Using Text Analysis Software in Schizophrenia Research

Contact: heczh@hotmail.com, www.ai.uga.edu/caspr

Congzhou He
University of Georgia

Sara Weinstein
University of British Columbia

Michael A. Covington
University of Georgia

The University of Georgia
ARTIFICIAL INTELLIGENCE CENTER

caspr
COMPUTER ANALYSIS OF SPEECH FOR PSYCHOLOGICAL RESEARCH

Computer language analysis and schizophrenia

Clinical studies have long established the close relationship between language abnormalities and schizophrenia. Recent literature reviews on speech disorder in schizophrenia (e.g. Covington et al. 2005, DeLisi 2001) gather evidence from a wide collection of studies to show that deviances from normal occur at all linguistic levels in schizophrenic language, ranging from phonetics, phonology, lexicon, syntax, semantics, to pragmatics and discourse.

In this study, we designed and tested three text analysis computer programs using cutting-edge natural language processing (NLP) technologies, which perform accurate, objective and speedy analysis of language without requiring substantial linguistic expertise or tremendous time from the researchers.

LINGUISTIC LEVELS	FOCUS OF STUDY	SELECTED LINGUISTIC DEVIANCES FROM NORMAL	OUR TOOLS TARGETING LINGUISTIC DEVIATIONS
Phonology	speech sounds and sound patterns	Lack of tonal inflections, pauses and hesitation	Fo Analysis Tool (He 2004)
Morphology, Lexicon	word formation, vocabulary	Word-finding difficulty, repetitiousness, neologism, stilted speech	Vocabulary Analyzer
Syntax	sentence structures	Normal but simplified syntax, fewer embedded structures	D-Level Rater
Semantics	meaning	Impaired semantic association, difficulty in organizing propositions	Idea Density Rater
Pragmatics	language use with world knowledge	Fewer cohesive devices, error-prone pronoun reference, difficulty in recognizing implicatures and in judging relevance and politeness	(Future work)
Discourse	discourse organization and coherence	Incoherent speech, derailment, loss of goal, tangentiality	(Future work)

Experiment

12 controls and 11 patients were recruited for the experiment at the University of British Columbia. All patients had a diagnosis of schizophrenia according to DSM-IV criteria and were stable outpatients with no recent changes to their medication. Controls were screened for a history of psychiatric illness, and all subjects were screened for a history of head injury, neurological disorder and substance abuse. Both the controls and the patients were right-handed native Canadian English speakers with no history of head injury or neurological disorder. Groups were matched for age, IQ as measured with the National Adult Reading Test (Nelson, 1982) and Quick Test (Ammons and Ammons, 1962), and parental socioeconomic status (Hollingshead Index, Hollingshead and Redlich, 1958).

All the subjects were recorded describing pictures from the Thematic Apperception Test (TAT; Murray 1971) using the administration procedure outlined in Liddle et al. (2002). These recordings were transcribed by typists unaware of each subject's psychiatric status. The transcripts served as the input files to all the software tools described in the study.

LEXICAL LEVEL: Vocabulary Analyzer

Vocabulary Analyzer (VA) is a computer program that computes the rarity of a speaker's vocabulary against general word frequencies obtained from large text corpora. It also calculates various kinds of type-token ratio. VA targets three language deviances: stilted speech (overuse of rare words, Andreasen 1979), neologism, and repetitiousness.

Vocabulary rarity is defined in terms of word frequency in large corpora. The less frequently a word is used by the general public, the rarer the word is. We use the word frequency table from the British National Corpus (BNC, Burnard 2000), and consider its first 500 base-form words (i.e. lemmas) common in English. We also consider any word rare if it is not in the corpus's most common 6000 words.

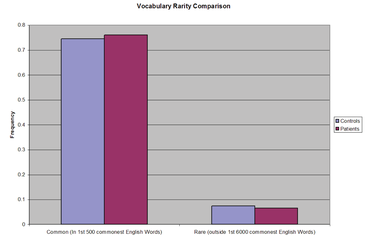
VA first tags each word as noun, verb, etc., using the OpenNLP Tagger (Baldridge and Morton 2004, VA's default tagger), then looks up each word in the widely used electronic dictionary WordNet (Miller et al. 2003) to reduce each word to a distinct lemma (dictionary entry). VA then checks against the BNC word frequency table to decide on the lemmas' rarity.

Results

Neologism and stilted speech are recognized symptoms of schizophrenia, but they are considered infrequent, florid symptoms. In milder form, they might show up as an increased tendency to use rare words vs. common ones, due to imprecise word retrieval.

In both groups, about 75% of the words were in the most common 500 words of the language, and about 5% were outside the most common 6000.

