USING TEXT ANALYSIS SOFTWARE IN SCHIZOPHRENIA RESEARCH

by

CONGZHOU HE

(Under the Direction of Michael A. Covington and Walter D. Potter)

ABSTRACT

Analysis of speech samples is crucial for schizophrenia research, since language is the outer representation of inner thoughts, and speech abnormality effectively reflects brain deterioration. Detailed psycholinguistic analysis, however, often requires substantial linguistic expertise and tremendous amount of time on the part of the analyst, which are not always available.

To demonstrate that sophisticated and useful psycholinguistic measures can be automated at various linguistic levels using cutting-edge NLP technologies, this dissertation describes the design of three NLP applications for schizophrenia research. They are Vocabulary Analyzer analyzing vocabulary rarity at the lexical level, D-Level Rater rating syntactic complexity at the syntactic level, and Idea Density Rater computing idea density at the semantic level. Speech samples from a schizophrenia experiment were used as a test bed for the usability of the software. Results show that, lexically, the schizophrenic patients in the experiment tend to use fewer rare words than the normal controls. Structure-wise, the patients' speech features lowered syntactic complexity as measured with D-Level Scale. And, semantically, no significant difference in idea density was found between the speech samples of the patients and those of the controls.

INDEX WORDS: schizophrenia, text analysis software, natural language processing, vocabulary rarity, D-Level, idea density

USING TEXT ANALYSIS SOFTWARE IN SCHIZOPHRENIA RESEARCH

by

CONGZHOU HE

B.A. Shanghai International Studies University, China, 1992

M.A. Shanghai International Studies University, China, 1995

Ph.D. Shanghai International Studies University, China, 2000

M.S. The University of Georgia, Athens, GA, USA, 2004

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2006

USING TEXT ANALYSIS SOFTWARE IN SCHIZOPHRENIA RESEARCH

by

CONGZHOU HE

Major Professors: Michael A. Covington
Walter D. Potter

Committee: John Stufken
John A. Miller
Liming Cai

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2006

ACKNOWLEDGEMENTS

I am deeply indebted to my major professors Dr Michael A. Covington and Dr Walter D. Potter for their guidance throughout my five years of study at the University of Georgia. Their keen insight, unwithholding support, and constant encouragement led me, right from the moment when they admitted me to UGA, from a linguist that barely knows anything about computers to a computer scientist who is able to design and implement useful software applications involving sophisticated psycholinguistic theories.

I am very thankful to my minor professor Dr John Stufken for always being supportive and for spending so much time in helping me to better understand experimental designs and methods of statistical analysis. I am very grateful to Dr John A. Miller and Dr Liming Cai for sharing with me their extensive knowledge in computer science. Doing projects under their supervision has been one of the best learning experiences that I have had at UGA. I also wish to thank my graduate coordinator Dr Hamid Arabnia, my other professors and staff, and my fellow students at the Department of Computer Science. Had it not been for their help, this dissertation would not have been possible.

I am grateful to my husband Huimin, a real theoretical linguist, who not only supports me emotionally but also took up the laborious job of collecting the testing data for me. I thank my lovely daughter Yinyu for showing me the joys of life through the hardship of dissertation writing, and my parents who offer their generous assistance with the greatest patience.

Special thanks go to Sara Weinstein at the Department of Psychiatry, University of British Columbia for providing speech samples from the schizophrenia experiment used in the dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

SCHIZOPHRENIA AND LANGUAGE

Schizophrenia is a chronic, disabling brain disease, which afflicts about 1% of the human population (Spearing 1999). MRI and functional MRI images provide unequivocal evidence that the brains of schizophrenia patients differ both structurally and functionally from normal, especially in the frontal and temporal cortices (Kuperberg and Caplan 2003, Dombeck 2006). Brain deterioration from schizophrenia results in assorted cognitive impairments, handicapping schizophrenic patients in different aspects of everyday life.

Unfortunately, despite extensive research, neither the cause nor the pathophysiology of the disease is understood thoroughly (Turner 2003). Schizophrenic patients are likely to be affected by the disease for the rest of their life, and complete recovery has been rare (Birchwood and Jackson 2001). Due to the many complicated factors involved, there is currently not and will not be, in the near future, a drug that cures the disease, although medications do help the patients remain stable (Turner 2003). In fact, so much about schizophrenia is still unknown that even its diagnosis can be fairly inconsistent (Strobel 2005, Flaum, Arndt, and Andreasen 1991), especially at early stages of the disease (Turner 2003).

This dissertation introduces an alternative approach to studying schizophrenic symptoms and diagnosing schizophrenia: automated analysis of speech is proposed both to simplify and speed up research process and to complement human evaluation of the disease. Appropriately used, machines are able to detect aspects of schizophrenic symptoms beyond human capacities, thereby bringing us closer to a better understanding, and earlier and more accurate diagnosis of the disease, and, hopefully, a sooner discovery of an effective cure.

In this introductory chapter, I will describe the relationship between schizophrenia and language, and explain how speech abnormalities have been exploited in the diagnosis of schizophrenia. I will briefly discuss the ways that machines can help in psychiatrists' research before I describe the layout of the dissertation in the last section.

1.1 Schizophrenia and Language

Like any brain disease, schizophrenia physically perturbs the Central Nervous System and the neurotransmitter systems (Hudson et al. 1993, Benes 2000), which are crucially involved in perception, memory and thinking. Profound cognitive malfunctions, however, start to exhibit themselves long before the damage can be established clinically (Chapman 1966, Klosterkötter 2001). Whether at the prodromal or later stages of schizophrenia, language is the overt reflection of cognitive disability, and is always one of the central objects of observation.

Diagnoses of schizophrenia are frequently partially based on a patient's abnormal use of language, as language is readily observable as a mirror of cognitive function. Both of the major diagnostic criteria for schizophrenia DSM-IV (American Psychiatric Association 1994) and ICD-10 (WHO 1992) rely heavily on speech disorder for classification purposes. For instance, DSM-IV includes "disorganized speech (e.g., frequent derailment or incoherence)" as one of its five characteristic symptoms. ICD-10 provides description of different speech disorders as symptoms for each subcategory of schizophrenia, regarding the quantity, quality and organization of language. Aside from the direct usage of speech abnormalities in diagnosis, DSM-IV and ICD-10 indirectly resort to speech for evidences of thought disorder and cognitive disturbance.

The reliance of schizophrenia diagnosis on language is based on decades of clinical study, which doubtlessly establishes the close relationship between language abnormalities and schizophrenia (e.g. Andreasen 1979a, 1979b, 1986, Rochester and Martin 1979, Chaika 1974, Docherty and Gottesman 2000, McKenna and Oh 2005). Observational experiments of various types have been conducted for psychiatrists to compare schizophrenic and normal speech. Linguistic details of both written and spoken language are studied, and strong correlations have been found between schizophrenia and many linguistic deviances. In particular, researchers agree on the recurrence of some abnormalities, such as disorganized speech and derailment, incoherence and illogicality, and poverty of speech, which have, as we have seen, made their way to the standard diagnostic criteria DSM and ICD.

In fact, it has been further argued that language, which calls for hemispheric lateralization, may lie at the center of the etiology of schizophrenia. As Crow puts it, schizophrenia is "the price *Homo sapiens* pays for language," when lateralization fails (Crow 1997a, b, 2000). Crow based his evolutionary origin of schizophrenia as an error in the evolution of language on various evidences, such as: the fact that schizophrenia is ubiquitous in all geographical areas with no indication of environmental or cultural influence (although it is more disabling in some settings than in others); the fact that schizophrenic patients do not develop the normal anatomical asymmetry in brain; the disadvantaged genetic disposition of schizophrenic patients, and so on. Although there are different opinions in the origin of schizophrenia (e.g. Pennisi, Plebe and Falzone 2004), and although there is no direct proof yet for Crow's theory, it does, in its unique way, reiterate the fact that language dysfunction is inherent to schizophrenia.

1.2 Traditional Analysis of Schizophrenic Speech

Schizophrenia is traditionally diagnosed or rated according to accepted psychiatric scales, where both linguistic and nonlinguistic symptoms are itemized and scored based on severity. The actual rating is often done by filling in a score sheet based on the psychiatrist's impression from clinical interviews and/or direct observation (American Psychiatric Association 2000). Because of the direct relationship between schizophrenia and language disorder, speech analysis has always been a crucial part of schizophrenic research or diagnosis. All of the very frequently used psychiatric scales have some or most of their items that rate verbal behavior, such as Brief Psychiatric Rating Scale (BPRS, Overall and Gorham 1962, Van Riezen and Vrijmoed-de Vries 2000, American Psychiatric Association 2000), Scale for the Assessment of Thought, Language and Communication (TLC, Andreasen 1979a), Scale for the Assessment of Positive Symptoms (SAPS, Andreasen 1984), Scale for the Assessment of Negative Symptoms (SANS, Andreasen 1982).

Take TLC for example. TLC, first published in Andreasen's classical paper "Thought, Language, and Communication Disorders" (1979a), is an 18-point scale often used to define “formal thought disorder” for disorganized symptoms of schizophrenia. All of the 18 items in TLC are speech related: 1) poverty of speech (poverty of thought, laconic speech); 2) poverty of content of speech (poverty of thought, empty speech, alogia, verbigeration, negative formal thought disorder); 3) pressure of speech; 4) distractible speech; 5) tangentiality; 6) derailment (loose association, flight of ideas); 7) incoherence; 8) illogicality; 9) clanging; 10) neologism; 11) word approximations (paraphasia, metonyms); 12) circumstantiality; 13) loss of goal; 14) perseveration; 15) echolalia; 16) blocking; 17) stilted speech; 18) self-reference.

Measures are taken to increase inter-rater reliability for ratings based on the psychiatric scales. Aside from the necessary training for the psychiatrist before the actual rating, the items in the scale are typically defined as precisely as possible, supplemented with adequate examples for better understanding. For example, in TLC, “tangentiality” is defined as “replying to a question in an oblique, tangential, or even irrelevant manner,” and is specifically restricted to “immediate response to a question” in contrast to the similar concept of “derailment.”

In addition to much research on the more mainstream and apparent linguistic deviances mentioned in psychiatric scales, there have also been scattered studies to discover other non-canonical and less obvious characteristics of schizophrenic language. These studies focus on one or a few linguistic qualities of schizophrenic language, such as vocabulary distribution (based on Zipf’s Law, for instance) (Ferrer i Cancho 2005, 2006), syntactic complexity (Morice and Igram 1982, 1983, Thomas et al. 1987, Thomas 1996, Barch and Berenbaum 1997), and conformity to pragmatic principles (De Decker and Van de Craen 1987, Corcoran and Frith 1996, Abu-Akel 1997, 1998, 1999). These studies normally draw upon a relatively small pool of subjects depending on availability. Some of these experiments approach the problem in an innovative way, and reach novel conclusions that serve as the inspiration for follow-up experiments. For example, Morice and Ingram (1982) were the first to measure syntactic complexity systematically and presented the then-very-interesting result that schizophrenic patients use less complex structures. Because such research requires well-designed linguistic analysis before a conclusion can be reached, the characteristics reported in the studies are often less noticeable simply through direct observation. Conclusions of original studies may be confirmed by follow-up studies, but it is not infrequent for follow-up studies to report conflicting results. Irigaray (1985) and Pennisi (1998), for instance, report respectively that syntactic capabilities were actually enhanced in some schizophrenic patients.

In short, while direct observation from numerous experiments results in general consensus on certain characteristics of schizophrenic language such as disorganization and incoherence, there are a comparatively small number of studies analyzing the less noticeable and the less agreed-upon features that require detailed linguistic analysis, such as lowered syntactic complexity.

1.3 Why Machines?

The conventional language analysis as we have seen in the last section for schizophrenia research is not without its problems. In this section, we look into these problems and discuss my motivations for introducing computers to schizophrenic language analysis.

Language is traditionally studied at several theoretical levels: phonetics, phonology, morphology, lexicon, syntax, semantics, pragmatics, and discourse. Very roughly, phonetics and phonology study speech sounds and sound patterns; morphology deals with word formation; lexicon is the vocabulary; syntax is the study of sentence structures; semantics studies the way words convey meaning; pragmatics relates world knowledge to language use, and studies how deixis, cooperation, relevance, and politeness are achieved as add-ons to literal meaning; and discourse analysis is the study of discourse organization and coherence. Among them, pragmatics and discourse are typically considered as higher levels than the other categories, for they are above "language proper," and involve much extralinguistic considerations such as world knowledge and cognitive abilities. While there is admittedly some overlap and interactions between the linguistic levels, these strata are considered to be fairly independent of each other. Word senses at the semantic level, for instance, can be studied with little reference to word pronunciation at the phonological level. The stratification into linguistic levels has made the study of language more systematic and comprehensive.

Psychiatrists have traditionally laid more emphasis on the pragmatic and discoursal levels of language. In TLC, for example, "tangentiality," "derailment," "incoherence," "illogicality," "circumstantiality," and "loss of goal" all belong to these two levels. It is natural for a psychiatric scale to contain such high linguistic level items, as these are errors most obvious to the human ear and easiest to perceive amongst all linguistic deviances.

There are, however, at least two difficulties associated with the assessment of pragmatic or discoursal properties. First, objective standards for assessment are hard to achieve. For example, as world knowledge varies from person to person, the degree of relevance or coherence of a piece of speech varies accordingly with different listeners. Even with the same person, extralinguistic knowledge changes over time, which is exactly the bridge over linguistic gaps. The same speech sample would, therefore, sound more coherent to a psychiatrist after he has worked with the patient for a while than when he first meets the patient. Because of the extralinguistic factors involved, the reliability of the ratings will be compromised by subjectivity, which, in turn, results in inconsistent diagnosis of the disease (Flaum, Arndt, and Andreasen 1991).

Second, many high level measures are not and cannot be entirely orthogonal to each other. TLC, for instance, goes to great lengths to distinguish the concepts of “tangentiality,” “derailment,” “incoherence,” “circumstantiality,” and “loss of goal.” Despite the author’s effort in providing definitions, constraints and examples, it is still very difficult to distinguish these concepts from one another. In fact, Andreasen herself acknowledges the co-occurrence of “derailment,” “incoherence” and “loss of goal” (Andreasen 1979a). Non-orthogonality is not only a challenge to later statistical analysis, but poses problems to precise rating based on these measures as well.

Language analysis for psychiatric research has been lopsided, with much less attention to lower linguistic levels which focus on language proper. This is understandable, as deviations at these levels cannot be reliably detected through direct observation, and low level linguistic measures normally require meticulous analysis on the part of the rater. Such analysis is often demanding, requiring professional training and thorough knowledge in linguistics as well as proficiency in the language; yet for many psychiatrists, going through linguistic jargons alone is a daunting task. Moreover, language analysis is intricate and time-consuming, with numerous measures computed on top of detailed analysis at various linguistic levels. Even for language specialists, dealing with just hours of recorded speech data would mean weeks or even months of analytic work.

Nevertheless, difficulties do not justify negligence, since a comprehensive analysis of a patient's language is required, on the one hand, to fully understand how brain diseases like schizophrenia affect the production and the processing of language and, on the other hand, to make inferences about brain damage (such as locations) based on sufficient linguistic evidence (Ojemann 1991).

The problems with traditional language analysis for schizophrenia research can, therefore, be very briefly summarized as: 1. rating with increased objectivity is a necessity; 2. deeper and larger scale research on language proper as opposed to language at cognitive level is much desired. These are the exact motivations to use computers to analyze schizophrenic language.

In contrast to human judgment, machines are objective to the greatest possible extent. Computers analyze texts rigorously according to pre-determined algorithms, and are not influenced by presumptions or expectations of the experiments. Unlike ratings from human raters, the output from machines is completely replicable provided that the same input is given. And such consistency in the results of the experiments is achieved regardless of who is using the machine where or when.

Furthermore, with well implemented language analysis software, psychiatrists do not need to resort to “simpler forms of language analysis” as suggested by Thomas et al. (1996), since the burden of fully mastering and accurately executing language analysis techniques is laid on the software developer. Ideally, for any user (including those with the least experience or training in linguistics), analytical results should be just a line of simple command or even a click away.

Machines are the ideal substitutes for humans when doing laborious computing. With the power of today’s personal computers, general problems can normally be solved within seconds. Despite the fact that dealing with natural language involves greater computational complexity than usual, better algorithms and heuristics are still able to reduce the computing time to minutes for most problems even with very long input. Running data through suitable software makes it possible to efficiently process large amounts of data and will in fact encourage the collection of data among larger populations.

With the advancement of computational linguistics, psychiatrists no longer have to settle for simpler analysis techniques or the large amount of time devoted to language data analysis in order to achieve thorough understanding of schizophrenic language. The computers on our desks have got us far past the stage of manually counting linguistic features; in fact, we are able to compute within a reasonable amount of time a huge variety of objective measures much more complicated than the traditional measures like sentence length and word length. Modern computer technology will doubtlessly bring human beings closer to a cure of schizophrenia.

1.4 Current Study at a Glance

This dissertation introduces a brand-new approach of using computers to systematically analyze schizophrenic language at multiple linguistic levels. In this introductory chapter, I have discussed the motivations, and the following chapters are detailed elaborations on the approach.

In Chapter Two, I review the major areas of natural language processing (NLP) research and their current development. I will go over some deviances in schizophrenic language as reported in literature, and I will discuss how NLP techniques can be used to measure such deviances and what NLP will bring to schizophrenia research.

While NLP techniques apply to human language in a general sense, i.e. both spoken and written language, this dissertation focuses more on the spoken side. Chapters Three, Four and Five each look into one linguistic level of schizophrenic speech: lexicon, syntax and semantics. One software package is implemented for each level, and each targets some potential language deviations reported in schizophrenia literature. The three software packages (Vocabulary Analyzer, D-Level Rater, and Idea Density Rater) are tested with the speech samples from a schizophrenia experiment, and the results are reported and discussed.

Chapter Six shows the areas where future work can be done relating to and based on automated analysis of schizophrenic speech, and concludes the dissertation.

