, with punctuation removed and text lowercased before semantic analysis.

Analyses

Our previous work (Elvevåg et al., 2007) led us to predict that we would be able to detect subtle differences between patients with schizophrenia and healthy control participants, and likely between probands and their unaffected relatives. We now sought to address why this is the case, namely which classes of language features make the models more or less sensitive to these familial and/or illness issues. In order to address this issue, we built models to discriminate between the most significant categories of participants resulting in four models. In all cases, linear discriminant analysis (LDA), often referred to as classical Fisher discriminant analysis (Fisher, 1936), was employed using the R statistical environment (R Development Core Team, 2008).

Models were constructed using a feature set based on each participant’s speech sample consisting of surface features, statistical language features, and semantic features. The semantic features associated the meaning of the discourse with whether the discourse was composed by a patient (P=Patient family participants + Patient non-family participants) or non-patient (Well family participants + Control non-family participants). A k-near feature first determines the set of semantically most similar transcripts to a given transcript. This set of nearest neighbors is then used to compute the fraction of nearest transcripts that were in the Patient category. This fraction provides a measure of how similar the given transcript is to patient transcripts. The rationale behind this feature is that it detects how well any response shares meaning with clusters of responses from the different categories and can then be used to help discriminate into which category a response falls.

By utilizing dimension reduction, LSA can reveal elusive similarities and differences of language semantics. Our earlier work (Elvevåg et al., 2007) has indicated that language differences between those in the Patient category and those without overt symptoms should separate in the lower dimension semantic space, so that the k-near transcripts of those in the Patient category should be comprised of mostly Patient transcripts, and the k-near set of those in the Control category, should mostly come from Control transcripts. The following excerpts from a Patient transcript indicate some of the semantics that LSA may be selecting in its multi-hundred dimensional representation.

> “I came from Lebanon I was born six four nineteen seventy-four I left my home in nineteen eighty-seven on April nine. Now I’m in the United States I have been here for several years, uh, I’m in uh I spent one month in the hospital for mental illness now I’m in day hospital and have been in day hospital for three weeks. Uh I’m doing better. I visited my country two years ago, I I stayed there for five months, then I came back. … I joined the ROTC Air Force in Philadelphia for two years and I also played for two years I was linebacker inside outside linebacker. I have a cousin who is in the Navy who is an officer. OK my sister got married in nineteen eighty-eight in California. … I visit Statue of Liberty. I worked at McDonalds in Florida I worked in Jiffy Lube in New York. I worked at Four Sons in New York. In Jiffy Lube I broke

> my hands so I brought my I saved it for two months but then I broke my hand again … I was going to go in the Air Force and be a pilot but I can't I'm too big, too tall. My favorite hobby is soccer. I used to play soccer in my country and and they would not let me to play football."

Two features of this text that LSA may be using to segregate Patient transcripts are the mention of medical treatment that occurs more frequently in Patient transcripts and the large range of topics covered in the transcript. There are certainly other semantic features that LSA is detecting, but these are two that stand out, and the strength of the LSA approach is in allowing the mathematics to detect the differences among the transcripts from different categories of participants.

After computing the values of the features, model selection is then used to determine a subset of the features that maximized the percentage correct prediction (the exact agreement) using stepwise linear discriminate function analysis (LDA). Model selection does not necessarily lead to a unique solution. The same level of overall performance can be obtained by trading off performance between each of the categories, and in cases of equal overall performance our preference is for good discrimination for each of the categories. If there are compelling reasons where misclassification in one category is deemed less desirable, it is possible to set up a cost function that would favor that category, though for the current study this avenue was not pursued.

The naïve classification rate computed as the exact agreement of the data used to build the LDA model to the model predictions (known as the resubstition estimator) is optimistic, that is, it is biased toward a higher classification rate (McLachlan, 1976). All predicted classification rates reported derive from cross-validated predictions. Cross-validation provides a more conservative, but better measure of the ability of the models to generalize to additional participants (for a general discussion of these issues, see for instance Yu, 2003). In the version of cross-validation used here, linear discriminate analysis models are built each using all the discourse samples except one, the sample to be predicted. From the hold-one-out models, the category of the excluded speech sample is predicted. The agreement number that is reported and used in the analysis is based on these held-out predictions. For each of the modeling tasks, we present summary statistics about the performance of the model, and a description of the model. All of the models required three or more features (independent variables) to obtain the reported discrimination making it difficult to visualize the separation between the two categories. Since for all models the data is imbalanced with different base-rates for the different categories, we report both the overall classification accuracy, and the accuracy for each category.

