

SOFTWARE TO MEASURE RAMBLING, COGNITIVE DIFFICULTY AND DEGREE
EXPRESSION IN SCHIZOPHRENIC SPEECH

by

LORINA NAÇI

(Under the Direction of Michael A. Covington)

ABSTRACT

This thesis develops new software to help distinguish schizophrenic speech from healthy speech. I used the Natural Language Analysis Tools (Covington 2002) and developed the Natural Language Feature Extraction program to collect linguistic features from forty four speech samples. Decision trees and neural networks use these linguistic features to classify the speech samples. The following linguistic features significantly distinguish schizophrenic speech from healthy speech: 1) mean sentence length, 2) degree expressions, and 3) difficulty words. Decision trees classify 85.7% of the speech samples correctly; neural networks classify 86% of the speech samples correctly. Schizophrenics' longer sentences may be a manifestation of lack of discourse planning, rambling, and running together of thoughts and sentences. The smaller number of degree expressions may be a manifestation of lack of theory of mind. The increased amount of difficulty words may be a sign of cognitive difficulty with interpreting information. Limitations and future research directions are discussed.

INDEX WORDS: Schizophrenia, Schizophasia, Thought disorder, Formal thought disorder, Computerized text analysis, Machine learning, Decision trees, Neural networks

SOFTWARE TO MEASURE RAMBLING, COGNITIVE DIFFICULTY AND DEGREE
EXPRESSION IN SCHIZOPHRENIC SPEECH

by

LORINA NAÇI

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

MASTER OF SCIENCE

ATHENS, GEORGIA

2004

© 2004

Lorina Naçi

All Rights Reserved

SOFTWARE TO MEASURE RAMBLING, COGNITIVE DIFFICULTY AND DEGREE
EXPRESSION IN SCHIZOPHRENIC SPEECH

by

LORINA NAÇI

Major Professor: Michael A. Covington

Committee: Zachary Estes
Donald Nute

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2004

DEDICATION

I dedicate this thesis to my Albanian parents Diana and Ylli Naçi, and to my American parents Fredericka and Elton Buesing, and Gail and Ben Fant who tremendously helped me, inspired me and encouraged me during my education.

ACKNOWLEDGEMENTS

I would like to especially thank my committee members. I thank my major professor Dr. Michael Covington for having involved me in The Speech Analysis project, where my thesis stems from, and for his unfaltering expectations of the highest academic standards. I am very thankful to Dr. Zachary Estes for his insightful help with the statistical niceties of the project. I also thank Dr. Donald Nute for his great help and support throughout the masters degree.

Further, I wish to acknowledge my friends Bill Hollingsworth, Shardul Vikram, Ramyaa, Fred Maier, and Chris Thomas for providing me with essential intellectual feedback on a day and night basis. I would also like to thank my colleagues Cati Brown and Congzhou He for contributing to an inspiring environment within the Speech Analysis research group.

Last but not least, I thank all my professors, especially Dr. Walter Potter, and friends at the AI Center for their valuable support.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER	
1 INTRODUCTION	1
1.1 Hypothesis	2
1.2 Background	3
1.3 About this study	5
1.4 The rest of this thesis	7
2 MATERIALS AND METHODS	8
2.1 Part one: linguistic analysis	8
2.2 Part two: machine learning	17
3 EXPERIMENTS	29
3.1 ANOVAs	29
3.2 Decision trees	29
3.3 Neural networks	32
4 RESULTS	36
4.1 Part one: results from linguistic analysis	36
4.2 Part two: results from machine learning	42

5	INTERPRETATION AND DISCUSSION	49
5.1	Interpretation of results	49
5.2	Discussion.....	51
	REFERENCES	55
	APPENDIX.....	61
	LISTS OF WORDS AND EXPRESSIONS	61

LIST OF TABLES

	Page
Table 2.1: The linguistic features used in this project	11
Table 2.2: Features measured by NLAT	13
Table 2.3: Features measured by ENLF	15
Table 2.4: Representation of <i>subject am</i> features for machine learning.....	18
Table 4.1: ANOVA results for seven ENLF features	37
Table 4.2: Tukey results for different comparison groups.....	37
Table 4.3: Decision trees' errors	43
Table 4.4: Neural networks' errors	45
Table 4.5: Cross-validation error	46

LIST OF FIGURES

	Page
Figure 2.1: Decision tree for healthy and non thought-disorder participants	20
Figure 2.2: See5 classifier design options.....	23
Figure 2.3: The graphical representation of a simple neural network	25
Figure 2.4: Neuroshell 2 neural network design window	27
Figure 2.5: Neuroshell 2 three layer back-propagation neural network	28
Figure 4.1: Mean sentence length in healthy and schizophrenic speech	38
Figure 4.2: Degree expressions in healthy and schizophrenic speech	39
Figure 4.3: Difficulty words in healthy and schizophrenic speech.....	40

CHAPTER 1

INTRODUCTION

The speech of schizophrenics, when disordered, is different from the speech of healthy people (Bleuler, 1950; Chaika, 1990; Andreasen, 1979a). As one of the distinguishing characteristics of schizophrenia, schizophrenic speech has been studied at length (Bleuler, 1950; Chaika, 1990; Andreasen, 1979a, 1979b; Crow, 2000; Docherty, DeRosa, & Andreasen, 1996). Many studies of schizophrenic language have been based on human judges' assessments of linguistic abnormalities. Computerized study of schizophrenic language can add accuracy and objectivity to existing findings, and also provide reliable tools for further investigation of the schizophrenic language. Machine learning techniques have been successfully employed to learn and extract linguistic information (Daelemans, Berck, & Gillis, 1997).

The goal of this thesis is two-pronged. First, it involves using existing software and developing new software to measure style and content features of schizophrenic speech. Second, it involves using these features to classify schizophrenic and healthy speech via computational classifiers, such as decision trees and neural networks. Hence, this study has two parts. In the first part, I computationally measure 14 linguistic features in schizophrenic and healthy speech samples. The stylistic features include the number of tokens, the number of types, the mean sentence length, the mean word length, repetitiousness etc. The content-based features include hedging words, degree words, degree expressions, cognitive difficulty words, cognitive difficulty expressions, and certainty words. I developed the Extraction of Natural Language Features

(ENLF) program to measure seven of the 14 linguistic features used¹; the other seven features I measured with the Natural Language Analysis Tools (NLAT²) program (Covington, 2002). In the second part, I used decision trees (Quinlan, 1986, 1993) and neural networks (McCulloch & Pitts, 1943; Rumelhart et al., 1986) to classify the participants into healthy and schizophrenic based on their speech samples.

1.1 Hypothesis

A single reading of a disturbed schizophrenic speech sample can reveal many linguistic peculiarities. Andreasen (1979a) reported the following characteristics of disturbed schizophrenic speech relevant to my study: 1) poverty of speech, derailment, and tangentiality (may cause skewed mean sentence length); 2) neologisms or new word formations (may cause skewed mean word length); and 3) perseveration or repetition of words and/or ideas (repetitiousness is one of the features I measure). Further, schizophrenics overtly express confusion and difficulty processing information, especially when performing cognitive tasks (Rosenberg & Tucker 1975, 1979). Some of the difficulty expressions are: *I can't see, it doesn't make sense, I dunno*, etc. Given schizophrenics' cognitive difficulty, I measured schizophrenics' level of certainty in their claims. I counted the occurrence of certainty words (i.e. *certainly, obviously, clearly, surely, evidently*, etc.). In addition, from manual word counts, I found that schizophrenics exhibit fewer hedging and degree terms³ than healthy participants. Hedges (i.e. *suggest, hint, imply, presume, conjecture, guess*, etc.) are used to make a statement more acceptable and to help strengthen the hearer's willingness to ratify the statement (Huebler, 1983). Degree words (i.e. *barely, hardly, scarcely, partially, partly*, etc.) modify a claim's strength.

¹ The features are explained in the materials and methods chapter.

² ENLF and NLAT are explained in detail in the materials and methods chapter.

³ I use the word 'term' to refer to either words and expressions, or both.

The hypothesis of this study has two parts. First, based on previous research and on my manual word counts on the speech samples, I expect: a) the mean sentence length, the mean word length, and the amount of repetition to be different in schizophrenic speech and healthy speech; and b) the number of hedging terms, the number of degree terms, of cognitive difficulty terms, and of certainty terms to be different in schizophrenic speech and healthy speech. I have not found any previous studies of hedging, expression of degree, and expression of certainty in schizophrenic speech. Second, I expect that the linguistic features measured by NLAT and ENLF will be sufficient in distinguishing schizophrenic from healthy speech using machine learning techniques (i.e. decision trees and neural networks).

1.2 Background

1.2.1 Language and schizophrenia

The schizophrenic linguistic disturbances have been related to thought disorder, which is the central symptom of schizophrenia (Carlson, 2003). The term *thought-disorder* has been used loosely in the schizophrenia literature to refer to different concepts (Andreasen, 1979a); narrowly, it has been used to denote a disorder of the *form* of thought (mistakenly equated with speech output), known as *formal thought disorder* (Andreasen, 1979a). Broadly, it has been used to refer to disordered *content* of thought, including hallucinations and delusions (Andreasen, 1979a). Based on the (rather underspecified) relation between speech and thought, a popular approach to measuring thought disorder has been by measuring the schizophrenic language disturbances. This thesis is concerned with schizophrenic speech and its relation to schizophrenic thought and cognition. Hence, schizophrenic verbal delusions and hallucinations are outside its scope.

1.2.2 Cognitive deficits and language disturbances in schizophrenia

Schizophrenics fail to control their thoughts and actions (Braver & Cohen, 1999). Braver, Barch, and Cohen (1999) suggest that the main three areas of schizophrenic cognitive control failures are: 1) “selective attention in the face of distraction;” 2) “inhibition of inappropriate responses;” and 3) “strategic regulation of behavior based on situational demands.” Cognitive control failures have been linked to impaired attention (Carter, Barch, & Cohen, 1997; Carter & Barch, 2000) and working memory (Docherty et al., 1996; Barch, Sheline, Csernansky, & Snyder, 2003). Schizophrenia patients show deficits on tasks involving the representation and maintenance of context (Barch, Carter et al., 1999a; Barch, Carter et al., 1998; Barch & Carter, 1998; Barch, Carter, Hatcher, & Cohen, 1999; Cohen, Barch, Carter, & Servan-Schrieber, 1999). A relationship has been shown between deficits in working memory and context processing and language disturbances in schizophrenia (Barch et al., 1999; Barch, Carter, Braver, & Cohen, 1997; Cohen, Barch, Carter, & Servan-Schrieber, 1999).

Frith and Corcoran (1996) and Corcoran, Mercer, and Frith (1995) suggest that schizophrenics lack theory of mind (i.e. are unable to represent the mental states of other people); the lack of a theory of mind may be behind schizophrenics’ communication failures. The apparent link between the representation and maintenance of context and the mastery of theory of mind remains yet to be investigated.