CHAPTER TWO

NATURAL LANGUAGE PROCESSING AND SCHIZOPHRENIC LANGUAGE

After briefly going over current NLP technologies and some characteristics of schizophrenic language at each linguistic level, this chapter shows where and how NLP techniques can be effectively employed in schizophrenic language analysis, and what benefits it will bring to schizophrenia research.

2.1 Natural Language Processing in General

NLP is using computers to study human language and perform human language related tasks. Significant advancements have been made in various aspects of NLP, covering all linguistic levels ranging from phonetics to syntax and from semantics to discourse (Mitkov 2003, Jurafsky and Martin 2000).

On the phonetic and phonological level, two major NLP research directions are speech recognition and text-to-speech synthesis. Whereas text-to-speech (TTS) synthesis technology has advanced to the stage of being able to generate fairly natural-sounding speech with human-like suprasegmental features (Dutoit 1997) (see AT&T Natural Voices Text-To-Speech Project and Bell Labs Text-To-Speech Synthesis websites for TTS demos), there is no speaker-independent, unrestricted speech recognition system yet with consistently acceptable error rate. Despite the fact we can reliably extract formant information from sound waves, high-performance applications still require much training for the systems to capture speaker specific parameters, so that the immense individual differences among speakers can be handled appropriately.

Like speech recognition, many NLP technologies are either too costly or too immature to be utilized in real-time natural language applications; nevertheless, a few other areas in NLP have reached the stage where reliable results are guaranteed. These areas include the fundamental operations in any natural language related applications: tokenization, tagging, and parsing.

Tokenization, a morphological and lexical level operation, is the process of breaking up input strings into meaningful units such as words and punctuation marks. For languages like English, tokenization can achieve near-perfect precision, as words are separated by blanks in written text.

Tagging, the process of assigning a Part-of-Speech (POS) tag to each token, is a syntactic level operation. Taggers often serve as the basic utility, upon which further processing of the text such as parsing can be done, since a large proportion of the words in our lexicon can function as more than one POS and need to be disambiguated. Taggers are among the easiest NLP applications to implement: a precision of over 90% can be easily reached simply by blindly assigning the most frequently used POS tag to each word (Manning and Schutze 1999). However, it is essential for a tagger to achieve much better precision in order for other applications to base their implementation on it. The state-of-the-art taggers typically use stochastic or rule-based algorithms and a variety of heuristics (ibid.), and a precision of over 97% is reported to be achieved by popular taggers such as CLAWS (Garside 1987) and OpenNLP tagger (Baldridge and Morton 2004).

Parsing, the process of assigning structures to an input string, is yet another common syntactic level operation. There are two main concerns with today's parsers: 1) precision: many parsing algorithms start with tagging, so very often errors in tagging percolate to parsing results. Moreover, since human language is inherently ambiguous in structure, it is typical that even very well implemented parsers have only a precision of around or slightly above 90%, which is significantly lower than the tagging precision. 2) speed: shallow parsing, which does not look into the internal details of phrasal constituents, is normally fast, whereas deep parsing aiming at structural details typically has much higher time and space requirements. For example, statistical parsing algorithms such as the commonly used stochastic CYK (Ney 1991) and Earley's (Stolcke 1995) algorithms try to find the optimal solution by computing all structural combinations.

Unlike the structural features of language which can be represented by some sort of grammar, anything related to meaning is difficult to represent and harder to process with a machine. Semantics is the lowest linguistic level that deals with meaning. The complexity of processing natural language at the semantic level varies, depending on whether the semantic analysis is syntax-driven or not. Syntax-driven semantic analysis is more or less an augmentation of syntactic analysis (Jurafsky and Martin 2000), which analyzes propositions and argument structures without a real understanding of the information contained in the sentences.

Syntax-driven semantic analysis can be done fairly reliably in the same way that syntax is analyzed.

Operations at a higher semantic level where computers actually start to “comprehend” the contents of language, and topics related to higher level semantic analysis, such as information extraction, have barely just become the center of NLP research. Entity extraction, event extraction, and time and space term extraction are, without exaggeration, the basis for natural language understanding. These operations are evaluated not only by precision, but by recall as well. A balance between high precision and high recall (and, if possible, high speed) is sought after by computational linguists, who have reported remarkable progress. However, it is impossible to compare published reports of precision, recall or the combined F-score for different implementations, which range from around 60% (e.g. Spasic et al. 2003, Callan and Mitamura 2002) to around 90% (e.g. Curran and Clark 2003), as the standards and experimental settings differ vastly. The unfortunate fact is that while some systems work reasonably well within certain domains (such as news broadcast), no existing information extraction system can be universally applied to deal with unrestricted domain and consistently maintain high precision and recall.

Word sense disambiguation is another important semantic level operation that is vital to many other applications like information retrieval and machine translation. Word sense disambiguation can be done via many approaches, of which the most often adopted are knowledge-base consultation (e.g. Navigli and Velardi 2005) and latent semantic analysis (Landauer and Dumais 1997). For both these approaches to work well, it is essential for them to be supported by huge databases, either in the form of ontology, electronic dictionaries, or in the form of huge corpora of texts, which, unfortunately, restricts the general employment of these approaches.

The high-level pragmatic and discoursal aspects of language are more problematic for today's computers to handle, for they deal with information beyond words themselves, such as speaker cooperation, idea organization and logic. The capturing of such frequently-covertly-conveyed information demands much more than linguistic knowledge alone; just to name a few, situational knowledge like formality and participant relationship, and background knowledge like traditions and social customs, etc. are all among the prerequisites (Levinson 1983).

Regrettably yet understandably, not much has been done in the realm which traditional pragmatic research focuses on, like cooperativeness, politeness and relevance. Cohesive devices (Halliday and Hasan 1976) and deixis such as pronoun reference are the closest NLP has got to at this level, and the most mature application to date is probably anaphora resolution, whose reported precision and recall vary mostly between around 60% (e.g. Soon et al. 2001) and over 80% (e.g. Muñoz and Palomar 2001) in different experiments, and are again impossible to be compared performance-wise.

In contrast to the little work done in pragmatics, computational linguists have invested tremendous effort in discourse-level research, driven by the enormous demand for practical software that does essay rating (e.g. Burstein et al. 2001, Miltsakaki and Kukich 2004), text summarization (e.g. Mani 2001, Hovy 2003), question answering (e.g. Ittycheriah et al. 2000, Harabagiu and Moldovan 2003), and so on. Coherence and discourse organization are allegedly difficult, if not impossible, for machines to approach directly; hence, indirect algorithms that make use of cue words and latent semantics are the most popular. It is difficult to evaluate the performance of discourse-related software, as even humans don't have a uniform standard for evaluating discourse features like coherence. Discourse-related software normally requires some form of human supervision, such as manual tagging and human co-rating.

To summarize, computational linguists have made remarkable progress and developed useful algorithms at all linguistic levels: phonetics, phonology, morphology, syntax, semantics, pragmatics and discourse. However, not all NLP research areas are equally developed due to the different complexities involved. The more accurate and reliable applications are low-level and structure-oriented, involving minimal extralinguistic knowledge. High-level applications concerning meaning, cohesion, and discourse organization are less reliable, yet still are remarkably useful and highly valuable especially when precision is not necessarily the top priority and when machines efficiently undertake otherwise labor-intensive tasks concerning huge amount of text processing. Table 2.1 summarizes the common applications at different linguistic levels.

LINGUISTIC LEVELS	SELECTED NLP APPLICATIONS
Phonology	Sound wave analysis, formant extraction, speech synthesis, speech recognition
Lexicon	High-precision tokenization, electronic dictionary
Syntax	High-precision tagging, parsing
Semantics	Argument structure analysis, information extraction, word sense disambiguation
Pragmatics	Anaphora resolution
Discourse	Discourse structure analysis, coherence rating

Table 2.1 Selected NLP Applications by Linguistic Levels

2.2 Characteristics of Schizophrenic Language

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 section, I selectively review some of the more well-known characteristics of schizophrenic language at each linguistic level as reported in important studies.

“Lack of vocal inflections” (Andreasen 1982), as a manifestation of “affective flattening,” is perhaps the most well-known symptom of negative schizophrenia at the phonetic and phonological level. Patients would speak with a monotonous intonation, with no stress on key words; furthermore, they are likely to talk with the same volume, with no regard to whether the topic is private or exciting (ibid.). Negative schizophrenia patients also tend to have longer pauses and more hesitations (Clemmer 1980, Andreasen 1982), which closely relate to “poverty of speech" of negative schizophrenia. The assumption is that pauses and hesitations arise from poverty of thoughts or impaired cognitive abilities.

Schizophrenia evidently exerts a negative effect on the lexical performance of the patient, especially those with thought disorder. Many studies suggest that the underlying problem is not a shrunken vocabulary size but disorganized storage and defective access and retrieval mechanism (e.g. Allen and Frith 1983, Allen et al. 1993, Paulsen et al. 1996). Schizophrenic patients typically perform much worse than normal controls in both producing and processing lexical tasks relating to semantic categories (e.g. Gurd et al. 1997, McClain 1983). Patients of thought disorder either fail to recognize categorical information presented in experiments, or produce items falsely related to a semantic category. This suggests that the lexicon is stored, accessed and retrieved based on some abnormal semantic network, and that the linkages between the lexical items are not established upon natural semantic relationships.

The above mentioned deficits result in difficulties in finding appropriate words, and, in turn, in lexical deviances in schizophrenic language such as the following: repetitiousness, neologism, and stilted speech. Neologism refers to the phenomenon where patients coin new words, sometimes without an understandable derivation (Andreasen 1979a), and use the word to describe a commonplace idea, like the word *bawk* in "So I sort of *bawked* the whole thing up" (ibid.). Stilted speech is the use of excessively formal, multi-syllabic words in common situations. Andreasen (ibid.) uses the following example to illustrate the phenomenon: *Whereas the attorney comported himself decorously, the physician behaved as is customary for a born gentleman.*

On the syntactic level, it has frequently been reported that although schizophrenic patients generally use normal and correct syntax, they exhibit both impaired comprehension and production as syntactic complexity increases (Hoffman et al. 1988, Anand et al. 1994). Simplified syntax is more noticeable in speech than in written language (Thomas et al. 1993), and syntactic simplicity in schizophrenic speech is represented by increased percentage of simple sentences and fewer embedded structures (Morice and Ingram 1982, 1983, Fraser et al. 1986). Other syntactic complexity measures used in schizophrenia experiments include mean number of words per sentence (Caplan et al. 1992), percentage of sentences with embedding (Fraser et al. 1986), mean number of coordinated sentences (ibid.), etc. A few researchers, however, do not consent to the conclusion of simplified syntax in schizophrenic speech, presenting experimental data showing either no change in syntactic complexity (Sanders et al. 1995) or increased complexity (Irigaray 1985, Pennisi 1998).

Much research has been done on the semantic level in conjunction with the lexicon, mainly about semantic association. Semantic network formed by semantic relationships such as synonymy, antonymy, hyponymy / hypernymy, etc. appears to be impaired in schizophrenic patients as was mentioned above, so they frequently perform worse than normal controls in studies testing word relations, such as priming or word recall experiments. Unfortunately, very little has been done on other interesting semantic aspects of schizophrenic language, which cover, for instance, argument structures or propositions, where deviances might be expected as well.

One of the research focuses in the pragmatic aspect of schizophrenic language has been cohesion. Many researchers (Rochester and Martin 1979, Hoffman et al. 1985, Docherty et al.

1988, Abu-Akel 1997) reach similar conclusions through various experiments that fewer cohesive devices or more err-prone pronoun references are used in schizophrenic speech. There are additional studies (Abu-Akel 1999) that show how patients have difficulties observing the Gricean Cooperative Principle and the Maxim of Relevance in particular.

As was mentioned in Chapter 1, many of the discoursal characteristics of schizophrenic speech have made their way to the standard diagnostic criteria and psychiatric rating scales, such as incoherence, derailment, loss of goal, tangentiality, and so on.

To sum up, researchers have noted deviances in schizophrenic language at all linguistic levels. Table 2.2 lists a selection of the often-reported characteristics. Sometimes a language phenomenon is ambiguously related to two or more linguistic levels, and labeling it as a deviance on a single level is difficult. While the overlap between linguistic levels is fully acknowledged, Table 2.2 shows a possible, if not the best, understanding of the stratification of linguistic deviations.

LINGUISTIC LEVELS	SELECTED LINGUISTIC DEVIANCES FROM NORMAL
Phonology	Lack of vocal inflections, pauses and hesitation
Lexicon	Word-finding difficulty, repetitiousness, neologism, stilted speech
Syntax	Normal but simplified syntax, fewer embedded structures
Semantics	Impaired semantic association, difficulty in organizing propositions
Pragmatics	Fewer cohesive devices, error-prone pronoun reference, difficulty in recognizing implicatures and in judging relevance and politeness
Discourse	Incoherent speech, derailment, loss of goal, tangentiality

Table 2.2 Selected Deviances in Schizophrenic Language by Linguistic Levels

2.3 What Can NLP Bring to Schizophrenia Research?

Equipped with the NLP techniques reviewed in section 2.1, numerous useful software applications have been implemented in various areas such as computer-assisted education, human-computer interaction, and machine translation (Mitkov 2003). However, there are areas where NLP is not necessarily considered as a priority, and psychiatric research is unfortunately one of these areas. This is not because there is no need for psychiatrists to analyze human language in their work; in fact, the analysis of natural language is so central to psychiatric research, researchers were already talking about "the diagnosis of schizophrenia by language analysis" two decades ago (Fraser et al. 1986). This is not because psychiatrists do not wish to take advantage of language analysis software either; very often, we meet psychiatrists who would like to run their large amount of speech data through software and obtain results very quickly.

There seem to be three reasons that explain the strange situation where there is a "demand" but almost no "supply" of natural language processing software for schizophrenia research.

First, NLP is an interdisciplinary task that requires expertise in both linguistics and computer science. NLP for schizophrenia research is more demanding, in that it further calls for knowledge of psychiatry. Interdisciplinary cooperation is necessary for useful NLP software for schizophrenia research. Such cooperation needs to involve experts in different areas: consultants in psychiatry, linguists, and software developers. Although many start to realize the importance of collaboration, there are, on most occasions, many practical difficulties to overcome before a successful cooperation can be achieved.

Second, to many psychiatrists, language disorder usually refers to abstract high-level concepts like “incoherence,” “derailment” and “illogicality” (Andreasen 1979a, 1979b, 1986, Lysaker et al. 2003), since high-level linguistic deviances can be more easily picked up by human ear than low-level linguistic deviances. For example, it would be much more realistic for a psychiatrist to detect unwarranted switch of topics than a slightly below-normal percentage of embedded clauses in the speech.

Such restricted understanding of speech disorder deters the development of psycholinguistic software in two ways. First, as was discussed in section 2.1, high-level NLP technologies are currently still very weak in dealing with pragmatic and discoursal properties such as coherence and logicality, and are incapable of providing the high precision and high recall required by medical procedures. It would be only naïve to focus solely on the pragmatic and discoursal aspects of language and to expect mature products to market within a short time. Second, with studies of higher-level deviances being much more prevailing, studies of language dysfunctions at other linguistic levels and, in turn, the development of software for such studies have been neglected. In psychiatrists’ attempt to avoid pure linguistic analysis of the data, where skills and tremendous effort are required, the value of the more precise low-level NLP technologies is disregarded, though they can be employed to reliably pick up various language characteristics.

LINGUISTIC LEVELS	SELECTED NLP APPLICATIONS	SELECTED LINGUISTIC DEVIANCES IN SCHIZOPHRENIC LANGUAGE
Phonology	Sound wave analysis, formant extraction, speech synthesis, speech recognition	Lack of vocal inflections, pauses and hesitation
Lexicon	High-precision tokenization, electronic dictionary	Word-finding difficulty, repetitiousness, neologism, stilted speech
Syntax	High-precision tagging, parsing	Normal but simplified syntax, fewer embedded structures
Semantics	Argument structure analysis, information extraction, word sense disambiguation	Impaired semantic association, difficulty in organizing propositions
Pragmatics	Anaphora resolution	Fewer cohesive devices, error-prone pronoun reference, difficulty in recognizing implicatures and in judging relevance and politeness
Discourse	Discourse structure analysis, coherence rating	Incoherent speech, derailment, loss of goal, tangentiality

Table 2.3 Combining NLP and Schizophrenic Language Analysis

Finally, with NLP appearing only very remotely related to psychiatric research, it takes insight and detailed analysis to figure out where and how exactly NLP can actually help in schizophrenia research. Sections 2.1 and 2.2 are meant to serve this purpose. Table 2.3 combines Table 2.1 “Selected NLP Applications by Linguistic Levels” and Table 2.2 “Selected Deviances in Schizophrenic Language by Linguistic Levels,” as the first step to bridge the gap between what NLP can do and what schizophrenic research needs.

It is obvious from Table 2.3 that on any linguistic level, there are NLP applications sharing the same or similar targets as psychiatrists do in schizophrenic language analysis. These applications may not aim to perform the same language analysis tasks as psychiatrists, yet they prove the existence of core NLP techniques, regardless of their precision and recall, which can be adopted to implement tools for analyzing schizophrenic language. For example, sound wave analysis and formant extraction on the phonological level can certainly help in many phonological applications like speech recognition, but they can also be employed to obtain crucial information about negative symptoms like "lack of vocal inflections, pauses and hesitation." By obtaining data of amplitude and F0 (Fundamental Frequency, which represents intonation contour) and computing their statistics of variance such as standard deviation, we acquire a reliable measure of vocal inflections. Lengths and frequencies of pauses and hesitations can also be computed using F0 data by defining, for instance, "pause" as any silent rupture in speech exceeding 200 milliseconds (Harel et al. 2004), represented by a consecutive sequence of zero F0's for over 200 milliseconds. A phonological software tool for psychiatric research with algorithms as described above, F0 Analysis Tool, was introduced in He (2004) and has been perfected subsequently.

