Syntactic complexity, error and the holistic evaluation of ESL student essays

John S. Sparks

Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

Part of the Bilingual, Multilingual, and Multicultural Education Commons, and the Syntax Commons

Let us know how access to this document benefits you.

Recommended Citation
https://doi.org/10.15760/etd.5714

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.
Title: Syntactic Complexity, Error and the Holistic Evaluation of ESL Student Essays.

APPROVED BY MEMBERS OF THE THESIS COMMITTEE:

Marjorie Terdal, Chairperson
Sandra Oster
Thomas Buell
Dirgham Sbait

This study was designed to test five hypotheses in order to answer the following general research questions: 1) Are measures of syntactic complexity valid indices of ESL writing quality as measured by the holistic rating of student essays? 2) Would a measure of frequency and seriousness of error reflect evaluators' perceptions of ESL writing quality?
Many researchers in language development have argued that greater syntactic complexity may be indicative of better writing quality. A number of studies have advocated the "T-unit," a measure of independent clauses, as a valid index of writing quality. Some researchers, however, have found the T-unit to be unreliable and have proposed alternative measures. In addition, studies of ESL writing have concluded that a measure of frequency and seriousness of error is a necessary index in the analysis of ESL student compositions.

A clause analysis technique designed by Louis Arena (1982) was used in this study to determine syntactic complexity in ESL student essays. A measure of error, the "Correctness Score," devised by Brodkey and Young (1981), was adapted for the purposes of this study.

The essay corpus analyzed in this study consisted of thirty ESL student compositions written for a practice Test of Written English (TWE) examination. The essays were rated holistically, according to the TWE grading scale, by two experienced ESL instructors. Each essay was then analyzed in terms of Arena's clause analysis technique, using three measures of syntactic complexity: sentence length, information block length, and embedding depth; and the Correctness Score. Statistical analysis, using Spearman's rank order correlation, was applied to determine the relationships between the three measures of the clause
analysis technique, the Correctness Score, and the holistic ratings.

The results revealed that no significant correlation existed between the three measures of the clause analysis technique and the holistic ratings. A significant, moderate correlation was, however, established between the Correctness Score and the holistic ratings of the thirty ESL student essays.

These results indicate that frequency and seriousness of error is a major factor in rater judgment of ESL student writing quality, but that complexity of syntax appears to have little relationship with evaluators' perceptions of what constitutes "good" writing. It must be noted, however, that the holistic ratings used in this study did not show a strong degree of reliability. In addition, while a measure of frequency and seriousness of error may be a valid index of evaluators' perceptions of ESL writing quality, the Correctness Score also appeared to possess inherent reliability problems.
SYNTACTIC COMPLEXITY, ERROR AND THE HOLISTIC EVALUATION OF ESL STUDENT ESSAYS

by

JOHN SHELDON SPARKS

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF ARTS
in
ENGLISH: TESOL

Portland State University
1988
TO THE OFFICE OF GRADUATE STUDIES AND RESEARCH:

The members of the committee approve the thesis of John S. Sparks presented January 27, 1988.

Marjorie Terdal, Chairperson

Sandra Oster

Thomas Buell

APPROVED:

John R. Cooper, Chair, Department of English

Bernard Ross, Vice Provost for Graduate Studies
ACKNOWLEDGEMENTS

I must gratefully acknowledge those who gave support and assistance in the completion of this study. My principal advisors, Dr. Marjorie Terdal and Dr. Sandra Oster, made themselves available and spent hours reading and criticizing the various drafts of this thesis. Their suggestions and advice were always forthright and greatly appreciated. Many thanks are also due to Dr. Thomas Buell and Dr. Dirgham Sbait who agreed to complete the committee for my oral examinations and offered trenchant criticisms of the work. Dr. Naguib Greis gave information about the Test of English as a Second Language and allowed me to use essay samples kept on record by the Center for English as a Second Language. Luis Colon completed the exhaustive task of copying these essays on his word processor and reproduced them for my use. Lisa Miletich copied the original copies. Jane Dresser, Ann Kirk and Tina English undertook the important tasks of rating the essays and completing reliability checks. To all of the above I am deeply indebted.
# TABLE OF CONTENTS

## ACKNOWLEDGEMENTS

PAGE iii

## LIST OF TABLES

PAGE vi

## LIST OF FIGURES

PAGE vii

## CHAPTER

### I INTRODUCTION

- Statement of Purpose 1
- Clarification of Terms 3
- Summary 6

### II REVIEW OF THE LITERATURE

- The T-Unit 8
- Alternatives to the T-Unit 15
- Clause Analysis 20
- Limitations of Measures of Syntactic Complexity 24
- The Measurement of Error in ESL Writing 25
- The Reliability of Holistic Ratings 29
- Summary 33

### III METHODS AND PROCEDURES

- Subjects 37
- Instrumentation 37
- Testing Conditions 38
Scoring .................................. 38
Reader Reliability .......................... 39
Analysis .................................... 39
Limitations in Applying the Analysis to ESL Writing .................... 45
Reliability of the Analysis ................. 48
Statistical Procedures ...................... 51
Summary .................................... 51

IV RESULTS AND DISCUSSION .............. 53
Results .................................... 53
Discussion .................................. 72
Summary .................................... 77

V CONCLUSION AND IMPLICATIONS .......... 79
Conclusion .................................. 79
Implications ................................. 84

REFERENCES .................................. 88
APPENDICES .................................. 95
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>50</td>
</tr>
<tr>
<td>II</td>
<td>54</td>
</tr>
<tr>
<td>III</td>
<td>64</td>
</tr>
<tr>
<td>IV</td>
<td>69</td>
</tr>
</tbody>
</table>

I. Reliability Coefficients for Clause Analysis Technique and Correctness Score

II. Essay Sample Data for Holistic Ratings, Clause Analysis Technique and Correctness Score

III. Spearman's Rank order Correlation (rho) of the Clause Analysis Technique and Holistic Ratings

IV. Spearman's Rank Order Correlation (rho) of Correctness Score and Holistic Ratings
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Distribution of Holistic Ratings in the Essay Corpus</td>
</tr>
<tr>
<td>2.</td>
<td>Distribution of Scores for Sentence Length in the Essay Sample</td>
</tr>
<tr>
<td>3.</td>
<td>Distribution of Scores for Information Block Length in the Essay Sample</td>
</tr>
<tr>
<td>4.</td>
<td>Distribution of Scores for Embedding Depth in the Essay Sample</td>
</tr>
<tr>
<td>5.</td>
<td>Distribution of Correctness Score Results in the Essay Sample</td>
</tr>
</tbody>
</table>
A perennial problem in first and second language instruction has been the construction of a valid objective index of writing quality. Beginning with Kellogg Hunt (1965) and the "T-unit," the search for a valid objective measure of writing quality has focused on measures of syntactic complexity. While much of this research has concentrated on growth and development in first language writing skills, significant attention has recently been given to the measurement of syntactic complexity in ESL writing. Several studies (Larsen-Freeman and Strom 1977, Gaies 1980, Perkins 1980, Homburg 1984) have concluded that a measure of syntactic complexity cannot be a valid index of ESL writing quality unless it accounts for the factor of error.

The purpose of this thesis was to investigate the relationship between ESL writing quality as determined by holistic ratings, syntactic complexity, and frequency and seriousness of error.

The two research questions this thesis sought to answer were: 1) Are measures of syntactic complexity valid
indices of ESL writing quality as measured by the holistic scoring of student essays?

2) Would a measure of frequency and seriousness of error reflect evaluators' perceptions of ESL writing quality?

This study was designed to address the following hypotheses:

1) that a significant correlation would not be evident between sentence length, here defined as the number of clauses per sentence, and the holistic ratings;

2) that a significant correlation would be found between information block length, or the number of clauses per main clause, and the holistic ratings;

3) that a significant correlation would be established between embedding depth, or the "value" of embedding, and the holistic ratings;

4) that there would be no significant correlation between a measure for frequency and seriousness of error and the holistic ratings;

5) that a positive correlation would be established, using a multiple correlation formula, between embedding depth, frequency and seriousness of error, and the holistic ratings.

The hypotheses assume that a measure of syntactic complexity which does not distinguish between main and subordinate clauses in a sentence, i.e. sentence length,
would not reflect writing quality, but that measures of subordination, such as information block length and embedding depth, would reflect ESL writing quality as determined by holistic ratings. In addition, it was assumed that a measure of error combined with an index of syntactic complexity, such as embedding depth, would reflect evaluators' perceptions of ESL student writing quality, but that a tabulation of error alone would not.

CLARIFICATION OF TERMS

Syntactic Complexity

Perron (1977) defined syntactic complexity as "the way a sentence becomes increasingly more elaborate via transformational processes which conjoin and embed full and reduced clauses and non-clause elements" (2). Thus, the more "elaborate" a sentence becomes in terms of embedding and conjoining, the greater its complexity. Researchers have argued over which measures of embedding and conjoining constitute the most "valid" indices of syntactic complexity. Some of these arguments will be discussed in the Review of the Literature chapter of this thesis.

In this study, three indices of syntactic complexity promoted by Louis Arena (1982) were applied. Collectively termed the "clause analysis technique," the three measures are: sentence length (number of clauses per sentence), information block length (number of clauses per main
clause), and embedding depth (the average "value" of embedded clauses).

Indices of syntactic complexity have been used by some researchers to determine syntactic "maturity." The term "maturity," however, used by Hunt (1965), O'Donnell, Griffin, and Norris (1967), O'Hare (1973), Loban (1976), and Daiker, Kerek, and Morenberg (1978), presents certain problems of definition. In these studies, "maturity" was taken to mean "increased complexity of syntax." But, as some critics of this definition have shown, increased complexity of syntax is not necessarily indicative of a "mature" style (Christensen 1968, Faigley 1980, Arena 1982). White (1985) has stated: "Some of the worst bureaucratic writing is 'syntactically mature,' ...while some of the best modern English writing consists of short and direct statements" (174). Hunt and others seem to have concluded that "maturity" in syntax was a result of increased complexity; some researchers would disagree. Ashida (1967) gives a more relative definition of maturity:

I suspect that when we speak of a student's relative written 'language maturity,' we are really speaking of the extent to which the formal, syntactic, and conceptual structures in his writing reflect the practices evident in the writing of the adult writers of a particular standard dialect (157).