(1) Differences in discourse between patient family participants and well family participants: Within family group—First we analyzed discourse within the family group in order to differentiate the probands from their unaffected relatives (Patient family participants versus Well family participants). Naturally, any modeling approach must be able to differentiate obviously ill people from those who appear well. (With reference to Table 1, approximately a similar amount of speech was produced by these two groups). With our modeling approach 87.5% (21 out of 24) patients and 81.8% (9 out of 11) well family members were correctly classified giving an overall cross-validated classification accuracy of 85.7%. The confusion matrix is shown in Table 2 demonstrating good categorization in both categories despite the smaller overall number of well family participants.

Five variables were necessary for the model to be sensitive to these diagnostic differences, three semantic variables and two surface level variables. Although this was a five dimensional model, a subset of two variables is sufficient to provide reasonable separation of the two groups.

As Figure 1 shows, with only two of the five variables from the model, there is not the full segregation provided by using all five variables, nonetheless the figure is quite informative. The y-axis is a feature from the statistical-semantics group, a k-near feature with larger normalized values indicating the nearest neighbors of a transcript are more similar in meaning to patient transcripts. The x-axis is a feature from the surface features group with higher normalized values indicating increased use of content words. Note that the figure shows the majority of Patient family participants' discourse samples are drawn toward the upper left, indicating the semantic closeness of these samples to other patient samples on the y-axis scale while showing relatively lower use of content words in contrast to the Well family participants group on the x-axis. The measure of the relative use of content words can indicate the use of more specific, focused discourse as contrasted with very general discourse that does not convey a lot of meaning. The remaining variables, two semantic variables measuring the relative amount of content in the transcript, and the last a surface feature helped fine-tune the model to better parse the discourse samples. Conceptually, a greater variety of both syntactic as well as more complex higher level semantic coherence measures were necessary in order to detect discourse differences between probands and their unaffected family members.

(2) Differences in discourse between patient non-family participants and control non-family participants: Not within a family group—Second we computed a similar comparison (for which there was approximately a similar amount of speech produced by the groups – see Table 1) within the non-family group, attempting to differentiate the controls from the patients (Control non-family participants versus Patient non-family participants), which again any modeling approach must be able to do. Our modeling approach was able to correctly classify 84.2% (16 out of 19) unrelated healthy controls and 82.8% (24 out of 29) patients giving an overall cross-validated classification accuracy of 83.3%. The confusion matrix is shown in Table 3 indicating that the model quite accurately separates the patient non-family participants from the control non-family participants. Seven variables were necessary for the model to be sensitive to these group differences. Despite this model only sharing one feature with the previous model (the semantic feature plotted in Figure 1), the model consists of three semantic features, two surface features, and two statistical language features while both this model and the previous model can separate the two groups, and use similar classes of features, it appears the different experimental conditions cause different sets of exact features within each of the groups.

(3) Differences in discourse: All Patients versus all non-Patients—Third we analyzed the entire sample of 83 transcripts (Well family participants + Control non-family participants versus Patient participants), by combining all the patient transcripts (n=53) and all the non-patient transcripts (n=30). With reference to Table 1, although the word count is not balanced across the groups, nonetheless both groups (Well family participants + Control non-family participants versus Patient participants) cut across speech sample type, and thus any success with parsing the groups cannot be simply an artifact of different experimental conditions. Our modeling approach was able to correctly classify 86.8% (46 out of 53) patients and 60% (18 out of 30) healthy individuals (well family members and unrelated healthy controls) giving an overall cross-validated classification accuracy of 77.1% The confusion matrix is shown in Table 4. Three variables were necessary for the model to be sensitive to these group differences. Two of these three variables are from the statistical language features set and the other was a surface level feature. Conceptually, it is interesting that in this case quite basic language features (more "surface" level) features were all that was needed to detect that the speech belonged to a patient, which fits well with clinical observation, namely that some very basic things in the speech of patients seem different (e.g., slightly unusual choice of words and shorter sentences). It is worth noting that the agreement for the healthy individual group underperforms the previous two models. One hypothesis is that each of the two previous models could take advantage of a more consistent semantic environment and use semantic features to better separate the two groups at the cost of more complex models supporting our concern that the multiple purposes of the original data gather goals may conflict with an attempt to model with all the data. In this case, because the data spans two different conditions, the semantic variables are unable to add much separation power, but we end up with a simpler model. This suggests that for coarse-grained separations, quite simple models might suffice, but for fine-grained analysis best results will be obtained with carefully focused probes likely yielding more complex models.