Contrary to expectation, the patients in our study showed a significantly lower percentage of rare words ($P<0.04$) and a slightly higher percentage of the first 500 most common English words ($P<0.07$) than the normal controls. From this we conclude that although the use of rare words is a well-known florid symptom of schizophrenia, it is not a common occurrence in ordinary cases.



SYNTACTIC LEVEL: D-Level Rater

D-Level Rater is a computer program that aims at providing psychiatrists with an informative and revealing measure of syntactic complexity.

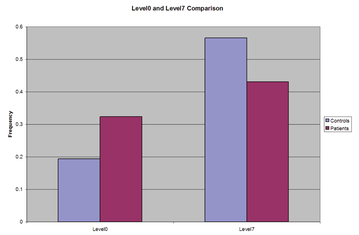
Syntactic complexity is an important topic in research on schizophrenia (Thomas 1996, DeLisi 2001), since deviant syntactic structures are often a direct reflection of brain damage. We believe that Developmental Level Scale (D-Level, Rosenberg and Abbeduto 1987; revised scale, Covington et al. 2006) is best suited for schizophrenia research among the more frequently used syntactic complexity scales, as it is based on evidence from child language and reveals much more structural/locational information than mere counts of grammatical forms.

D-Level	Structure	Example
0	simple sentence	*The dog barked.*
1	non-finite object clause without overt subject	*Try to brush her hair.*
2	coordinate structure	*John and Mary left.*
3	finite object clause, object with clausal modifier, subject extraposition	*John knew that Mary was angry.*
4	small clause as object, comparative	*I want it done today.*
5	finite or non-finite adjunct clause	*They will play if it does not rain.*
6	clausal subject	*The man who cleans the room left early.*
7	more than one structure of levels 1-6	*John decided to leave when he was told the truth.*

D-Level Rater is built upon cutting-edge stochastic parsing technologies implemented in the OpenNLP Parser (Baldridge and Morton 2004), which determines the structure of each sentence. Sentences are rated based on significant structural features specifically mentioned in D-Level scale from each parse returned from the parser. A brief internal experiment showed that D-Level Rater agrees well with levels assigned by human raters.

Results

Our analysis focused on levels 0 and 7. Schizophrenia patients used proportionally more level 0 sentences ($P<0.001$) and fewer Level 7 sentences ($P<0.02$). Our results agree with related studies conducted without computer aid (e.g. Morice and Ingram 1982, DeLisi 2001). Since Level 7 comprises sentences with substructures at multiple D-Levels, D-Level Rater also provides the option to look into the frequency of substructures as defined for Level 1 to Level 6 sentences. Such analysis showed that in our experiment the decreased complexity was not only due to the lowered percentage of relative clauses as reported in the literature (e.g. Morice and Ingram 1982), but also due to fewer occurrences of other complex structures like adjuncts and coordinations.



SEMANTIC LEVEL: Idea Density Rater

Idea Density Rater is a computer program that determines the complexity of language at the semantic level.

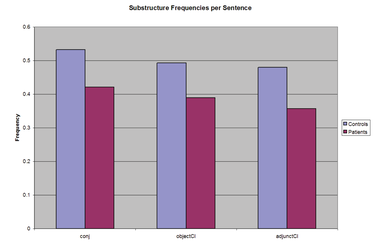
Idea density is a well-defined measure for semantic complexity in continuous speech, which is defined as the number of ideas (or propositions) per N words (N=100 in this implementation). The concept of "proposition" in the definition of idea density originates in Kintsch's Propositional Theory (1974), which states that a proposition "contains a predicator and *n* arguments ($n \geq 1$)." Experiments (e.g. Kintsch and Keenan 1973) have demonstrated that idea density, rather than sentence length, is the key to semantic complexity.

PROPOSITION TYPE	DEFINITION	EXAMPLE
Predication	statement or assertion about the subject of a proposition	Sentence: *Betty bought a balloon.* Proposition: (BUY, BETTY, BALLOON)
Modification	qualifiers, quantifiers, partitives, negatives, etc.	Sentence: *Milton is fat.* Proposition: (QUALITY OF, MILTON, FAT)
Connection	relationship between propositions	Sentence: *Gil caught a cab and went home.* Proposition: (CONJUNCTION: AND, (CATCH, GIL, CAB), (GO, GIL, HOME))

Snowdon et al. (1996) note that the number of ideas or propositions is roughly approximated by the number of words that encode predication or relationships: verbs, adjectives, adverbs, prepositions, coordinators and subordinators (not nouns or pronouns). Idea Density Rater approximates idea density by computing the percentage of these parts of speech based on results from the OpenNLP tagger.