In this thesis, increased complexity was hypothesized to be indicative of evaluators' perceptions of better writing quality. Given the controversy surrounding the
term, the word maturity will be avoided altogether to the studies of the specific researchers who have used the term.

**Writing Quality**

The term writing quality refers to the overall quality of a piece of writing. It includes all the factors which may influence the effectiveness of an essay: content, organization, selection of vocabulary, grammar, spelling, frequency and seriousness of error, variety and complexity of syntax, etc. Measures of syntax and error will be applied in this study in order to determine whether one, both, or neither of these characteristics are indices of evaluators' perceptions of ESL writing quality.

**Holistic Rating**

Three major scoring methods are used by evaluators to measure writing quality. These are the systems of analytical, primary trait, and holistic scoring. Analytical scoring judges a writer's performance according to several established categories, such as vocabulary, grammar, content, organization. Primary trait scoring measures only one aspect of a piece of writing, such as grammar. Holistic rating is the only system which gives a single score for an overall impression of writing quality. The principal advantage of holistic rating is its efficiency in terms of consumption of time. Cooper (1977) describes the holistic rating procedure:
Holistic evaluation of writing is a guided procedure for sorting or ranking written pieces. The rater takes a piece of writing and either (1) matches it with another piece in a graded series of pieces or (2) scores it for the prominence of certain features important to that kind of writing or (3) assigns it with a letter or number grade. The placing, scoring, or grading occurs quickly and impressionistically, after the rater has practiced the procedure with other raters. The rater does not make corrections or revisions in the paper. Holistic evaluation is usually guided by a scoring guide which describes each feature and identifies high, middle, and low quality levels for each feature (Cooper 1977: 3).

The scoring guide used in this study was that devised by the TWE (Test of Written English). The essay sample was rated by trained, experienced raters.

SUMMARY

This study was designed to address five hypotheses derived from two basic research questions involving the relationships between syntactic complexity, as well as frequency and seriousness of error, and evaluators' perceptions of ESL writing quality as determined by holistic ratings. The key terms syntactic complexity, writing quality, and holistic rating were defined.
CHAPTER II

REVIEW OF THE LITERATURE

The quality of a piece of writing is the measure of a variety of characteristics. Some of these characteristics are more objectively quantifiable than others. In recent years, the measurement of syntactic complexity has received much attention. In 1965, Kellogg Hunt developed the T-unit, which has been used in a number of important studies investigating syntactic complexity in the development of written language and its relationship to writing quality. Some critics of the T-unit have developed other more elaborate measures of syntactic complexity.

This review of the literature will present some of the various methods of analysis which have been devised to measure syntactic complexity in student writing in both native and second languages. Also included in this chapter are a discussion of objective measurements of error and a review of the reliability factors concerned in the holistic evaluation of writing.
The T-unit in First Language Research

Any discussion of research in syntactic complexity must begin with Kellogg Hunt's classic study in the development of syntactic structures at three grade levels among American students in public schools (1965). Strongly influenced by Chomsky's (1957) theories of transformational grammar, Hunt experimented with several measures of syntactic complexity—sentence length, clause length, ratio of subordinate clauses to all clauses, and the "T-unit." Hunt concluded that the most efficient and reliable measure of syntactic complexity was the "T-unit" or "minimal terminable unit." He defined the T-unit as "one main clause with all the subordinate clauses attached to it" (1965: 21). Hunt found that syntactic complexity, as measured by the length of T-units, increased with the ages of his student sample. He proposed that the T-unit was a valid measure of syntactic complexity because, in mature writing, a T-unit was often the same as a sentence; immature writing, however, was usually characterized by an over-abundance of coordination, resulting in many short T-units per sentence. Hunt posited, therefore, that it was not sentence length, but T-unit length and clause length, which were the best measures of "mature" writing. Hunt preferred the measure of T-unit length over clause length because he found the former less time-consuming to apply and thus more
convenient for the purposes of research. T-unit length is computed by multiplying mean words per clause by the number of clauses per T-unit.

Hunt's findings were confirmed in a small pilot study by Potter (1967). In addition, the measure of T-unit length was found to be the best measure of syntactic complexity in studies by O'Donnell, Griffin, and Norris (1967) and Loban (1976). O'Donnell et al., in research on children's syntax, found positive correlations between an increase in T-unit length and the number of sentence-combining transformations per T-unit. O'Donnell et al. concluded that T-unit length should be the established measure for syntactic complexity, since it was easier to compute than the number of transformations per T-unit. In another study, Loban (1976) measured the growth of children's language throughout their schooling, from kindergarten to the twelfth grade. In this longitudinal study, Loban found a high correlation between increased T-unit length and the overall quality of the student essays.

Hunt urged that a new pedagogy be based on his findings: curricula should be designed for the writing class to help students combine short, simple sentences into longer sentences, the objective being that the natural development of syntactic complexity should be augmented by classroom instruction in sentence-combining. Studies by Mellon (1969), O'Hare (1973), and Daiker, Kerek, and
Morenberg (1978) demonstrated that sentence-combining exercises produced increases in syntactic complexity over many grade levels.

Hunt (1970), having found that syntactic complexity increased with age, also showed significant correlations between students of good writing ability and higher scores in T-unit length and clause length. This finding was supported by independent research (San Jose 1972, Loban 1976).

These findings led to the proposal of two important hypotheses: 1) that T-unit length and clause length may show a significant correlation with ratings of student essays, and 2) that sentence-combining exercises may significantly improve the quality of student writing (Crowhurst 1983: 2). Mellon (1985) stated that the T-unit length and clause length measures were the "best single variable indices of syntactic maturity" (12) on the evidence of their applicability to first language writing development research.

The T-unit, then, has been forcefully promoted as a valid index of student writing quality. However, critics such as Ney (1966) have questioned the ultimate validity of T-unit analysis. Ney wrote a review of Hunt's 1965 study, arguing that the T-unit measure imposed unnatural barriers in syntax at odds with the writer's intentions:

... by repunctuating sentences and establishing the T-unit, [Hunt] abolishes co-ordination between
main clauses in student writing. Although Hunt does count the occurrences of co-ordination between main clauses, the establishment of the T-unit and the emphasis given to this as a unit of measurement tend to detract from the co-ordination between main clauses as an index of maturity (235).

As an example, Ney cited a pair of sentences:

Then the rain falls and spring comes.

So he went on through the woods and pulled the feather out of his hat from the partridges and put a flower in his hat and walked on through the woods.

Ney found the decision to call the first sentence two T-units and the second sentence one T-unit "arbitrary" and "not very attractive" (234-5).

Additional criticism of T-unit length as an index of writing quality has considered the effect of modes of discourse on syntactic complexity in student writing. Perron (1977) found that mean T-unit length was greatest in argument, followed by compositions in the modes of exposition, narration and description. Crowhurst and Piché (1979) found no significant difference between Grades Six and Ten in the mode of narration as measured by T-unit length, and that "variations in mode produced differences in syntactic complexity at each of the two grade levels" (107). Crowhurst and Piché also examined the effect of audience on student writing styles, and discovered that clause length and T-unit length were greater in compositions written for "teacher" than in compositions written for "best friend."

T-unit length has also been criticized as an unreliable measure of writing quality at higher grades and at the
college level (Nold and Freedman 1977, Gebhard 1978, Stewart and Grobe 1979). Hake and Williams (1985) have contradicted the assumption that increased T-unit length denoted more "mature" writing. In a study of 212 students of different levels of competence, they concluded that growth in writing "may be associated not with longer T-units, but with shorter ones" (86). Hake and Williams' subjects who passed a writing test wrote shorter T-units than those students who failed.

Furthermore, some researchers have questioned the value of the sentence-combining instruction advocated by Hunt (Crowhurst 1983, de Beaugrande 1985, White 1985). In a review of the practical implications of syntactic complexity research, Strong (1985) addressed the dangers inherent in applying Hunt's theory: "So far, no one has articulated a real rationale for [Sentence Combining] a rationale related to language learning, writing process, and 'naturalistic' classroom pedagogy" (745). Prescriptive attempts to improve student writing quality through the sentence combining pedagogy based on T-unit research could be seen as impediments to the development of an individual's "natural" style. The dangers of the pedagogy are that greater complexity could be seen as an end in itself and, as a result, other aspects of writing, such as clarity, individual style and comprehensibility, could suffer.
The T-unit in Second Language Research

Despite recent criticisms of the validity of the T-unit for first language research, it has remained the most widely used instrument for measuring syntactic complexity in ESL student writing. Flahive and Snow (1980) found only two discriminant functions in their analysis of syntactic complexity in ESL student writing: length of T-unit and the clause per T-unit ratio. Kameen (1979) found that the number of long T-units (21 + words) may be a powerful index of writing quality among ESL students at the college level.

The T-unit has also been applied as an index of writing development in other second languages. Monroe (1975) and Cooper (1976) found T-unit length to be an effective measure of language acquisition for students of French and German, respectively. Monroe wrote that "studies in foreign and native language development have shown beyond doubt that the student's ability to combine short, simple sentences into longer and more complex ones has much to do with his language maturity" (Monroe 1975: 1031).