1.2.3 The study of language disorder

The schizophrenic language disorders provide a way to study schizophrenic cognition. Scales of schizophrenic thought disorder have been built based on analysis of schizophrenic speech (Andreasen, 1979a; Chen et al., 1996; Liddle et al., 2002; Docherty, DeRosa, &

Andreasen, 1996). The Scale for the Assessment of Thought, Language and Communication (TLC) (Andreasen, 1979a) is the first to define the schizophrenic speech disturbances.

One type of language analysis employed to study schizophrenic language, and of particular relevance to my thesis, is content analysis. Differently from other language studies, content analysis studies draw conclusions about a sample's style and content based on its word frequencies. The following are some findings from content analysis of schizophrenic language. Schizophrenics tend to have lower adjective-verb quotient, a measure of description in verbal expression (Mahler, 1972; Harrow & Quinlan, 1977; Whitehorn & Zipf, 1943; Fairbanks, 1944). The same authors find decreased type-token ratio, a measure of vocabulary size, in schizophrenic speech. Thematic content studies (Rosenberg & Tucker, 1975, 1979; Mete et al., 1993; Oxman, Rosenberg, Schnurr, Tucker, & Gala, 1988) find concern with one's own thought processes, cognitive difficulty, deviance, distress, and hostility to be predominant themes in schizophrenic language. Except for the theme of cognitive difficulty, the results concerning the other themes have not been replicated (Mete et al., 1993). In addition, content analysis has been used to classify schizophrenics based on their speech. Mete et al. (1993) report an accuracy of above 76% in distinguishing schizophrenics from other diagnostic groups and healthy participants.

1.3 About this study

The first part of this thesis can be categorized as a content analysis study. It measures linguistic features by counting the frequencies of pre-specific terms. As a quantitative text analysis study, it employs the assumption that words can reveal much about a speaker's mental state and even physical health (Pennebaker & Lee, 2002; Pennebaker, Mehl, & Niederhoffer, 2003).

Content analysis studies are easy to computerize. Hence, they possess a few practical advantages; they are less expensive (once automated) and more reliable, accurate, and objective than studies employing human judges (Rosenberg, Schnurr, & Oxman, 1990; Rosenberg & Tucker, 1979). Oxman, Rosenberg, Schnurr and Tucker (1988) used computerized content analysis to classify patients into their respective diagnostic groups with higher accuracy than human judges in four out of five cases. A more subtle advantage of computerized studies is being able to manipulate large data patterns and find new relationships in the data. For example, back-propagation neural networks are able to invent new features that are not given in the input, and use them to learn the target function (Mitchell, 1997). I will show that equipped with the intuitions of human experts, computerized quantitative text studies make promising research tools. My approach to content analysis differs in three ways from most existing content analysis studies.

1. ENLF measures the occurrences of pre-specified words and phrases that make up a small portion of the whole text. Previous content analysis studies have matched all the words of text into pre-defined psychological categories and themes (Rosenberg, Schnurr, & Oxman, 1990; Tucker & Rosenberg, 1975; Rosenberg & Tucker 1979; Oxman, Rosenberg, Schnurr & Tucker, 1988; Mete et al., 1993). For this purpose, the Gottschalk-Gleser scales (Gottschalk & Gleser, 1969), The Dartmouth Adaptation of the General Inquirer/Harvard III Psychosocial Dictionary (Stone et al., 1966), and the General Inquirer and Harvard III Psychosocial Dictionary (Kelly & Stone, 1975) have been used. One problem with this approach is that the themes of schizophrenic language vary from one culture to the next (Mete et al., 1993).

2. Most previous computerized content analysis studies (Rosenberg, Schnurr, & Oxman, 1990; Tucker & Rosenberg, 1975; Rosenberg & Tucker, 1979; Oxman, Rosenberg, Schnurr & Tucker, 1988; Mete et al., 1993) measure word frequencies by using adaptations of the General Inquirer Computer Content Analysis Program (Stone et al., 1966). Instead, this study employs NLAT and ENLF for statistical analysis of text.
3. Most previous content analysis studies (Rosenberg, Schnurr, & Oxman, 1990; Tucker & Rosenberg, 1975; Rosenberg & Tucker, 1979; Oxman, Rosenberg, Schnurr & Tucker, 1988; Mete et al., 1993), investigate either style or content but not both. This thesis looks at both the style and content of a text. NLAT and ENLF together measure 14 linguistic features, of which 10 are stylistic and four are content-based.

1.4 The rest of this thesis

Chapter two presents the materials and methods I used. Chapter three introduces the machine learning experiments, specifying the design choices for the decision trees and neural networks. Chapter four presents the results of the experiments. Chapter five presents plausible interpretations of the results, as well as the limitations, strengths, and the future directions of the research presented in the thesis.

CHAPTER 2

MATERIALS AND METHODS

2.1 Part one: linguistic analysis

Part one of this thesis consists of a corpus study where I measured both novel and previously studied linguistic features of schizophrenic speech. I developed the Extraction of Natural Language Features (ENLF) program, to measure seven pre-specified linguistic features. I ran ANOVA tests on the linguistic features measured by ENLF. In addition, I used Natural Language Analysis Tools (NLAT) (Covington, 2002) to measure seven additional linguistic features.

2.1.1 The speech sample data

I used the following three data sets containing healthy and schizophrenic (non thought-disorder and thought-disordered⁴) speech samples: the “GSK speech samples 2002,” the “SB speech samples,” and the “10 Healthy Controls.” All three data sets have received human subjects’ approval from The University of Georgia. Philip McGuire, at the Institute of Psychiatry in King’s College London collected the “GSK speech samples 2002” and the “SB speech samples.” The conditions under which these two datasets were collected have not been disclosed. The data suggest that the patients in the two sets suffer from different stages of schizophrenia and may be different in other aspects as well, such as age, medication level, etc.

⁴ The presence or absence of thought disorder is determined with a battery of cognitive tests. Both non thought-disordered and thought-disordered schizophrenics can exhibit language disturbances.

GlaxoSmithKline collected the “10 Healthy Controls” data set. The conditions of collecting the “10 Healthy Controls” have not been disclosed either. In total, I have 44 four speech samples: 14 healthy participants, 12 non thought-disordered schizophrenics, and 18 thought-disordered schizophrenics.

There are two main limitations pertaining to the data. First, the amount of data is very limited. Any result of this thesis will have to be tested on more data. Second, there is much linguistic variation between the schizophrenic speech samples. The variation can be due to the different manifestations of the schizophrenic language disturbances.

The speech samples in all three datasets have been gathered under similar conditions. Each participant describes the same four pictures from the Thematic Apperception Test (Murray, 1943). The Thematic Apperception Test is administered by having a participant verbally describe one or more of a series of ambiguous and emotionally complex pictures. The sound recordings of the descriptions are transcribed into speech samples. I eliminated the interviewer speech and considered valid for analysis a speech sample including all four picture descriptions, before and after interviewer’s promptings.

2.1.2 Introducing noisy data

Noisy data is data that is slightly incorrect due to erroneous measurement. If the noise is small, noisy data is not problematic for a classifier. In fact, slightly noisy data can be introduced to augment the amount of data in a data poor problem. The noise can be introduced by replicating an existing data instance and slightly changing the values of each attribute. The change can be done by adding or subtracting 1/100th of an existing value.

I introduced noisy data to tackle the following two problems. First, the amount of data is limited. The addition of noisy instances enlarges the data set. I added 10 data instances, making a total of 54 data instances. Second, in the original data set, there are more than twice as many schizophrenic speech samples as there are healthy speech samples. Given the limited data, a classifier (decision tree or neural network) may become biased in its classification towards the larger set. I added noisy data to make the participants' subsets equal. Thus, with the addition of noisy instances, there are a total of 18 healthy, 18 non thought-disordered and 18 thought-disordered speech samples.

2.1.3 Linguistic features

Table 2.1 lists the 16 linguistic features used in this project. NLAT measured features one through seven, whereas ENLF measured features eight through 16. Features three (type-token ratio) and five (stem-token ratio) are similar in that they measure vocabulary size. They differ in that the stem-token ratio is a stricter measure of a speaker's vocabulary size. The reason is the following: A type is a word form. The tokens of a type are the different occurrences of that word form in a text (Covington, 2002). Hence, the type-token ratio is a measure of the number of different types as compared to how many times those word forms are repeated in a text. A stem is the base form of a word. For example, the stem *play-* is part of *plays*, *playing*, etc. Since the same stem is the base form of many words, there are fewer stems than there are types. A speaker using more word types with different stems than word types with the same stem has a truly larger vocabulary than one who uses the same number of word types but a smaller number of stems. Also, features six (repetitiousness) and 16 (text repetition) are similar in that they measure the degree of repetitiousness for the entire text.

Table 2.1 The linguistic features used in this project

Linguistic Features	Motivation For Using Each Feature
1. Number of Tokens	Provides basic statistical information about speech
2. Number of Types	Provides basic statistical information about speech
3. Type-Token Ratio	A measure of speaker's vocabulary size
4. Number of Stems	Provides basic statistical information about speech
5. Stems-Token Ratio	A stricter measure of the speaker's vocabulary
6. Repetitiousness	Characteristic of schizophrenic speech
7. Repetitiousness-Words Ratio	Measures repetitiousness as independent from the total number of words. The amount of speech output of positive and negative symptom schizophrenics varies greatly.
8. Mean Word Length	Provides basic statistical information about speech
9. Mean Sentence Length	Schizophrenics ramble; sentences may run together
10. Hedging Words	Schizophrenics appear to* use fewer hedging terms
11. Degree Words	Schizophrenics appear to use fewer degree terms
12. Degree Expressions	Schizophrenics appear to use fewer degree terms
13. Difficulty Words	Previous research shows expressed cognitive difficulty in schizophrenic language
14. Difficulty Expressions	Previous research shows expressed cognitive difficulty in schizophrenic language
15. Certainty Words	Cognitive difficulty may reduce certainty terms
16. Text Repetition	Characteristic of schizophrenic speech

* This became evident from manual word counts performed on the speech samples.

Feature six, takes into account the number of words within which a word repeats itself. The basic principle is that the shorter the interval within which a word reoccurs, the higher the repetitiousness index. Hence, a word reoccurring after a short interval is more repetitious than a word reoccurring after a long interval (Covington, 2002). Feature 16 does not take into account the speech interval within which a word repeats itself. It just counts the number of times a word is repeated regardless of how recently that word was last uttered. Both features will be useful for decision trees and neural networks. When using machine learning techniques for data classification it is valuable to have as many features to learn from as possible.

2.1.4 Processing the data

2.1.4.1 Natural Language Analysis Tools (NLAT)

I measured the first seven linguistic features by using NLAT (Covington, 2002). Written in C Sharp, NLAT is a program for statistical analysis of natural language texts (Covington, 2002). Table 2.2 lists the linguistic features measured by NLAT. These features are not measured with reference to a text unit.