(4) Differences in Discourse between Unaffected Relatives and Unrelated Controls—Fourth, we compared Well family participants versus Control non-family participants and the model was highly accurate with 89.5% (17 out of 19) of unrelated healthy controls and 90.9% (10 out of 11) well family members being correctly classified giving an overall cross-validated classification accuracy of 90.0%. The confusion matrix is shown in Table 5. Three variables were necessary for the model to be sensitive to the group, two semantic and one surface measure. This model shares the semantic feature and the surface feature with the first model that are plotted in Figure 1. While this is a simple, well performing model, due to the data we can not distinguish if the performance is due to the differing experimental conditions or characteristics of the participants. We present this result because it is interesting, but can only look to better datasets to see if this result holds.

Discussion

We have opportunistically reused an existing dataset to understand speech differences among patients with schizophrenia, their well relatives and controls by using methods from automated communication analysis. Despite limitations of an experimental design not specifically intended for communication analysis, overall, our modeling approach clearly demonstrates that it is possible to obtain an accurate discrimination of the groups based on using three types of measures, namely measures of statistical language features, measures based on the semantic similarity of a discourse sample to patient or control discourse sample, and surface features of the discourse such as sentence length or variability as measured by numbers of words or syllables. The goal of the present study was to illustrate that modeling approach can perform such discrimination and show that these classes of features can be computationally derived. Future studies will be able to fine-tune these measures and better indentify and classify the interactions of the communication features that best distinguish classes of participants based upon larger and more ‘custom-designed’ language samples.

As in previous findings (e.g., Elvevåg et al., 2007), semantic measures contributed the most to the disciminant models. From a psychological perspective, such measures are assessing aspects of the content being discussed in a manner similar to other members of the same group. Language use in terms of what participants choose to speak of in open-ended questions and their specific word choice tend to be more similar than the discourse generated by participants in the other group. The fact that such measures discriminate indicates that subtle semantic aspects of language are reflected between groups of patients and controls and within family members. While some of this effect possibly could be attributed to life experiences, socio-economic status, and education, the fact that it can discriminate between groups, both between and within family, suggest that the measures can detect the subtle psychological differences that are reflected in language use, and that have been reported using other methods (e.g., Docherty et al., 2004, 2006).

While semantic features played a prominent role, surface features contribute as well. Concerning the use of sentence length measures, these were either number of words per sentence, syllables per sentence or a count of capitalized words which counted use of beginnings of sentences and proper nouns. Given that the current study employed transcribed data, the use of measures representing structure imposed by the transcriber in models is certainly not ideal. However, we suggest that even transcribers often are able to distinguish one idea from the next and so are annotating some information about the size of "idea units" (e.g., semantic propositions, Kintsch 1988) that the current assay uses. Future studies involving speech recognition may not have punctuation and certainly not at human reliability and thus will likely want to include some variant of this current measure such as measuring of pauses or clauses that seem to hang together. However, for current purposes we consider sentence length useful since the transcriptions did look fairly regular in how idea units were broken up. Nonetheless, sentence length variables did not enter as the first or best variable in the models, but for some models they did help improve them, thus indicating that as a measure sentence length does indicate that size of idea unit can play a role in discriminating groups.

Observations on a few transcripts that were misclassified will help delimit the boundaries of this technique. In the model that separated Unaffected Relatives and Unrelated Controls, the following excerpt is from a participant in the Control category that was incorrectly predicted to be in the Relative category.

> "Ok, um, well, I like to, uh, play softball when I get a chance to Um, I belong right now to the, uh, Men's Club Softball league and they play every Sunday. We, uh, it's a short season. I used to belong to two other leagues and I had played for my former company's team … I like to do is, uh, working with computers which is what I do for a living. Uh, I have a PC at home. And uh, I've been doing some consulting work, uh, which, uh, is helping out with the new kid since my new wife is, uh, stepped, uh, uh, working. She's a special education teacher, um, but uh, so I've been pretty fortunate. I've been able to work at home, on the computer while, uh, my wife's been home with the baby … and, uh, mixed in with that I have other games on the computer, uh, I just bought a Star Trek game which is pretty wild. It um, uh, could say the twenty-fifth anniversary edition and it's uh, a, um, kind of a recreation of the old series. … I've been telling' my brother about that; he's enjoyed it, uh, from what I told him, he, uh, likes Star Trek as well. Uh, I know they just had the, uh, convention in Nassau Coliseum. But, uh, wasn't able to go down there. I heard it was pretty wild because it was the first time they had, uh, William Shatner and Leonard Nimoy together at one of these conventions in, uh, quite a while. …, and. Let's see. what else do I like to do.. uh Let's see um. Well, I've done a lot of work around the house. … We've uh when we first moved in, uh, we re- did the kitchen, uh, we totally gutted everything with the help of my, uh, father-in-law and, uh"