Other researchers, however, have pointed to specific characteristics of ESL writing, and have contended that T-unit length may not be a valid measure of ESL writing quality:

Since adult learners of English often write long T-units by translating sentences from their native language word for word into English or by using interlanguage rules to combine English words, longer T-units in the writing of foreign adults may not indicate a more sophisticated control of the

Research in the measurement of ESL writing quality, therefore, has turned to the consideration of error in T-unit analysis. In a study of compositions written by Arab students of ESL, Scott and Tucker (1974) measured only the percentage of error-free T-units. Larsen-Freeman and Strom (1977) used the average length of error-free T-units per composition as a measure of ESL writing quality. Larsen-Freeman (1978) promoted the percentage of error-free T-units and the average length of error-free T-units as discriminating measures of ESL proficiency. Homburg (1984) found that the number of error free T-units per composition was a discriminating measure of writing quality. On the other hand, Arthur (1979) discovered that average length of T-units and average length of error-free T-units both decreased slightly in his short-term study of college-level ESL compositions.

A central concern with the use of the error-free T-unit as an index of ESL writing quality, however, has to be that the essay being analyzed ought to be of a sufficiently high standard to contain a number of error-free T-units. Scott and Tucker (1974) defined "error-free" to be those units correct in all "syntax and function words." For Larsen-Freeman and Strom (1977), however, an error-free
T-unit had to be correct in all aspects, including spelling. ESL compositions at beginning and intermediate levels of proficiency would not contain many error-free T-units by either of these criteria and thus would have to be excluded from this type of analysis. Gaies (1980) has argued this point, stating that learners at low levels of proficiency would have such a high number of errors in their written language that the tabulation of accurate data would be difficult.

ALTERNATIVES TO THE T-UNIT

Crowhurst (1983) advocated that "a measure is needed which will be sensitive to different levels of linguistic maturity in the structures within the T-unit or clause" (13). In an effort to devise more discriminating measures of syntactic complexity, alternatives to the T-unit have been proposed by Botel, Dawkins, and Granowsky (1973), Endicott (1973), and Golub and Kidder (1974).

The syntactic complexity formula devised by Botel, Dawkins, and Granowsky is a complicated scale in which various structures are given different point "counts" based upon their "syntactic difficulty." For example, Subject-Verb sentences, Subject-Verb-Object sentences, Subject-Verb-be-Complement and Subject-Verb-Infinitive are 0-count structures; adjectives, possessive nouns and participles are examples of 1-count structures; passive sentences and
dependent clauses are 2-count structures; clauses used as subjects are 3-count structures. Syntactic complexity of a composition is the average of the complexity of the sentences measured.

The very intricacy of the Botel et al. syntactic complexity formula makes its application tedious and time-consuming. Forty-eight categories and sub-categories of various linguistic structures are listed in the Botel formula, divided between the four "count" levels. In addition, the allocation of structures to 0, 1, 2, or 3-count levels appears to be rather arbitrary in some cases, a factor which the authors themselves acknowledged (Botel et al. 1973: 78). For example, a gerund is listed as a 1-count structure when it is the subject of a sentence (as in "Running is fun"), but in all other contexts, the gerund is rated a regular noun as part of a sentence. When an adverbial begins a sentence (as in "Quickly, he ran to the store"), it is scored as a 2-count structure. An adverbial in any other position in the sentence is counted as a 1 count structure.

In addition, Botel, Dawkins, and Granowsky warned that the scale should not be used on its own, but "in conjunction with a measure of vocabulary" (78), thus adding a further dimension to a measure of already cumbersome proportions. The authors themselves do not offer such a measure. Perkins (1980) applied the Botel formula, along
with six other objective measures of writing quality, to ESL student compositions and did not find it to be a significant discriminator when correlated with holistic ratings of the same compositions.

Endicott (1973) devised a system which proposed "to advance a theoretic model which will allow us to define units of syntactic complexity in psycholinguistic terms" (6). This system was based on a "transformational and morphemic" analysis of language. Endicott defined five units in his syntactic scale:

Co-meme: A unit of complexity in language consisting of four sub categories: The base co-meme, the syntactic co-meme, the compression co-meme and the morphemic co-meme.

Base Co-meme: Those morphemes expressed at a level of language which has a one morpheme per-word ratio.

Syntactic Co-meme: A theoretical syntactic operation by which sentences are combined or altered to achieve efficiency or variation of purpose beyond that achieved at a minimal level of language.

Compression Co-meme: The theoretical morphemic bundle of deep structure which is compressed into surface structure through combination or deletion transformation.

Morphemic Co-meme: Morphemes other than those expressed by base co-memes, i.e., 'The productivity was low.' 'Productivity' represents one base co-meme: 'product,' and two morphemic co-memes 'ive' and 'ity' (7).

In the sentences "The cloud was ominous" and "The cloud was black," there is a "minimal level of language" - one co-meme per word, with a complexity ratio, according to Endicott's scoring system, of 4/4. However, in the sentence "The black cloud was ominous," there is a complexity ratio of 9/5 (9 co-memes per five words) since, in Endicott's words: "The
weighting now represents the full eight basic co-memes of the deep structure plus one syntactic co-meme" (8). In another example, the sentences "The man walked a mile" and "The boy walked a mile" are compounded to give "The man and the boy walked a mile." In this sentence, each of the eight words is a base co-meme, walked a mile represents three compression co-memes, and the conjunction and is a syntactic co-meme. The complexity ratio for the sentence is 12/8.

The Endicott scale is less complicated than the Botel et al. syntactic complexity formula, but is still a rather cumbersome apparatus when the analysis of a large number of essays is concerned. The scale also produces, in some cases, somewhat arbitrary definitions of "complex" language, for example:

1 5 1 1 1 5 1
The brown dog was a good dog. Complexity ratio: 15/7

1 5 1 1 1
The brown dog was good. Complexity ratio: 9/5

The first sentence, given a higher rating than the second in Endicott's system, cannot be said to represent more complex language.

In a review of measures of syntactic complexity, O'Donnell (1976) stated that Endicott's scale had "potential capacity to reflect differences in degree of complexity" (33), but that it was untested and could prove to be unreliable, since raters could be confused as to what was
supposed to constitute a separate morpheme. Flahive and Snow (1980) adapted Endicott's scale to formulate an Index of Complexity. They found it to be the least discriminating of the four objective measures they used to evaluate ESL compositions.

Another purportedly "objective" measure of syntactic complexity, the Syntactic Density Score (SDS), was devised by Golub and Kidder. The SDS takes into account the use of "complex verb phrase expansions, use of some advanced structures of time, and reductions and embeddings that take the form of prepositional phrases" (Golub and Kidder 1974: 1130). The SDS is based on an analysis of ten indices, including words per T-unit, subordinate clauses per T-unit, words per main clause, prepositional phrases, and adverbs of time. Each of these indices is given a different "loading" in the final calculation, according to its degree of significance. Golub and Kidder also drew up a Grade Level Conversion Table, in which SDS scores were supposed to predict proficiency levels: for example, a SDS of 10.9 was assumed to predict a high school grade level of 10.

It is immediately obvious that some of the ten indices may replicate each other. For example, a measure of words per T-unit may, in many instances, produce the same figure as a measure of words per main clause. O'Donnell (1976) argued that Golub and Kidder's ten indices have a "high degree of redundancy in what they measure," and
stated that the weighting of different items was arbitrary. For example, participial phrases carry a .10 greater "loading" than prepositional phrases (37). In addition, the selection of only ten items in order to produce an index of complexity to be used as a predictive tool leaves this method open to the criticism that it fails to address a full scope of linguistic structures. Golub and Kidder's Syntactic Density Score was designed to be calculated on a computer and in this regard can hardly be seen as an alternative to the simple T-unit.

The syntactic complexity scoring systems devised by Botel et al., Endicott, and Golub and Kidder all attempt to provide more accurate analysis of writing quality, but all suffer from the same criticisms of validity which have been leveled at the T-unit: that the selection and weighting of items is arbitrary. In addition, none of these systems has been tested to any great degree, all are somewhat cumbersome and none can be seen to approach the simple and swift application procedures of the T-unit.

CLAUSE ANALYSIS

Hunt (1965) found that clause length, as well as T-unit length, was a powerful index of syntactic "maturity," but wrote: "It seems unlikely that depth of clause subordination will turn out to be as useful as an index of maturity" (27). However, Moffett (1968) proposed that
clause reduction would be a valid indicator of writing quality: "'clause reduction'" refers not only to some sentence transformations but also to a psychological process of language maturation" (173). O'Donnell, Griffin, and Norris (1967) argued for the "computation of the relative frequency of all sentence-combining transformations, including subordinate clauses" (98). Mellon (1969) found that an increase in writing skills came with the increased use of relative transforms, nominalized sentences, and recursive embeddings at increasingly deeper levels.

Bever (1972) defined the clause as "the primary perceptual unit." Bever carried out a number of experiments in which subjects listened to spoken language which was interrupted by a series of simple "clicks." From the results of these experiments, Bever determined that:

1. reaction time to 'clicks' is faster at clause boundaries;
2. 'clicks' are accurately located in the language sample when they occur at clause boundaries; and
3. 'clicks' are not located accurately when they occur at other than clause boundaries.

Bever concluded that "sentence comprehension is the joint result of the segregation of clauses and phrases on the basis of specific external cues" (1972: 105). In Bever's experiments, the final clause heard was the one more clearly remembered, although the surface structure of clauses was forgotten after a few clauses were heard.
Arena (1975b, 1982) concurred with Bever's assertion that the clause was the "primary perceptual unit" and argued for the clause analysis of student writing based upon the tagmemic model of grammar advocated by W.A. Cook (1969):

Since we are analyzing the surface structure of sentences in students' compositions, and since clauses operate at the sentence level, the tagmemic model of grammar is quite appropriate as a basis for providing a description of the student's writing in terms of number, type and frequency of occurrence of the clause types, and the embedding depth employed by the student in his/her compositions (Arena 1975b: 284).

Cook used the term "reduction" to describe the process of breaking up sentences into clauses: "In the reduction process, the corpus is considered to be a body of sentences, made up of simple clauses. These clauses may be dependent or independent" (1969: 43). For example, Cook reduced the compound sentence "She rode and I walked home" to two main clauses, symbolized by the letter "A." The complex sentence "There is the lady who stole my horse" would be analyzed as one main and one subordinate clause, symbolized as "A + B."