Table 2.2 Features measured by NLAT

Linguistic Features	Brief Explanation of Each Feature
1. Number of Tokens	Number of different words used
2. Number of Types	Number of different kinds of words
3. Type-Token Ratio	Gives size of speaker's vocabulary
4. Number of Stems	A stricter estimate of kinds of words
5. Stems-Token Ratio	A stricter estimate of the speaker's vocabulary
6. Repetitiousness	Amount of repetition in the speech sample
7. Repetitiousness-Words Ratio	Repetition per word

2.1.4.2 The Extraction of Natural Language Features (ENLF) Prolog program

ENLF is a Prolog program that measures stylistic and content-based linguistic features of natural language text (Table 2.3). My motivations for choosing each feature are summarized below.

- 1) **Mean Word Length:** Schizophrenics routinely make up new words (Andreasen 1979a). In addition, this feature provides basic statistical information about speech.
- 2) **Mean Sentence Length:** Verbosity is a well recognized feature of schizophrenic language (Andreasen, 1979a; Chaika, 1990). Schizophrenics ramble, derail from the topic, and are easily distracted by tangential stimuli (Andreasen, 1979a). If suffering from negative symptoms, schizophrenics speak very little (Andreasen, 1979a), or speak normally but utter little content. All of these findings point to a possibly skewed mean sentence length.
- 3) **Hedging Words:** I performed manual word counts on the data. These preliminary tests seem to indicate that schizophrenics, especially those suffering aggravated thought disorder, use fewer hedging terms than healthy participants.
- 4) **Degree Words/Expressions:** Preliminary tests seem to indicate that schizophrenics use fewer degree terms than healthy participants. ENLF looks for degree words and expressions (both part of degree terms) by different techniques. Hence, the results for each are treated individually when performing statistical tests.

Table 2.3 Features measured by ENLF

Linguistic Features	Brief Explanation of Each Feature
1. Mean Word Length	Mean length of words in speech sample
2. Mean Sentence Length	Mean length of sentences in speech sample
3. Hedging Words	Number of hedging words
4. Degree Words	Number of degree words
5. Degree Expressions	Number of degree expressions
6. Difficulty Words	Number of hardship words
7. Difficulty Expressions	Number of hardship expressions
8. Certainty Words	Number of certainty words
9. Text Repetition	An estimate of a text's repetitiousness

5) **Difficulty Words/Expressions:** Previous research has shown that schizophrenics overtly express difficulty, or concern with their own thought processes when performing cognitive tasks (Rosenberg & Tucker, 1975, 1979; Mete et al., 1993; Oxman, Rosenberg, Schnurr, Tucker, & Gala, 1988). ENLF looks for difficulty words and expressions (both part of difficulty terms) by different techniques. Difficulty words and difficulty expressions are treated individually when performing statistical tests.

6) **Certainty Words:** I measured certainty terms because the expression of certainty may be effected by cognitive difficulty and diminished confidence in one's analytic abilities.

7) **Text Repetition:** Schizophrenics abnormally persevere on one topic, idea or word, once they get fixated on it (Andreasen, 1979a). Schizophrenic speech may have a high repetitiousness index.

ENLF performs four main tasks for each speech sample.

1. It converts the text into a list of tokens, or a list of all the words used in the speech samples.

2. It removes suffixes creating a list of stems.
3. It looks for specific expressions. ENLF picks out expressions pre-specified in the expression lists (Appendix) such as *to some extent*, *more or less*, *kind of*, etc.
4. It measures seven linguistic features (Table 2.3). The total number of words varies widely from one speech sample to the next. To neutralize the effect of text length in each feature's value, ENLF measures features three to seven with reference to a text unit. One text unit is arbitrarily chosen equal to 100 words.
 - A. Features one and two are measured simply by finding the averages of the words and sentences in a text.
 - B. Features three, four, five, and six are measured by adding together the number of occurrences of specific (for each feature) words or expressions in the speech sample. ENLF uses different strategies to measure the occurrences of words and expressions. A word's occurrences are measured by counting and adding together each time the word occurs in the token list; an expression's occurrences are measured by counting and adding together each time the expression occurs in the expression list derived from parsing the text. The words and expressions for each feature are derived from literature on hedging, degree, and modality expression in the English language (Huebler, 1983; Palmer, 2001; Channell, 1994). I used the Merriam-Webster Thesaurus (Merriam-Webster Inc., 1989) to expand the lists provided in the literature. The lists contain the most used words or expressions for that feature; they are not exhaustive and can be expanded to include less frequent words or expressions.
 - C. Feature seven is measured by adding together the amount of repetition for each word in the text.

2.2 Part two: machine learning

Part two of this thesis consists of a classification study. I use machine learning techniques, namely decision trees and neural networks, to classify schizophrenic and healthy speech based on linguistic features.

Both decision trees and neural networks are widely used automated classifiers. Decision trees and neural networks have many aspects in common. Each of them has slightly different strengths. The following aspects are common between decision trees and neural networks.

- A. In both techniques, the data instances are represented as a set of attribute-value pairs matched with a classification, i.e. *healthy* or *schizophrenic* (Table 2.4).
- B. Both decision trees and neural networks are powerful computational learners. They automatically learn the target function, or the function which maps the inputs to the output(s) in a data set (Mitchell, 1997).
- C. Both decision trees and the back-propagation neural networks learn by seeing the inputs and the outputs in the training data. They learn by supervision.
- D. The performance of both techniques is tested by presenting the inputs of new data instances and predicting the output(s) or the participant's classification. The accuracy is measured by comparing the predicted classification (outputted by the classifier) with the actual classification of the data.

Table 2.4 Representation of *subject am* features for machine learning

ATTRIBUTES	1. Number of Tokens	617
	2. Number of Types	328
	3. Type-Token Ratio	0.53
	4. Number of Stems	310
	5. Stems-Token Ratio	0.5
	6. Repetitiousness	85.55
	7. Repetitiousness-Words Ratio	0.14
	8. Mean Word Length	4.54
	9. Mean Sentence Length	14.61
	10. Hedging Words	0.79
	11. Degree Words	0.32
	12. Degree Expressions	0.16
	13. Difficulty Words	1.11
	14. Difficulty Expressions	0.32
	15. Certainty Terms	0
	16. Text Repetition	220.52
TARGET ATTRIBUTE	Classification	Thought-Disordered

One major difference between neural networks and decision trees is that a decision tree's output reflects the reasoning behind its classification process, which a human observer can decipher. The output of a decision tree is simply a collection of steps. Each step (a node in the tree) corresponds to a decision point, where the data instances are divided into different groups based on the value of the attribute at that node. On the other hand, neural networks output only a number that can be interpreted as a classification. There is no information about how the neural network arrived at that classification. A human observer cannot tell how the strength of the network's connections or how the hidden nodes' activation level may represent the networks' classification decisions.

Despite the poor readability of their process, neural networks do indicate the features they find most important in the classification process. Neuroshell 2 considers the features with the highest weight to have the most say in classification. Decision trees indicate the features they find most important as well. The features with highest information gain, or those that classify the largest part of the data (Mitchell, 1997), are the most important in a decision tree's classification process. Hence, by using both decision trees and neural networks, we can see which features are important for more than one technique. This proves especially useful if the most important features are the same for both decision trees and neural networks.

2.2.1 Decision trees

Decision trees classify instances by sorting them down a tree of attributes, starting with the root node and ending with the leaf nodes. In the graphical representation of a decision tree (Table 2.1), the root node is at the top of the chart and the leaf nodes at the bottom.

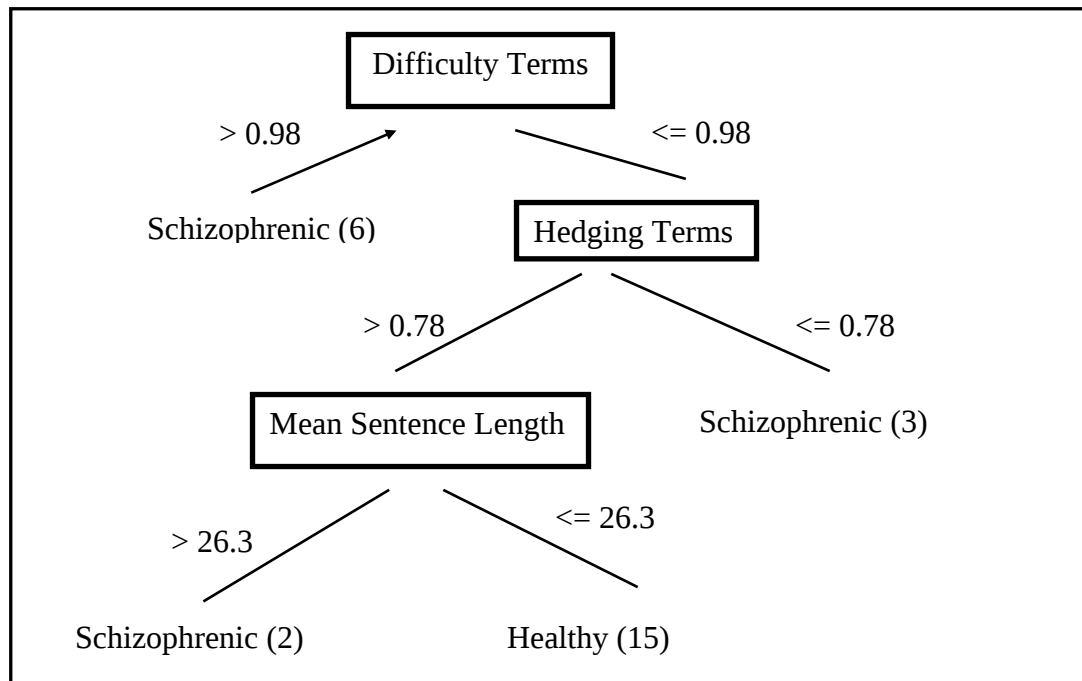


Figure 2.1 Decision tree for healthy and non thought-disordered participants

Each node in the tree is a test of some attribute, in this case, the linguistic features. The attribute with the highest information gain represents the root node. A leaf node represents a classification (Mitchell, 1997).

Once the tree has been built from the training data, it is used to classify new data instances. The classification of a new data instance proceeds as follows. Starting from the root node, the instance is queried regarding the value of the attribute at that node in the tree. From each node descend as many branches as data groups derived from this attribute. The data instance to be classified moves down the tree branch corresponding to its attribute value at that node. This process is repeated at each node until the leaf node is reached (Mitchell, 1997).

I used the demo version of See5⁵ software-package to construct 40 decision trees. See5 provides automated design options that are very useful in building the decision trees (Figure 2.2). Please see the experiments and results chapters for a listing of the decision trees using these options.

1. *Winnowing* is technique that cuts the attributes which do not contribute to the predictive power of a decision tree. If the superfluous attributes are winnowed, the predictive power of a tree increases (Mitchell, 1997).
2. *Rulesets* is a set of “if-then” rules that represent the classification process. The rulesets option produces classifiers that are highly human readable, but it does not affect the classification accuracy. I do not find the rulesets option particularly useful for this project. The reason is that the features used in this project are meaningful and a tree based on these features is already easy to read (Figure 2.1).

