

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 used<sup>1</sup>; the other seven features I measured with the Natural Language Analysis Tools (NLAT<sup>2</sup>) 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 terms<sup>3</sup> 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.

---

<sup>1</sup> The features are explained in the materials and methods chapter.

<sup>2</sup> ENLF and NLAT are explained in detail in the materials and methods chapter.

<sup>3</sup> 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, Hachen, & Cohen, 1999; Cohen, Barch, Carter, & Servan-Schieber, 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-Schieber, 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-disordered<sup>4</sup>) 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.

---

<sup>4</sup> 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**

<b>Linguistic Features</b>	<b>Motivation For Using Each Feature</b>
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**

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

<b>Linguistic Features</b>	<b>Brief Explanation of Each Feature</b>
<b>1. Mean Word Length</b>	Mean length of words in speech sample
<b>2. Mean Sentence Length</b>	Mean length of sentences in speech sample
<b>3. Hedging Words</b>	Number of hedging words
<b>4. Degree Words</b>	Number of degree words
<b>5. Degree Expressions</b>	Number of degree expressions
<b>6. Difficulty Words</b>	Number of hardship words
<b>7. Difficulty Expressions</b>	Number of hardship expressions
<b>8. Certainty Words</b>	Number of certainty words
<b>9. Text Repetition</b>	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**

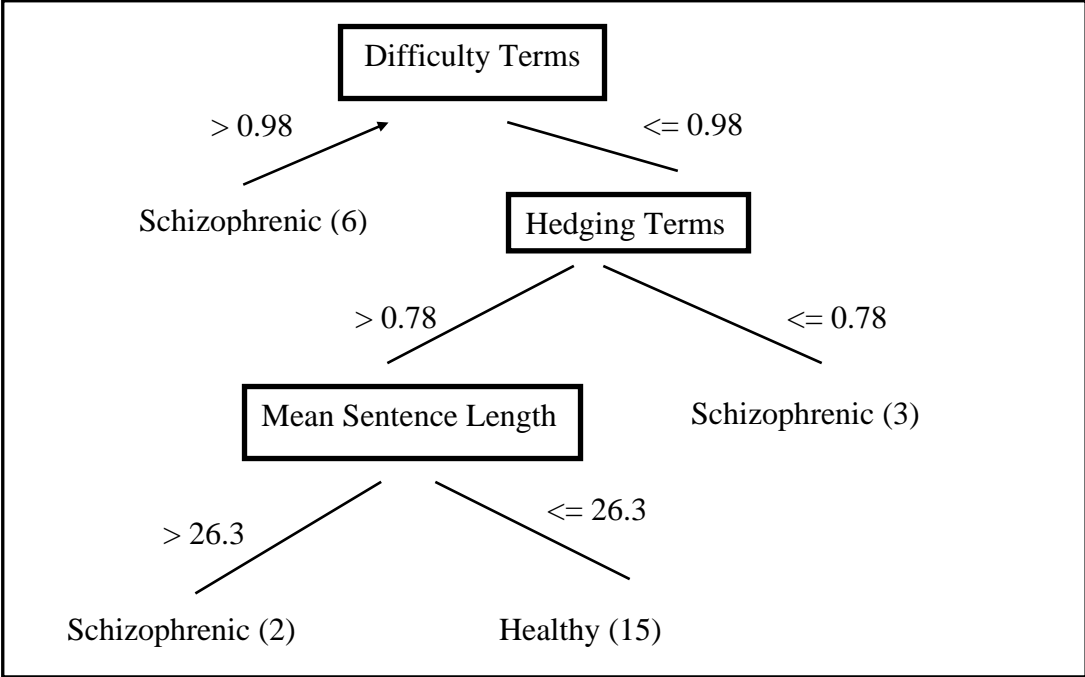
<b>ATTRIBUTES</b>	1. Number of Tokens	617
	2. Number of Types	328
	3. Type-Token Ratio	0.53
	4. Number of Stems	310
	5. Stems-Token Ratio	0.5
	6. Repetitiousness	85.55
	7. Repetitiousness-Words Ratio	0.14
	8. Mean Word Length	4.54
	9. Mean Sentence Length	14.61
	10. Hedging Words	0.79
	11. Degree Words	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
<b>TARGET ATTRIBUTE</b>	Classification	<b>Thought-Disordered</b>

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 See5<sup>5</sup> 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).

---

<sup>5</sup> 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.