Arena (1982) adapted Cook's system of symbols to devise a clause analysis technique, based on reduction and depth of embedding, to analyze syntactic complexity in the writing of U.S. attorneys. The clause analysis technique included three measures: sentence length, which Arena defines as the number of clauses in a sentence; information block length, or the number of clauses per main clause; and embedding depth, or the average "value" of embedded clauses.
These measures will be further described in the Methods and Procedures chapter of this thesis.

Arena proposed that information block length and embedding depth would be found to be valid indices of writing complexity, but that sentence length would not, since it did not entail a measure of subordination.

Arena's clause analysis technique was used in this thesis is to measure the syntactic complexity of ESL student writing, since this measure offers a simple but detailed system for calculating subordination ratios. However, it must be noted that Arena did warn about the possible inadequacies of this technique as a measure of ESL writing complexity:

Embedding at increased depth . . . is one of the structural characteristics which indicates an improvement in the writing skills of American students. However... I feel it prudent to caution those researchers and teachers who may not be aware that for a large number of foreign students studying in America, the process of sentence embedding is almost non-existent as a feature of their native language. To include recursive embedding as a criterion of the writing performance of such students may be somewhat frustrating for both student and teacher (1975b: 290).

Arena's argument is all very well, but the fact is that foreign students are expected, in the ESL classroom, to work towards a standard of American English, regardless of their language background. Indeed, Homburg (1984) found that measures of dependent clauses per composition and the number of dependent clauses per main clause were more
discriminating as measures of syntactic complexity in ESL writing than length of T-unit.

Bever's assertion that the clause is the 'primary perceptual unit' points to a measure of subordination as a valid method of analyzing the complexity of student writing. Arena's clause analysis technique, based on the work of Cook and Bever, was designed to provide the means with which to conduct this analysis. Although Arena was reluctant to recommend clause analysis for ESL research, Homburg's (1984) results have shown that the clause analysis technique may, indeed, be a valid measure of ESL student writing complexity.

LIMITATIONS OF MEASURES OF SYNTACTIC COMPLEXITY

Gaies (1980) wrote that "there is no necessary correlation between syntactic complexity and quality or effectiveness--in other words, communicative competence--in speech or writing" (56). Perkins (1983) criticized objective measures of syntactic complexity on the same count, that they did not quantify "cohesion, coherence, organization, ... idiom, diction, tone, relevance or focus--all factors which contribute to good writing" (662). Hatch (1983) contended:

There is no one-to-one relationship between complexity and comprehension. More complexity can mean greater comprehensibility, though we soon reach a point of diminishing returns if we continue to add more and more information to each utterance (76).
Similar objections to the use of syntactic complexity measures as indices of writing quality have been voiced by Faigley (1980), Witte and Faigley (1981), and Hake and Williams (1985). O'Donnell (1968, 1976) asserted that there was no evidence to suggest that measures of syntactic complexity, such as clause length, T-unit length, and number of clauses per T-unit, as well as complexity scales devised by Endicott (1973) and Golub and Kidder (1974), "measure the structural complexity of an individual student's writing in various situations" (1968: 6).

On the other hand, intensive research, starting with Hunt (1965), has determined that there may be strong correlations between complexity of syntax as measured by various indices and the quality of student essays at various levels of proficiency. Some (Arena 1975b, Gaies 1980) have questioned the validity of applying these measures to the writing of students of English as a Second Language. Others, however, have found various measures of syntactic complexity to show a strong correlation with independent ratings of ESL compositions (Larsen-Freeman 1978, Kameen 1979, Flahive and Snow 1980, Homburg 1984).

THE MEASUREMENT OF ERROR IN ESL WRITING

Gaies (1980) addressed the fundamental problem encountered when using a measure of syntactic complexity, such as the T-unit, for second language research:
The fact that errors have to be accounted for by a modified index of language development in adult second language acquisition has an important implication: it amounts to a concession that there is a qualitative difference in the developmental process of first and second language acquisition (56).

It is this "qualitative difference" which must be taken into account when measuring the writing quality of ESL students. Arthur (1979) found frequency of error to be a discriminating factor in determining ESL writing quality. Kaczmarek (1980) used a count of error-free words in "correct and meaningful sequences" to rate ESL student essays. Others (Larsen-Freeman and Strom 1977, Scott and Tucker 1974) found error-free T-units to be valid indices of ESL writing proficiency.

The measurement of error, however, raises a critical issue, that of "determining what constitutes an error and in deciding which errors should be included in a measure of ESL writing proficiency" (Homburg 1984: 93). Hirsch (1977) also expressed reservations about the scoring of error, citing problems of reliability, "which will vary according to pedagogical circumstances" (1987).

Various research studies have reinforced these reservations. Schachter (1974) found that frequency of error varied according to the nationality of her subjects. Students with native language structures similar to those in English tended to use those structures and commit more errors, while those whose native languages had completely different structures avoided committing errors by avoiding
the use of the English structures which were alien to them. Maimon and Nodine (1978) demonstrated that frequent errors made while attempting complex structures may, in fact, be an indication of writing "maturity."

Such conclusions brought White (1985) to propose that the relationship between error and writing quality in the first language "beyond the lower grades" was a tenuous one (175). But if error analysis is applicable only at low levels of proficiency in the first language, then it may, indeed, be a valid assessment tool in second language acquisition. The error-free T-unit has been shown to be a useful index of ESL writing quality at higher levels of proficiency. However, the number of error-free T-units in poor ESL writing may be too low, or even non-existent, rendering this measure inappropriate for ESL writers of low proficiency.

Much research has focused on the linguistic classification of errors, but Burt (1975) has argued for an analysis from the "reader's point of view" (54). Burt devised a two-tier differentiation scale of "global" and "local" errors, which categorized errors according to their communicative importance (Burt and Kiparsky 1972). "Global" errors, which affect the comprehensibility of an entire sentence or passage, are considered more serious than "local" errors, which affect only the word or phrase in which they are located.
Nas (1975) used a scale which distinguished between first-degree, second-degree, and third-degree errors, the latter corresponding to Burt and Kiparsky's "global" category. Brodkey and Young (1981) also devised a three-point scale, the "Correctness Score," which correlated highly with holistic ratings of ESL student essays and differentiated among different levels of proficiency. In the design of their scoring method, Brodkey and Young took into consideration the following features:

1) it had to be a reliable and valid measure of writing performance for college work; 2) it had to be easy and natural for any English teacher to administer and score; 3) it needed to provide a numerical index rather than an impressionistic comment; 4) it had to be a particularly fine-toothed discriminator among levels of instruction which often deceive the naked eye of the grader; and 5) it should help discriminate between ESL students and native speaker students within our very heterogeneous population of vari-linguals (1981: 160).

In recent research, the Brodkey and Young Correctness Score has been used by Robb, Ross, and Shortreed (1986) as one of eighteen objective measures in a study on the relationship between error feedback and improvement in ESL student writing. The reliability of the Correctness Score has been strongly supported by Brodkey and Young, who have used it for placement purposes at the University of New Mexico (Brodkey and Young 1981: 161).

Since the frequency and seriousness of error is a major concern for evaluators of ESL student writing, a measure which takes errors into account ought to be an
integral component of any analysis of ESL student writing quality. Brodkey and Young's Correctness Score, which records both frequency and seriousness of error in a single figure, was chosen for this study as the index of error in the analysis of ESL student essays.

THE RELIABILITY OF HOLISTIC RATINGS

A principal concern in the use of holistic ratings as an evaluating tool has been their reliability. Stiggins (1982) has addressed the central problems of test reliability: "Test scores must be stable over time, across parallel test forms, across different parts of the same test and across raters" (109). With the holistic scoring of student essays, it is the issue of inter-rater reliability which has received the most attention.

The holistic ratings used for this study conform to the six-point scale set for the Test of Written English (TWE). The TWE guidelines outline a number of procedures in order to ensure reader reliability (Stansfield and Webster 1986). However, a number of studies have questioned the reliability of holistic evaluations of student writing quality (Remondino 1959, Diederich, French, and Carlton 1961, Diederich 1974).

Diederich et al (1961) asked sixty professionals, including college English teachers, social science teachers, natural science teachers, writers and educators, lawyers,
and business executives, to grade three hundred college freshman essays by sorting them into nine score levels. Ninety-four percent of the essays received at least seven different grades, pointing to extreme reliability problems. While the majority of the graders were not English teachers, there was no consistency in the sample graded by the English teachers either.

Coffman, (1966) found that the reliability of holistic ratings increased with the number of readers (up to five) and the number of topics (up to five). In the most commonly used technique of rating holistically, one topic with two readers, Coffman estimated a scoring reliability of only 0.38. Breland and Gaynor (1978) disputed Coffman's results. In a study of the holistic rating of 2,000 freshman compositions, they found that inter-rater reliability for one topic with two raters was 0.51. Breland and Gaynor concluded that inter-rater reliability could be further improved with the application of standard rating guidelines.

Hirsch (1977) noted that reliability could only be achieved when there was consensus among the graders on the qualities of good writing. When this consensus was present, several studies have shown that holistic rating can achieve a high degree of reliability. Consensus can only be attained, however, through group training of raters. White (1986) has outlined some of the problems concerned with this
kind of training: the "calibration" of readers may all too often be subject to the "coercion" of one member of the group; frequent re-training is necessary to prevent reader "drift" up or down; and a sense of community and professional "collegiality" must prevail to ensure genuine responses on the part of those being trained (69-70).

Since holistic ratings are done impressionistically, readers spend only a short time, perhaps two minutes, on an essay written for a thirty-minute test such as the TWE. This leaves the readers open to distraction by some features to the exclusion of others. Research has supported this concern. Freedman (1979) found that raters of freshman compositions were influenced most by content and organization, rather than by sentence structure or mechanics. Charney (1984) has warned that if holistic ratings are to be considered valid, readers must not be unduly influenced by superficial features (78). McColly (1970) recommended that essay papers be typed before rating, in order to reduce undue influence on the reader from the effects of handwriting and neatness.