⁵ See5 is the commercial version of the C4.5 decision tree algorithm developed by Ross Quinlan. It is distributed by RuleQuest Research Pty Ltd. The See5 demo version is available for free from the internet: <http://www.rulequest.com/download.html>. It allows the classification of only 400 instances. Given this project’s limited data amount, the demo software is usable.

3. *Boosting* is an iterative technique for developing complementing decision trees based on the same data. The instances that are classified incorrectly in one iteration are taken forward in the next iteration, until all the instances are classified correctly. In boosting, many classifiers pull their forces together to classify the data (Witten & Frank, 1990).
4. *Cross-validation* is often used to derive a reliable error estimate when the dataset is small. During cross-validation, the data are partitioned into training and testing groups many times, each time using different parts of the data. For each partition of the data a different classifier is built; for example, a total of ten classifiers are built for the ten-fold cross-validation. The errors from all classifiers are averaged, deriving a relatively solid error estimate for that data set (Mitchell, 1997).
5. The *pruning* technique monitors the predictive power derived from developing each branch in a decision tree. Pruning safeguards against overtraining, or learning the training data too well, at the cost of learning the irrelevant information instead of focusing on the general trends that can be extrapolated beyond the training set. As the predictive power derived from new branches becomes small, that branch is pruned and the tree stops growing in that direction.

Classifier Construction Options

☐ Winnow attributes

☐ Rulesets

☐ Sort by utility ☐ bands

☐ Boost ☐ trials

☐ Subsets of values

☐ Use sample of ☐ %

☐ Lock sample

☐ Cross-validate ☐ folds

☐ Ignore costs file

Advanced options

☐ Fuzzy thresholds

☒ Global pruning

Pruning CF %

Minimum cases

Figure 2.2 See5 classifier design options

2.2.2 Neural networks

Artificial neural networks are very effective learning methods. They can approximate any real-valued, discrete-valued, and vector-valued target functions (Mitchell, 1997). A simple neural network has an input layer, one or more hidden layer(s), and an output layer, all of which can have a varying number of nodes (Figure 2.3).

Once the neural network has been created from the training data, it is used to classify data instances that were not used in training. The classification process proceeds as follows. The different attributes' values for a data instance are fed to the input nodes. The connections between the input and the hidden nodes create a weighted sum of these inputs. Based on the input's weighted sum, the hidden nodes' activation function determines the hidden nodes' output. Further, the connections between the hidden and the output nodes create a weighted sum of the hidden nodes' output. Finally, based on the weighted sum of the hidden nodes' output, the activation of the output nodes determines the network's output. The output of the neural network is a numerical value that represents the classification of that data instance. Based on a pre-specified threshold, it can be determined whether or not the network classified the instance correctly.

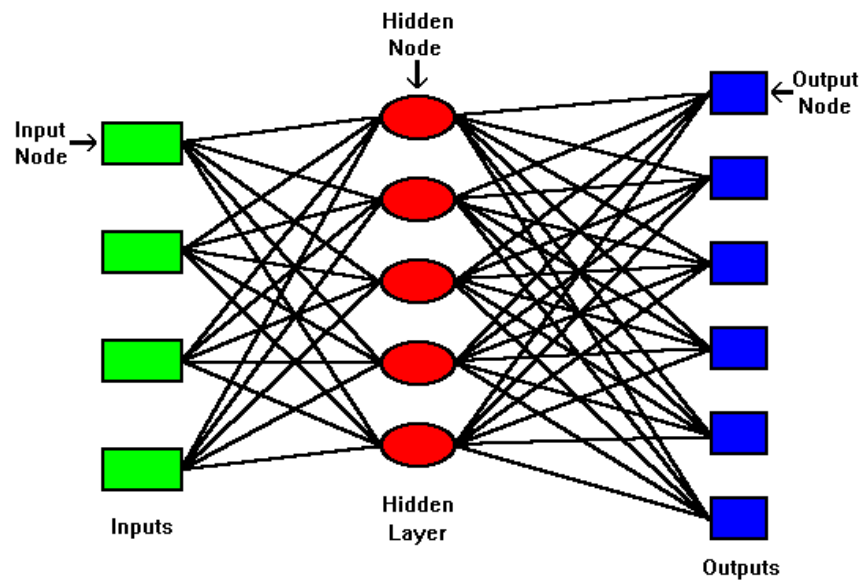


Figure 2.3 The graphical representation of a simple neural network

I used Neuroshell 2⁶ release 3.0 to build 19 neural networks. The process of building a neural network has four main stages (Figure 2.4).

- 1) **Defining inputs and outputs:** The inputs to the network are the values of the linguistic features for each data instance. The output is a number(s) corresponding to the classification of that data instance. Whether there are one or more outputs depends on the output design schema. The different design choices are explained in detail in the experiments chapter.
- 2) **Partitioning the data into subsets:** The data is partitioned by random extraction in three sets, the *training*, *testing*, and *validation* sets. The training set (50% of data) is used to train the network. The testing set (25% of data) is used to stop training. The validation set (25% of data) is used to test the network.
- 3) **Designing the network:** I used the three layer back-propagation architecture for all the neural networks (Figure 2.5). The design specifications for all the networks are the following.
 - A. Input Nodes: 14
 - B. Hidden Nodes: 5 (the number is arbitrarily chosen)
 - C. Output Nodes: 1, 2, or 3 depending on the network
 - D. Activation Function: Logistic
- 4) **Training the network:** I used the following specifications for training each network
 - A. Momentum⁷: 0.1 (suggested by Neuroshell 2)
 - B. Learning Rate⁸: 0.1 (suggested by Neuroshell 2)
 - C. Initial Weights: 0.3 (suggested by Neuroshell 2)
 - D. Criteria for stopping the training: 200,000 epochs subsequent to the best error

⁶ Neuroshell II is copyrighted by Ward Systems Group, Inc.

⁷ The momentum helps the network to keep in line with what it has previously learned (Smith, 1993).

⁸ The learning rate helps the network to explore all the space of weight possibilities (Smith, 1993).

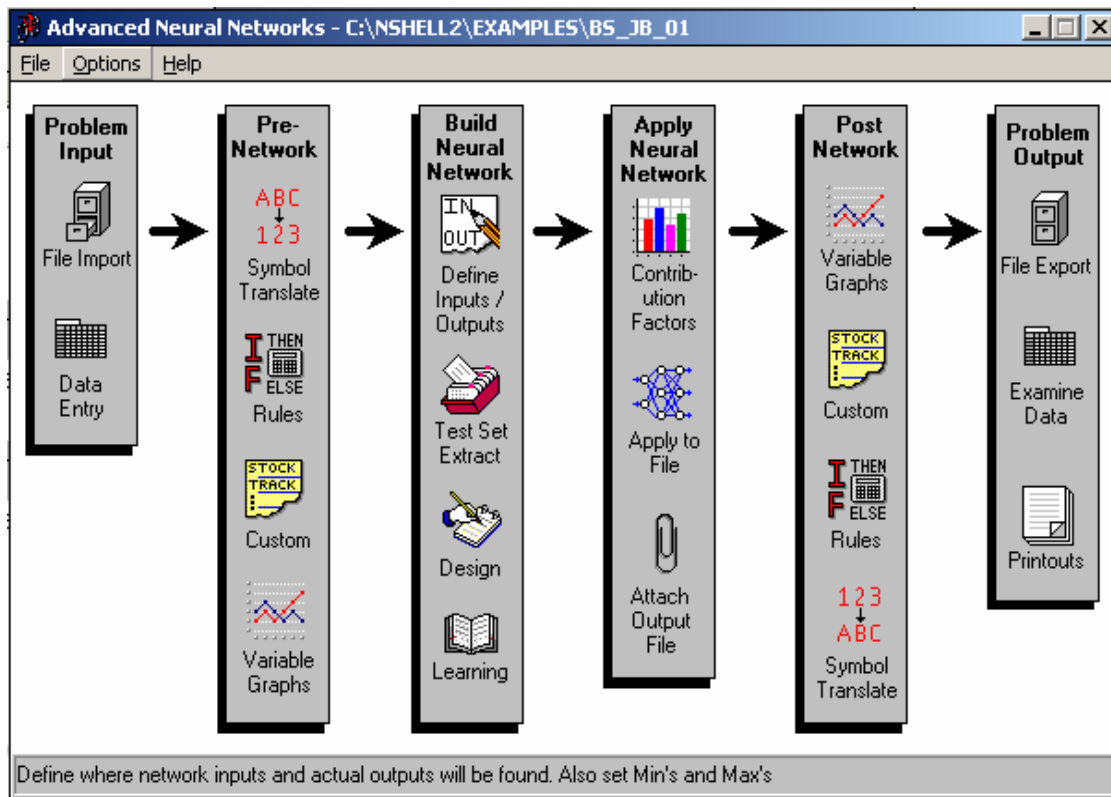


Figure 2.4 Neuroshell 2 neural network design window

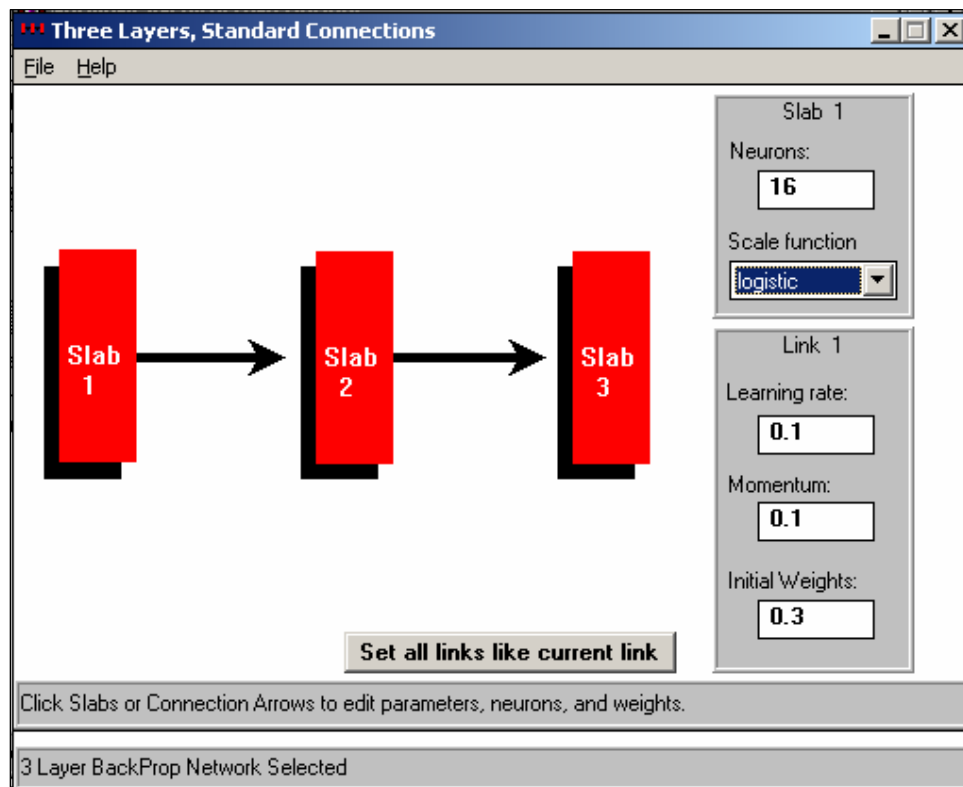


Figure 2.5 Neuroshell 2 three layer back-propagation neural network

CHAPTER 3

EXPERIMENTS

3.1 ANOVAs

I use SPSS (SPSS Inc., 2004) to run nine univariate ANOVA (analysis of variance) tests, one for each features measured by ENLF. An ANOVA looks for significant differences between a linguistic feature's mean value in schizophrenic speech and that feature's mean value in healthy speech (StatSoft Inc., 2004). The Tukey test, is a post hoc test that compares different participants' groups together. The Tukey test makes the following comparisons: 1) healthy vs. non thought-disordered; 2) healthy vs. thought-disordered; and, 3) non thought-disordered vs. thought-disordered. I run ANOVA and Tukey tests only for the nine features measured by ENLF, the software that I developed.