Similarly, NLP technologies at other linguistic levels can serve as the base of language analysis tools for psychiatric research. An important objective of this dissertation is to introduce useful implementations at lexical, syntactic and semantic levels. Very briefly, at lexical level, word frequency lists obtained from large corpora together with electronic dictionary provide significant information on lexical dysfunction; cutting-edge parsers can be utilized to compute more elaborate syntactic complexity measure; and at the semantic level, the very time-consuming proposition analysis can also be efficiently done with some heuristic and appropriate NLP techniques as we will see in Chapter 5.

Apparently, NLP applications can be used in implementations that target high-level deviances in schizophrenic language too, despite their less maturity and less precision. In fact, there is no doubt that once software for information extraction, reference resolution and discourse analysis is perfected to an acceptable precision level, they can be widely applied to analyze patients' speech (Cohen 1987), in the same way that they are being used to rate essays for coherence and discourse planning (Burstein 1998, Miltsakaki 2004). Currently, researchers have also designed some ingenious measures to tackle high-level problems like rating coherence in an indirect way, such as using Latent Semantic Analysis (Elvevaag et al. 2004).

Fortunately, compared to high-level NLP techniques, low-level techniques are more advanced and reliable. Given that purely linguistic features are easier for computers to deal with and that low-level NLP applications have reached high precision, we are at the stage where numerous measures can be automated reliably, whether phonetically, syntactically or semantically.

Other than speeding up the processing time, language analysis software for psychiatric research will be valuable in several ways. First and the most important, computers work as a perfect complement to the psychiatrist to pick up linguistic deviations not obvious to the human ear. Schizophrenic language is seldom so obviously deviant as common people would imagine; in fact, even trained linguists cannot tell just by listening to the speech samples or by reading the transcripts from the experiment used in this dissertation whether the speaker is schizophrenic. Besides, healthy speakers tend to make similar types of error in their speech, such as repetitiousness and incoherence, only to a lesser degree. Very often, this degree may be too small for humans to notice, but large enough for machines. Second, testing comparatively psycholinguistic measures for their sensitivity and validity becomes a simpler job with NLP software, as we are now in a position to conduct massive experiments using a large number of measures without much effort. Third, we will be better equipped to uncover previously undetected linguistic patterns. It is not infrequent that we acquire “by-products” from the output of the software, which turn out to be the gold in the mine, and this will be demonstrated in the following chapters. Last, unlike humans, NLP software is consistently an unbiased rater, which will never be affected by expectations. This doubtlessly adds reliability and consistency to schizophrenia research, which is traditionally built upon human judgment and which is subject to high degree of inconsistency (Strobel 2005).

LINGUISTIC LEVELS	SELECTED TOOLS TARGETTING LINGUISTIC DEVIATIONS
Phonology	F0 Analysis Tool
Lexicon	Vocabulary Analyzer
Syntax	D-Level Rater
Semantics	Idea Density Rater
Pragmatics	Reference Analysis Tool
Discourse	Coherence Rater

Table 2.4 NLP Tools for Schizophrenia Research

Table 2.4 is a summary of selected NLP tools for schizophrenia research, some of which I will describe in the following chapters. They are not just some novel tools, but a new method of approaching the problem of language analysis for psychiatric research.

In the following three chapters, I show, through the cases of Vocabulary Analyzer, D-Level Rater, and Idea Density Rater, how more sophisticated linguistic analysis can be done at lexical, syntactic, and semantic level using text analysis software, and how significant linguistic features of schizophrenic speech can be unveiled with further statistical analysis on the data obtained from the software.

I have been, till now, using the general term “language” to refer to written and spoken language inclusively, and NLP tools should be able to handle both. Except for phonetic / phonological level analysis, the input to software at all other levels is text files, as speech will be transcribed before lexical, syntactic, semantic, pragmatic or discourse analysis can be done. However, speech is undoubtedly the dominant object of study in schizophrenia research, although much work is based on written material as well (Manschreck et al. 1980, 1987, Thomas et al. 1993). I will be using the term “speech” from now on, both because the software I implemented is mainly intended for measuring speech deviances, and because the experiments in this dissertation are done on spoken material. The change in the wording certainly does not rule out the possibility of applying my software to written data.

CHAPTER THREE

VOCABULARY ANALYZER

Vocabulary Analyzer is an NLP implementation at the lexical level to compute the rarity of a speaker's vocabulary against general word frequencies obtained from large text corpora and to calculate various types of type-token ratio.

3.1 Background

Lexicon is the most often studied object in schizophrenia experiments concerning language; yet many experiments are designed to study isolated words (e.g. Johnson and Shean 1993, Faber and Reichstein 1981), and lexicon in spontaneous speech is very poorly analyzed. Researchers design word-association (Moran et al. 1964, Mefferd 1979, Jung 1981), repetition and recall (Gold et al. 1992, Brebion et al. 1997, Mesure et al. 1998) tasks to study the categorization, storage, and retrieval of certain controlled words in schizophrenic patients; even when words are studied in context, experiments typically investigate the performance of repetition or recall based on certain cues. Lexicon in unrestricted, continuous speech receives much less attention, since most characteristics of lexical usage are subtle and hard to define, and the degree of the markedness of these characteristics is even more difficult to measure.

As discussed in section 2.2, schizophrenic patients experience difficulties in retrieving the appropriate words, which would sometimes result in “neologism” and “stilted speech.” “Neologism” and “stilted speech” are among the very few characteristics of spontaneous schizophrenic speech, and, hence, are crucial manifestations of lexical deviance that deserve further study. But it is also important to understand that “neologism” and “stilted speech” are florid symptoms, which do not occur in most speech samples from the patients.

It is, therefore, crucial, to design a lexical measure, which provides a scale to mark the deviance in vocabulary choice from normal that is not only able to note florid symptoms but also sensitive to small changes in common vocabulary as well. This is exactly the motivation for the design of the measure “vocabulary rarity” in this chapter.

Vocabulary rarity is defined against “word frequency.” The less frequently a word is used by the general public, the rarer the word is. The whole lexicon in a language is divided into G groups (mostly of equal size of N words), respectively marked as the most common, the second most common, the third most common, …, and the least common (equivalent to the rare word group, or the G^{th} most common group) based on accepted frequency list from a relevant corpus. Vocabulary rarity is then defined as a list of G numbers which are the percentages of words from each of the groups in a speech sample.

For example, we may define rare English words as any word not contained in the first 4,000 most common word list. Assuming that 2000-word is the size chosen for each word group, we will divide the whole English lexicon into 3 groups: Group 1 contains the first 2,000 most common English words, Group 2 contains the second 2000 most common English words, and Group 3 contains all the rest of the English lexicon. Suppose we have a speech sample of 100 words, of which 60 belong to Group 1 – the group of the most common words, 30 belong to Group 2 – the group of the less common words, and 10 belong to Group 3 -- the rare words. Then the vocabulary rarity of the speech sample is represented by the list [60%, 30%, 10%].

The definition of vocabulary rarity is flexible, since it is left for the researcher to decide what is considered common in what task and how many groups he would like to divide the lexicon into. There are several reasons for the flexibility. First, the word frequency list used as the gold standard for measuring word rarity could differ on various occasions. A simple example is that we would prefer a frequency list from an American English corpus over a British English corpus (provided that both lists are accessible), if the speech samples are from American English speakers. Second, it is infeasible to define some universal cutting point of word rarity for all experiments. For instance, because vocabulary development depends on age and education, what count as common words for some age or education level may be rare words for some other age or education level. Third, it is also infeasible to fix the word group size N and in turn, the number of word groups G, without knowing how rare words are defined in a specific experiment. Naturally, N could be slightly larger if rare words are defined as words outside of the first 6,000 most common words than if rare words are defined as words outside of the first 1,000 most common words.

Equipped with the concept of vocabulary rarity, it is now possible not only to capture the frequency of the use of rare words, but also to observe the general tendency of vocabulary choices in any speech sample even when the sample lacks florid symptoms like neologism.

As was also mentioned in section 2.2, word-finding difficulties in schizophrenic patients would give rise to increased repetitiousness too. Unless evidently demonstrated, increased repetitiousness is another linguistic characteristic which easily eludes the human ear and which is easier for machines to pick up.

The most commonly (though not necessarily the most effective) quantitative measure of repetitiousness adopted in clinical and linguistic research is type-token ratio (TTR). TTR is the ratio of the total number of different words (types) over the total number of words (tokens). For example, the sentence *The cat is chasing the dog which chased the cats.* has eight types (*the, cat, is, chasing, dog, which, chased, cats*) and ten tokens (*the, cat, is, chasing, the, dog, which, chased, the, cats*), and a TTR of *4/5*. Manschreck et al. (1981) report that TTR statistically differentiates thought-disordered schizophrenic patients from both non-thought-disordered patients and normal controls.

There are different kinds of TTR based on slightly different definitions, and I will, in this chapter, be using the terms “normal TTR” as defined above, “vanilla TTR” and “lemmatized TTR.” Vanilla TTR counts not only words as types or tokens, but also punctuations and symbols. According to this definition, the sentence *The cat chases the dog.* has five types (*the, cat, chases, dog,* .) and six tokens (*the, cat, chases, the, dog,* .), and a TTR of *5/6*. Like normal TTR, lemmatized TTR does not count punctuations and symbols either. It does, however, count types differently from normal TTR: types are the number of distinct root forms of words, or lemmas. According to this definition, the sentence *The cat is chasing the dog which chased the cats.* has six types (*the, cat, be, chase, dog, which*) and ten tokens (*The, cat, is, chasing, the, dog,* *which, chased, the, cats*) and a TTR of *3/5*. Though the three TTR's differ only slightly, one TTR could be more sensitive than another in some experiment.

The Vocabulary Analyzer introduced in the next section computes vocabulary rarity and all the three kinds of type-token ratio mentioned above.

3.2 Implementation

Vocabulary Analyzer is a text analysis tool at the lexical level targeting three language deviances: repetitiousness, neologism, and stilted speech. It is implemented in Java 1.5, and, therefore, independent of platforms.

The input to Vocabulary Analyzer is text files, which could be both written samples or transcribed speech samples. As with most NLP applications, the first step is to tokenize the text. Tokenization for the English language is relatively simple, but tokenization is not sufficient for Vocabulary Analyzer. Although this is an application dealing with tokens in the input text, a tagger needs to be employed to disambiguate between homographs. For instance, in order to tell whether an *'s* represents *is*, *has*, or possessive, it is crucial to know the grammatical function of the specific *'s*. A tagger will also conveniently distinguish punctuation marks and symbols from words, which is essential for computing normal TTR and lemmatized TTR.

Hence, Vocabulary Analyzer invokes a Penn Treebank (Bies et al. 1995) style tagger as a base utility. Although taggers nowadays generally have fairly high precision, they are not perfect, and different parsers perform better on different occasions. Therefore, the tagger used by Vocabulary Analyzer is conveniently implemented as a Java plugin, so that the user can have a choice between different taggers simply by specifying the name of the tagger plugin at run time, without the need for re-compilation. The user can even write a tagger of his own and easily plug it into Vocabulary Analyzer with no change in any other part of the implementation. To demonstrate the use of multiple tagger plugins, Vocabulary Analyzer provides the choice between Stanford Tagger (Toutanova et al. 2003) and OpenNLP Tagger (Baldridge and Morton 2004).

The central implementation of Vocabulary Analyzer includes two parts: a vocabulary rarity rater and a TTR calculator.

The vocabulary rarity rater is an implementation based on the concept of vocabulary rarity discussed in section 3.1. As was discussed, vocabulary rarity is a flexible definition, so flexibility is an absolute requirement for the software, which should be able to accept user defined parameters to accommodate each specific experiment condition. In order to use the vocabulary rarity rater, three questions need to be answered.

First, what is the gold-standard word frequency list to be used? Depending on the source corpus, two kinds of word frequency lists may be used. Frequency lists published by general corpora is a natural choice. Large corpora, such as British National Corpus (BNC, Burnard 2000) and American National Corpus (ANC, Ide and Sutherman 2004), release word frequency lists, sometimes of different genres, for public use. Vocabulary Analyzer chooses to use as default “a lemmatized frequency list for the 6,318 words with more than 800 occurrences in the whole 100M-word BNC” (Kilgarriff 1997). This BNC frequency list is comprised of lemmas rather than unlemmatized tokens such as those containing non-words / typos (e.g. *was15th*), inflections (e.g. *vodkas*) or even punctuation marks and symbols (e.g. *way?"*) as found in ANC frequency lists. Lemmatized frequency lists are a more applicable standard especially for speech, where spellings should not matter. For psychiatric studies, frequency lists built from corpora for specific tasks are another choice. For instance, describing pictures from the Thematic Apperception Test (TAT; Murray, 1971) is a common task for psychiatric experiments. A small corpus can be built based on descriptions for each picture. Frequency lists extracted from such corpora could be more valuable than general frequency lists for analyzing speech samples from TAT experiments. For such specific-purpose frequency lists to work, however, speech samples from a large number of participants are required.

Second, what is to be considered as rare words? The user needs to specify an appropriate vocabulary size for the subjects in the experiments. By default, Vocabulary Analyzer uses the 6,000-word cutting point, which means that anything outside the 6,000 most common English words is considered rare. Hoover (2005) suggests that 6,000 words is the appropriate size for statistical stylistics and authorship attribution, a point at which almost all the words in texts are included.

Third, what is the desired group size, or how many groups should the lexicon be divided into? The user needs to specify an appropriate number N for word group size. For example, if the cutting point is 6,000 words and N=3000, then Vocabulary Analyzer divides the whole lexicon into G=3 groups, which are respectively: the first 3,000 most common words, the second 3,000 most common words, and rare words not in the previous 6,000 words.

In order to check with a lemmatized frequency list, the vocabulary rarity rater needs to consult an electronic dictionary first, to convert the words in the input to lemmas. It looks up the widely used electronic dictionary WordNet (Miller et al. 2003) developed by the Cognitive Science Laboratory at Princeton University, through Java WordNet Library (JWNL, Didion et al. 2003), an open source API developed at Stanford University for accessing WordNet-style relational dictionaries.

To sum up, the procedure for the vocabulary rarity rater is as follows:

1. Read the input files, and tokenize them;
2. Tag the input files to resolve homographs;
3. Keep only the word tokens (as opposed to punctuations and symbols) for further analysis;
4. Read the BNC lemmatized frequency list or a user-specified frequency list;
5. Divide the whole lexicon into *G* groups of size *N* specified by the user;
6. For each word in the input files, look up in WordNet for its lemma; for words not found in WordNet, use the word as its lemma;
7. For each lemma, check with the frequency list, and decide which of the *G* lexical groups it belongs to;
8. Count the total number of words (including repetitions) in each of the *G* groups;
9. For each file, compute the percentage of words belonging to each group against the total number of words.

Then there is the TTR calculator. As was mentioned in 3.1, the TTR calculator computes three types of TTR: normal TTR, vanilla TTR and lemmatized TTR, as I call them. The algorithms to compute the three TTR's are similar, except that normal TTR works on word tokens only, and lemmatized TTR needs to work on the lemmas retrieved from WordNet on top of word tokens.

There is, nevertheless, one widely known limitation of TTR in measuring repetitiousness and vocabulary variability: TTR inevitably decreases as the text length increases (Scott 2004, Malvern and Richards 1997). The comparison between TTR's of texts of different length is, therefore, meaningless. The conventional strategy to deal with the problem is to compute the TTR for the first *W* words (Scott 2004). Yet, due to the great variability of speech sample lengths, it is impractical to guarantee that all samples would contain at least a pre-specified number *W* of tokens for TTR analysis. The TTR calculator solves the problem by first finding out dynamically the number of tokens *M* in the shortest of all the input files and then computing the TTR's for the first *M* number of tokens.

Briefly, the procedure for the lemmatized TTR calculator is as follows:

1. Read the input files, and tokenize them;
2. Keep only the word tokens (as opposed to punctuations and symbols) for further analysis;
3. Find the shortest of the input files and count its total number of tokens *M*;
4. For each word in the first *M* tokens of the input files, look up in WordNet for its lemma; for words not found in WordNet, use the word as its lemma;
5. Count the total number of distinct lemmas *L* for each file;
6. Compute the lemmatized TTR as *L*/*M* for each file.

Neither vocabulary rarity nor TTR can be computed manually without real hard work; it is a different story when the right tool is readily available. In the next three sections, I show how Vocabulary Analyzer has been used in a schizophrenia experiment, and discuss the results from the experiment.

3.3 Experiment

In this section, I first describe a schizophrenia experiment in detail (also in Covington et al. 2006a), which is used as a test bed for all the software tools developed in this dissertation.

12 controls (5 female, 7 male) (mean ± S.D. age=32.6 ± 12.4 years) and 11 patients (5 female, 6 male) (mean ± S.D. age=34.8 ± 12.6 years) were recruited for the experiment at the University of British Columbia. The controls were screened for a history of psychiatric illness, and all the 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.

All the patients had a diagnosis of schizophrenia according to DSM-IV criteria and the diagnosis was confirmed using the Signs and Symptoms of Psychotic Illness (SSPI) rating scale (Liddle et al. 2002) (mean ± S.D. SSPI=9.27 ± 6.3). The controls were evaluated for similarities to psychosis using the scales for Magical Ideation (Eckblad and Chapman 1983), and Perceptual Aberration (Chapman et al. 1978).

The controls and the patients 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). The mean ± S.D. years of education was 15.1 ± 2.5 for the schizophrenics and 17.9 ± 2.7 for the comparison subjects. There were no significant differences between patients and controls in terms of age (p=0.38) or years of education (p=0.11).

All the participants provided informed written consent and were screened for MRI compatibility. All experimental procedures were approved by the University of British Columbia’s Clinical Research Ethics Board, and use of the data was approved by the Human Subjects Office of the University of Georgia.

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 serve as the input files to all the software tools described in the dissertation.

One important fact about the experiment that needs special attention is that the patients were stable outpatients with no recent changes to their medication. Of the 11, 3 (27%) were receiving typical and 6 (55%) atypical antipsychotic medication; 2 (18%) were receiving both. Additionally, 5 patients (45%) were also taking antidepressants, all of which were serotonin reuptake inhibitors (Weinstein et al. 2005).