Syntactically, this participant's use of "uh", connectives and function words is closer to that of the Relative category than the Control category. That coupled with the range of topics, which causes the k-near semantic feature to have more Patient transcripts as nearest transcripts cause the model to predict that this participant is in the Relative category. For the other case where the model incorrectly predicts a Control as a Relative, the semantic variables again have a larger proportion of Patient transcripts. In this case, the participant discusses an upcoming operation which for which the medical content may have caused the transcript to be nearer to Patient transcripts than a Control transcript should. The third and last misclassified transcript from this model was a Relative incorrectly identified as a Control. Excerpts from that transcript do not clearly indicate why the model generated an incorrect prediction.

> "The reason why I'm in this I'm in this country is because I have to be here. Uh we faced a lot of troubles back in my country we lost the house, my parents lost their jobs and we had that's all we had to move to to Beirut which is is the city and we lived like for two years and a half … After we lost the house the properties and everything so at that tune my dad decided to go to Cyprus to get us the visa the American visa

> in which was which was he did so and finally decided to come … So, um this country I mean I'm not saying I don't like this country I do like this country this country in this country there is a lot of challenge and a lot of opportunities to do and uh and I appreciate that and still everyday I still have that someday I will go back to my country to my own, house and I was born there so, that's it... Alright when me and my sister first came to this country that was in nineteen eighty seven and the first week we got here the next week her um husband proposed to her and in one week she said yes. I mean that was like I consider it as a arranged marriage but my parents didn't force her to do that they they bought the guy to her and they let her talk with her and go out with turn and, but it was fast I mean in one week everything was was fast. … but um I um my mom's health couldn't help her keep on with my brother … I didn't think of myself I thought of my brother's situations the one who got sick and uh came to New York and we start taking him to a special treatments and I think things start he's he's OK compared to the old days that we you know …"

The model incorrectly predicted that this transcript was from a Control participant. Notice that despite mentioning health issues, which tend to be associated with Patient transcripts, the model features, both syntactic and semantic predict a Control participant. A conjecture is that there is a range of expression within the Relatives group and this participant is at the Control end, but any attempt to measure such a range awaits a more controlled dataset.

In our previous study (Elvevåg et al., 2007) we found that with the appropriate choice of question types, LSA could predict the presence of schizophrenia from the responses about as well as trained clinicians. This issue of question sensitivity was formally assessed by comparing the accuracy of our automated approach to predicting scores by raters of content and the accuracy of predicting group membership on responses to three questions that we a priori expected to result in especially varied and unusual content (Elvevåg et al., unpublished observations). In order to elicit expository discourse, participants (23 patients with schizophrenia and 22 healthy controls) were asked to describe color to a blind person, sound to a deaf person and to describe an especially strange personal experience. LSA-based measures were derived to generate a predictive model of the human ratings of patients and controls (of organizational structure, tangentiality and content) of these transcripts using stepwise regression. Using discriminant function analyses to predict group membership (patient or control), we found that LSA was able to correctly classify 83.7% (81.4% with cross-validation) of the responses to the strange experience questions, 82.9% (80.5% with cross-validation) to the color questions and 72.1% (69.8% with cross-validation) to the sound questions. The classification results in the present study therefore are not greatly different from those in the previous study using different questions and populations. Indeed, in line with previous results that we took to suggest that different kinds of probes may be more suitable for doing more fine-grained versus coarse grained detection, here too in our current study we find that for coarse-grained separations simple models suffice, but for fine-grained analysis more carefully focused probes will yield more complex models and hence better precision. 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 detecting both between and within group differences. Indeed, choice of question may increase sensitivity both for automated analyses and clinical raters alike.

As discussed above there were notable limitations with the current study. Additionally, we were not able to systematically explore the possible role of medication on the current results. Rather, what we have provided here is a framework for future analyses where issues such as the role of medication if of primary interest can be studied where there is sufficient statistical power to do so properly. Additionally, we were not able to examine the relationship of the current findings (i.e., models) to the neuropsychological profile of the various groups or medication status. This is likely to be of enormous relevance (e.g., Docherty et al., 2006), and the framework we have presented is well suited for such a future study.