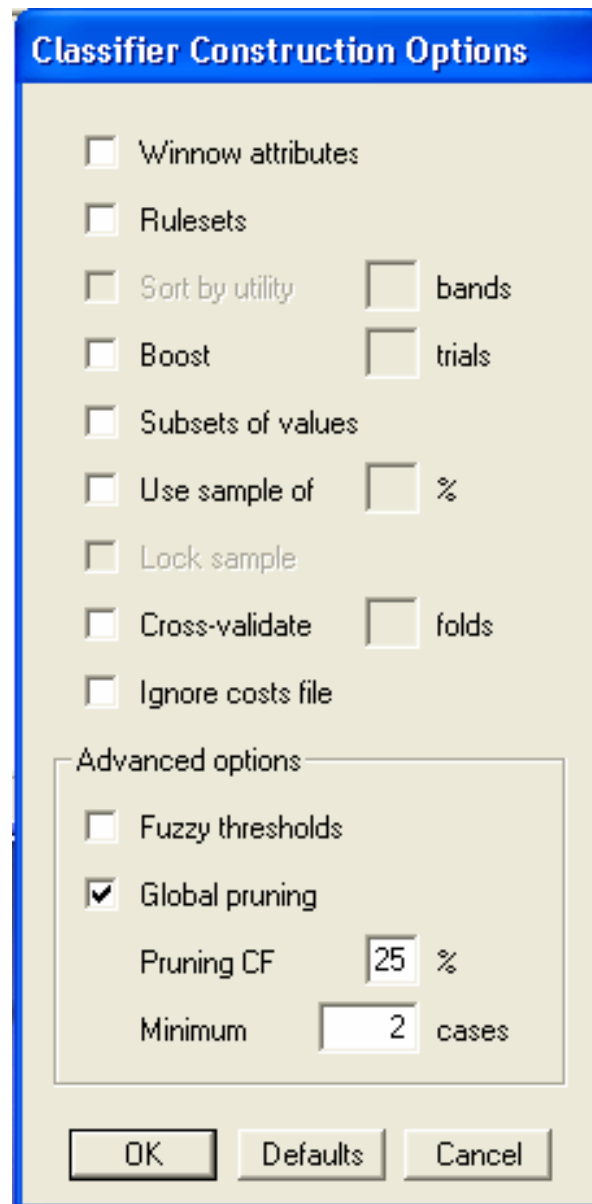
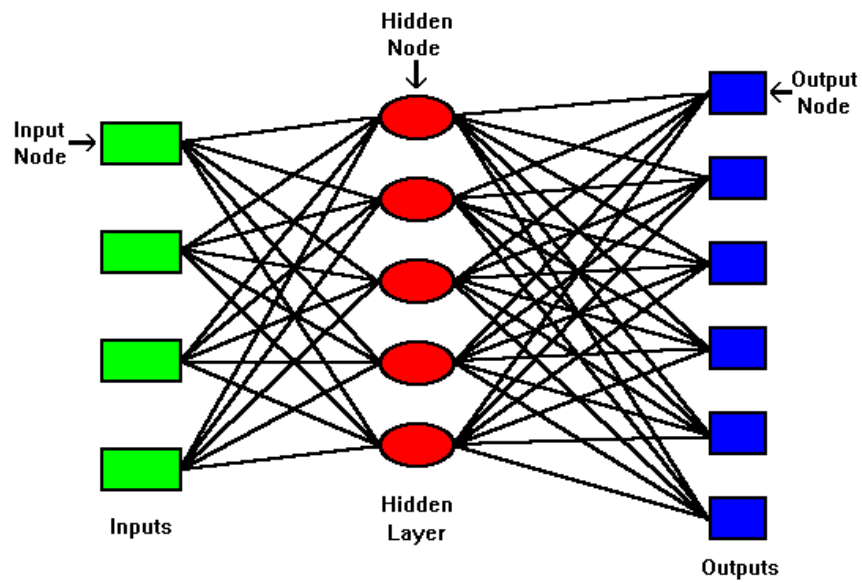


Figure 2.2 See5 classifier design options

### 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 2<sup>6</sup> 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. Momentum<sup>7</sup>: 0.1 (suggested by Neuroshell 2)
  - B. Learning Rate<sup>8</sup>: 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

---

<sup>6</sup> Neuroshell II is copyrighted by Ward Systems Group, Inc.