The stability of the patients is crucial, since schizophrenic symptoms were significantly reduced by the medication. It is not surprising that there is no florid symptom in the speech samples and that there is almost no apparent linguistic difference between the speech samples of the patients and those of the controls that can be detected simply by reading the transcripts.

As will be shown in the following chapters, the linguistic differences between the control and the patient group are relatively small. With only 12 and 11 subjects in each of the two groups, the power of our experiment is sometimes not sufficiently high to detect the differences. Hence, the purpose of applying the software to these speech samples is more to prove the usability of NLP tools in psychiatric research than to make new psycholinguistic discoveries concerning schizophrenic speech, though interesting results from the experiments will be discussed.

For the experiment in this chapter concerning the lexical deviances of schizophrenic speech, the transcripts from the above-mentioned schizophrenia experiment were fed into the current version of Vocabulary Analyzer, i.e. version 0.5. The default values were used as the extra arguments:

1. By default, Vocabulary Analyzer used OpenNLP Tagger;
2. By default, Vocabulary Analyzer used the lemmatized BNC word frequency list;
3. By default, Vocabulary Analyzer used the first 6,000 words in the lemmatized BNC word frequency list as common vocabulary, and considered anything outside that as rare words;
4. By default, the common vocabulary was divided into 500-word groups;

The output was saved and later statistically analyzed. Unpaired one-tailed t-tests were done to see if there was any lexical difference between the patient group and the control group. Power analysis was done to determine the power of the experiment.

3.4 Results

Neologism and stilted speech are some accepted symptoms of schizophrenic speech, but they are considered florid symptoms, which are infrequent. It would be interesting to see whether there is a general tendency of schizophrenic patients to use less common words than normal, and whether rare words are what patients resort to when they have difficulties retrieving the right words.

Therefore, the null hypothesis for the word rarity test is that there is no difference between the patient group and the normal group in their use of rare words. A null hypothesis about the use of the most common words is also tested: there is no difference between the patient group and the normal group in their use of the first 500 most common English words. Table 3.1 is the results concerning vocabulary rarity of the speech samples generated by Vocabulary Analyzer together with the results from the t-tests.

The results from our speech samples do show a significant difference ($p<0.04$) between the two subject groups in their use of rare words defined as words outside the 6,000 most common English words. The first null hypothesis is therefore rejected. Very surprisingly, however, the results show that the patients use significantly lower percentage of rare words than the normal controls. The power of this experiment is about 0.4 at the observed variance to detect the observed group difference.

Although we are able to see that the patients use slightly higher percentage of the first 500 most common English words than the normal controls in this experiment, at the 0.05 significance level, the difference between the two groups is only close to, but not exactly, significant. Statistical analysis shows that with the variance observed in the experiment, the power of this experiment is low at only about 0.3, and that in order to reach a power of 0.8 so that a group difference of over 0.02 can be more reliably detected, 25 subjects for each group will be needed.

SUBJECT	PERCENTAGE OF WORDS IN THE FIRST 500 MOST COMMON	PERCENTAGE OF RARE WORDS
Normal Control 1	0.787739	0.064368
Normal Control 2	0.714908	0.066376
Normal Control 3	0.768506	0.072678
Normal Control 4	0.70765	0.093807
Normal Control 5	0.771995	0.078067
Normal Control 6	0.758621	0.070155
Normal Control 7	0.747619	0.075238
Normal Control 8	0.746997	0.063063
Normal Control 9	0.738574	0.067642
Normal Control 10	0.742701	0.07208
Normal Control 11	0.737274	0.082102
Normal Control 12	0.717272	0.086909
Patient 1	0.773556	0.047112
Patient 2	0.739884	0.07948
Patient 3	0.725709	0.067068
Patient 4	0.776507	0.068607
Patient 5	0.753131	0.075134
Patient 6	0.780723	0.06506
Patient 7	0.747262	0.076664
Patient 8	0.779519	0.073145
Patient 9	0.722461	0.060086
Patient 10	0.780549	0.071072
Patient 11	0.782396	0.046455
1-tailed t-test	0.070685	0.037494
Control Average	0.744988	0.074374
Patient Average	0.760154	0.066353

Table 3.1 Word Rarity Results from Vocabulary Analyzer

The second part of the experiment concerns TTR. As was mentioned, in order for the TTR’s between the two subject groups to be comparable, Vocabulary Analyzer adopts the strategy of first counting the number of tokens *F* in the shortest transcript and then analyzing only the first *F* tokens in each of the input files, which is, in the case of this experiment, around 400. (Note that the number of tokens varies with different definitions of TTR.)

The null hypotheses for this part of the experiment are: there is no significant difference between the patients’ and the controls’ TTR’s. Table 3.2 is the results concerning Vanilla TTR, normal TTR, and lemmatized TTR generated by Vocabulary Analyzer together with statistical analysis of the data.

There is no evidence to reject the null hypotheses based on the results in Table 3.2. No significant difference in the degree of repetitiousness or in vocabulary variability can be detected in the experiment.

Unfortunately, with the number of subjects available and with the small TTR differences between the two groups, the power of the TTR experiment is below 0.1. In order to reach a power of 0.8 so that a group difference of 0.006 can be detected more reliably, over 200 subjects for each group will need to be enrolled in the experiment.

SUBJECT	VANILLA TTR	NORMAL TTR	LEMMATIZED TTR
Normal Control 1	0.342672	0.378973	0.349633
Normal Control 2	0.37931	0.422983	0.405868
Normal Control 3	0.37931	0.432763	0.400978
Normal Control 4	0.364224	0.415648	0.393643
Normal Control 5	0.353448	0.396088	0.371638
Normal Control 6	0.396552	0.447433	0.422983
Normal Control 7	0.359914	0.403423	0.374083
Normal Control 8	0.346983	0.383863	0.364303
Normal Control 9	0.364224	0.405868	0.391198
Normal Control 10	0.355603	0.400978	0.386308
Normal Control 11	0.37069	0.398533	0.374083
Normal Control 12	0.381466	0.422983	0.403423
Patient 1	0.318966	0.356968	0.334963
Patient 2	0.377155	0.435208	0.413203
Patient 3	0.327586	0.364303	0.327628
Patient 4	0.400862	0.464548	0.440098
Patient 5	0.407328	0.447433	0.415648
Patient 6	0.368534	0.405868	0.378973
Patient 7	0.351293	0.386308	0.364303
Patient 8	0.387931	0.432763	0.403423
Patient 9	0.31681	0.364303	0.349633
Patient 10	0.392241	0.437653	0.405868
Patient 11	0.342672	0.381418	0.354523
1-tailed t-test	0.378532	0.43283	0.32312
Control Average	0.3662	0.409128	0.386512
Patient Average	0.362853	0.406979	0.380751

Table 3.2 TTR Results from Vocabulary Analyzer

3.5 Discussion

In this experiment, 23 speech transcripts were fed to Vocabulary Analyzer to process for word rarity and three different kinds of TTR's. The lengths of the transcripts range from around 400 words to over 1,300 words. Manually mapping so many words to a word frequency list or manually computing any kind of TTR would be nothing short of inconceivable. Vocabulary Analyzer, however, finishes all the computation within a minute. Moreover, changing the values of the user-specified parameters such as word group size and common vocabulary size is simply effortless.

The results from the experiment are interesting. As was mentioned, the patients in the experiment in this dissertation are stable patients under medication, and their speech samples do not contain florid symptoms. Although there is no human-detectable deviance in their vocabulary, their choice of words does differ from normal as expected. Quite surprisingly, however, the patients use significantly fewer rare words than do the normal controls, and there is an insignificant tendency for the patients to use more of the most common words in the lexicon. The results cannot be attributed to IQ, education, or age, as the two groups were matched for those items. Brain deterioration which causes a decrease in vocabulary size may be an explanation, and this calls for further research on the correlation between vocabulary rarity and vocabulary size.

The fact that TTR's do not significantly differentiate the subject groups does not necessarily mean that schizophrenic patients exhibit the same degree of repetitiousness as normal speakers. It could well be that TTR is not a highly sensitive measure, and there may be other tests more suitable for measuring repetitiousness in speech (See Covington et al. 2006a for another measure of repetitiousness). It could also be that the speech of the stable patients in this experiment indeed does not contain so much repetition as that of most other schizophrenic patients.

To sum up, vocabulary is the level of language study that involves a particularly large amount of counting and matching – tasks easy for the machines but which can be unreasonably onerous for researchers to perform manually. With facilities made available by modern corpus linguistics, such as word frequency lists, and the constantly improving relational electronic dictionaries, many of the traditional measures concerning lexicon and semantic relations can be implemented and applied in psychiatric research to efficiently obtain results.

CHAPTER FOUR

D-LEVEL RATER

D-Level Rater is an NLP implementation that aims at providing psychiatrists with an informative and revealing measure of syntactic complexity. In the following sections, I introduce the background of the original D-Level scale and a revised D-Level scale, explain why D-Level is one of the more effective scales to rate syntactic complexity in psychiatric research, and show how computing D-Level leads to finding out some interesting structural patterns in schizophrenic speech.

4.1 Background

Syntactic complexity is an important topic in research on brain diseases such as aphasia (Haarmann and Kolk 1994, Thompson et al. 2003), Alzheimer's (Kempler and Curtiss 1987, Tomoeda et al. 1990, Kemper et al. 1993), and schizophrenia (Thomas 1996, DeLisi 2001), since deviant syntactic structures are often a direct reflection of brain damage. A number of syntactic complexity scales have been adopted in psychiatric studies. Among the most frequently used are mean length of utterance (MLU, Brown 1973), Developmental Sentence Scoring (DSS, Lee 1974), the Index of Productive Syntax (IPSyn, Scarborough 1990), and Developmental Level Scale (D-Level, Rosenberg and Abbeduto 1987).

Brown (1973: 53-54) describes the procedure to compute MLU by counting the number of morphemes. Developmental Sentence Scoring (Lee 1974: 136) is the sum of points earned by using grammatical forms in eight categories: indefinite pronoun or noun modifier, personal pronoun, main verb, secondary verb, negative, conjunction, interrogative reversal in questions and wh-questions. IPSyn (Scarborough 1990: 14-21) counts up to two instances of each of the 56 grammatical types regarding noun phrases, verb phrases, questions and sentence structures. And D-Level, as originally developed by Rosenberg and Abbeduto (1987), is a 7-point sentence complexity scale based on the authors' survey on child language development patterns. By definition, the later a syntactic structure is acquired by children, the more complex the structure is.

There are several reasons why D-Level is considered better suited for schizophrenia research.

First, D-Level Scale has been proven suitable as a syntactic complexity scale for psychiatric research (Kemper et al. 1995, Snowdon et al. 1996, Garrard et al. 2005). D-Level was adopted in the famous Nun Study (see section 5.1 for more information on the Nun Study) to experimentally demonstrate the relationship between early low syntactic complexity and later development of Alzheimer's (Snowdon et al. 1996, Kemper et al. 2001). Studies of schizophrenic speech are similar in nature to the Nun Study, in that both aim to predict the occurrences of brain diseases through linguistic deviances.

Second, D-Level, when slightly revised (Covington et al. 2006b), is applicable to most corpora regardless of the size of the corpora, while scales like DSS and IPSyn require a certain number of utterances before an analysis can be done. DSS, for instance, needs a collection of 50 utterances, each with both a subject and a verb. In schizophrenia studies, such a requirement is sometimes impractical, as patients (especially those with negative symptoms), and even some normal controls, may not be so productive in their speech.

Third, D-Level takes into consideration more factors than just the number of occurrences of certain grammatical structures. For example, both IPSyn and D-Level note the occurrence of relative clauses; yet, D-Level scores very differently when the relative clause is used to modify the object of the sentence from the occasion when the relative clause is used to modify the subject of a sentence (Level 3 vs. Level 6 as per Covington et al. 2006b). D-Level scores reveal much more structural / locational information than mere counts of grammatical forms.

4.2 Revised D-Level Scale

The original D-Level is, however, not ready to be used for schizophrenia research as it is. Covington et al. (2006b) refines the traditional D-Level scale. The main purpose of the revision is to generalize the original scale so that all sentence types can be evaluated. For example, the revised scale now rates simple sentences, elliptical sentences and wh-questions, which are common structures in speech but left unrated in the original scale. Such generalization is absolutely necessary to make D-Level a better research tool, because it is normal to study the complexity level of speech in psychiatric studies by averaging the complexity of each sentence, rather than by noting the maximum complexity level reached by the speakers as Rosenberg and Abeduto did with their subjects.

In addition, a few structures have also been re-leveled in the revised scale based on evidence from current psycholinguistic experiments. In particular, the original Level 5 and Level 6 structures are switched. According to the original D-Level, the extremely rare subject-with-clause structure is rated Level 5, while the relatively very common subordinate clause is rated Level 6. Based on results from studies on child-corpora (Bowerman 1979, Miller 1981) and related psychiatric experiments, the revised D-Level corrected the misplacement of the two levels.

Table 4.1 is a brief overview of the revised D-Level scale, including a brief description of the structures and a sample sentence for each level (see Covington et al. 2006b for detailed specifications). The D-Level Rater introduced in this chapter is an NLP implementation built upon the revised D-Level.

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.

Table 4.1 Revised D-Level Scale (Covington et al. 2006b)

4.3 Implementation

D-Level is appropriate and informative as a syntactic complexity scale for schizophrenia research; nevertheless, an automatic D-Level Rater is more difficult to implement than raters for other syntactic scales such as DSS or IPSyn, which count surface grammatical structures. D-Level, on the other hand, scores sentences based on the deep structures (see Chomsky 1965 for the distinction between "surface" and "deep" structures).

D-Level Rater is built upon cutting-edge stochastic parsing technologies. Sentences are rated on top of the parsing results by either of the two default parsers: Stanford Parser (Klein and Manning 2002) or OpenNLP Parser (Baldridge and Morton 2004). The users, however, can choose to plug in other high accuracy parsers for D-Level rating without re-compilation.

D-Level Rater follows a four-step procedure:

1. Deep-parse a speech transcript with a Penn Treebank style parser. The preference for Penn Treebank style is based on the popularity of the tree style.
2. Analyze the parser output and extract significant structural features from each parse which are specifically mentioned in D-Level scale. The grammatical functions of clauses and phrases are determined based on word order, clause types, verb forms, and government relations based on the parse tree.
3. Rate each sentence according to the revised D-Level scale.
4. Calculate the percentage of each level for the transcript.

Even with a high precision parser, automatic rating of D-Level is not trivial mainly due to three difficulties:

1. Simple tag/word search does not work. For example, even the simple coordinate structures cannot be determined only based on the existence of "CC" (conjunction, coordination) tags or terminals like *and*. This is because the revised D-Level scale does not regard sentences starting with *and* as a coordinate structure, which are very common in spontaneous speech and where *and* works like a pause filler rather than a coordinator. So Figure 4.1 does not contain a Level 2 coordinate structure, while Figure 4.2 does, where the coordinate structure is at the end of the sentence.