Finally, we have implicitly assumed that language-use (‘conventional’ or ‘unconventional’) is attributable to differences in underlying biology, but we have not explicitly examined to what extent putative differences can be ascribed to other factors (e.g., cultural, social or psychological). Given the centrality of communication in the diagnoses and treatment of psychiatric patients, this methodological issue illustrates the critical importance of a multidisciplinary approach to understanding communication patterns in order to be able to establish to what extent these factors are attributable to biology and illness.

Acknowledgments

This research was supported in part by the NIMH Intramural Research Program.

References

Aukes MF, Alizadeh BZ, Sitskoorn MM, Selten JP, Sinke RJ, Kemner C, Ophoff RA, Kahn RS. Finding suitable phenotypes for genetic studies of schizophrenia: Heritability and segregation analysis. Biological Psychiatry 2008;64:128–136. [PubMed: 18295748]

Berry, MW.; Browne, M. Understanding Search Engines: Mathematical Modeling and Text Retrieval. 2. SIAM Books; 2005.

Bilder RM, Goldman RS, Robinson D, Reiter G, Bell L, Bates JA, Pappadopulos E, Willson DF, Alvir JMJ, Woerner MG, Geisler S, Kane JM, Lieberman JA. Neuropsychology of first-episode schizophrenia: initial characterization and clinical correlates. American Journal of Psychiatry 2000;157:549–559. [PubMed: 10739413]

Bokat CE, Goldberg TE. Letter and category fluency in schizophrenic patients: a meta-analysis. Schizophrenia Research 2003;64:73–78. [PubMed: 14511803]

Cannon TD. The inheritance of intermediate phenotypes for schizophrenia. Current Opinion in Psychiatry 2005;18:135–140. [PubMed: 16639165]

David AS, Malmberg A, Brandt L, Allebeck P, Lewis G. IQ and risk for schizophrenia: a population-based cohort study. Psychological Medicine 1997;27:1311–1323. [PubMed: 9403903]

DeLisi LE. A prospective follow-up study of brain morphology and cognition in first-episode schizophrenia patients: preliminary findings. Biological Psychiatry 1995;38:349–360. [PubMed: 8547454]

DeLisi LE. Speech disorder in schizophrenia: review of the literature and exploration of its relation to the uniquely human capacity for language. Schizophrenia Bulletin 2001;27:481–496. [PubMed: 11596849]

DeLisi LE, Sakuma M, Kushner M, Finer DL, Hoff AL, Crow TJ. Anomalous cerebral asymmetry and language processing in schizophrenia. Schizophrenia Bulletin 1997a;23:255–271. [PubMed: 9165636]

DeLisi LE, Sakuma M, Tew W, Kushner M, Hoff AL, Grimson R. Schizophrenia as a chronic active brain process: A study of progressive brain structural change subsequent to the onset of schizophrenia. Psychiatry Research: Neuroimaging 1997b;74:129–140.

DeLisi LE, Sherrington R, Shaw S, Nanthakumar B, Shields G, Smith AB, Wellman N, Larach NW, Loftus J, Razi K, Stewart J, Vita A, De Hurt M, Crow TJ, Sherrington R. A genome-wide scan of 382 affected sibling-pairs with schizophrenia suggests linkage to chromosomes 2cen and 10p. American Journal of Psychiatry 2002;159:803–812. [PubMed: 11986135]

Docherty NM, Gordinier SW, Hall MJ, Cutting LP. Communication disturbances in relatives beyond the age of risk for schizophrenia and their associations with symptoms in patients. Schizophrenia Bulletin 1999;25:851–862. [PubMed: 10667753]

Docherty NM, Gordinier SW, Hall MJ, Dombrowski ME. Referential communication disturbances in speech of nonschizophrenic siblings of schizophrenia patients. Journal of Abnormal Psychology 2004;113:399–405. [PubMed: 15311985]

Docherty NM, Strauss ME, Dinzeo TJ, St-Hilaire A. The cognitive origins of specific types of schizophrenic speech disturbances. American Journal of Psychiatry 2006;163:2111–2118. [PubMed: 17151162]

Eastman, C.; Weiss, S. A tree algorithm for nearest neighbor searching in document retrieval systems; Proceedings of the ACM-SIGIR International Conference on Information Storage and Retrieval; 1978. p. 131-149.