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

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

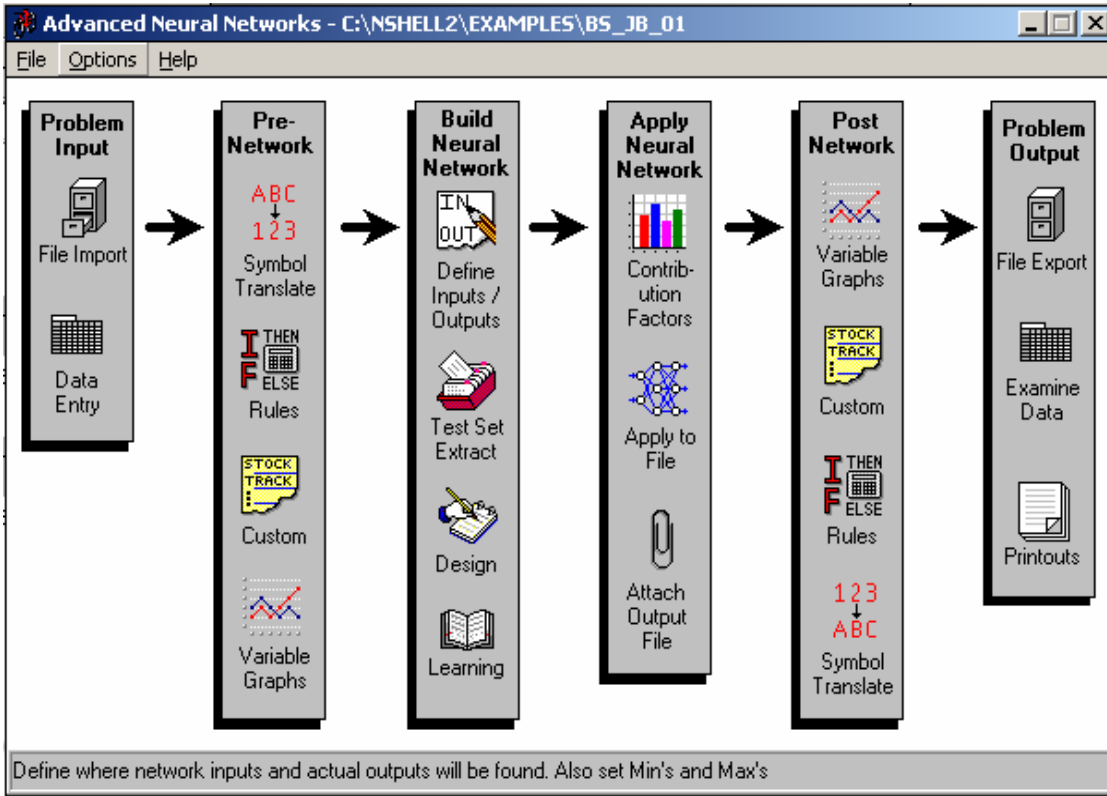


Figure 2.4 Neuroshell 2 neural network design window

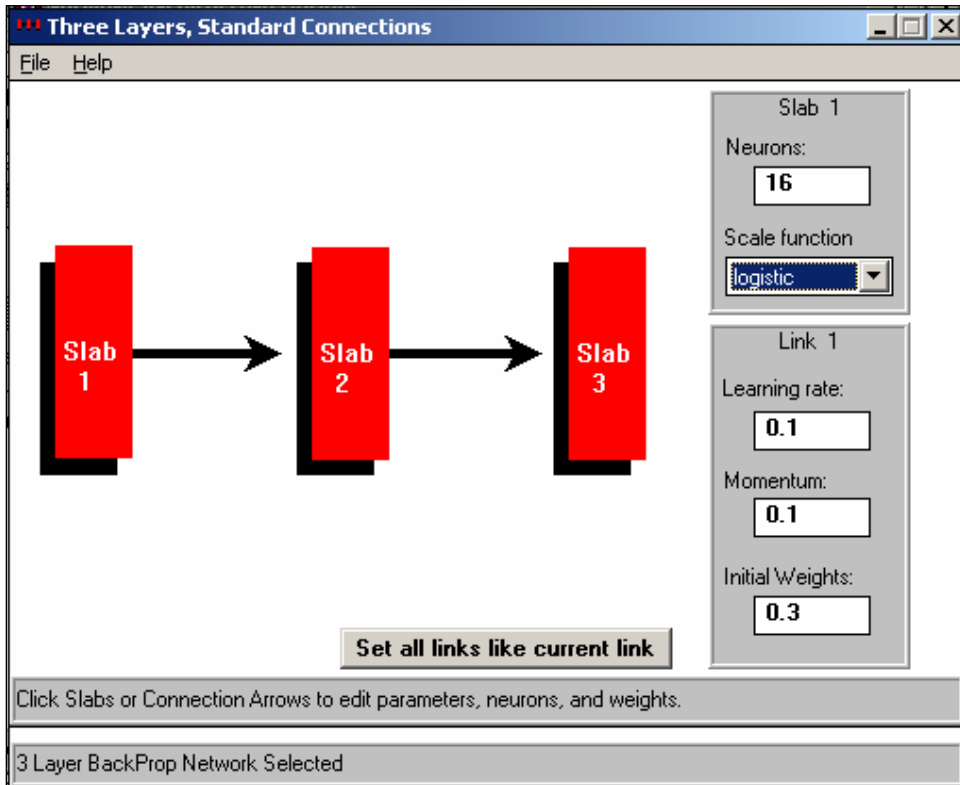


Figure 2.5 Neuroshell 2 three layer back-propagation neural network

## **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.









































