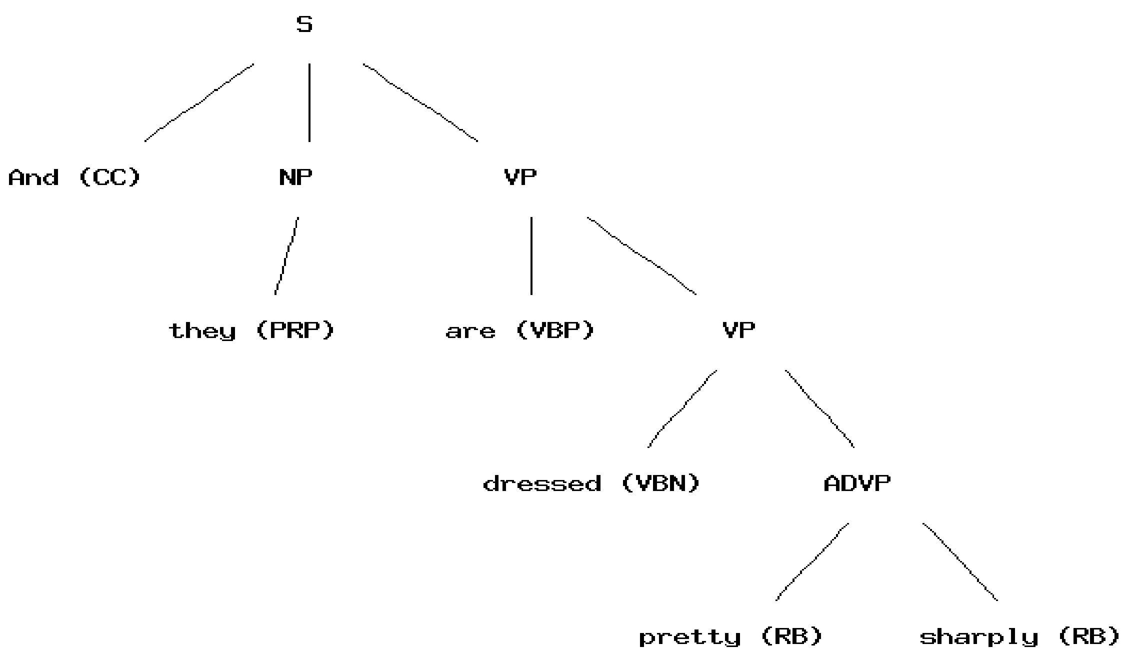

Figure 4.1 "and" not in coordinate structure

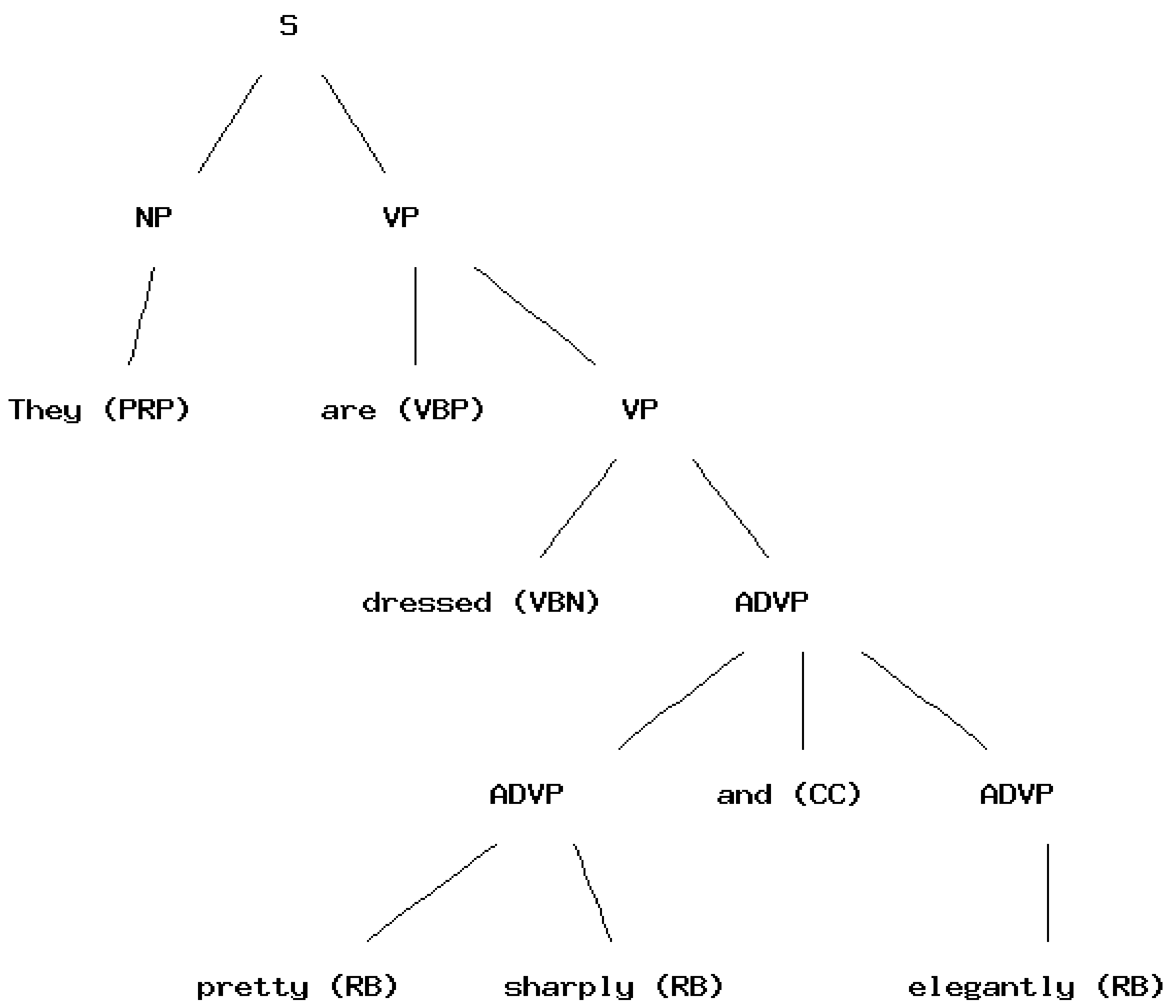

Figure 4.2 “and” in coordinate structure

2. There are occasions when the same surface structure and the same parse tree do not guarantee the same syntactic structure. For example, the parse trees generated by Stanford Parser for sentences *He talked about dancing with the stars* (Figure 4.3) and *He talked after dancing with the stars* (Figure 4.4) are exactly the same. However, the two sentences actually vary in their deep structures: while the PP (prepositional phrase)

about dancing with the stars is the complement of the verb *talked*, the PP *after dancing with the stars* serves as the adjunct of *talked*.

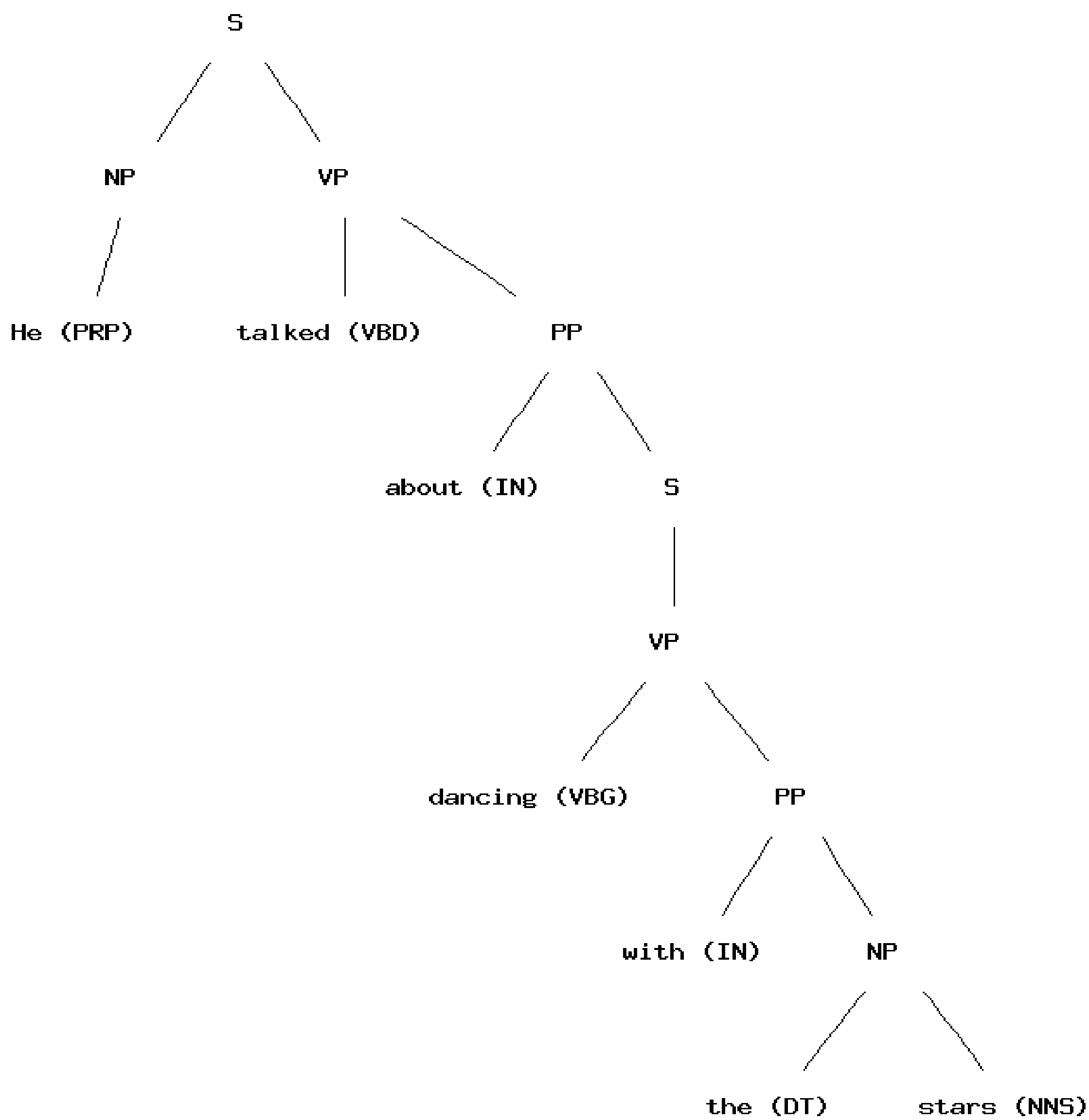

Figure 4.3 PP as Complement

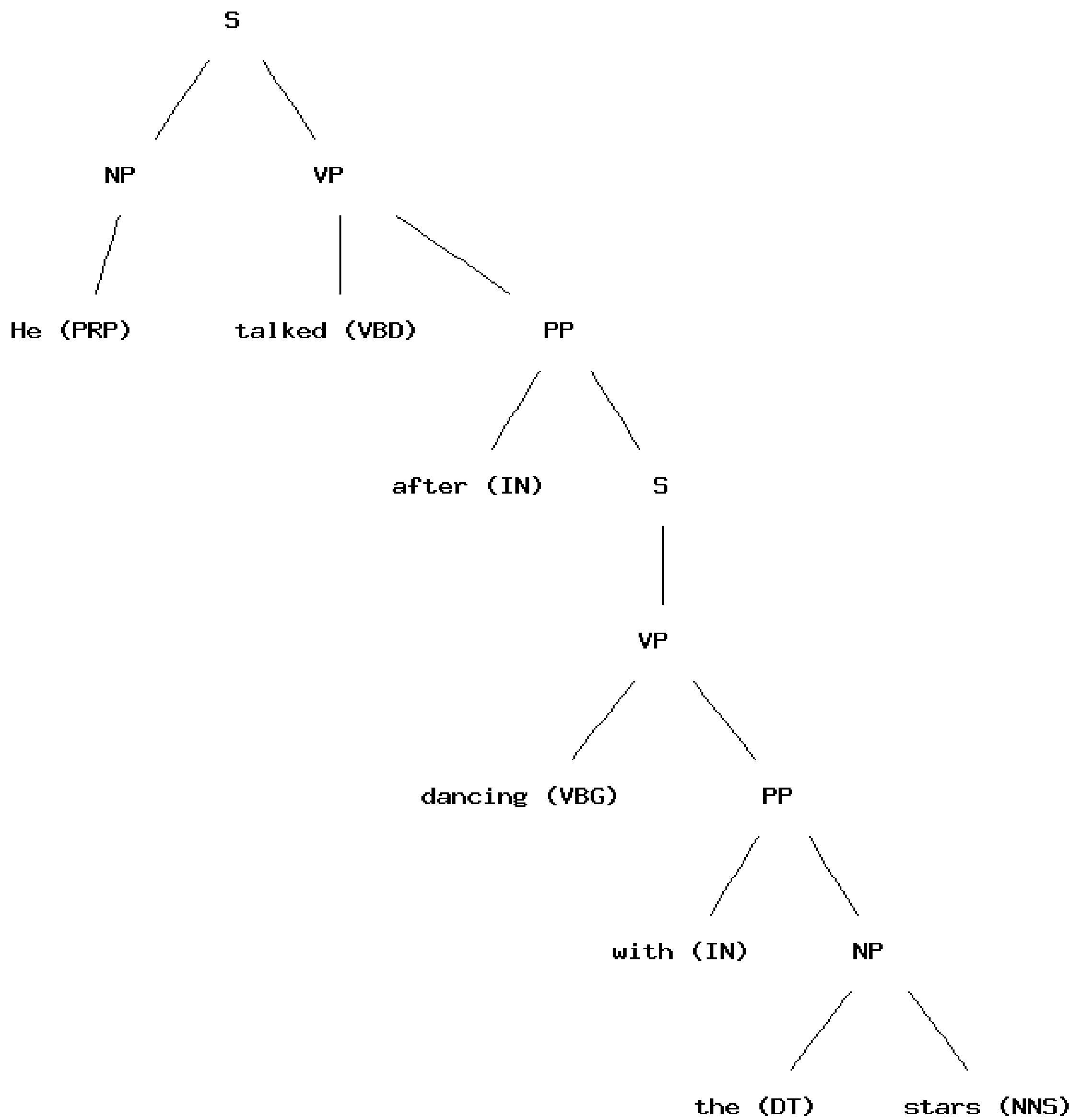

Figure 4.4 PP as Adjunct

3. Parser errors cannot be neglected and need to be counteracted if possible. As was mentioned in section 2.1, precision is still a concern today even with the best parsers. Figure 4.5 is a sample error for the sentence *I need some more samples when things*

become so complicated. from Stanford Parser: *when things become so complicated* is wrongly parsed as attached to the NP node rather than the VP or the S node. This basically means, incorrectly, that the *when*-clause works as a relative clause for the NP rather than an adverbial clause for the sentence. This wrong attachment, nevertheless, does not render the whole parse useless, as most of the substructures such as the SBAR subtree are correctly parsed. In order to counteract such parser errors, D-Level Rater takes into consideration the fact that certain SBAR structures, such as those beginning with *when*, have a much higher probability to be an adjunct clause than a relative clause, and override the structure represented by the parse tree.

Other measures that D-Level Rater takes to increase its reliability and precision include the hot swap of parsers to facilitate comparison of results from different parsers and taking into consideration global information of the parse trees such as tree depth. These measures are acknowledgement of the fact that parser errors are inevitable and have been proven helpful in increasing the precision of D-Level Rater.

Last but not least, ease of use, as a major objective of D-Level Rater, is also achieved by providing a *verbose* option to psychiatrists with little linguistic knowledge, so they can see the justifications for each of the ratings, and incorrect results can be overridden where necessary.

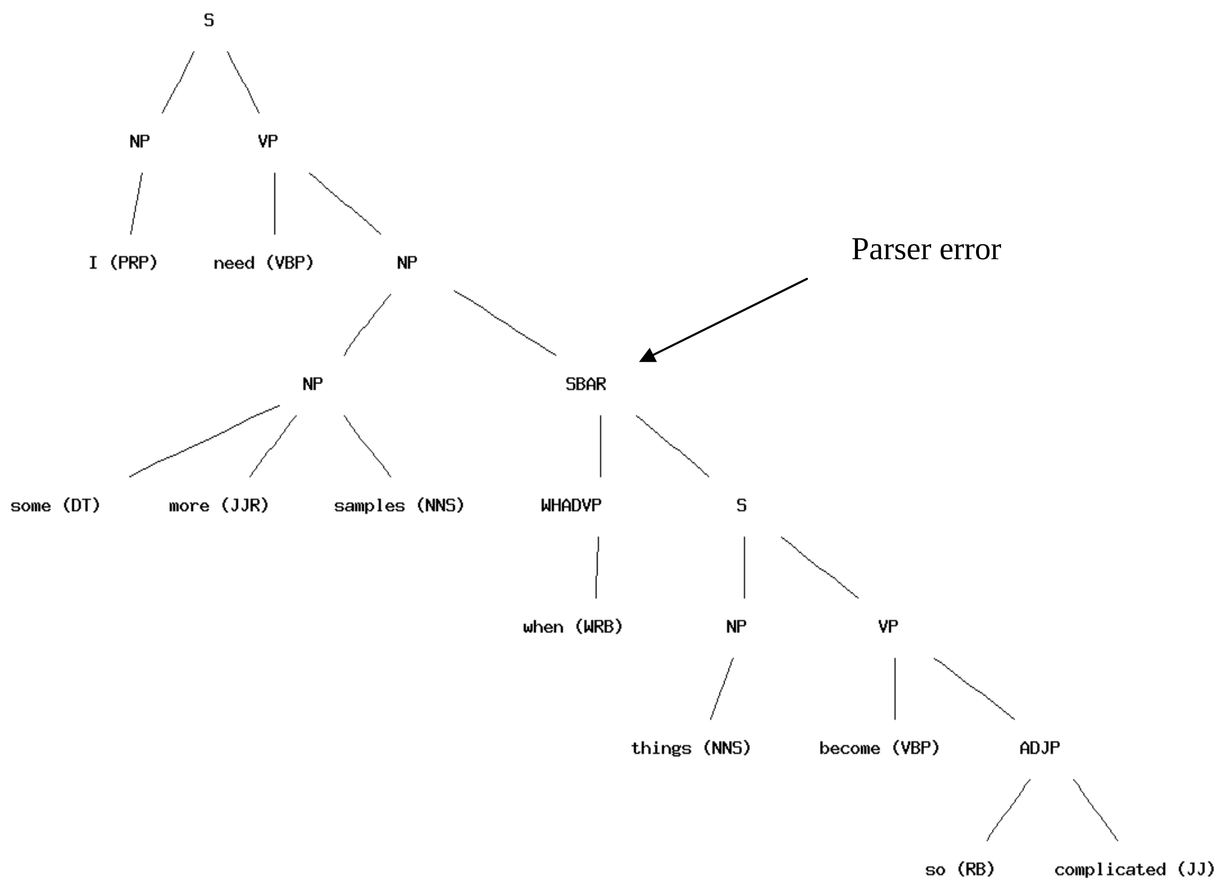

Figure 4.5 A Sample Parser Error

4.4 Reliability of D-Level Rater

Although D-Level Rater has its core technology based on parsers, it consistently achieves a slightly better precision than its base parser utilities in several tests on randomly collected sentences. This is due to two reasons. First, measures have been taken to override parser errors as described in section 4.3. Second, as some of the complexity levels comprise several different structures, D-Level Rater does not require a correct distinction between these structures for an accurate rating. For instance, as both objective clauses and relative clauses modifying objects are rated Level 3, D-Level Rater would still come up with a correct rating, even if the parser mistakes a relative clause for an objective clause.

To ensure that the performance of D-Level Rater is reasonably reliable and accurate, a small scale experiment was designed and conducted to test both the reliability and the precision of D-Level Rater version 0.5. In addition to D-Level Rater version 0.5, four human raters participated in the experiment independently. The human raters included both linguists and non-linguists, both native speakers of English and non-native speakers, all of whom had been trained to rate sentences with D-Level scale. Five different paragraphs were chosen for the experiment, which consisted of 37 sentences from different styles of literature ranging from newspapers to children's books, thus covering all levels of syntactic complexity. Both the machine and the human raters rated the sentences based on the revised D-Level scale. The parser D-Level Rater version 0.5 used is the default OpenNLP parser. For the human raters, there was no restriction on the time used for rating the sentences, and it generally took about an hour to finish the rating.

The results from the experiment are satisfactory. Of the 37 sentences, D-Level Rater agreed with at least one human rater on 35 sentences (94.6%), and its ratings concurred with at least half of the four human raters on 33 sentences (89.2%). Inter-rater reliabilities were also computed based on Robert Ebel's algorithm (Ebel 1951) using the Inter-rater Reliability Calculator available on the website of *Medical Education Online* (http://www.med-ed-online.org/rating/reliability.html). The inter-rater reliability between the human raters was 0.93, and the inclusion of D-Level Rater version 0.5 effectively raised the inter-rater reliability to 0.95.