Egan MF, Goldberg TE, Gscheidle T, Weirich M, Rawlings R, Hyde TM, Bigelow L, Weinberger DR. Relative risk for cognitive impairments in siblings of patients with schizophrenia. Biological Psychiatry 2001;50:98–107. [PubMed: 11527000]

Elvevåg B, Goldberg TE. Cognitive impairment in schizophrenia is the core of the disorder. Critical Reviews in Neurobiology 2000;14:1–21. [PubMed: 11253953]

Elvevåg, B.; Weinberger, DR. Neuropsychology in context of the neurodevelopmental model of schizophrenia. In: Nelson, CA.; Luciana, M., editors. Handbook of Developmental Cognitive Neuroscience. Cambridge, Mass.: MIT Press; 2001. p. 577-595.

Elvevåg B, Foltz P, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophrenia Research 2007;93:304–316. [PubMed: 17433866]

Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics (London) 1936;7:179–188.

Foltz PW, Kintsch W, Landauer TK. The measurement of textural coherence with Latent Semantic Analysis. Discourse Processes 1998;25:285–307.

Foltz, PW.; Martin, MJ.; Abdelali, A.; Rosenstein, M.; Oberbreckling, R. Automated team discourse modeling: test of performance and generalization. Proceedings of the 28th Annual Conference of the Cognitive Science Society; Vancouver, BC. 2006.

Gernsbacher M, Tallent K, Bolliger C. Disordered discourse in schizophrenia described by the structure building framework. Discourse Studies, 1 1999;3:355–372.

Graesser AC, McNamara DS, Louwerse M, Cai Z. Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers 2004;36:193–202.

Hoff AL, Svetina C, Maurizio AM, Crow TJ, Spokes K, DeLisi LE. Familial cognitive deficits in schizophrenia. American Journal of Medical Genetics Part B (Neuropsychiatric Genetics) 2005;133B:43–49.

Jurafsky, D.; Martin, J. Speech and Language Processing. 2. Prentice-Hall; 2008.

Kern RS, Nuechterlein KH, Green MF, Baade LE, Fenton WS, Gold JM, Keefe RS, Mesholam-Gately R, Mintz J, Seidman LJ, Stover E, Marder SR. The MATRICS consensus cognitive battery, part 2: Co-norming and standardization. American Journal of Psychiatry 2008;165:214–220. [PubMed: 18172018]

Kintsch W. The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. Psychological Review 1988;95:163–182. [PubMed: 3375398]

Kuperberg G, Heckers S. Schizophrenia and cognitive function. Current Opinion in Neurobiology 2000;10:205–210. [PubMed: 10753790]

Landauer TK, Dumais ST. A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 1997;104:211–240.

Landauer TK, Foltz PW, Laham D. Introduction to Latent Semantic Analysis. Discourse Processes 1998;25:259–284.

Li X, Branch CA, Bertisch HC, Brown K, Szulc KU, Ardekani BA, DeLisi LE. An fMRI study of language processing in people at high genetic risk for schizophrenia. Schizophrenia Research 2007a;91:62–72. [PubMed: 17306963]

Li X, Branch CA, Ardekani BA, Bertisch H, Hicks C, DeLisi LE. fMRI study of language activation in schizophrenia, schizoaffective disorder, and in individuals genetically at high risk. Schizophrenia Research 2007b;96:14–24. [PubMed: 17719745]

Manning, CD.; Raghauan, P.; Schütze, H. Introduction to Information Retrieval. Cambride University Press; 2008.

Manning, CD.; Schütze, H. Foundations of Statistical Natural Language Processing. MIT Press; Cambridge, MA: 1999.

Marder SR, Fenton WS. Measurement and treatment research to improve cognition in schizophrenia: NIMH MATRICS initiative to support the development of agents for improving cognition in schizophrenia. Schizophrenia Research 2004;72:5–10. [PubMed: 15531402]

McGlashan TH, Miller TJ, Woods SW. Pre-onset detection and intervention research in schizophrenia psychosis: Current estimates of benefit and risk. Schizophrenia Bulletin 2001;27:563–570. [PubMed: 11824483]

McGlashan, TH.; Miller, TJ.; Woods, SW.; Rosen, JL.; Hoffman, RE.; Davidson, L. Structured Interview for Prodromal Syndromes. Version 4.0. Prime Clinic, New Haven, Connecticut: Yale University School of Medicine; 2003a.