3.2 Decision trees

A decision tree automatically classifies a participant into healthy or schizophrenic based on the characteristics of his or her speech. In order to find trees with high performance, I built a total of 40 decision trees with different specifications. I calculated each tree's classification accuracy based on the number of its correct classifications.

3.2.1 Inputs and outputs

The input (for each participant) consists of 14 attribute-value pairs corresponding to the linguistic features. The values of the attributes were provided by NLAT and ENLF. The output consists of the value of the target attribute or the participant's classification. The target attribute is either *schizophrenic* with values: *yes* (for schizophrenic) and *no* (for healthy), or *state* with values: *td* (for thought-disordered), *ntd* (for non thought-disordered), and *h* (for healthy).

The decision tree handles the inputs and the outputs differently in the training and testing phase. In the *training phase* the decision tree sees both the inputs and the output for the training set speech samples. In the *testing phase*, the decision tree sees only the inputs. Subsequently, the tree predicts a classification of each participant. The output is correct if it matches the actual label (*yes/no* or *h/ td/ntd*) for that speech samples.

3.2.2 Design specifications for the decision trees

I built a set of five different decision trees for each of eight classification schemas. The eight classification schemas are the following.

1. A binary target-attribute tree distinguishing healthy speech from non thought-disordered speech based on noise-free data.
2. A binary target-attribute tree distinguishing healthy speech from non thought-disordered speech based on noisy data.
3. A binary target-attribute tree distinguishing healthy speech from thought-disordered speech based on noise-free data.
4. A binary target-attribute tree distinguishing healthy speech from thought-disordered speech based on noisy data.

5. A binary target-attribute tree distinguishing healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noise-free data.
6. A binary target-attribute tree distinguishing healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noisy data.
7. A three-valued target-attribute tree distinguishing healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noise-free data.
8. A three-valued target-attribute tree distinguishing healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noisy data.

The eight different design specifications are the following. (Please refer back materials and methods chapter for an explanation of the design options mentioned below.)

- I. A decision tree with global pruning.
- II. A decision tree with winnowing.
- III. A decision tree with 10-fold cross-validation.
- IV. A decision tree with 10-fold cross-validation and winnowing.
- V. A decision tree with boosting.

3.2.3 Classification and error measurement for decision trees

The See5 software automatically reports the classification error for a decision tree. The See5 default error measure gives an optimistic estimate of the decision tree's predictive power. The reason is that, in See5, the network's error is measured based on the training data, the same data used to build the decision tree. This constitutes a problem because the tree was developed to

fit the training data as best as it can. We are interested in how well this tree will classify instances not used in training. Thus, we need to test the tree on data it has not used in training.

The cross-validation technique provides a more realistic estimate of the network's predictive power over examples not used in training. The cross-validation classification error is much higher than the training data error. It is worth noting that the cross-validation error estimation may be an overly pessimistic one. Cross-validation partitions the data in small sub-groups. Given our initially small data set, and the large variation between the speech samples, it is possible that a) the network is learning less from a smaller training sample, and b) it is committing more errors on a smaller testing sample, which may not be representative of what the network has learned. The dilemma about what to consider the most accurate performance measure would not arise if the classifier had a large data set to work with.

3.3 Neural networks

Like a decision tree, a neural network automatically classifies a participant into healthy or schizophrenic based on the characteristics of his or her speech. I built a total of 19 neural networks. The reason for building fewer neural networks than decision trees is that there are fewer reliable optimization options for neural networks. Much less is known about the classification process in neural networks than in decision trees. The process of building an efficient neural network resembles the process of trial and error.

3.3.1 Input and outputs

The input consists of 14 numerical values (corresponding to the linguistic features). The output consists of a numerical value(s) corresponding to a participant's classification. The networks have different outputting designs including one, two, or three output nodes.

Like decision trees, neural networks handle the inputs and the outputs differently in the training and testing phase. In the *training phase* both inputs and outputs are provided. An output node is on or off depending on the participant's classification. In the *testing phase* the network sees only the inputs, and, based on them, it predicts the participant's classification. Depending on the outputting design, an output node can behave as follows. If the network has only one node, a correct classification for a schizophrenic is an output > 0.1 and < 0.5 ; a correct classification for a healthy participant is an output ≥ 0.5 and < 0.9 . If the network has more than one node, the active node determines a participant's classification, based on a pre-specified labeling of the nodes. The node with the highest activation is considered active and determines the network's classification. Consider a two-nodded network. If the node labeled for the healthy classification outputs a 0.75 and the node labeled for the schizophrenic classification a 0.43, the classification is *healthy*.

3.3.2 Design specifications for the neural nets

For comparison purposes, the neural networks partition the data into the same participants' subgroups as the decision trees. I built the following networks.

1. A two output nodes network to distinguish healthy speech from non thought-disordered speech based on noise-free data.

2. A two output nodes network to distinguish healthy speech from non thought-disordered speech based on noisy data.
3. A two output nodes network to distinguish healthy speech from thought-disordered speech based on noise-free data.
4. A two output nodes network to distinguish healthy speech from thought-disordered speech based on noisy data.
5. A two output nodes network to distinguish healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noise-free data.
6. A two output nodes network to distinguish healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noisy data.
7. A three output nodes network to distinguish healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noise-free data.
8. A three output nodes network to distinguish healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noisy data.
9. A one output node network to distinguish healthy speech from schizophrenic (including both non thought-disordered and thought-disordered) speech based on noise-free data.
10. 10 two output node networks to distinguish healthy speech from thought-disordered for cross-validation.

3.3.3 Classification and error measurement for neural networks

Neuroshell 2 tests the network's performance on data it has not used in training. This data is contained in the validation set. The error on the validation set is computed by comparing a network's classification with the actual classification and by calculating an error/miss every time

the two are different. All the errors are added together and divided by the total number of participants in the validation set. This yields the network's percentage error.

Testing the network's prediction on the validation set provides a relatively unbiased performance measure. However, the original data set is very small. The validation set is a mere 25% of that data set. Simply, the validation set is very small (containing 9-16 participants). Hence, the error on the validation set is not a completely robust measure. For example, a different error may be derived if a different set of participants (the data extraction from the main data pool is random) are chosen for the validation set.

I used cross-validation error as a more reliable error measure. Given that in Neuroshell 2 cross-validation is not automated, I performed manual cross-validation. Manual cross-validation is not only very time consuming, but, most importantly, introduces the problem of human error. I performed cross-validation only for one network to give a general idea of the amount of cross-validation error for the neural networks.

CHAPTER 4

RESULTS

4.1 Part one: results from linguistic analysis

Six of the ENLF's features do not significantly distinguish schizophrenic speech from healthy speech (Table 4.1). This may be in part due to measurement. For example, it is well-known that schizophrenic speech is highly repetitious (Andreasen 1979). However, the feature 'Text Repetition' does not reach a significant level of difference between schizophrenic and healthy speech. Measuring the repetitiousness of every word in the text may not be a good measure of repetition in schizophrenic speech. It is possible that healthy speech is overall repetitious enough to not be distinguishable (via this method) from schizophrenic speech. In hindsight, a measure of the repetition of specific words/ideas may be a more effective measure of repetition in schizophrenic speech.

The non-statistically significant results may offer valuable insight in their own right. For example, based on this study, schizophrenics do not use significantly more certainty words than healthy people. However, based on this study, schizophrenics use significantly more cognitive difficulty words than healthy people. These two results may suggest that, unlike healthy people, schizophrenics do not use caution in their claims even in the face of cognitive difficulty.

The features 'Mean Sentence Length,' 'Degree Expressions,' and 'Difficulty Words' reach a significant level of difference between schizophrenic and healthy speech.

Table 4.1 ANOVA results for the ENLF features

FEATURES	1. Mean Word Length	F(2,41) = .59, p = .56
	2. Mean Sentence Length	F (2,41) = 4.77, p = .014
	3. Hedging Words	F (2,41) = 2.58, p = .088
	4. Degree Words	F (2,41) = .37, p = .69
	5. Degree Expressions	F (2,41) = 5, p = .011
	6. Difficulty Words	F (2,41) = 4.02, p = .025
	7. Difficulty Expressions	F (2,41) = .74, p = .482
	8. Certainty Words	F (2,41) = .17, p = .842
	9. Text Repetition	F(2,41) = .17, p = .845

Table 4.2 Tukey results for different comparison groups

FEATURES	H. vs. N.T.D.⁹	H. vs. T.D.	N.T.D. vs. T.D.
1. Mean Word Length	p = .67	p = .99	p = .56
2. Mean Sentence Length	p = .03	p = .03	p = .96
3. Hedging Words	p = .89	p = .09	p = .27
4. Degree Words	p = .67	p = .8	p = .96
5. Degree Expressions	p = .04	p = .01	p = .98
6. Difficulty Words	p = .02	p = .12	p = .43
7. Difficulty Expressions	p = .79	p = .45	p = .88
8. Certainty Words	p = .87	p = .86	p = 1
9. Text Repetition	p = .99	p = .84	p = .92

⁹ H.: Healthy; N.T.D.: Non thought-disordered; T.D.: Thought-disordered.