4.5 Experiment

D-Level Rater was used to rate the syntactic complexity of the speech samples from the schizophrenia experiment described in section 3.3.

The transcripts of the speech samples were input into D-Level Rater version 0.5. After each sentence was rated based on parsing results from the default OpenNLP Parser, the frequency of each complexity level was computed for each of the transcripts. The results were then statistically analyzed using t-tests to compare the performance between the patient group and the control group. Power analysis was done to test the power of the experiment.

4.6 Results

Syntactic complexity has been repeatedly reported to be compromised in schizophrenia speech (Morice and Ingram 1982, Morice and McNicol 1985, 1986, DeLisi 2001). Two unpaired one-tailed t-tests were done on the data computed by D-Level Rater version 0.5 with the following null hypotheses: 1. schizophrenic speech would contain at most an equal percentage of Level 0 sentences as normal speech; 2. schizophrenic speech would contain at least an equal percentage of Level 7 sentences as normal speech. The results from the experiment show a significant increase in the patients' use of Level 0 simple sentences ($p<0.001$) and a marked decrease in Level 7 sentences ($p<0.02$), thus rejecting the null hypotheses. Table 4.2 shows the detailed results from the two subject groups, and Figure 4.6 is a bar graph on the group average differences.

SUBJECT	LEVEL 0 PERCENTAGE	LEVEL 7 PERCENTAGE
Normal Control 1	0.146341	0.682927
Normal Control 2	0.222222	0.527778
Normal Control 3	0.230769	0.487179
Normal Control 4	0.196078	0.509804
Normal Control 5	0.232558	0.534884
Normal Control 6	0.238095	0.5
Normal Control 7	0.183673	0.489796
Normal Control 8	0.136364	0.530303
Normal Control 9	0.242424	0.454545
Normal Control 10	0.173913	0.608696
Normal Control 11	0.090909	0.787879
Normal Control 12	0.230769	0.666667
Patient 1	0.315789	0.631579
Patient 2	0.162162	0.513514
Patient 3	0.136364	0.636364
Patient 4	0.192308	0.461538
Patient 5	0.431034	0.258621
Patient 6	0.314286	0.571429
Patient 7	0.333333	0.577778
Patient 8	0.425926	0.407407
Patient 9	0.392857	0.321429
Patient 10	0.38961	0.285714
Patient 11	0.463415	0.073171
1-tailed t-test	0.000817	0.018053
Control Average	0.193676	0.565038
Patient Average	0.323371	0.430777

Table 4.2 Results on Level 0 and Level 7 Percentages

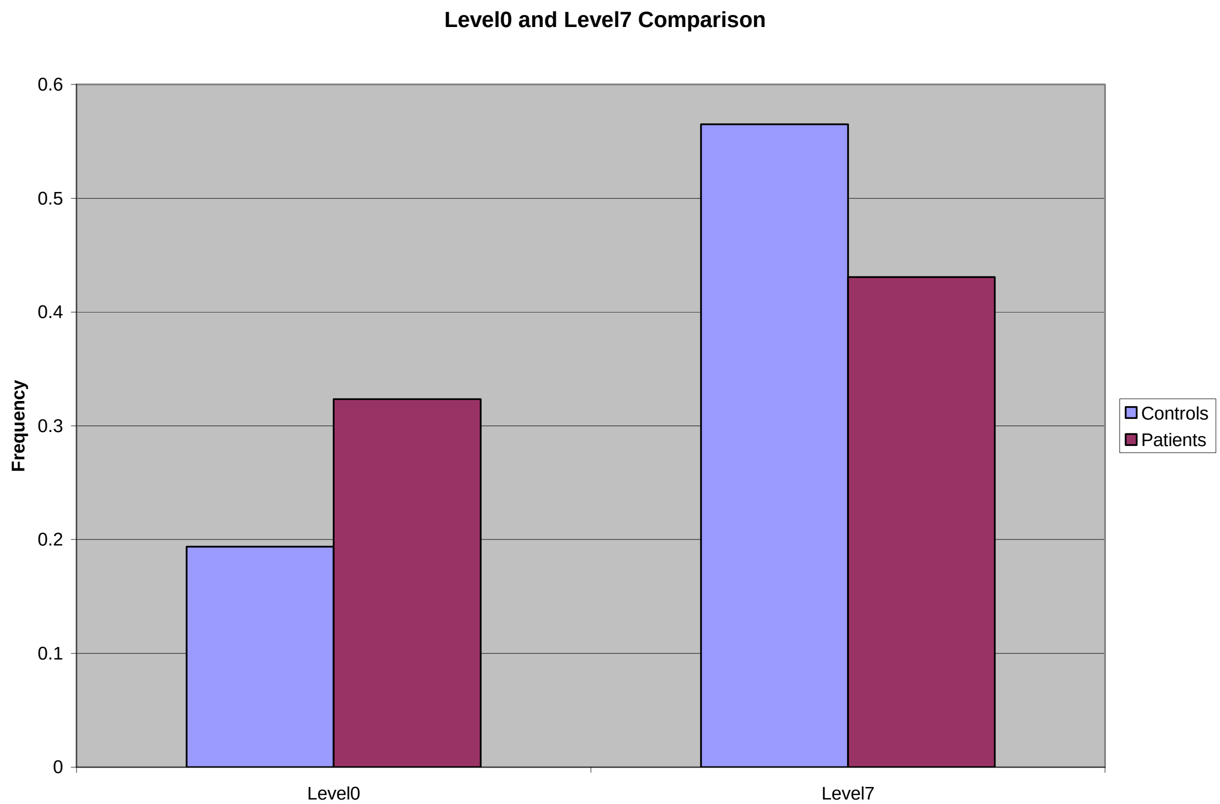

Figure 4.6 Level 0 and Level 7 Averages of Controls and Patients

Since Level 7 is a label that wraps around sentences with substructures at multiple D-Levels, D-Level Rater also provides the option to look into the probability of the occurrence of substructures as defined for Level 1 to Level 6 per sentence. For instance, a sentence that is rated Level 7 because of its substructures of Level 2 and Level 3 is now counted as one occurrence of Level 2 and one occurrence of Level 3. The probability of the occurrence of structures between Level 1 and Level 6 per sentence is defined as the total number of occurrences of each structure divided by the total number of sentences.

Because there are actually six null hypotheses being tested, I chose to use the very conservative Bonferroni method for multiple comparisons, and the significance level is raised from the usual 0.05 to 0.01. Such analysis shows that in our experiment, increased syntactic complexity in the normal controls is due to the fact that they use adjunct clauses ($p<0.01$) more often than the schizophrenic patients, and that the increase in coordinate structures ($p<0.02$) and that in "clauses in objects" ($p<0.02$) are also close to being significant. Table 4.3 shows the detailed results on the probabilities of substructure occurrences per sentence.

Power analysis of the experiment shows a quite decent power of about 0.8 for the Level 0 test and a power of about 0.5 for all the other tests. For these tests, increasing the number of subjects to 22 for each group will raise the power to 0.8, so that the group differences at the current level can be reliably detected.

SUBJECT	COORDINATE STRUCTURE	OBJECTIVE/RELATIVE CLAUSE	ADJUNCT CLAUSE
Normal Control 1	0.707317	0.487805	0.585366
Normal Control 2	0.5	0.388889	0.555556
Normal Control 3	0.487179	0.487179	0.487179
Normal Control 4	0.529412	0.431373	0.411765
Normal Control 5	0.418605	0.534884	0.418605
Normal Control 6	0.452381	0.47619	0.47619
Normal Control 7	0.367347	0.612245	0.408163
Normal Control 8	0.575758	0.409091	0.439394
Normal Control 9	0.515152	0.515152	0.454545
Normal Control 10	0.608696	0.413043	0.565217
Normal Control 11	0.606061	0.69697	0.575758
Normal Control 12	0.615385	0.461538	0.384615
Patient 1	0.631579	0.526316	0.473684
Patient 2	0.459459	0.459459	0.351351
Patient 3	0.613636	0.590909	0.545455
Patient 4	0.442308	0.326923	0.480769
Patient 5	0.241379	0.344828	0.189655
Patient 6	0.542857	0.485714	0.342857
Patient 7	0.466667	0.488889	0.466667
Patient 8	0.296296	0.37037	0.481481
Patient 9	0.410714	0.232143	0.321429
Patient 10	0.285714	0.363636	0.194805
Patient 11	0.243902	0.097561	0.073171
1-tailed t-test	0.018672	0.023452	0.009593
Control Average	0.531941	0.492863	0.480196
Patient Average	0.421319	0.389704	0.356484

Table 4.3 Results on Substructure Frequencies

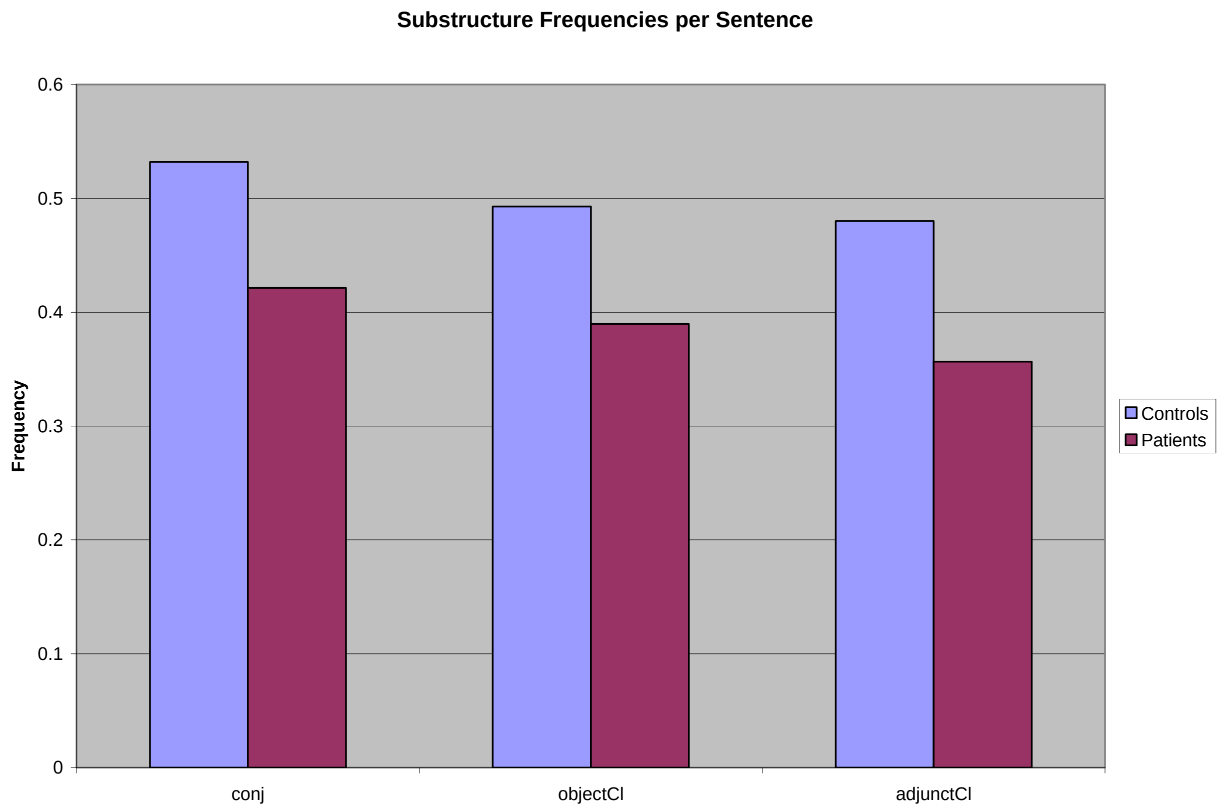

Figure 4.7 Averages of Substructure Frequencies Per Sentence

4.7 Discussion

The lowered sentence complexity measured with D-Level found in the Nun Study for Alzheimer's patients is replicated in schizophrenic patients in our experiment. Our results also conform to the manual analysis of sentence complexity in similar schizophrenia studies (e.g. Morice and Ingram 1982, DeLisi 2001).

What is more interesting is that D-Level Rater is able to assist in finding out that the decreased complexity is 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. Clausal subject does not seem to be a distinguishing category, because even the normal controls rarely use clausal subjects. Counting the number of substructures is simply a by-product in rating D-Level for machines; if the same thing should be done manually, it will be much more time-consuming than just rating D-Level for each sentence.

In short, D-Level Rater is a syntactic level NLP application that automates the demanding and tedious process of rating each sentence in the text input. D-Level Rater is not only fast and easy-to-use, but it provides the possibility of discovering significant psycholinguistic features that elude manual rating.

CHAPTER FIVE

IDEA DENSITY RATER

Idea Density Rater is an NLP application that analyzes language at the semantic level. In this chapter, I introduce the concept of idea density, explain how a heuristic can be used to simplify the implementation of Idea Density Rater, and show the results of an experiment done with the Rater.

5.1 Background

Semantic abnormality has long been an important field of study in schizophrenia research, but many more studies focus on semantic relationships between words than on proposition-level semantic contents, and more experiments are done over isolated words than over spontaneous continuous speech. Unlike single words without context, the semantic meaning of sentences in continuous speech are much harder to formalize and quantify in order to be represented by some measure that corresponds to psychological units rather than words themselves.

Idea density is exactly the kind of measure designed to address the above problem by standardizing the quantification of psychological processing of texts. Idea density is defined as the number of ideas (or propositions) per *N* words. In this dissertation and the implementation of Idea Density Rater, *N* is set to be 100.

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 ($n \geq 1$)." Turner and Greene (1977) created a practical manual for proposition analysis, where propositions are further defined to be "a representation of a conceptual unit" and a relation between a set of arguments (ibid.). Turner and Greene classify propositions into three types: Predication, Modification, and Connection. With Turner and Greene's refinement, the concept of proposition becomes more formalized and procedural, and, in turn, more suitable for content evaluation in psychiatric experiments. Table 5.1 is a very brief summary of Turner and Greene's three types of propositions (ibid.).

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))

Table 5.1 Turner and Greene's Proposition Types

Experiments (e.g. Kintsch and Keenan 1973) have demonstrated that idea density, rather than sentence length, is the key to semantic complexity. Kintsch and Keenan (1973) show that for sentences of the same length, the processing time of the sentences significantly correlates with the number of propositions in the sentences.

Idea density has been an important measure of text complexity widely used in psychiatric studies (e.g. Oller et al. 1995, Snowdon et al. 1996), among which the Nun Study is probably the most famous one. The Nun Study is a pilot longitudinal study on Alzheimer's disease, using data collected from 678 elderly nuns from all over the United States (Snowdon 2006). The nuns, with an average age of 85, have handwritten autobiographies from their early life (around 50 years ago), which have been studied to reveal relationships between Alzheimer's disease and cognitive disabilities as reflected by various text features. One of the most significant discoveries the Nun Study has made is low linguistic ability (including low idea density and low sentence complexity) in early life is a strong predictor of Alzheimer's disease in late life (e.g. Snowdon 1996, Kemper et al. 2001). Note that measures used in the Nun Study have been computed manually.

Like Alzheimer's disease, schizophrenia causes cognitive disabilities, which may exhibit themselves through low linguistic ability. It would, therefore, be interesting to see whether low idea density is a characteristic of schizophrenic speech. Proposition analysis, as we have seen, is not simple, and Idea Density Rater is exactly the tool needed for speedy processing of speech samples.

5.2 Implementation

Implementing an idea density rater strictly according to Turner and Greene's manual (1977) would be very challenging, if not impossible at all, as this would amount to a general understanding of literal meaning of human speech by the computer. A heuristic, however, has been briefly mentioned in Snowdon et al. (1996) on how to calculate idea density without actually going through the process of finding each relation between arguments. I call this approach "counting idea words."

Snowdon et al. (ibid.) note that the number of ideas or propositions is roughly approximated by the number of words of certain parts of speech, which I call "idea words." Idea words typically include verbs, adjectives, adverbs, prepositions, coordinators and subordinators. Very roughly, verbs are most likely to form Predications; adjective, adverbs and prepositions Modifications; and coordinators and subordinators Connections.

Take the sentence *John wants a small dog for a pet.* for example. Proposition analysis of the sentence results in three propositions:

1. Predication: (WANT, JOHN, DOG)
2. Modification: (QUALITY OF, DOG, SMALL)
3. Modification: (ROLE OF, DOG, PET)

If, on the other hand, we tag the words in the sentence first, we have:

John wants a small dog for a pet.

(NOUN, VERB, DETERMINER, ADJECTIVE, NOUN, PREPOSITION, DETERMINER, NOUN)

There is one verb, one adjective, and one preposition in the sentence, totaling three idea words, which is the exact number of propositions for the sentence. In both cases, the idea density of the sentence is 3 ideas over 8 total words, i.e. 37.5 ideas per hundred words.

The alternative to strict proposition analysis is, therefore, using the number of idea words to approximate the number of ideas when computing idea density. Below is the procedure for Idea Density Rater:

1. Tag the input files with a high-precision tagger;
2. For each file, count the total number of idea words *I* and the total number of words *T*;
3. Compute idea density as *100* I*/*T*.

Since Idea Density Rater first tags speech samples, it is capable of generating many other useful measures as by-products. Much can be done, in fact, with a tagged speech sample, such as computing the percentage of various parts of speech. Simple measures like verb-noun ratio can be very revealing with regard to brain damage (e.g. Druks 2002, Silveri et al. 2003, Bak et al. 2001, Kim and Thompson 2004, Berndt et al. 2002). Research has shown that patients with brain damage often have more difficulties dealing with words of more complex argument structure and less imageability, such as verbs, rather than words that are semantically simpler, such as nouns (Harris et al. 2005, Shapiro and Caramazza 2003).

5.3 Experiment

Idea Density Rater was used to rate the idea density of the speech samples from the schizophrenia experiment described in section 3.3.

The transcripts of the speech samples were input into Idea Density Rater version 0.5. After the transcripts were tagged with the default OpenNLP Tagger, idea density was computed for each transcript, as well as percentages of various POS tags.