McGlashan TH, Zipursky RB, Perkins D, Addington J, Miller TJ, Woods SW, Hawkins KA, Hoffman R, Lindborg S, Tohen M, Breier A. The PRIME North America randomized double-blind clinical trial of olanzapine versus placebo in patients at risk of being prodromally symptomatic for psychosis I study: Rationale and Design. Schizophrenia Research 2003b;61:7–18. [PubMed: 12648731]

McLachlan GJ. The bias of the apparent error rate in discriminant analysis. Biometrika 1976;63:239–244.

Murray RM, Lewis SW. Is schizophrenia a neurodevelopmental disorder? British Medical Journal 1987;295:681–682. [PubMed: 3117295]

Nuechterlein KH, Green MF, Kern RS, Baade LE, Barch DM, Cohen JD, Essock S, Fenton WS, Frese FJ 3rd, Gold JM, Goldberg T, Heaton RK, Keefe RS, Kraemer H, Mesholam-Gately R, Seidman LJ, Stover E, Weinberger DR, Young AS, Zalcman S, Marder SR. The MATRICS Consensus Cognitive Battery, Part 1: Test selection, reliability, and validity. American Journal of Psychiatry 2008;165:203–13. [PubMed: 18172019]

Pennebaker JW, Mehl MR, Niedergoffer KG. Psychological aspects of natural language use: Our words or selves. Annual Review of Psychology 2003;54:547–77.

Pierce, J. An Introduction to Information Theory. Dover Publications; 1980.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2008.

Shannon CE. A mathematical theory of communication. Bell System Technical Journal 1948;27(3):379–423.

Shedlack K, Lee G, Sakuma M, Xie SH, Kushner M, Pepple J, Finer DL, Hoff AL, DeLisi LE. Language processing and memory in ill and well siblings from multiplex families affected with schizophrenia. Schizophrenia Research 1997;25:43–52. [PubMed: 9176926]

Shenton ME, Solovay MR, Holzman PS, Coleman M, Gale HJ. Thought disorder in the relatives of psychotic patients. Archives of General Psychiatry 1989;46:897–901. [PubMed: 2489936]

Snitz BE, MacDonald AW, Carter CS. Cognitive deficits in unaffected first-degree relatives of schizophrenia patients: a meta-analytic review of putative endophenotypes. Schizophrenia Bulletin 2006;32:179–194. [PubMed: 16166612]

Sommer IE, Ramsey NF, Kahn RS. Language lateralization in schizophrenia: an fMRI study. Schizophrenia Research 2001;52:57–67. [PubMed: 11595392]

Sommer IE, Ramsey NF, Mandl RC, Kahn RS. Language lateralization in female patients with schizophrenia: an fMRI study. Schizophrenia Research 2003;60:183–190. [PubMed: 12591582]

Weinberger DR. Implications of normal brain development for the pathogenesis of schhizophrenia. Archives of General Psychiatry 1987;44:660–669. [PubMed: 3606332]

Woods SW, Brier A, Zipursky RB, Perkins DO, Addington J, Miller TJ, Hawkins KA, Marquez E, Lindborg SR, Tohen M, McGlashan TH. Randomized trial of olanzapine versus placebo in the symptomatic acute treatment of the schizophrenic prodrome. Biological Psychiatry 2003;54:453–464. [PubMed: 12915290]

Yu, Chong Ho. Resampling methods: concepts, applications, and justification; Practical Assessment, Research & Evaluation. 2003. p. 8Retrieved February 3, 2007 from http://PAREonline.net/getvn.asp?v=8&n=19

Zhang L, Zhu J, Yao T. An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing 2004;3(4):243–269.