Carlson and Bridgeman (1986) have noted that cultural differences may play a part in evaluating ESL writing quality. Readers who are not "culturally sensitive" may not make allowances for specific cultural traits, such as digression in the writing of French and Spanish speakers. On the other hand, Carlson and Bridgeman warned that
"culturally sensitive" ratings may not be appropriate for ESL students at U.S. colleges (145).

Jacobs et al. (1981) supported a holistic approach to composition evaluation as long as subjectivity in the ratings was reduced by following a set of seven steps. Similarly, Charney (1984) stated that holistic rating could only be reliable as long as a set of preparation criteria were followed. Charney stated that holistic ratings could be reliable only if the following five conditions were observed:

- if the design of the training and rating sessions takes the factors necessary for reliability into account;
- if the readers are qualified and come from similar backgrounds;
- if the readers are 'calibrated,' that is, trained to conform to agreed upon criteria of judgment;
- if the criteria, which either are supplied to the readers in the form of a ratings guide, or are decided upon by the readers as a group, are appropriate; and
- if readers work quickly, usually under supervision (Charney 1984: 69).

Cooper (1977) found that "Spending no more than two minutes on each paper, raters, guided by ... holistic scoring guides ... can achieve a scoring reliability as high as .90 for individual writers" and confidently asserted that holistic evaluation remained "the most valid and direct means of rank-ordering students by writing ability" (3). Perkins (1983) stressed that "holistic scoring has much to recommend it as a tool for certification, placement, proficiency, and research testing" (653). In research on

Whenever holistic rating is used as a measure of writing quality, it should always be noted that the holistic scoring scheme is subjective and impressionistic by its very nature, and that even the most controlled "calibration" of raters would not prevent certain inconsistencies and disagreements. Despite these shortcomings, this impressionistic form of grading is a frequently used method of evaluating writing. If certain procedures are followed, such as those advocated by Charney (1984), then holistic rating of student essays can be a quick and reliable method of assessment.

SUMMARY

Many researchers in language development have argued that greater syntactic complexity may be indicative of better student writing quality. The T-unit, promoted by Kellogg Hunt (1965), has been supported by some researchers as the most useful measure of syntactic complexity in the writing of native speakers of English. Critics, such as Ney (1966), have argued that use of the T-unit ignores certain
structures, such as conjunctions, which could be seen as indicative of greater complexity.

In second language research, the error-free T-unit has been promoted by some, for example Larsen-Freeman and Strom (1977), as a valid index of ESL writing quality. Alternatives to the T-Unit have been proposed by Botel, Dawkins, and Granowsky (1973), Endicott (1973), and Golub and Kidder (1974). Although some of these have been adapted for ESL research, none have managed to replace the T-unit as the standard measure of syntactic complexity in language development research.

Arena (1982) devised a clause analysis technique to measure syntactic complexity, based on conclusions drawn from the experiments of Bever (1972) and a tagmemic model of clause analysis advocated by Cook (1969). Although Arena (1975b) questioned the applicability of clause analysis to the writing of ESL students, Homburg (1984) found that measures of subordination may indeed by valid indices of ESL writing quality.

While the relationship between syntactic complexity and writing quality should not be seen as a one-to-one correlation (Gaies 1980, Perkins 1983, Hatch 1983), many researchers have found various measures of syntactic complexity to be valid indices of evaluators' perceptions of writing quality.
In addition, evaluators of ESL writing are bound to be influenced in no small measure by the frequency and seriousness of error in an essay, an aspect which is not as crucial to the evaluation of first language writing. Various measures (Burt and Kiparsky 1972, Nas 1975, Brodkey and Young 1981) have been designed to quantify error in ESL writing. Some researchers, however, have noted reliability problems in the scoring of error (Hirsch 1977, Schachter 1974).

The most widely-used method of evaluating student writing is a holistic rating scale. While some (Cooper 1977) have promoted the use of holistic scoring as a reliable and valid measure of evaluating writing, others (Diederich, French, and Carlton 1961, Coffman 1966, Charney 1984) have demonstrated and voiced doubts about its potential unreliability. However, holistic rating remains the quickest and most efficient method of evaluating student writing and has been recommended as a valuable research tool (Perkins 1983).
CHAPTER III

METHODS AND PROCEDURES

The essays of thirty college-level ESL students who wrote a practice TWE (Test of Written English) examination were rated holistically and analyzed with regard to syntactic complexity and incidence of error.

The syntactic complexity of the student essays was tabulated according to the clause analysis technique designed by Louis Arena (1982). Three measures were used as indices of syntactic complexity: sentence length, information block length, and embedding depth. The Correctness Score, devised by Brodkey and Young (1981), was adapted and used in this analysis to rate the occurrence of error in the student essays.

The data obtained from the three indices of syntactic complexity and the adapted version of the Correctness Score were correlated with the holistic ratings of the essays in order to determine whether there existed a significant relationship between syntactic complexity and/or frequency and seriousness of error, and the evaluators' perceptions of student writing quality as determined by the holistic ratings.
SUBJECTS

The subjects of this study were thirty college-level ESL students who wrote a practice TWE examination in May, 1987. The thirty student essays used in this analysis were randomly selected from an original body of fifty written for the examination. The students who wrote the test came from three proficiency levels, the levels being identified with TOEFL (Test of English as a Foreign Language) scores of below 460, 460-490, and 490+ respectively. Sixty seven percent of the students whose essays were analyzed in this study came from the highest level of proficiency (TOEFL score of 490+).

INSTRUMENTATION

The TWE (Test of Written English) was used to provide essay samples for this analysis. The TWE was initiated in July, 1986, following proposals to include an integrative-type writing test as part of the TOEFL examination. Standards for selection of topic, scoring, and analysis of TWE results have been described by Stansfield and Webster (1986). The topic for the particular essay used in this study was given as follows:

Some people believe that the best way of learning about life is by listening to the advice of family and friends. Other people believe that the best way of learning about life is through personal experience. Compare the advantages of these two different ways of learning about life. Which do you think is
preferable? Use specific examples to support your preference.

TESTING CONDITIONS

The fifty students who originally wrote the test were seated in a large room and instructed to write on the topic for a maximum of thirty minutes. All the students had had some previous classroom experience in the writing of TWE-type exercises, although the actual topic of this particular test was not revealed until the time of the examination. No particular audience was identified in the instructions given for the test. Students were advised to make notes and organize their essays before beginning to write.

SCORING

The thirty essays used in this analysis were rated holistically by two experienced ESL teachers according to the six-point scale provided for the TWE. The scoring guidelines give the following broad definitions to identify each of the six levels of writing competence:

6  Clearly demonstrates competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.
5  Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will have occasional errors.
4  Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.
3  Demonstrates some developing competence in writing, but it remains flawed on either
the rhetorical or syntactic level, or both.
2 Suggests incompetence in writing.
1 Demonstrates incompetence in writing.

Detailed criteria describing each of the six levels are presented in the TWE scoring guide (see Appendix A).

READER RELIABILITY

The essay corpus used in this study was rated by two experienced ESL teachers. In order to maintain their reliability in the rating of TWE-type tests, the raters had recently undergone at least three hours of instruction and calibration in the rating of the TWE. This instruction involved the practice rating of sample essays and comparison with the grades given by other teachers. The rating of the essay sample was carried out with reference to the TWE scoring guidelines.

A reliability study (see the Results and Discussion section for further discussion) gave a reliability coefficient of .90 for the two sets of ratings used in this study.

ANALYSIS

In order to facilitate analysis, each essay was typed using a word processor and several copies were produced. The typed versions were faithful to the originals in all respects, except for handwriting and neatness. Errors,
deletions, insertions, and misspellings were carefully reproduced using a system of symbols (see Appendix B).

At the time of the analysis, student names and proficiency levels, as well as holistic ratings for the essays, were not known, or marked on the essay copies. Each essay was identified by a code number in the top right-hand corner on the first page of the typed version.

Clause Analysis Technique

The clause analysis technique designed by Louis Arena involves three measures: sentence length, information block length, and embedding depth. In his book, *Linguistics and Composition* (1975a), Arena defined the clause as "a potential group of words which has one and one only verbal in the string and which typically operates on the sentence level" (48). In this study, "verbals" were taken to mean finite and non-finite verbs, infinitives, and active and passive participles.

In accordance with the examples given by Arena in his article "The Language of Corporate Attorneys" (1982), exemplification, denoted by the markers "such as" and "for example," was identified as a reduction from a relative clause. Comparative structures after "than" and "as" also signal subordination, as in "John is bigger than Joe," although the second verbal, "is," is missing in the written sentence.
The first measure of the clause analysis technique was sentence length, which was computed by averaging the number of clauses per sentence. The second measure, information block length, was tabulated by calculating the average number of clauses per main clause. The third measure, embedding depth, was computed by assigning a "value" to each clause in an "information block" or main clause; the main clause received a value of 1, the first subordinate clause a value of 2, the second subordinate clause a value of 3, and so on. The total of the embedding values for each essay was then divided by the number of clauses in the essay to arrive at a figure for "embedding depth."

The following symbols, after Arena (1982), were used in the analysis:

A = a main clause
B = a subordinate clause embedded into A
C = a subordinate clause embedded into B
D = a subordinate clause embedded into C, etc.

# = a boundary sign for the beginning and end of a sentence
+ = a clause block initial boundary inside of a sentence

In addition to Arena's symbols, the following marker was devised for use in this analysis:

Θ = a clause block initial boundary inside of a sentence, erroneously punctuated by the ESL writer, but clearly a separate sentence: a "run-on."

An analysis of the following example of writing taken
from *Newsweek* magazine illustrates the application of the clause analysis technique:

The high priest of the new look is Gigli, whose austere style has earned him near cult status since his first collection two years ago. His avant-garde clothes proved a best-seller for American department stores and sales of the Gigli line jumped more than 50 percent to $12 million last year. Gigli's success on the fringes of Italy's mainstream fashion establishment, populated by such master tailors as Giorgio Armani, Gianfranco Ferre and Gianni Versace, has opened the way for young designers who share his spare approach.