Results

We found no significant difference between the idea densities of the controls' speech and that of the patients' speech. In this respect, the cognitive impairment of schizophrenia is quite different from that of Alzheimer's disease, which produces a significant drop in idea density (Snowdon et al., 1996).



Conclusions

It is obvious that much can be done in automated analysis of schizophrenic speech. With the speed of today's computers, large-scale experiments are now feasible and, very often, effortless on psycholinguistic features rarely tested before due to various difficulties. High-precision automatic speech analysis at various linguistic levels, such as phonology, lexicon, syntax and semantics, is feasible with today's technology, and would greatly facilitate data analysis for large-scale experiments with its objectivity, precision, speed and ease of use. Automated analysis also lays the foundation for further research, such as disease prediction and classification.

References

Ammons, R. B., and Ammons, C. H. (1962) "The Quick Test (QT): Provisional Manual," *Psychological Reports 11*: 111–161.

Andreasen, Nancy C. (1979) "Thought, Language, and Communication Disorders: Clinical Assessment, Definition of Terms, and Assessment of Their Reliability," *Archives of General Psychiatry 36*:1315–1321.

Baldridge, J. and Morton, T. (2004) OpenNLP, http://opennlp.sourceforge.net/.

Burnard, Lou (2000) *Reference Guide for the British National Corpus (World Edition)*, http://www.natcorp.ox.ac.uk/docs/userManual/.

Chapman, James (1966) "The Early Symptoms of Schizophrenia," *British Journal of Psychiatry 112*: 225-251.

Covington, M. A., He, C., and Brown, C. (2006) "How Complex is that Sentence? A Proposed Revision of the Rosenberg and Abbeduto D-Level Scale," Research Report 2006-01, CASPR Project, Artificial Intelligence Center, the University of Georgia.

Covington, Michael A., He, C., Brown, C., Naci, L., McClain, J. T., Fjordbak, B. S., Semple, J., and Brown, J. (2005) "Schizophrenia and the Structure of Language: the Linguist's View," *Schizophrenia Research 77*: 85-98.

DeLisi, Lynn E. (2001) "Speech Disorder in Schizophrenia." *Schizophrenia Bulletin 27*: 481-496.

Eckblad M. and Chapman, L. J. (1983) "Magical Ideation as an Indicator of Schizotypy," *Journal of Consulting and Clinical Psychology 51*: 215–255.

He, Congzhou (2004) Computer-aided Analysis of Ketamine-influenced Speech, Thesis, M.S., The University of Georgia.

Hollingshead, A. B., and Redlich, F. C. (1958).*Social class and mental illness.* New York: Wiley.

Kintsch, W. (1974) *The Representation of Meaning in Memory*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Kintsch, W. and Keenan, J. M. (1973) "Reading Rate and Retention as a Function of the Number of the Propositions in the Base Structure of Sentences," *Cognitive Psychology 5*: 257-274.

Liddle, P. F., Ngan, T. C., Duffield, G., Kho, K., and Warren, A. J. (2002) "Signs and Symptoms of Psychotic Illness (SSPI): A Rating Scale," *British Journal of Psychiatry 180*: 45–50.

Miller, George A. et al (2007) WordNet. http://wordnet.princeton.edu/.

Morice, Rodney, and Ingram, J. C. L. (1982) "Language Analysis in Schizophrenia: Diagnostic Implications," *Australian and New Zealand Journal of Psychiatry 16*:11–21.

Murray, H. A. (1971) *Thematic Apperception Test: Manual*. Cambridge, MA: Harvard University Press.

Nelson, H. E. (1982) *The National Adult Reading Test (NART) Manual*. NFER-Nelson, Windsor, Berks., UK.

Rosenberg, S., and Abbeduto, L. (1987) "Indicators of Linguistic Competence in the Peer Group Conversational Behavior of Mildly Retarded Adults." *Applied Psycholinguistics 8:* 19-32.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996) "Linguistic Ability in Early Life and Cognitive Function and Alzheimer's Disease in Late Life: Findings from the Nun Study," *Journal of the American Medical Association 275(7)*: 528-532.

Thomas, P., Kearney, G., Napier, E., Ellis, E., Leudar, I., and Johnston, M. (1996) "The Reliability and Characteristics of the Brief Syntactic Analysis," *British J. of Psychiatry 168*: 334-343.

Weinstein, S, Woodward, J. S., Werker, J. F., Ngan, E. T. (2005) "Functional Mediation of the Association between Structural Abnormality and Symptom Severity in Schizophrenia," *Schizophrenia Bulletin 31*: 439–440.

International Congress on Schizophrenia Research
Colorado Springs
2007