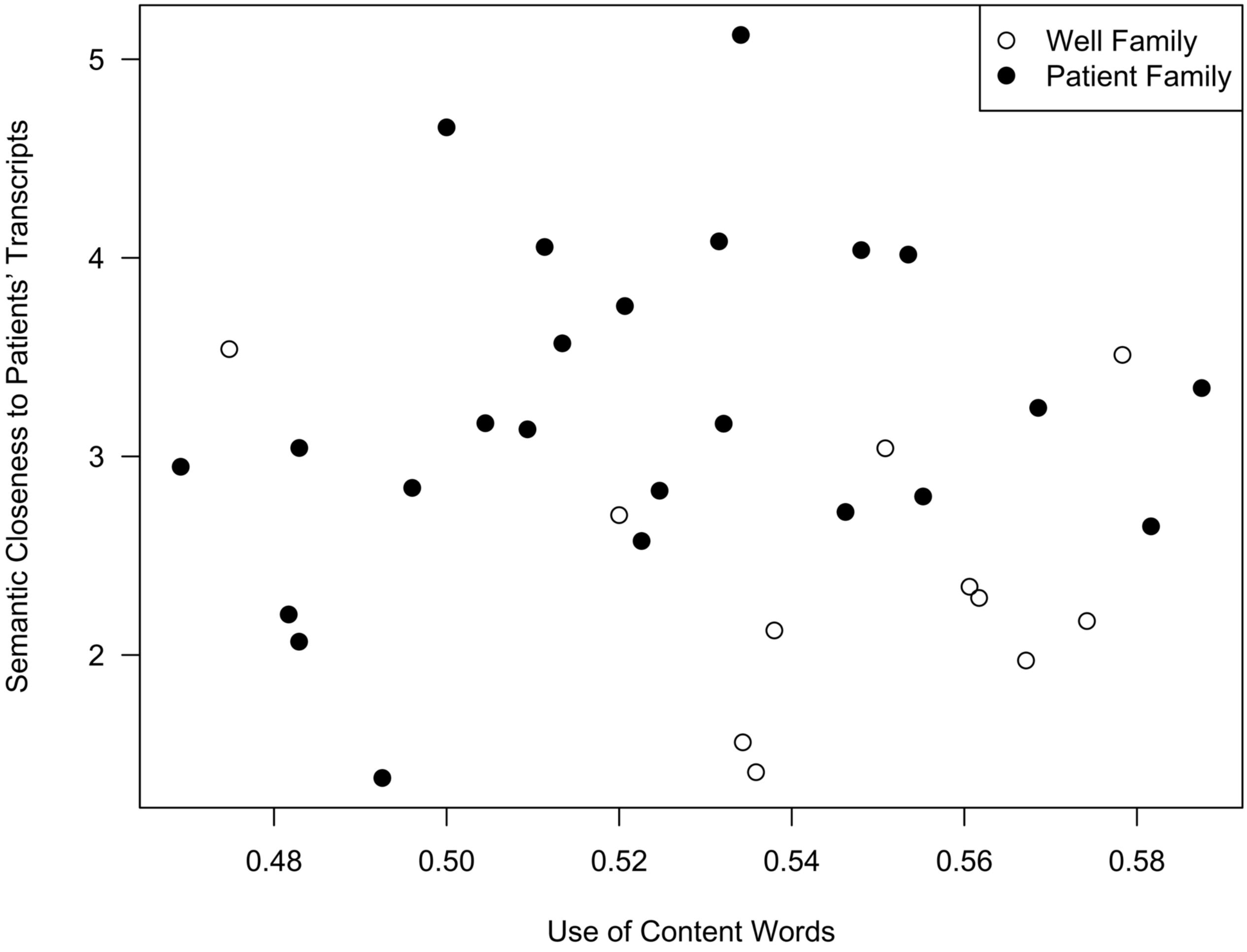

Figure 1.

Clustering of Well versus Patient Family participants along a k-near similarity to patient semantic dimension and a use of content words syntactic dimension.

Table 1

Contingency table of transcript data as a function of family and illness status with mean word count (standard deviation).

		Patient		**Total**
		Yes	**No**	

Family	Yes	Patient family
		N=24 (28.9%)
		WC=695 (287)
			Well family
			N=11 (13.3%)
			WC=780 (159)
				N=35 (42%)

	No	Patient non-family
		N=29 (34.9%)
		WC=1034 (478)
			Control non-family
			N=19 (22.9%)
			WC=1083 (541)
				N=48 (58%)

Total		N=53 (64%)	N=30 (36%)	N=83 (100%)

Table 2

Confusion Matrix of cross-validated model predictions between patient family participants and well family participants.

	Predicted		
Actual	**Patient family**	**Well family**	**Total**
Patient family	21	3	24
Well family	2	9	11
Total	23	12	35

Table 3

Confusion Matrix of cross-validated model predictions between patient non-family participants and control non-family participants.

	Predicted		
Actual	**Control non-family**	**Patient non-family**	**Total**
Control non-family	16	3	19
Patient non-family	5	24	29
Total	21	27	48

Table 4

Confusion Matrix of cross-validated model predictions between all patient participants and all non-patient participants.

	Predicted		
Actual	**Patient**	**Non-patient**	**Total**
Patient	46	7	53
Non-patient	12	18	30
Total	58	25	83

Table 5

Confusion Matrix of cross-validated model predictions between well family participants and control non-family participants.

	Predicted		
Actual	**Control non-family**	**Well-family**	**Total**
Control non-family	17	2	19
Well-family	1	10	11
Total	18	12	30