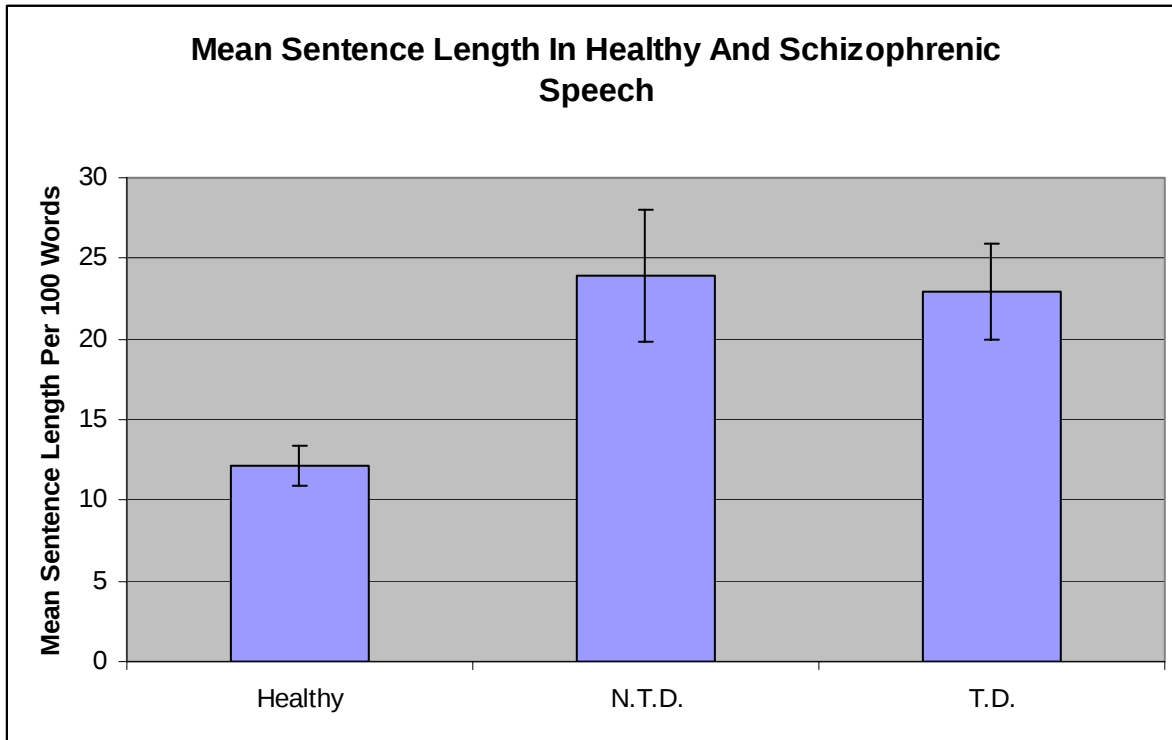


Figure 4.1 Mean sentence length in healthy and schizophrenic speech

1. Schizophrenics use **longer sentences** than healthy participants. $F(2, 41) = 4.77, p = .014$.

Figure 4.1 shows that healthy participants, as a group, use significantly shorter sentences than schizophrenic participants.

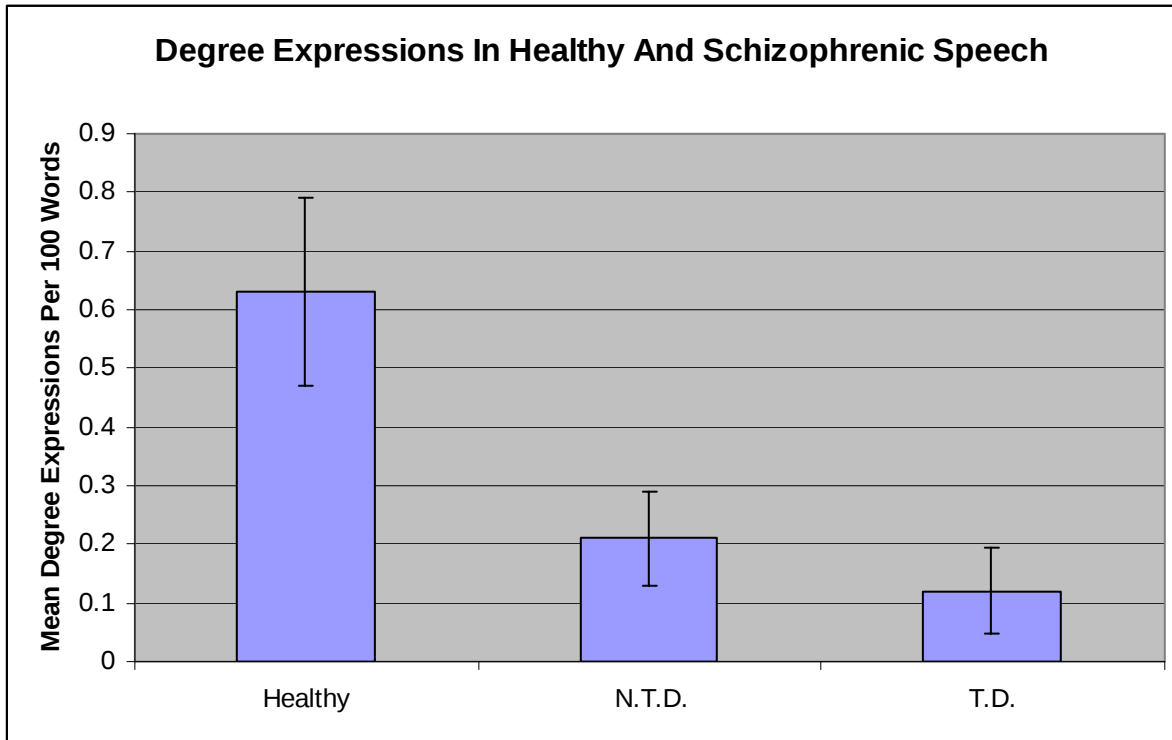


Figure 4.2 Degree expressions in healthy and schizophrenic speech

2. Schizophrenics use **fewer degree expressions** than healthy participants. $F(2, 41) = 5, p = .011$. Figure 4.2 shows that healthy participants, as a group, use significantly more degree expressions than schizophrenic participants.

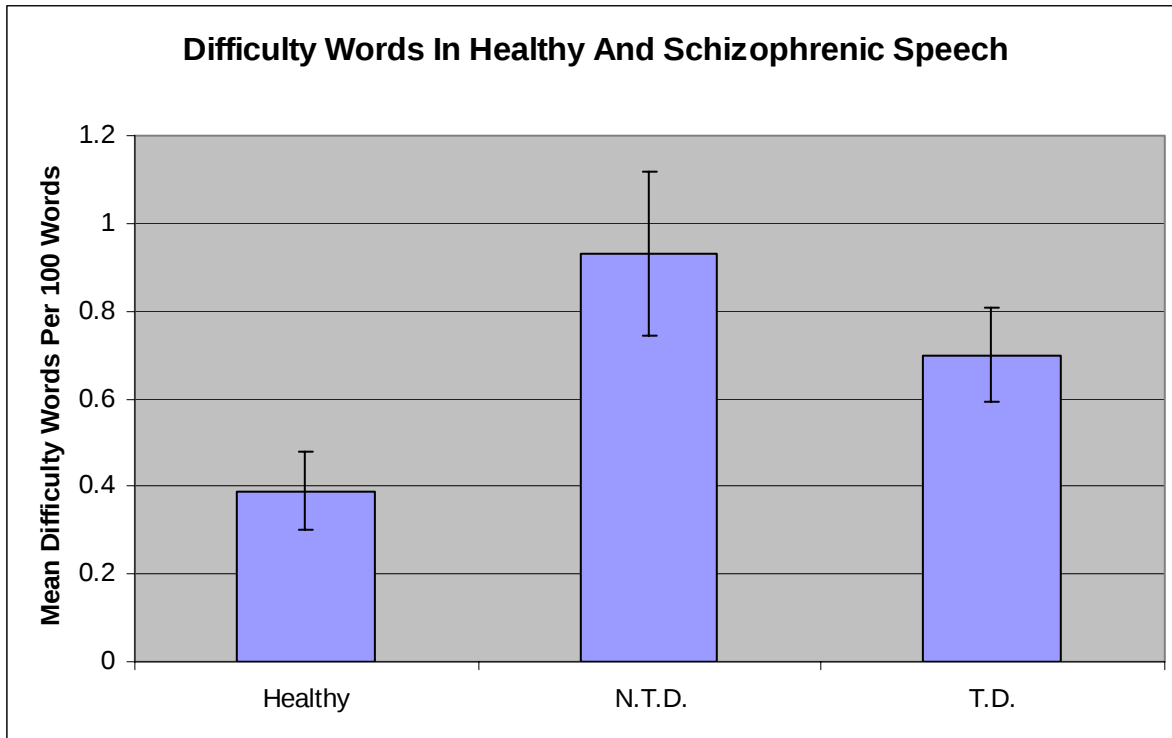


Figure 4.3 Difficulty words in healthy and schizophrenic speech

3. Schizophrenics use **more difficulty words** than healthy participants: $F(2, 41) = 4.02$, $p = .025$. Figure 4.3 shows that healthy participants, as a group, use significantly fewer difficulty words than schizophrenic participants.

The feature ‘Hedging Words’ comes close to reaching a significant level of difference in schizophrenic and healthy speech ($F(2,41)$; $p = .088$). Based only on the “SB speech samples” (12 speech samples), the feature ‘Hedging Words’ distinguishes schizophrenic speech from healthy speech at a significant level ($p = .001$). However, the feature ‘Hedging Words’ does not reach a significant level of difference ($p = .088$) on the whole data (44 speech samples). Linguistic variation may play a large role on the level of significance reached by ‘Hedging Words’ over the three data sets. The schizophrenic speech samples can vary greatly from one data set to the other. For example, the schizophrenic speech in the “SB speech samples” appears to reflect more aggravated thought disorder than the schizophrenic speech in the other two data sets. For example, the speech of Subject AM (thought-disordered), from the “SB speech samples” data set, sounds like the following:

Where does she come from, that Bridget Bardot, of Bridget Bardot existence. She looks in stain-like angst upon the heavy pyramids which flow.....the light in the background. They look, to an extent, barriered and outside the space-like, clown-like warrior of E-so-e-ski – born. Heaps upon the fettered, ploughed fields which come to the Eastern ottoman manor house of Fiefdom cake.

In contrast, the speech of Subject JJ (thought-disordered), from the “GSK speech samples 2002” data set, sounds like this:

They are farming, the farmer is ploughing his land while his wife stands by and his daughter may go to church or she reads a book or something, or the bible, may be she is religious or something, you know. The wife is watching her husband ploughing the field. I don’t know what else to say about that really.

The speech of Subject AM is patently more disturbed than the speech of Subject JJ.

4.2 Part Two: results from machine learning

4.2.1 Results from decision trees experiments

Table 4.3 summarizes the results of the decision tree experiments. The decision tree comparing healthy speech to non thought-disordered speech (with the winnowing feature selected) achieves 100% accuracy on the training data. This simply means that a decision tree can learn the training data perfectly. As expected, the cross-validation error is higher for most trees. A low cross-validation error indicates that the tree performs well on data it has not been trained on.

The tree with the lowest cross-validation error (16.7%) is a two-valued target attribute tree that compares healthy speech to all schizophrenic speech. This tree uses noisy data. I introduced noisy data in order to augment the data amount. The high performance on noisy data indicates that increasing the number of data patterns enhances the performance of a decision tree.

In addition, two-valued target-attribute trees classify the data more accurately than three-valued target attribute trees. This result may also be due to the amount of data used by each of these classifiers.