---

**Sentence Patterns (3 sentences, 8 clauses)**

1. A B C
2. A + A
3. A B C

Sentence length = clauses/sentences
= 8/3
= 2.67 clauses per sentence

**Information Blocks (4 blocks)**

1. A B C
2. A
3. + A
4. A B C

Block length = clauses/main clauses
= 8/4
+ 2.0 clauses per information block
III. Embedding Depth

4 "A" clauses = 4 x 1 = 4
2 "B" clauses = 2 x 2 = 4
2 "C" clauses = 2 x 3 = 6

Clause depth = total "value" of clauses/number of clauses
= 14/8
= 1.75 average value of embedding

The figures of 2.67 for sentence length, 2.0 for information block length, and 1.75 for embedding depth are the syntactic complexity data for the passage.

Correctness Score

A means for computing the incidence of error in the essay sample was adapted from the guidelines established by Brodkey and Young (1981) for the Correctness Score. The Correctness Score is derived from a three-point scoring system, described by Brodkey and Young as follows:

A score of 3 is a severe distortion of readability or flow of ideas which throws the reader off the sense of the message through the intrusion of an erroneous linguistic element; 2 is a moderate distortion; 1 is a minor error that does not affect readability in any significant way (Brodkey and Young 1981: 160).

The procedure devised by Brodkey and Young for calculating the Correctness Score was: 1) the first 250 words of an essay were counted; 2) each error was assigned a score according to the three-point scale; 3) the Correctness Score was computed by dividing the number of words (250) by the total "score" for error.
For the purposes of this analysis, two modifications were made to Brodkey and Young's original procedure. First, since many of the essays in the sample were considerably shorter than 250 words, only the first 120 words of each essay were analyzed for error; such a restriction is not altogether desirable, since longer passages would provide better sampling, but scores produced would approximate those of Brodkey and Young's analysis. Secondly, Brodkey and Young's description of their scoring system is rather brief, being only the paragraph quoted above. Thus, a more detailed definition of the criteria was considered for the purpose of maintaining reliability. The descriptions given by Nas (1975) for his three-level error tabulation system appear to be consistent with Brodkey and Young's criteria, and were therefore used as the guidelines for the evaluation of error in this analysis:

First-Degree Errors

Spelling: deviation from correct spelling is minor, a reader has no trouble recognizing the word. Lexical: deviation from meaning is so minor that a reader has no trouble substituting correct word. Grammatical: 1) occurring in a form that is an exception to a grammatical rule; 2) form or structure would be correct in partly different context, no problem in understanding; 3) form used is correct only in intermediate context; 4) error can be explained as the use of the wrong register.

Second-Degree Errors

Spelling: serious deviation from correct spelling; word interpretable in context. Lexical: so serious that item is only interpretable with the help of context.
Grammatical: 1) results in alien word combination or word order, but sentence is still interpretable; 2) would be fatal to communication, except that rest of sentence is interpretable even without wrong words; 3) results in a form that can only be interpreted in context.

Third-Degree Errors

Spelling: makes it impossible to be certain about the word that is meant.
Lexical: makes it impossible to be certain about the meaning, except with the help of context.
Grammatical: makes it impossible to be certain about the meaning of the sentence, even with the help of context (Nas 1975: 16-21).

In this analysis, first-degree, second-degree, and third-degree errors are considered synonymous with the scoring levels of 1, 2, and 3 in Brodkey and Young's system.

In an essay with an error total of 36 for the first 120 words, the Correctness Score would be 120/36, or 3.3. The lower the Correctness Score, the higher the incidence of error.

LIMITATIONS IN APPLYING THE ANALYSIS TO ESL WRITING

Clause Analysis Technique

Analysis of the essay sample demonstrated some significant practical limitations of the clause analysis technique when applied to ESL student writing.

The clause analysis technique requires the measurement of two structures: the sentence and the clause. A few practical problems were met when measuring sentences. Some passages contained no punctuation between different
"sentences." In these cases, the symbol "©" was used to denote a "run-on," although run-ons were considered part of the same sentence for the purpose of the analysis. For example, the sentence:

In every matter and problem about life I will depend on my family I will wait for their advice.

is really two separate sentences.

In some other cases, there was an over-use of periods; if the first word after a period began with a lower-case letter, and if the "sentence" which this word initiated was, in conventional English usage, a part of the previous sentence, then it was noted as a subordinate clause. In the case:

so He has no any experience about how to resister to school. how to choose clase.

the second "sentence" was counted as subordinate to the first.

It is clear that the clause analysis technique may have to undergo further refinement if it is to be applied to ESL writing. However, difficulties in defining sentence boundaries in the essays of the sample were few and far between, and problems were only to be encountered in essays with below-average ratings.

Problems in defining clause boundaries were, however, more frequent. At the clause level, frequent and serious errors in syntax and punctuation required subjective
judgments about the writer's intent. Missing verbals were common, as in:

I think that not true always.

This sentence was analyzed as having two clauses.

In other cases, a grammatical error would be overlooked and a structure was scored in terms of the "obligatory context." In the following sentence:

so He could not go to scool for studing

the structure "stud[y]ing" ought to be "in order to study," and was counted as a subordinate.

In general, even though many essays contained convoluted and erroneous syntax, an analysis could be effected based on Arena's technique. For example, the following passage:

for instece Ensetan who was outstanding scintist in the world when He sas child his family was very poor

was analyzed as:

A for instece Ensetan (C,B) his family was very poor
B when He sas child
C who was outstanding scintist in the world

This sentence could be reorganized in several ways, but it is clear that there are two subordinate structures. Such writing was an exception, however, and the majority of essays presented few problems for accurate analysis.

**Correctness Score**

While there were no problems in identifying individual errors, subjective judgments were required, with
the aid of Nas' scoring guide, in allotting errors a "score" according to Brodkey and Young's three-point scoring scheme. In cases where the same error was repeated, such as a misspelling, the error was usually rescored at a lower value; for example, an error scored as a 3 may be scored as a 2 when it is repeated.

The Correctness Score system in general is vulnerable to inconsistencies in judgment, and this is readily acknowledged by its developers:

A given error in a very poor paper laced with numerous substantial errors is usually graded more lightly than the same error appearing in an almost perfect paper. In the same way, no generic category of errors receives the identical rating throughout a given paper. A spelling error, morphology error, error of verb tense, of preposition, etc. may receive one or even two points difference from the next error of the identical linguistic category. Thus we continue to stress individualism and relativity in error rating, and we eschew the coding of error types into fixed ratings (Brodkey and Young 1981: 165).

RELIABILITY OF THE ANALYSIS

Statistics for the three measures of the clause analysis technique and the Correctness Score were derived from totals for the number of sentences, number of clauses, number of main clauses, embedding "value," and error score in each essay.

Fifteen essays, representing 50 percent of the sample, were randomly selected and photocopied. The essays
were then analyzed in order to determine inter-rater and intra-rater reliability for the analysis.

An experienced ESL teacher analyzed the essays for the inter-rater reliability check. The rater underwent two practice sessions in tabulation of the clause analysis and Correctness Score. Both the inter-rater and intra-rater reliability checks were done two months after the original analysis.

For the purposes of calculating reliability, the following margins of difference from the original analysis figures were permitted: 1 for the total number of sentences in each essay; 3 for clauses and main clauses; 5 for embedding "value" and error score. The rationale for these margins was based on the size of the scores involved: a larger margin was allowed for error score because the size of most error scores left greater margin for error. Reliability was calculated by adding the number of scores for each category which fell within the margin of difference and dividing by 15.

Table I presents the reliability coefficients for the inter-rater and intra-rater reliability checks on the five basic measurements used in the clause analysis technique and Correctness Score.
TABLE I
RELIABILITY COEFFICIENTS FOR CLAUSE ANALYSIS TECHNIQUE AND CORRECTNESS SCORE

<table>
<thead>
<tr>
<th>Item</th>
<th>Intra-rater Reliability</th>
<th>Inter-rater Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sentences</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Number of clauses</td>
<td>1.0</td>
<td>0.93</td>
</tr>
<tr>
<td>Number of main clauses</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Embedding value</td>
<td>0.87</td>
<td>0.67</td>
</tr>
<tr>
<td>Error score</td>
<td>0.80</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Reliability statistics for the number of sentences, number of clauses, and number of main clauses were satisfactory. The lower coefficient for embedding value indicates that raters may have had some questions about the actual depth of embedding in the essays. This confusion may have resulted from the fact that, in ESL writing, erroneous structures may give rise to different interpretations of their relationships within the sentence.

Reliability coefficients for the error score were disappointingly low. The descriptions offered by Brodkey and Young (1981) and Nas (1975) may not have been detailed enough to produce the desirable consistency in different ratings. Indeed, the Correctness Score may require intensive "calibration" sessions before it is applied for assessment purposes. Brodkey and Young's emphasis on
"individualism and relativity in error rating" (1981: 165) does not appear, in this analysis, to have produced consistent results in the cases of some essays.

STATISTICAL PROCEDURES

Spearman's rank order correlation was used to compute the correlations between the three measures of the clause analysis technique, the Correctness Score, and the holistic ratings of the thirty essays in the sample.

Spearman's rank order correlation is a statistical index considered useful for measuring ordinal data and sets numbering thirty items or less (Crocker 1969: 58).

SUMMARY

The research corpus for this study consisted of thirty ESL student essays written for the Test of Written English (TWE). The essays were rated by two experienced ESL teachers trained in the procedures involved in rating TWE essays. A grade was given to each essay according to the TWE Scoring Guidelines.

Each essay was typed and a blind analysis was conducted using the criteria established by Arena's clause analysis technique and Brodkey and Young's Correctness Score. The clause analysis technique involves three measures of subordination: sentence length, information block length, and embedding depth; the Correctness Score
measures frequency and seriousness of error. Some limitations were noted when applying these measures to ESL student writing.