Statistical analysis was done on the results obtained. Unpaired one-tailed *t*-tests were conducted to compare the patient and the control group on idea density ratings and verb-noun ratio. Power analysis was done to test the power of the experiment.

5.4 Results

Till this day, there has been no published experiment formally studying idea density of schizophrenic speech, although other semantic aspects of schizophrenic speech, such as word association and categorization, have been empirically proven deviant to a certain degree, as was discussed in section 2.2. The null hypothesis for this experiment is that there is no difference between the idea density of the controls' speech and that of the patients' speech. The results show that there is no significant difference between the idea densities of the two groups.

Also tested in the experiment is the null hypothesis that the verb-noun ratio of the controls' speech does not differ from that of the patients' speech. The results, though not significant at the 0.05 significance level, are somewhat within expectation, as they do show a slightly higher verb-noun ratio in the control's speech. Table 5.2 is the results from the experiment and the t-test results.

SUBJECT	IDEA DENSITY	VERB-NOUN RATIO
Normal Control 1	52.64256	0.64988
Normal Control 2	48.33006	0.573574
Normal Control 3	56.91158	0.873874
Normal Control 4	45.76412	0.556787
Normal Control 5	55.61927	0.932806
Normal Control 6	50.79365	0.719298
Normal Control 7	47.3251	0.714697
Normal Control 8	51.65144	0.808933
Normal Control 9	48.21429	0.732955
Normal Control 10	50.79225	0.623596
Normal Control 11	48.378	0.575829
Normal Control 12	53.38809	0.66065
Patient 1	48.07453	0.680365
Patient 2	45.40881	0.5
Patient 3	49.38744	0.729268
Patient 4	51.54265	0.68997
Patient 5	47.52	0.617925
Patient 6	51.62304	0.587097
Patient 7	49.03418	0.566449
Patient 8	52.24857	0.747159
Patient 9	50.06605	0.694323
Patient 10	52.60181	0.670455
Patient 11	43.94299	0.607692
1-tailed t-test	0.118358	0.0953
Control Average	50.81753	0.701907
Patient Average	49.22273	0.644609

Table 5.2 Results on Idea Density and Verb-Noun Ratio

The power of the idea density test is low at slightly over 0.2 with the variance observed in the experiment. In order to obtain a power of 0.8 so that a group difference of over 2 in idea density can be detected reliably, at least 40 subjects will be needed for each group. The power of the verb-noun ratio test is also low at 0.23 with the variance observed in the experiment. In order to obtain a power of 0.8 so that a group difference of over 0.06 in verb-noun ratio can be detected reliably, about 50 subjects will be needed for each group.

5.5 Discussion

Idea density is a well-defined measure for semantic complexity in continuous speech. Computing idea density manually is, however, a tedious and challenging task even with the "counting idea words" shortcut, which prevents the measure from being more widely adopted in psycholinguistic studies, including studies on schizophrenic speech. Idea Density Rater uses high-precision taggers, takes the "counting idea words" shortcut, and makes it possible to quickly obtain useful statistics for a large amount of text data.

Our experiment failed to show, between schizophrenic and normal speech, the kind of significant difference regarding idea density as reported in Alzheimer's experiments, which might indicate the diverse nature and distinct brain deterioration of the two diseases. The experiment is, however, not conclusive. On the one hand, the stable status of the patients may make their speech less representative; on the other hand, the power of our experiment on idea density is very low.

Many mental disorders exhibit similar linguistic tendency of increased use of nouns and decreased use of verbs, such as dementia (Silveri et al. 2003, Bak et al. 2001), Alzheimer's disease (Kim and Thompson 2004), and aphasia (Berndt et al. 2002). The difference between the accessibility of verbs and nouns lies in the fact that verbs have inherent complex argument structures, while nouns normally serve as arguments, which also gives rise to the fact that texts with higher verb-noun ratio may be syntactically more complicated. Therefore, besides being a measure of semantic complexity, verb-noun ratio can also be a primitive measure of syntactic complexity. There is, however, not enough evidence in the schizophrenic speech samples in our experiment to prove that there is a significant decrease in verb-noun ratio in schizophrenic speech. In order for a reliable conclusion to be reached, a more powerful experiment with many more subjects needs to be done.

CHAPTER SIX

FUTURE WORK AND CONCLUSION

Previously, I discussed the implementation of three NLP applications dealing with psycholinguistic measures respectively at lexical, syntactic, and semantic level. These applications are only the first steps toward great possibilities to advance research on schizophrenic speech, where much is left to be discovered. In this chapter, I discuss the areas where future work is to be done, and conclude the dissertation.

6.1 Future Work

It is obvious that much can be done in automated analysis of schizophrenic speech. Let us start with the already-existing applications like the ones introduced in this dissertation. It is vital to conduct extensive experiments to test the validity and sensitivity of the implemented psycholinguistic measures. 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. Sensitive measures are invaluable, because they can not only help detect the nature of brain diseases and prove or disprove previous theories, but also serve as diagnostic criteria well defined as in software implementations. And tools on sensitive measures deserve special efforts to be refined and made more sophisticated.

Second, a wide range of new tools need to be designed and implemented at various linguistic levels, which will doubtlessly lead to the discovery of unknown patterns and regularities. It is essential to keep up-to-date on cutting-edge NLP technologies, both to improve the reliability and precision of the software and to gain inspiration and insight on new implementations. As demonstrated by the applications in this dissertation, there is normally no apparent direct relation between NLP techniques and psycholinguistic measures, yet extensive knowledge of both can often establish the connection and combine them seamlessly in software design as was done in Chapter 2.

Moreover, although low-level applications based on mature NLP techniques have been shown to be useful and easier to implement, efforts on high-level applications, such as those on discoursal and pragmatic level, should be encouraged. In spite of the fact that algorithms for direct computation of high-level features are not available, and the fact that indirect solutions have not reached the desired performance level, high-level tools may still be able to provide important information, especially under human supervision.

Lastly, predictors or classifiers based on speech samples are to be implemented on the basis of speech analysis tools. Significant measures are not just to be used solely for psycholinguistic research purposes; ultimately, their value is proven by their capability to classify or predict.

The relationship between speech analyzers and classifiers for schizophrenia is self-evident: first, since classifiers accept as input only structured data like numericals and not unstructured data like natural language, obtaining structured data with the help of speech analysis tools is the absolute first step for classifiers; second, statistics from speech analyzers directly influence the performance of classifiers, as accurate rating of sensitive measures fundamentally determines the success of a classifier.

Equipped with various speech analyzers, we are ready for all kinds of classifiers: linear classifiers, neuronetworks, decision trees, etc. It would be interesting to compare the performance of standard or hybrid classifiers utilizing the significant psycholinguistic features extracted with speech analysis software.

6.2 Conclusion

Doing psycholinguistic analysis automatically rather than manually is the inevitable trend in psychiatric research. Computer-assisted analysis of schizophrenic speech is not simply a new interdisciplinary research area; it introduces a brand-new methodology.

The advantages of computer-assisted over manual speech analysis can be summarized as follows:

1. While manual text analysis requires professional training in linguistics, very little is expected from the user of text analysis software: very little linguistic knowledge and very little computer skill.
2. While manual analysis can turn out to be a laborious and complicated procedure that combines analysis, counting and computing, text analysis software is designed by software developers to be easy for end users.
3. There is a huge difference between the time and effort invested in manual analysis and

those in computer-assisted analysis. Seconds or minutes of work by the computer could very likely mean days or even weeks for human analysts.

4. Whereas subjectivity is more or less involved with human raters, text analysis software always outputs the same result with the same input.
5. For any experiment to be scientific, it needs to be replicable. Analysis results, or even errors, from text analysis software are easily replicable, while exact replication of results is seldom possible with human raters.

Current NLP technologies have made it possible for more sophisticated measures than the traditional simple counts to be rated automatically, as demonstrated by the implementations in this dissertation. High-precision automatic speech analysis at various linguistic levels, especially at lower levels such as phonology, lexicon, syntax and semantics, is feasible with today's technology, and would greatly facilitate data analysis for large-scale experiments. Automated analysis also lays the foundation for further research, such as disease prediction and classification.

It is worth mentioning, nevertheless, that automated psycholinguistic analysis assumes tremendous effort on the part of software designers, due to the interdisciplinary nature of the task. Capabilities in the field of software development, linguistics, or psychiatry, if not well integrated, do not guarantee useful end products for psychiatric research. It is crucial to base software design on solid understanding of psychiatric research problems, and to be able to see the often-hidden connection between psycholinguistic measures and NLP technologies is the key to success.

BIBLIOGRAPHY

Abu-Akel, Ahmad (1997) “A Study of Cohesive Patterns and Dynamic Choices Utilized by Two Schizophrenic Patients in Dialog, Pre- and Post-medication,” *Language and Speech 40:* 331–351.

Abu-Akel, Ahmad (1998) “The Study of Cohesion in Schizophrenia: Theory and Application,” *Issues in Applied Linguistics 9*:37–60.

Abu-Akel, Ahmad (1999) “Impaired Theory of Mind in Schizophrenia,” *Pragmatics and Cognition 7*:247–282.

Allen, H. A., and Frith, C. D. (1983) “Selective Retrieval and Free Emission of Category Exemplars in Schizophrenia,” *British Journal of Psychiatry 144*: 611-617.

Allen, H. A., Liddle, P. F., and Frith, C. D. (1993) “Negative Features, Retrieval Processes and Verbal Fluency in Schizophrenia,” *British Journal of Psychiatry 163*: 769-775.

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR).* 4th ed., text revision. Washington, D. C.: American Psychiatric Association.

American Psychiatric Association (2000) *Handbook of Psychiatric Measures*. Washington, D.C.: American Psychiatric Association.

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

Anand, A. Wales, R. J., Jackson, H. J., and Copolov, D. L. (1994) “Linguistic Impairment in Early Psychosis,” *The Journal of Nervous and Mental Disease 182(9)*: 488-493.

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

Andreasen, Nancy C. (1979b) “Thought, Language, and Communication Disorders: Diagnostic Significance,” *Archives of General Psychiatry 36*:1325–1330.

Andreasen, Nancy C. (1982) “Negative Symptoms in Schizophrenia: Definition and Reliability,” *Archives of General Psychiatry 39*: 784-788.

Andreasen, Nancy C. (1984) *Scale for the Assessment of Positive Symptoms (SAPS).* Iowa City, IA: University of Iowa, 1984.

Andreasen, Nancy C. (1986) “Scale for the Assessment of Thought, Language, and Communication (TLC),” *Schizophrenia Bulletin 12*: 473–482.

Bak, T. H., O’Donovan, D. G., Xuereb, J. H., Boniface, S. and Hodges, J. R. (2001) “Selective Impairment of Verb Processing Associated with Pathological Changes in Brodmann Areas 44 and 45 in the Motor Neurone Disease-Dementia-Aphasia Syndrome,” *Brain 124*: 103-120.

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

Barch, D. M. and Berenbaum H. (1997) “The Effect of Language Production Manipulations on Negative Thought Disorder and Discourse Coherence Disturbances in Schizophrenia,” *Psychiatry Reseach 71(2)*:115-27.

Benes, F. M. (2000) “Emerging Principles of Altered Neural Circuitry in Schizophrenia,” *Brain Research. Brain Research Reviews 31(2-3)*: 251-69.

Berndt, R. S., Burton, M. W., Haendiges, A. N., and Mitchum, C. C. (2002) “Production of Nouns and Verbs in Aphasia: Effects of Elicitation Context,” *Aphasiology 16(1/2)*: 83-106.

Bies, A., Ferguson, M., Katz, K., and MacIntyre, R. (1995) “Bracketing Guidelines for Treebank II Style Penn Treebank Project,” ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz.

Birchwood, M. and Jackson, C. (2001) *Schizophrenia.* Philadelphia, PA: Psychology Press Ltd.

Bowerman, M. (1979). “The Acquisition of Complex Sentences,” in Fletcher, P. and Garman, M. (Eds.), *Language Acquisition,* pp. 285-306. Cambridge: Cambridge University Press.

Brebion, G., Amador, X., Smith, M. J., and Gorman, J. M. (1997) “Mechanisms Underlying Memory Impairment in Schizophrenia,” *Psychological Medicine 27(2)*: 383-393.

Brown, R. (1973) *A First Language*, Cambridge, MA: Harvard University Press.

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

Burstein, J., Kukich, K., Wolff, S., and Lu, C. (1998) “Automated Scoring Using a Hybrid Feature Identification Technique,” *Proceedings of ACL*, pp. 206-210.

Burstein, J., Marcu, D., Andreyev, S., and Chodorow, M. (2001) “Towards Automatic Classification of Discourse Elements in Essays,” *ACL 2001*, pp. 90-97.

Callan, J.and Mitamura, T. (2002) "Knowledge-based Extraction of Named Entities," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02)*, pp. 532-537. McLean, VA: ACM.

Caplan, R., Guthrie, D., Shields, W. D., and Mori, L. (1992) “Formal Thought Disorder in Pediatric Complex Partial Seizure Disorder,” *Journal of Child Psychology and Psychiatry 33*: 1399-1412.

Chaika, Elaine (1974) “A Linguist Looks at ‘Schizophrenic’ Language,” *Brain and Language 1*: 257–276.

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

Chapman L. J., Chapman J. P., Raulin M. L. (1978) “Body-image Aberration in Schizophrenia,” *Journal of Abnormal Psychology 87*: 399–407

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Clemmer, Edward J. (1980) “Psycholinguistic Aspects of Pauses and Temporal Patterns in Schizophrenic Speech,” *Journal of Psycholinguistic Research 9*: 161–185.

Cohen, Robin (1987) “Analyzing the Structure of Argumentative Discourse,” *Computational Linguistics 13(1-2)*: 11-24.

Corcoran, Rhiannon, and Frith, Christopher D. (1996) “Conversational Conduct and the Symptoms of Schizophrenia,” *Cognitive Neuropsychiatry 1*: 305–318.

Covington, M. A., Riedel, W.J., Brown C., He, C., Morris, E., Weinstein, S., Semple, J., and Brown, J. (2006a) “Does Ketamine Mimic Aspects of Schizophrenic Speech?” submitted to *Journal of Psychopharmacology*.

Covington, M. A., He, C., and Brown, C. (2006b) "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.

Crow, T. J. (1997a) “Is Schizophrenia the Price that *Homo sapiens* Pay for Language,” *Schizophrenia Research 28*: 127-141.

Crow, T. J. (1997b) “Schizophrenia as Failure of Hemispheric Dominance for Language,” *Trends in Neurosciences 20*: 339-343.

Crow, T. J. (1998) “Nuclear Schizophrenic Symptoms as a Window on the Relationship between Thought and Speech,” *British Journal of Psychiatry 173*: 303-309.

Crow, T. J. (2000) “Schizophrenia as the Price that *Homo sapiens* Pays for Language: A Resolution of the Central Paradox in the Origin of the Species,” *Brain Research Reviews 31*: 118-129.

Curran, J. R. and Clark, S. (2003) “Language Independent NER Using a Maximum Entropy Tagger,” in *Proceedings of the Seventh CoNLL Conference*, HLT-NAACL, pp. 164-167.

Cutting, John (1985) *The Psychology of Schizophrenia*. Edinburgh: Churchill Livingstone.

De Decker, Brigitte, and Van de Craen, Pete (1987) “Towards an Interpersonal Theory of Schizophrenia,” in *Neurotic and Psychotic Language Behaviour*, Ruth Wodack and Pete Van de Craen (eds.), pp. 249–265. Clevedon, England: Multilingual Matters.

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

Docherty, N., Schnur, M., and Harvey, P. D. (1988) “Reference Performance and Positive and Negative Thought Disorder: A Follow-up Study of Manics and Schizophrenics,” *Journal of Abnormal Psychology 97(4)*: 437-442.

Docherty, N. M. and Gottesman, I. (2000) “A Twin Study of Communication Disturbances in Schizophrenia,” *Journal of Nervous and Mental Disorders 188*: 395-401.

Dombeck, M. (2006) “Schizophrenia,” http://mentalhelp.net/poc/center_index.php?id=7&cn=7.

Druks, J. (2002) “Verbs and Nouns – A Review of the Literature,” *Journal of Neurolinguistics 15*: 289-315.

Dutoit, Thierry (1997) *An Introduction to Text-to-Speech Synthesis (Text, Speech and Language Technology),* Springer.

Ebel R.L. (1951) “Estimation of the reliability of ratings,” *Psychometrika 16*: 407-424.

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

Elvevaag, B., Foltz, P. W., Weinberger, D. R., and Goldberg, T. E. (2004) “Quantifying Incoherence in Speech,” presented at SANE-POWIC Conference, Oxford, UK.

Faber, R., and Reichstein, M. B. (1981) “Language Dysfunction in Schizophrenia,” *British Journal of Psychiatry 139*: 519-522.

Ferrer i Cancho, R. (2005) “Zipf's Law from a Communicative Phase Transition,” *The European Physical Journal B 47(3)*: 449 – 457.

Ferrer i Cancho, R. (2006) “When Language Breaks into Pieces: A Conflict between Communication through Isolated Signals and Language,” *Biosystems 84(3)*: 242-53.

Flaum, M., Arndt, S., and Andreasen, N. C. (1991). “The Reliability of ‘Bizarre’ Delusions,” *Comprehensive Psychiatry 32*: 59–65.

Fraser, W. I., King, K. M., Thomas, P. and Kendell, R. E. (1986) “The Diagnosis of Schizophrenia by Language Analysis,” *Bristish Journal of Psychiatry 148*: 275-278.

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2005) “The Effects of Very Early Alzheimer’s Disease on the Characteristics of Writing by a Renowned Author,” *Brain 128*: 250-60.

Garside, R. (1987). “The CLAWS Word-tagging System,” in R. Garside, G. Leech and G. Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Gold, J. M., Randolph, C., Carpenter, C. J., Goldberg, T. E., and Weinberger, D. R. (1992) “Forms of Memory Failure in Schizophrenia,” *Journal of Abnormal Psychology 101(3)*: 487-494.