Table 4.3 Decision trees' errors

Decision Trees Experiments	Global pruning	Winnowing	Cross - Validation	Winnowing & Cross-Validation	Boost
1. Two-Valued Healthy vs. Non Thought-Disordered Noise-Free	3.8%	0%	41.7%	26.7%	0%
2. Two-Valued Healthy vs. Non Thought-Disordered Noisy Data	2.8%	13.9%	24.2%	23.3 %	0%
3. Two-Valued Healthy vs. Thought-Disordered Noise-Free	3.1%	9.4%	37.5%	22.5%	0%
4. Two-Valued Healthy vs. Thought-Disordered Noisy Data	2.8%	19.4%	26.7%	50%	0%
5. Two-Valued Healthy vs. All Schizophrenics Noise-Free	2.3%	15.9%	25%	27%	0%
6. Two-Valued Healthy vs. All Schizophrenics Noisy Data	1.9%	13%	14.3%	16.7%	0%
7. Three-Valued Healthy vs. All Schizophrenics Noise-Free	11.4%	25%	60%	42.5%	0%
8. Three-Valued Healthy vs. All Schizophrenics Noisy Data	13%	46.6%	44.3%	56.3%	0%

4.2.2 Important features for decision trees

For each tree, I list below the two attributes with the highest information gain. The topmost important attributes for all the decision trees are:

- 1) Difficulty Words
- 2) Hedging Words
- 3) Mean Sentence Length
- 4) Degree Expressions

4.2.3 Results from neural networks experiments

Table 4.4 summarizes the results of the neural networks experiments. Table 4.5 summarizes the cross-validation experiment's results, including the errors for each of the 10 networks involved. The best performing neural network (network 2 in Table 4.4) (0.00% error) compares healthy speech to non thought-disordered speech based on noisy data. Such good performance is achieved on data from the validation set or data not used in training. However, the validation set is small and may not produce a robust error measure. Judging from the cross-validation experiment, we can expect high cross-validation errors for the other networks as well. This result is similar to the cross-validation results for decision trees. Given that the cross-validation error is a more robust error measure, we can expect neural networks to have an error measure of at least 14%.

Table 4.4 Neural networks' errors

Neural Networks	Validation Set Error
1. Two Output Nodes Healthy vs. Non Thought-Disordered Noise-Free	22.2%
2. Two Output Nodes Healthy vs. Non Thought-Disordered Noisy Data	0%
3. Two Output Nodes Healthy vs. Thought-Disordered Noise-Free	40%
4. Two Output Nodes Healthy vs. Thought-Disordered Noisy Data	10%
5. Two Output Nodes Healthy vs. All Schizophrenics Noise-Free	7.1%
6. Two Output Nodes Healthy vs. All Schizophrenics Noisy Data	25%
7. Three Output Nodes Healthy vs. All Schizophrenics Noise-Free	64.3%
8. Three Output Nodes Healthy vs. All Schizophrenics Noisy Data	45%
9. One Output Nodes Healthy vs. All Schizophrenics Noise-Free	8%

Table 4.5 Cross-validation error

Two output nodes Healthy vs. Thought-Disordered Schizophrenics Noisy Data	Networks	Validation Set Error
	1	20%
	2	0%
	3	0%
	4	20%
	5	0%
	6	50%
	7	25%
	8	0%
	9	0%
	10	25%
	Total (averaged error)	14%

Another similar characteristic for decision trees and neural networks is the high variability of error between the classifiers involved in a cross-validation experiment (usually 10). From 10 classifiers (in one cross-validation run), some have a high error (i.e. 25%), whereas others have a 0.00% error. This variability means either that the networks built on different portions of the data are very different, or that the networks are similar but tested on very different data. The error variability in the cross-validation networks is a measure of the data variability. The variability of the data is due to the inherent variability in natural language and to the variability of the schizophrenic language disorder manifestations. The larger the data set the smaller the impact of the data variability on a classifier's accuracy.

Again, similarly to decision trees, the networks based on noisy data are more accurate. This effect is more pronounced in the neural networks than in the decision trees. For example, a network's error decreases from 22% on noise-free data to 0% on noisy data, or from 40% on noise-free data to 10% on noisy data. This result indicates that increasing the amount of data dramatically improves the computerized data classification.

Further, the networks with two output nodes perform better than those with three output nodes. This result is the same for decision trees. It suggests that it is more effective for classification to compare healthy speech to schizophrenic speech than to compare healthy speech to any subgroup of schizophrenic speech individually.

4.2.4 Important features for neural networks

For each network, Neuroshell 2 ranks the features based on their weights, indicating that the features with highest weight have larger impact on classification. I select the two features with

the highest weights for each network. The topmost important features (out of 14 features) for all the networks are:

- 1) Mean Sentence Length
- 2) Difficulty Words
- 3) Difficulty Expressions
- 4) Degree Expressions
- 5) Hedging Words

CHAPTER 5

INTERPRETATION AND DISCUSSION

5.1 Interpretation of results

5.1.1 Significant features

1. Schizophrenics use *longer sentences* than healthy participants. This result seems to lend support to existing findings of schizophrenics' lack of discourse planning, derailment and rambling (Andreasen, 1979a). Schizophrenics' longer sentences may be due to thoughts/sentences running together.
2. Schizophrenics use *fewer degree expressions* than healthy participants. This result may be due to schizophrenics making strong claims and failing to modify their discourse to the same extent as healthy people. It appears that a schizophrenic is not preoccupied with whether or not the hearer will ratify his or her statements. This inference supports the claim that schizophrenics have an impaired theory of mind and do not understand well the thoughts, intentions and expectations of others (Frith, 1992; Frith & Corcoran, 1996).
3. Schizophrenics use *more difficulty words* than healthy participants. This result supports prior findings that schizophrenics encounter cognitive difficulty in performing cognitive tasks (Rosenberg & Tucker, 1975, 1979). It suggests that schizophrenics are aware of their cognitive difficulty. However this inference is not as straightforward as it appears. We expect people that experience cognitive difficulty and exhibit healthy levels of self-awareness to be cautious in their claims. Based on this study, schizophrenics do not use less certainty words

than healthy people. In addition, schizophrenics use significantly fewer degree expressions than healthy people. It appears that schizophrenics are aware of their own cognitive difficulty, but cannot modify their claims accordingly. This conclusion would be in line with findings of schizophrenics' cognitive control failures (Braver, Barch, & Cohen, 1999b; Braver & Cohen, 1999).

5.1.2 Insights from machine learning

Decision trees and neural networks have many results in common.

1. Both machine learning techniques perform better when comparing schizophrenic speech to healthy speech than when comparing any subgroup of schizophrenic speech to healthy speech. The 'non thought-disordered' subgroup is highly correlated to the 'thought-disordered' subgroup for all the linguistic features measured by ENLF (Table 4.2). In other words, schizophrenics as a group are, with regards to the ENLF features, more dissimilar from healthy participants, than each schizophrenics' subgroup compared individually to the healthy participants.
2. The same effect may also be due to the fact that, when healthy speech is compared to all schizophrenic speech, all the original data is utilized and there are more speech samples to learn from. Both techniques' increased performance on noisy data (noise augments the data) shows that the size of the data set has a tangible effect on a classifier's accuracy. This is a very optimistic result. It shows that the linguistic features used in this project are effective for classification. In fact the networks perform relatively well, even by conservative performance measures (86% accuracy in cross-validation), in the face of the paucity of data.

3. Another similarity is that both techniques rank the same features as topmost for classification; neural networks rank as very important one feature ('Difficulty Expressions') that decision trees don't. The common features are:

- 1) Mean Sentence Length ($p = .01$)
- 2) Difficulty Words ($p = .02$)
- 3) Hedging Words ($p = .09$)
- 4) Degree Expressions ($p = .01$)

Three out of four features found important from both techniques distinguish healthy speech from schizophrenic speech at a statistically significant level. The non-significant feature ('Hedging Words') has the next best distinguishing level on the feature list. Thus, decision trees and neural networks are finding on their own (based on the provided features) the most salient information for classifying schizophrenic and healthy speech.

5.2 Discussion

5.2.1 Limitations

The small data sample is this project's biggest limitation. The classifiers based on noisy data perform better than the classifiers based only the original data. Also, classifiers using all the data perform better than the classifiers that use only subgroups of the data.

Another limitation is the variability of the speech samples. Part of this variability is inherent in natural language and part of it is inherent in the manifestation of schizophrenia. The schizophrenic participants do not make homogenous groups based on their symptoms. The linguistic abnormalities vary greatly from one participant to the next, making it difficult to group participants together. Thus, the labels 'non thought-disordered' and 'thought-disordered' group

together patients with very different linguistic features. A larger amount of speech samples would be a partial remedy to the problem of the inter-data variability.

A limitation in ENLF's applicability to classify schizophrenic speech lies in the fact that other diagnostic groups, including unipolar and bipolar depression patients, manifest similar language disorders, potentially including features measured by ENLF (Andreasen, 1979). This limitation may be sidestepped by fine-tuning the program and testing it with the speech of other diagnostic groups.

Last but not least, a possible limitation of this thesis' text analysis stems from using text derived from manual transcriptions of sound recordings. Manual transcription can introduce inaccuracies and subjectivity in finding the boundaries of utterances, interpreting a speaker's intonation (i.e. question markers), and in interpreting other speech elements, which subsequently are transcribed into the text sample. One way to address this limitation would be to use a computerized parser of natural language speech.

5.2.2 What has this thesis achieved?

This thesis finds three linguistic features that distinguish schizophrenic speech from healthy speech at a statistically significant level. I did not find any previous study that reports the same findings. Further, in the light of the above-mentioned limitations, the classifiers built here perform promisingly well, with a conservative accuracy measure of 86% correct. This level of accuracy is 10% higher than the highest classification accuracy of 76% for the schizophrenic group (Mete et al. 1993) that I was able to find in the content analysis literature.

Machine learning techniques, decision trees and neural networks have natural advantages over systems based on human judges. They are accurate, objective and reliable. Given their

natural strengths and the good classifying accuracy shown in this project, decision trees and neural networks are promising techniques for distinguishing healthy speech from schizophrenic speech. Future research could focus on finessing the types of features, and on addressing the linguistic variation between the speech samples.

5.2.3 Directions for future research

The current approach to content analysis studies can be taken further in several directions. The immediate next step would be to apply this program to more schizophrenic and healthy speech samples as they become available. This would be a good way to further test ENLF and to potentially notice new results. Along the same lines, ENLF could be applied to the ketamine-influenced speech samples. It is hypothesized that the drug ketamine induces schizophrenia-like symptoms in speech. The results of running ENLF on ketamine data would be one viable test of this hypothesis.

Further, ENLF could be enhanced to measure other features of natural language. For a further stylistic analysis, it would be easy to measure the frequencies of the parts of speech in schizophrenic and healthy language. Pennebaker and Lee (2002) find that the occurrence of particles – parts of speech that include pronouns, articles, prepositions, conjunctives, and auxiliary verbs, can tell much about a person’s psychological state. Another stylistic element to investigate would be the ratio of content words to function words. Other measurable stylistic elements are filler words, such as *um*, etc., and new word formations or neologisms, a known characteristic of schizophrenic speech (Andreasen, 1979a).