A reliability check showed satisfactory dependability for the three measures of the clause analysis technique. However, reliability for the Correctness Score was not altogether encouraging, suggesting that the measure lacks the objective characteristics to ensure confidence in detailed analysis. Correlations between the three measures of the clause analysis technique, the Correctness Score, and the holistic ratings were computed using Spearman's rank order correlations.
CHAPTER IV

RESULTS AND DISCUSSION

RESULTS

Thirty ESL student essays were rated holistically by two experienced ESL teachers. The essays were then analyzed to determine their syntactic complexity using the three measures of a clause analysis technique (sentence length, information block length, and embedding depth). In addition, the essay corpus was analyzed to determine frequency and seriousness of error, using the Correctness Score as an index of measurement. The data resulting from the analysis are presented in Table II (page 54).

Spearman's rank order correlation was used to compute the relationship between sentence length, information block length, embedding depth, frequency and seriousness of error, and the holistic ratings of the essays.

Reliability of the Holistic Ratings

The thirty student essays constituting the research corpus were rated holistically by two experienced ESL teachers, according to the six-point grading scale established by the TWE (Test of Written English).
For the purpose of calculating reliability, a one-point difference was permitted between each pair of ratings. This margin of difference is compatible with the TWE guidelines (Stansfield and Webster 1986).
Three pairs of ratings showed a greater than one-point difference. Twenty-seven pairs of ratings came within the margin permitted. This gave a reliability coefficient of .90 for the holistic ratings. Independent ratings were sought for the three essays which elicited disagreement among the raters. The remaining pairs of scores were averaged to produce the ratings in Table II. Figure I presents the distribution of essay scores, showing symmetrical alignment around the mean of 3.5.

Figure 1. Distribution of holistic ratings in the essay corpus.
Two important qualifications must be noted when reviewing the analysis with regard to the holistic ratings.

Firstly, the reliability coefficient of .90 for the ratings does not reflect the degree of disagreement among the raters. Although 27 out of 30 pairs of ratings showed no greater than a one-point difference, only eight pairs of scores were identical. On a six-point scale, this type of disagreement leaves substantial doubts about reliability. A further check on reliability was carried out, comparing the averaged ratings of the two raters used in this analysis with the scores given for the essays after the students had completed the TWE examination. The reliability coefficient for this comparison, allowing a one-point difference, was .87, but only seven pairs of ratings were identical. Only a greater number of identical ratings could give the holistic scoring scheme unqualified approval as a reliable instrument of measurement.

Secondly, 17 of the 30 essays were given ratings of 3.0 or 3.5, this high concentration around the mean accounting for the leptokurtic shape of the curve in figure 1. Such a concentrated distribution has some effect on the results of the analysis, particularly in applying the Spearman's rank order correlation, since such a high proportion of the sample achieved identical rankings according to the holistic scale. The holistic scale, based as it is on a six-point rating scheme, does not possess the
fine discriminatory characteristics which may be necessary for accurate analysis.

Examples of the Analysis

The analysis was conducted on typed copies of the student essays. Working copies of the analysis are shown in Appendix D. Detailed examples of the application of the analysis in two essays, rated 4.5 and 3.0 respectively, are given below, beginning with the clause analysis technique and followed by the Correctness Score.

Clause Analysis Technique:

A "4.5" essay:

I think one of advantages of learning about life through the advice of family and friends is that it's easy to get lots of information we don't know. We can learn lot of things that we have never experienced or will never experience. We may learn the life we are not able to experience any more from grandparents or anyone else like them, while personal experience is limited.

On the other hand, I would like to say learning about life through personal experience is more practical, realistic and clearer. As old proverb says "Seeing is believing" We can make realize what's taking place completely and deeply. If we learn the thing from our experience, we will remember it strongly.

As for me, I prefer to learning about life through personal experience. Whenever I hear an interesting story from my family or friends I really want to experience the same thing. It's better to know something just through my experience, even though I sometimes have to take a risk.

For example, I heard a lot about American life from some friends who had lived in the U.S. before I came here. So I thought I already knew a lot about American way of thinking, living and so on. However sometimes I face the reality that is totally different from my friends' advices or stories, although their advices have been a big help.
I think one of advantages of learning about life through the advice of family and friends is that it's easy to get lots of information we don't know. We can learn lots of things that we have never experienced or will never experience. We may learn the life we are not able to experience any more from grandparents or anyone else like them, while personal experience is limited. On the other hand, I would like to say learning about life through personal experience is more practical, realistic and clearer. As old proverb says "Seeing is believing" we can make realize what's taking place completely and deeply [20] we will remember it strongly. If we learn the thing from our experience, As for me, I prefer to learning about life through personal experience [25] I really want to experience the same thing. Whenever I hear an interesting story from my family or friends it's better to know something just through my experience even though I sometimes have to take a risk. For example, I heard a lot about American life from some friends who had lived in the U.S. before I came here. So I thought I already knew a lot about American way of thinking, living and so on. However sometimes I face the reality that is totally different from my friend's advices or stories. although their advices have been a big help.

I. Sentence Patterns (13 sentences, 36 clauses)

1. A B C D E
2. A B + B
3. A B C
II. Information Blocks (14 blocks)

1. A B C D E
2. A B + B
3. A B C
4. A B C
5. A B
6. @ A B
7. A B
8. A B
9. A B C
10. A B C
11. A B C
12. A B
13. A B
14. A

Sentence length = clauses/sentences
= 36/13
= 2.77 clauses per sentence

II. Information Blocks (14 blocks)

1. A B C D E
2. A B + B
3. A B C
4. A B C
5. A B
6. @ A B
7. A B
8. A B
9. A B C
10. A B C
11. A B C
12. A B
13. A B
14. A

Block length = clauses/main clauses
= 36/14
= 2.57 clauses per information block

III. Embedding Depth

14 "A" clauses = 14 x 1 = 14
14 "B" clauses = 14 x 2 = 28
6 "C" clauses = 6 x 3 = 18
1 "D" clause = 1 x 4 = 4
1 "E" clause = 1 x 5 = 5
TOTAL = 69

Clause depth = total "value" of clauses/number of clauses
= 69/36
= 1.75 average value of embedding
A "3.0" essay:

Learning about life is to get skill to make personality and to make money. Therefore learning about life is how to survive through my life. We listen to the advice of family and friends or we get it through personal experience. These category depend on what I want to learn. I believe both category but on whole my life I think personal experience is better than advising of the other person.

We usually learn education, job or how to do from the other person. Then it must be good way and I can get informations which aren't know before. We should listen the advice of the other person and we should judge by myself if information fit to me.

Also, learning about life through personal experience have good advantage. Because we are not able to buy personal experience. Therefore my experience is so different from the other person and it is the most important to make personality. It must be difficult to get perfect. But we are able to get it by effort.

I will choose learning life through personal experience. That is my personality and I am able to proud it the other person. If I don't have personal experience, it will seem to be robot and not human.

# A 1. Learning about life is
   B 2. to get skill
   + C 3. to make personality
      C 4. and to make money.
# A 5. Therefore learning about life is how
   B 6. to survive through my life.
# A 7. We listen to the advice of family and friends
   + A 8. or we get it through personal experience.
# A 9. These category depend on what
   B 10. I want
      C 11. to learn.
# A 12. I believe both category
   + A 13. but on whole my life I think
      B 14. personal experience is better
      C 15. than advising of the other person [is].
# A 16. We usually learn education, job [17] from the other person.
   B 17. or how to do
# A 18. Then it must be good way
   + A 19. and I can get informations
      B 20. which aren't know before.
# A 21. We should listen the advice of the other person
   + A 22. and we should judge by myself
if information fit to me.

Also, learning about life through personal experience have good advantage.

Because we are not able to buy personal experience.

Therefore my experience is so different from the other person and it is the most important to make personality.

It must be difficult to get perfect.

But we are able to get it by effort.

I will choose learning life through personal experience.

That is my personality and I am able to proud it the other person.

[40] it will seem to be robot and not human.

If I don't have personal experience,

I. Sentence Patterns (16 sentences, 40 clauses)

1. A B C + C
2. A B
3. A + A
4. A B C
5. A + A B C
6. A B
7. A + A B
8. A + A B
9. A
10. A B
11. A + A B
12. A B
13. A B
14. A
15. A + A B
16. A B C

Sentence length = clauses/sentences
= 40/16
= 2.50 clauses per sentence

II. Information Blocks (22 blocks)

1. A B C + C
2. A B
3. A
4. + A
5. A B C
III. Embedding Depth

22 "A" clauses = 22 x 1 = 22
13 "B" clauses = 13 x 2 = 26
 5 "C" clauses = 5 x 3 = 15
TOTAL = 63

Clause depth = total "value" of clauses/number of clauses
= 63/40
= 1.58 average value of embedding

Correctness Score:
A "4.5" essay (first 120 words):

I think one of the advantages of learning about life through the advice of family and friends is that it's easy to get lots of information we don't know. We can learn a lot of things that we have never experienced or will never experience. We may learn the life we are not able to experience any more from grandparents or
anyone else like them, while personal experience is limited.

On the other hand, I would like to say learning about life through personal experience is more practical, realistic and clearer. As old proverb says "Seeing is believing". We can make realize what's taking place completely and deeply. If we learn the thing from our experience, we will remember it strongly.

Correctness Score = number of words/error score
= 120/20
= 6.0

A "3.0" essay (first 120 words):

Learning about life is to get skill to make personality and to make money. Therefore learning about life is how to survive through my life. We listen to the advice of family and friends or we get it through personal experience. These category depend on what I want to learn. I believe both category but on whole my life I think personal experience is better than advising of the other person.

We usually learn education, job or how to do from the other person. Then it must be good way and I can get informations which aren't know before. We should listen the advice of the other person and we should judge by myself if information fit to me.
Also, learning about life through personal experience have good advantage. Because. . .