Gurd, J. M., Elvevaag, B., and Cortina-Borja, M. (1997) “Semantic Category Word Search Impairment in Schizophrenia,” *Cognitive Neuropsychiatry 2(4)*: 291-302.

Haarmann, H. J., and Kolk, H. H. (1994) “On-line Sensitivity to Subject-Verb Agreement Violations in Broca's Aphasics: The Role of Syntactic Complexity and Time,” *Brain and Language 46(4)*: 493-516.

Halliday, M. A. K. and Hasan, R. (1976) *Cohesion in English*, London: Longman.

Harabagiu, S., and Moldovan, D. (2003) “Question Answering,” in Mitkov (2003), pp. 560-582.

Harel, Brian T., Michael S. Cannizzaro, Henri Cohen, Nicole Reilly and Peter J. Snyder (2004)

"Acoustic Characteristics of Parkinsonian Speech: A Potential Biomarker of Early Disease Progression and Treatment," *Journal of Neurolinguistics 17*: 439-453.

Harris, J. R., Moss, H. E., Randall, B., and Tyler, L. K. (2005) "Noun and Verb Homophones: Important Predictors of Picture Naming Latency and Implications for Aphasia," *Brain and Language 95(6)*: 67.

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

Hoffman, R. E., Hogben, G. L., Smith, H., and Calhoun, W. F. (1985) "Message Disruptions During Syntactic Processing in Schizophrenia," *Journal of Communication Disorders 18*: 183–202.

Hoffman, R. E., and Sledge, W. (1988) "An analysis of Grammatical Deviance Occurring in Spontaneous Schizophrenic Speech," *Journal of Neurolinguistics 3*: 89-101.

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

Hoover, David (2005) "Word Frequency, Statistical Stylistics, and Authorship Attribution," AHRC ICT Methods Network Expert Seminar on Linguistics, Lancaster, UK.

Hovy, Eduard (2003) "Text Summarization," in MitKov (2003), pp. 583-598.

Hudson, C. J., Young, L. T., Li, P. P., and Warsh, J. J. (1993) "CNS Signal Transduction in the Pathophysiology and Pharmacotherapy of Affective Disorders and Schizophrenia," *Synapse 13(3)*: 278 – 293.

Ide, N., Suderman, K. (2004). "The American National Corpus First Release," *Proceedings of the Fourth Language Resources and Evaluation Conference* (LREC), Lisbon, 1681-1684.

Irigaray, L. (1985). Parler n'est jamais neutre, Minuit, Paris.   Cited in Pennisi et al. 2004.

Ittycheriah, A., Franz. M., Zhu, W. J., and Ratnapakhi, A. (2000) "IBM's Statistical Question-answering System," *Proceedings of the Text Retrieval conference (TREC-9)*, pp. 229-237.

Johnson, D. E., and Shean G. D. (1993) "Word Associations and Schizophrenic Symptoms," *Journal of Psychiatric Research 27(1)*: 69-77.

Jung, C. G. (1981) "Reaction time Ratio in the Association Experiment," in *The Collected Works of C. G. Jung*, Read, H., Forman, M., and Alder, G. (eds.), pp. 227-265, Princeton: Princeton University Press.

Jurafsky, D., and Martin, J. H. (2000) *Speech and Language Processing*, Upper Saddle River, New Jersey: Prentice Hall.

Kemper, S., LaBarge, E., Ferraro, F. R., Cheung, H., Cheung, H., and Storandt, M. (1993) "On the Preservation of Syntax in Alzheimer's Disease: Evidence from Written Sentences," *Archives of Neurology 50(1)*: 81-86.

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K., and Mitzner, T. L. (2001). "Language Decline across the Life Span: Findings from the Nun Study," *Psychology and Aging 16(2)*: 227-239.

Kempler, D., and Curtiss, S. (1987) "Syntactic Preservation in Alzheimer's Disease," *Journal of Speech and Hearing Research 30*: 343-350.

Kilgarriff, A. (1997) “Putting Frequencies in the Dictionary,” *International Journal of Lexicography 10 (2)*: 135--155.

Kim, Mikyong and Cynthia K. Thompson (2004) “Verb Deficits in Alzheimer’s Disease and Agrammatism: Implications for Lexical Organization,” *Brain and Language 88(1):* 1-20.

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.

Klein, D., and Manning, C. D. (2002) “Fast Exact Inference with a Factored Model for Natural Language Parsing.” In *Advances in Neural Information Processing Systems 15* (NIPS 2002).

Klosterkötter, J., Hellmich, M., Steinmeyer, E. M., and Schultze-Lutter, F. (2001) “Diagnosing Schizophrenia in the Initial Prodromal Phase,” *Archives of General Psychiatry 58:* 158-164.

Kuperberg, G., and Caplan, D. (2003) “Language Dysfunction in Schizophrenia,” in Schiffer, R. B., Rao, S. M., and Fogel, B. S. (Eds.), *Neuropsychiatry* (2 ed.), pp. 444-466. Philadelphia: Lippincott Williams and Wilkins.

Landauer, T. K., and 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.

Lee, L. (1974) *Developmental Sentence Analysis*, Evanston, IL: Northwestern University Press.

Levinson, S. C. (1983) *Pragmatics*, Cambridge, UK: Cambridge University Press.

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.

Lysaker, P. H., Wickett, A. M., Wilke, N., and Lysaker, J. (2003) “Narrative Incoherence in Schizophrenia: The Absent Agent-protagonist and the Collapse of Internal Dialogue,” *American Journal of Psychotherapy 57(2)*: 153-166.

Malvern, D. D., and Richards, B. J. (1997) “A New Measure of Lexical Diversity,” in Ryan, A. and Wray, A. (Eds.), *Evolving Models of Language. Papers from the Annual Meeting of the British Association of Applied Linguists held at the University of Wales, Swansea, September 1996*, pp. 58-71, Clevedon: Multilingual Matters.

Mani, I. (2001) *Automatic Summarization*, Amsterdam: John Benjamins.

Manning, C. D. and Schutze, H. (1999) *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Manschreck, T. C., Maher, B. A., and Rucklos, M. E. (1980) “Cloze Procedure and Written Language in Schizophrenia,” *Language and Speech 23(4)*: 323-328.

Manschreck, T. C., Maher B. A., and 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.

Manschreck, T. C., Ames D., Maher, B. A., and Schneyer, M. L. (1987) “Impoverished Written Responses and Negative Features of Schizophrenia,” *Perceptual and Motor Skills 64 (2)*: 1163-1169.

McClain, L. (1983) “Encoding and Retrieval in Schizophrenics’ Free Recall,” *The Journal of Nervous and Mental Disease 171(8)*: 471-479.

McKenna, Peter, and Oh, Tomasina (2005) *Schizophrenic Speech*, Cambridge, UK: Cambridge University Press.

Mefferd, R. B. Jr. (1979) “Word Association: Grammatical, Semantic, and Affective Dimensions of Associates to Homonyms,” *Psychological Reports 45(2)*: 359-374.

Mesure, G., Passerieux, C., Besche, C., Widlocher, D., and Hardy-Bayle, M. C. (1998) “Impairment of Semantic Categorization Processes among Thought-disordered Schizophrenic Patients,” *Canadian Journal of Psychiatry 43(3)*: 271-278.

Miller, J. F. (1981). *Assessing Language Production in Children*. Baltimore: University Park Press.

Miltsakaki, E., and Kukich, K. (2004) “Evaluation of Text Coherence for Electronic Essay Scoring Systems,” *Natural Language Engineering 10(1)*: 25-55.

Mitkov, Ruslan (2003) *The Oxford Handbook of Computational Linguistics*, New York: Oxford University Press.

Moran, L. J., Mefferd, R. B., Kimble, J. P. (1964) “Idiodynamic Sets in Word Association,” *Psychological Monographs: General and Applied 78*: 1-22.

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

Morice, Rodney, and Ingram, J. C. L. (1983) “Language Complexity and Age of Onset of Schizophrenia,” *Psychiatry Research 9(3)*: 233-42.

Morice, Rodney, and McNicol, Don (1985) “The Comprehension and Production of Complex Syntax in Schizophrenia,” *Cortex 21*: 567-580.

Morice, Rodney, and McNicol, Don (1986) “Language Changes in Schizophrenia: A Limited Replication,” *Schizophrenia Bulletin 12*: 239-251.

Muñoz, R. and Palomar, M. (2001) “Semantic-driven Algorithm for Definite Description Resolution,” *Proceedings of Recent Advances in Natural Language Processing, RANLP2001*, pp. 180-186.

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

Navigli, R. and Velardi, P. (2005) “Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*PAMI*) *27(7)*: 1063-1074.

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

Ney, H. (1991) “Dynamic Programming Parsing for Context-free Grammars in Continuous Speech Recognition.” *IEEE Transactions on Signal Processing 39 (2)*: 335-340.

Ojemann, George A. (1991) “Cortical Organization of Language,” *The Journal of Neuroscience 11(8)*: 2281-2287.

Oller, D. K., Eilers, R. E., Basinger, D., Steffens, M. L., and Urbano, R. (1995) “Extreme Poverty and the Development of Precusors to the Speech Capacity,” *First Language 15*: 167-187.

Overall, John E. and Gorham, Donald R. (1962) “The Brief Psychiatric Rating Scale,” *Psychological Reports 10*: 799-812.

Paulsen, J. S., Romero, R., Chan, A., Davis, A. V., Heaton, R. K., and Jeste, D. V. (1996) “Impairment of the Semantic Network in Schizophrenia,” *Psychiatry Research 63(2-3)*: 109-121.

Pennisi, A. (1998) “Psicopatologia del linguaggio,” Storia, analisi, filosofie della mente, Carocci, Roma. Cited in Pennisi et al. 2004.

Pennisi, A., Plebe, A., and Falzone, A. (2004) “What Has *Homo Sapiens* Really Bought at the Cost of Schizophrenia? Evolutionism Seen through a Philosophy of Language,” Evolution of Language: Fifth International Conference, Leipzig, Germany.

Rochester, S. R. and Martin, J. R. (1979) *Crazy Talk: A Study of the Discourse of Schizophrenic Speakers*. New York: Plenum.

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.

Sanders L.M., Adams J., Tager-Flusberg H., Shenton M.E., and Coleman M. A. (1995) “A Comparison of Clinical and Linguistic Indices of Deviance in the Verbal Discourse of Schizophrenics,” *Applied Psycholinguistics 16*: 325–338.

Scarborough, H. S. (1990) “Index of Productive Syntax,” *Applied Psycholinguistics 11*: 1-22.

Scott, M. (2004) *WordSmith Tools*, version 4, Oxford: Oxford University Press.

Shapiro, K. and Caramazza, A. (2003) “The Representation of Grammatical Categories in the Brain,” *Trends in Cognitive Sciences 7(5)*: 201-206.

Silveri, M. C., Salvigni, B. L., Cappa, A., Vedova, C. D., and Puopolo, M. (2003) "Impairment of Verb Processing in Frontal Variant-Frontotemporal Dementia: A Dysexecutive Symptom," *Dementia and Geriatric Cognitive Disorders 16(4):* 296-300.

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.

Snowdon, D. A. (2006) "The Nun Study FAQ," http://www.mc.uky.edu/nunnet/faq.htm.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001) "A Machine Learning Approach to Coreference Resolution of Noun Phrases," *Computational Linguistic 27(4)*: 521-544.

Spasic, I., Nenadic, G. and Ananiadou S. (2003) "Using Domain-Specific Verbs for Term Classification," In *Proceedings of NLP in Biomedicine, ACL 2003*, pp. 17-24. Sapporo, Japan.

Spearing, M. K. (1999) "Overview of Schizophrenia," http://www.schizophrenia.com/family/sz.overview.htm.

Stolcke, Andreas (1995) "An Efficient Probabilistic Context-free Parsing Algorithm that Computes Prefix Probabilities." *Computational Linguistics 21(2)*: 165-202.

Strobel, Gabrielle (2005) "SRF Interviews Robin Murray," http://www.schizophreniaforum.org/for/int/default.asp.

Thomas, Philip (1996) "Syntactic Complexity and Negative Symptoms in First Onset Schizophrenia," *Cognitive Neuropsychiatry 1(3):* 191 – 200.

Thomas, P., King, K, and Fraser, W. I. (1986) “Positive and Negative Symptoms of Schizophrenia and Linguistic Performance,” *Acta Psychiatrica Scandinavica 76(2)*: 144-51.

Thomas, P., King, K., and Fraser, W. I. (1990) “Linguistic Performance in Schizophrenia: A Comparison of Acute and Chronic Patients,” *British Journal of Psychiatry 156*: 204-210.

Thomas, P.; Leudar, I.; Newby, D.; and Johnston, M. (1993) “Syntactic Processing and Written Language Output in First Onset Psychosis,” *Journal of Communication Disorders 26*: 209–230.

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

Thompson, C. K., Shapiro, L. P., Kiran, S., and Sobecks, J. (2003) “The Role of Syntactic Complexity in Treatment of Sentence Deficits in Agrammatic Aphasia: the Complexity Account of Treatment Efficacy (CATE),” *Journal of Speech, Language, and Hearing Research 46(3)*: 591-607.

Tomoeda, C. K., Bayles, K. A., Boone, D. R., Kaszniak, A. W., and Slauson, T. J. (1990) “Speech Rate and Syntactic Complexity Effects on the Auditory Comprehension of Alzheimer Patients,” *Journal of Communication Disorders 23(2)*: 151-161.

Toutanova, K., Klein, D., Manning, C., and Singer Y. (2003) “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network,” In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

Turner, A. and Greene, E. (1977) *The Construction and Use of a Propositional Text Base*, University of Colorado Psychology Department, Boulder, CO.

Turner, Trevor (2003) *Schizophrenia: Your Questions Answered.* London, UK: Elsevier Science Ltd.

Van Riezen, Henk and Vrijmoed-de Vries, Maria C. (2000) *Rating Scales for Psychoneuropharmacology* (CD-ROM edition), Elsevier.

Weinstein, S, Woodward, I. 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.

World Health Organization (1992) *The ICD-10 Classification of Mental and Behavioural Disorders*. Geneva: World Health Organization.

APPENDIX A

VOCABULARY ANALYZER README

VocabAnalyzer version 0.5

Released on Feb 27, 2006

```
usage: edu.uga.caspr.posresearch.VocabAnalyzer
         [-saveToTextFile textOutputFilename]
         [-tagger taggerClassName]
         [-freqList freqFilename]
         [-firstHowMany numberOfCommonestWords]
         [-interval interval]
         [-TTR]
         -files textFilenames
```

No graphical user interface has been implemented for this version.

On a Windows machine where Java 1.5 is installed,

simply use the run.bat file to test-run it .

On a Unix/Linux machine where Java 1.5 is installed,

first chmod run.sh to executable,

and then type "run.sh" to test-run it.

Change the arguments in the .bat/.sh file as you wish.

1) The default output is standard output,

but you may choose to save the output to a text file.

2) There are two tagger plugins to choose from:

the default is

edu.uga.caspr.ideaDensity.implementations.OpenNLPTagger,

but you can choose to use

edu.uga.caspr.ideaDensity.implementations.StanfordTagger,

or to write a tagger class yourself which implements

edu.uga.caspr.dlevel.PennTreeStyleTagger.

3) The default frequency list is "models/BNC_lemma.txt",

but you can choose to supply any other word frequency list in plain text format with one lemma each line.

4) The default size of common vocabulary is 6000 words,

but you may choose to specify any other number that is smaller than or equal to the size of the frequency list.

5) The default group size for word rarity is 500 words, but you

may choose to use any number that is smaller than

or equal to the common vocabulary size.

6) By default, the TTR calculator is turned off.

You can turn it on by using the option -TTR.

7) The input filenames are mandatory.

Thank you for using VocabAnalyzer 0.5.

APPENDIX B

D-LEVEL RATER README

```
DLevelRater version 0.5

Released on Jan 17, 2006


usage: edu.uga.caspr.dlevel.TextRater

         [-verbose]

         [-saveToTextFile textOutputFilename]

         [-parser parserClassName]

         -files textFilenames


No graphical user interface has been implemented for this version.


On a Windows machine where Java 1.5 is installed,

simply use the run.bat file to test-run it.


On a Unix/Linux machine where Java 1.5 is installed,

first chmod run.sh to executable,
```

and then type "run.sh" to test-run it.

Change the arguments in the .bat/.sh file as you wish.

1) Choose the -verbose option to see the justifications for the ratings.

2) The default output is standard out,

but you can choose the –saveToTextFile option

to save the output to an output text file.

3) There are two parser plugins to choose from:

the default is

edu.uga.caspr.dlevel.implementations. OpenNLPParser,

but you can choose to use

edu.uga.caspr.dlevel.implementations.StanfordParser,

or to write a parser class yourself which implements

edu.uga.caspr.dlevel.PennTreeStyleParser.

4) The input filenames are mandatory.

Thank you for using DLevelRater 0.5.

APPENDIX C

IDEA DENSITY RATER README

```
IdeaDensityRater version 0.5
Released on Feb 27, 2006

usage: edu.uga.caspr.ideaDensity.IdeaDensityRater
         [-saveToTextFile textOutputFilename]
         [-tagger taggerClassName]
         -files textFilenames

No graphical user interface has been implemented for this version.

On a Windows machine where Java 1.5 is installed,
simply use the run.bat file to test-run it.

On a Unix/Linux machine where Java 1.5 is installed,
first chmod run.sh to executable,
and then type "run.sh" to test-run it.
```

Change the arguments in the .bat/.sh file as you wish.

1) The default output is standard out,

but you can choose the –saveToTextFile option

to save the output to an output text file.

2) There are two tagger plugins to choose from:

the default is

edu.uga.caspr.ideaDensity.implementations.StanfordTagger,

but you can choose to use

edu.uga.caspr.ideaDensity.implementations.OpenNLPTagger,

or to write a parser class yourself which implements

edu.uga.caspr.dlevel.PennTreeStyleTagger.

Thank you for using IdeaDensityRater 0.5.