The measurement of content in novel ways is more involved. It has been noted that schizophrenic discourse lacks cohesion (Rochester & Martin, 1979; Abu-Akel, 1998). One way

to measure cohesion would be to devise a measure of semantic distance and or relatedness between content words.

REFERENCES

Abu-Akel, Ahmad (1998) The study of cohesion in schizophrenia: theory and application. *Issues in Applied Linguistics* 9, 37–60.

Andreasen, N. 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, N. C. (1979b). Thought, language, and communication disorders: diagnostic significance. *Archives of General Psychiatry*, 36, 1325-1330.

Barch, D. M., & Carter, C. S. (1998). Selective attention in schizophrenia: Relationship to verbal working memory. *Schizophrenia Research*, 33, 53-61.

Barch, D. M., Carter, C. S., Braver, T. S., & Cohen, J. D. (1997). The effects of D-amphetamine on working memory and language deficits in schizophrenia. *Schizophrenia Research*, 24, 129.

Barch, D. M., Carter, C. S., Braver, T. S., MacDonald, A., Sabb, F., Noll, D. C., et al. (1999). Prefrontal cortex and context processing in medication-naïve first-episode patients with schizophrenia. *Schizophrenia Research*, 36, 217-218.

Barch, D. M., Carter, C. S., Hatcher, P. C., Cohen, J. D. (1999). The "benefits" of distractibility: The mechanisms underlying increased stroop effects in schizophrenia. *Schizophrenia Research*, 24(4), 749-762.

Barch, D. M., Carter, C. S., MacDonald, A. W., III, Braver, T. S., Cohen, J. D. (1998). Context processing deficits in schizophrenia. *Journal of Abnormal Psychology*, 105, 592-601.

Barch, D. M., Carter, C. S., Perlstein, W., Baird, J., Cohen, J. D., Schooler, N. (1999). Increased stroop facilitation effects in schizophrenia are not due to increased automatic spreading activation. *Schizophrenia Research*, 39(1), 51-64.

Barch, D. M., Sheline, Y. I., Csernansky, J. G., & Snyder, A. Z. (2003). Working memory and prefrontal cortex dysfunction: Specificity to schizophrenia as compared to major depression. *Biological Psychiatry*, 53, 376-384.

Bleuler, E. (1950). *Dementia praecox or, the group of schizophrenias* (J. Zinkin, Trans.). New York: International Universities Press. (Original work published 1911).

Braver, T. S., Barch, D. M., & Cohen, J. D. (1999b). Mechanisms of cognitive control: Active memory, inhibition, and the prefrontal cortex (Tech. Rep. No. PDP.CNS.99.1). Pittsburgh PA: Carnegie Mellon University.

Braver, T. S. & Cohen, J. D. (1999). Dopamine, cognitive control, and schizophrenia: The gating model. *Progress in Brain Research*, 121, 327-349.

Carlson, N. R. (2003). *The biological foundations of behavior (7th ed.)*. Needham Heights, MA: Allyn and Bacon.

Carter, C. S. & Barch, D. M. (2000). Attention, memory and language disturbances. In C. Andrade (Ed.), *Advances in Psychiatry* (Vol. 3, pp. 45-72). New Delhi, India: Oxford University Press.

Carter, C. S., Barch, D. M., & Cohen, J. D. (1997). Disturbed language processing, disorganization, and attentional impairment in schizophrenia. *Schizophrenia Research*, 24, 130.

Chaika, E. (1990). *Understanding psychotic speech: Beyond Freud and Chomsky*. Springfield, IL: Charles C Thomas Publishers.

Channell J. (1994). *Vague language*. Oxford, United Kingdom: Oxford University Press.

Chen, E. Y. H., Lam, L. C. W., Kan, C. S., Chan, C. K. Y., Kwok, C. L., Nguyen, D. G. H., et al. (1996). Language disorganization in schizophrenia: validation and assessment with a new clinical rating instrument. *Hong Kong Journal of Psychiatry*, 6(1), 4-13.

Cohen, J. D., Barch, D. M., Carter, C. S., Servan-Schrieber, D. (1999). Schizophrenic deficits in the processing of context: Converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology*, 108, 120-133.

Corcoran, R., Mercer, G., & Frith, C. D. (1995). Schizophrenia, symptomatology and social inference: Investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research*, 17, 5-13.

Covington, M. (2002a). Natural Language Analysis Tools (NLAT). Unpublished manuscript.

Covington, M., Brown, C., He, C., McLain, J., & Naçi, L. (2003). Schizophrenia and the structure of language: A selective review of the literature. Unpublished manuscript.

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.

Daelemans, W., Berck, P., Gillis, S. (1997) Data Mining as a Method for Linguistic Analysis: Dutch Diminutives, *Folia Linguistica* XXXI/1-2, 57-75.

Docherty, N. M., DeRosa, M., & Andreasen, N. C. (1996). Communication disturbances in schizophrenia and mania. *Archives of General Psychiatry* 53, 358-364.

Docherty, N. M., Hawkins, K. A., Hoffman, R. A., Quinlan, D. M., Rakfeldt, J., & Sledge, W. H. (1996). Working memory, attention, and communication disturbances in schizophrenia. *Journal of Abnormal Psychology*, 105, 212-219.

Fairbanks, H. (1944). Studies in Language Behavior: II. The quantitative differentiation of samples of spoken language. *Psychological Monographs*, 56, 19-38.

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. Hove, East Sussex, England: Psychology Press.

Frith, C. D. & Corcoran, R. (1996). Exploring "theory of mind" in people with schizophrenia. *Psychological Medicine*, 26(3), 521-530.

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley, CA: University of California Press.

Harrow, M. & Quinlan, D. (1977). Is disordered thinking unique to schizophrenia? *Archives of General Psychiatry* 34, 15-21.

Huebler, A. (1983). *Understatements and hedges in English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kelly, E. F., & Stone, P. J. (1975). *Computer recognition of English Word Senses*. Amsterdam: North-Holland.

Liddle, P. F., Ngan, E. T. C., Caissie, S. L., Anderson, C. M., Bates, A. T., Quesed, D. J., et al. (2002). Thought and language index: An instrument for assessing thought and language in schizophrenia. *British Journal of Psychiatry*, 181, 326-330.

Mahler, B. (1972). The language of Schizophrenia: A review and interpretation. *British Journal of Psychiatry*, 120, 3-17.

McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 7, 115 - 133.

Merriam-Webster Thesaurus (1989, October). Springfield, MA: Merriam-Webster Inc.

Mete, L., Schnurr, P. P., Rosenberg, S. D., Oxman, T. E., Doganer, I., & Sorias, S. (1993, November). Language content and schizophrenia in acute-phase turkish patients. *Social Psychiatry And Psychiatric Epidemiology*, 28 (6), 275-280.

Mitchell, T. M. (1997) *Machine Learning*. McGraw-Hill.

Murray, A. H. (1943). *Thematic Apperception Test*. Boston: Harvard University Press.

Oxman, T. E., Rosenberg, S. D., Schnurr, P. P., & Tucker, G. J. (1988). Diagnostic classification through content analysis of patients' speech. *American Journal of Psychiatry*, 145, 464-68.

Oxman, T. E., Rosenberg, S. D., Schnurr, P. P., Tucker, G. J., & Gala, G. (1998, July). The language of altered states. *Journal Of Nervous And Mental Disease*, 176 (7), 401-408.

Palmer, R. F. (2001). *Mood and Modality*. Cambridge, England: Cambridge University Press.

Pennebaker, J.W., & Lee, C. H. (2002). The power of words in social, clinical, and personality psychology. *The Korean Journal of Thinking and Problem Solving*, 12, 35-43.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology*, 54, 547-77.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1, 81-106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Rochester, S., & Martin, J. R. (1979). *Crazy talk: A study of the discourse of schizophrenic speakers*. New York: Plenum.

Rosenberg, S. D., Schnurr, P. P., & Oxman, T. E. (1990). Content analysis: A comparison of manual and computerized systems. *Journal of Personality Assessment*, 54(1 & 2), 298-310.

Rumelhart, D. E., & McClelland, J. L., (Eds.). (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Smith, M (1993) *Neural networks for statistical modeling*. New York: Van Nostrand Reinhold.

SPSS for Windows, Rel. 11.0.1. 2001. Chicago: SPSS Inc.

StatSoft, Inc. (2004). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: <http://www.statsoft.com/textbook/stathome.html>.

Stone, P. J., Dunphy, D, Smith, M. S., Ogilvie, D. M. (Eds.). (1966). *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Tucker, G. J., & Rosenberg, S. D. (1975). Computer content analysis of schizophrenic speech: a preliminary report. *American Journal of Psychiatry*, 132, 611-16.

Whitehorn, J., & Zipf, G. (1943). Schizophrenic language. *Archives of General Psychiatry*, 49, 831-851.

Witten, I. H., & Frank, E. (1999). *Data mining: Practical machine learning tools and techniques with java implementations*. Morgan Kaufmann.

APPENDIX

LISTS OF WORDS AND EXPRESSIONS

The following lists were derived from literature on hedging, degree, and modality expression in English (Huebler, 1983; Palmer, 2001; Channell, 1994), as well as by using the thesaurus (Merriam-Webster Inc., 1989).

Hedging Words:

appear,

emerge,

loom,

show,

seem,

suggest,

hint,

imply,

presume,

conjecture,

guess,

pretend,

suppose,
surmise,
think,
mean,
interpret,
presuppose,
assume,
premise,
imagine,
conceive,
fancy,
realize,
believe,
expect,
gather,
reckon,
suspect,
understand,
evidence,
imagination,
explanation,
look,
impression,

appearance,
presumably,
assumably,
likely,
presumptively,
possibly,
maybe,
perchance,
perhaps,
necessarily,
probably,
supposedly,
seemingly,
conceivably,
may,
might,
could,
can,
whether,
if

Degree Words:

some,

certain,

various,

kind,

sort,

thing,

any,

ever,

barely,

hardly,

scarcely,

partially,

partly,

comparatively,

enough,

relatively,

basically,

practically,

technically,

virtually,

quite,

rather,

mildly,

moderately,

pretty,

slightly,

almost,

nearly

Degree Expressions:

in many respect,

in some respect,

in part,

to some extent,

kind of,

sort of,

more or less,

a little,

a bit

Difficulty Words:

cannot,

don't,

can't,

doesn't,

dunno,

couldn't,

why¹⁰,

where,

what,

when,

who,

which,

whose,

wonder,

question

Difficulty Expressions:

for some reason,

doesn't look,

doesn't match,

doesn't make sense,

not know,

not sure,

not clear,

not obvious

¹⁰ Question words (including wh-words) are taken to express cognitive hardship. The interviewer, discouraging dialogue, asks the subject to briefly describe the picture to the best of his/her ability. Most questions appear to be rhetoric and to express the speaker's confusion.

Certainty Words:

certainly,

obviously,

clearly,

surely,

evidently,

must,

should,

would,

will,

doubtless