Correctness Score = number of words/error score = \( \frac{120}{49} \) = 2.4

The analysis was rechecked before final figures for sentence length, information block length, embedding depth, and the Correctness Score were established for each of the thirty essays. However, it must be noted that reliability data for the Correctness Score was not satisfactory (see Chapter III, page 36) and its objectivity is open to question.

The Correlation between Syntactic Complexity (as measured by the Clause Analysis Technique) and the Holistic Ratings

The data presented in Table III show the correlations established between the three measures of the clause analysis technique (sentence length, information block length, and embedding depth) and the holistic ratings of the thirty ESL student essays.

### TABLE III

SPEARMAN'S RANK ORDER CORRELATION (rho) OF THE CLAUSE ANALYSIS TECHNIQUE AND HOLISTIC RATINGS

<table>
<thead>
<tr>
<th>Index</th>
<th>rho</th>
<th>significance (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence length</td>
<td>-.158</td>
<td>ns</td>
</tr>
<tr>
<td>Information block length</td>
<td>+.252</td>
<td>ns</td>
</tr>
<tr>
<td>Embedding depth</td>
<td>+.296</td>
<td>ns</td>
</tr>
</tbody>
</table>
The first hypothesis to be addressed in this study, that sentence length, or number of clauses per sentence, was not indicative of ESL writing quality as determined by the holistic ratings, was supported by the results. The correlation between sentence length and the holistic ratings of -.158 was not statistically significant. This result indicates that the relationship between length of sentences in the student essays and the evaluators' perceptions of their writing quality was negligible. Indeed, many of the "long" sentences in the sample resulted from punctuation errors or lack of sophistication in style. Figure 2 gives the distribution of sentence length statistics, showing a generally even spread, with a few essays containing a remarkably high average number of clauses per sentence.

The second and third hypotheses, that information block length and embedding depth would be found to be indicative of ESL writing quality as determined by the holistic ratings, were not supported by the results. Correlations of .252 and .296, for information block length and embedding depth respectively, were not statistically significant.

An analysis of the figures for embedding depth in Table II (page 54) reveals that the highest score was 2.27 and the lowest was 1.41. Deviation from the mean was 0.54 for the highest score and only 0.32 for the lowest. Thus,
Figure 2. Distribution of scores for sentence length in the essay sample.

The majority of the essays fell within the lower range of complexity, and the sample as a whole fell within the same general range: a difference of less than 1.0 between the highest and lowest embedding values. A similar lack of spread, signified by the leptokurtic curves in figures 3 and 4, was evident for information block length.
Thus, the two measures of subordination do not appear to be valid indices of writing quality as determined by the holistic ratings of the essay sample.

Lack of complexity may have been a factor, however, in the rating of the five poorest essays, those with holistic scores of 2.0 – 2.5 (see Table II). All five of these essays ranked below the mean in embedding depth, perhaps reflecting a general lack of writing ability. Of the five best essays, with holistic ratings of 4.5 and

---

Figure 3. Distribution of scores for information block length in the essay sample.
Figure 4. Distribution of scores for embedding depth in the essay sample

above, only three ranked above the mean in embedding depth, indicating that raters were not necessarily concerned with greater syntactic complexity as a mark of better writing quality. The fact that most of the essays exhibited the same general degree of complexity in subordination, as shown by figures 3 and 4, supports this contention.
The Correlation and Seriousness of Error (as measured by the Correctness Score) and the Holistic Ratings

The data presented below in Table IV shows the correlation established between frequency and seriousness of error, as measured by the Correctness Score, and the holistic ratings.

TABLE IV

SPEARMAN’S RANK ORDER CORRELATION (rho) OF CORRECTNESS SCORE AND HOLISTIC RATINGS

<table>
<thead>
<tr>
<th>Index</th>
<th>rho</th>
<th>significance (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctness Score</td>
<td>+.644</td>
<td>.01</td>
</tr>
</tbody>
</table>

The fourth hypothesis, that frequency and seriousness of error would not be found to be indicative of ESL writing quality as determined by holistic ratings, could not be supported by the results. The correlation of +.644 between the Correctness Score and the holistic ratings indicates a "substantial relationship," according to the guidelines set by Guilford (1956: 145). The high significance level of .01 supports this conclusion.

The results show that evaluators' perceptions of ESL writing quality may be influenced rather strongly by the frequency and seriousness of error. Figure 5 shows that there was a wide range of Correctness Scores for the essay sample, indicating that error, rather than complexity of syntax, was a more viable measure for holistic
discrimination among the essays rated. The reliability coefficients for the Correctness Score are not entirely encouraging (see Reliability of the Analysis section in Chapter III), but the results indicate that a measure of
frequency and seriousness of error may be a useful gross discriminator in the evaluation of ESL student essays.

The data in Table II (page 54) show that four of the five most highly rated essays, those with holistic scores of 4.5 or above, received high Correctness Scores (6.0 +). The five essays with the lowest ratings (2.5 and below) received Correctness Scores below the mean (4.42 and below). The eight essays with the highest Correctness Scores (4.5 and above) received holistic ratings above the mean. The seven essays with the lowest Correctness Scores (less than 2.9) received holistic ratings below the mean.

The Correctness Score, then, exhibited a general ability to discriminate between essays of good and poor quality, as determined by the evaluators' ratings. However, the usefulness of the Correctness Score as a "fine-toothed discriminator" was not supported by the analysis, as indicated by the significantly high correlation of +.644 with the holistic ratings. This conclusion must be qualified by questions concerning the reliability of both the holistic ratings and the Correctness Score analysis.

The Correlation between Embedding Depth, Frequency and Seriousness of Error, and the Holistic Ratings

The fifth hypothesis, that a positive correlation would be established between a combination of results for embedding depth and the Correctness Score, and the holistic ratings, could not be supported by the results.
The multiple correlation formula used to determine correlations between more than two sets of figures requires that all predictors, in this case embedding depth and the Correctness Score, exhibit a high correlation with the criterion, here the holistic ratings (Crocker 1969: 84). The results of this analysis give only the Correctness Score a significant predictive quality. A multiple correlation which included the index of embedding depth would, therefore, be considered invalid.

DISCUSSION

Clause Analysis

Arena's (1982) assertion that sentence length was not "a true measure of style complexity" (147) was supported by the analysis. A high number of clauses per sentences was not necessarily indicative of greater complexity, since the measure of sentence length did not show any close relationship with the two measures of subordination, information block length and embedding depth. As Arena states, sentence length "neutralizes the difference between embedding and conjoining" and "does not permit an objective observation of the facility and difficulty in processing simple and complex sentences, respectively" (150).

The two measures of subordination in the clause analysis technique, information block length and embedding depth, appear to possess some potential as reliable indices
of ESL writing complexity, in the light of the reliability coefficients reported in Chapter III (page 36). Reliability for information block length, representing "quantity" of subordination, was very high, belying Gaies' (1980) warning that ESL writing was too error-prone for accurate syntactic analysis. Reliability for embedding depth, representing "quality" of embedding by assigning each subordinate clause a "value," was reasonably high, suggesting that further refinement of this measure may be required in order to ensure accurate analysis of ESL writing.

The focus of this study was, however, to attempt to provide support for the theory that increased complexity of syntax reflects evaluators' perceptions of better writing quality. The hypotheses that two measures of syntactic complexity, information block length and embedding depth, would be indicative of better writing quality, as determined by holistic rating of ESL student essays, could not be supported by the results.

Studies by Hunt (1965) and his followers have shown that American composition teachers regard "well-written" compositions to be those which demonstrate "the ability to 'say more,' on the average, with every statement" (Arena 1975b: 282). The results, in this analysis, for information block length and embedding depth, indicate that complexity of syntax, as portrayed by these two indices, has little relationship with evaluators' perceptions of college-level
ESL student writing quality. These findings do not support Homburg's (1984) contention that a measure of dependent clauses would be a powerful discriminator among different levels of ESL writing proficiency.

The results of this study may, however, have been influenced by the nature of the data. First, there remain questions about the reliability of the holistic ratings; secondly, the narrow range of scores for information block length and embedding depth may reflect the fact that two-thirds of the essays were written by students from one proficiency level, that identified with TOEFL scores of 490 and above. Only a study which included numbers of students from a variety of proficiency levels would be able to give some indications of the relationship between complexity in writing and language proficiency level. Since the student sample exhibited only a narrow range of complexity scores, in other words there was little differentiation among complexity levels in the writing sample as a whole, the evaluators could have taken into account other, more obvious, characteristics to differentiate among the essays.

The TWE guidelines stress variety of syntax, not complexity, as a factor in rating the essays. Essays which exhibited variety, with an equal distribution of simple, compound, and complex sentences, would not have received especially high scores for information block length or embedding depth. Indeed, some students who attempted
greater complexity may have forced themselves into making more errors. This may be reason why some essays in the sample which ranked high in embedding depth showed low Correctness Scores.

Research by Schachter (1974) has supported the contention that ESL students attempting greater complexity in their writing may be more error-prone. Schachter found that students whose first language contained some syntactic structures similar to those in English may attempt to transfer these structures to their English writing and commit more errors in the process. For example, the structure of the relative clause is similar in both English and Arabic, the principal difference being that the Arabic relative clause often carries an object pronoun after the verb. Arab students may, therefore, write in English:

There is the man who I saw him yesterday.

Students without such similar structures in their first language, for example Chinese and Japanese, may not attempt such levels of complexity, keeping their prose relatively error-free.

Evaluators' perceptions of ESL writing quality may, therefore, be related more to errors in syntax than to complexity of syntax per se. Indeed, Arena (1975b) cautions against the use of clause analysis as an index of ESL writing quality, citing the transfer of "rhetorical modes" from the first language to English: "many American
university teachers of freshman English may seriously misjudge the writing abilities of students whose native languages do not contain the embedding process " (290). Moreover, Arena emphasizes correctness and comprehensibility in ESL writing over any sophisticated display of complexity in syntax. The results of this study would appear to support this view.

The Factor of Error

Inter-rater and intra-rater reliability coefficients for the Correctness Score show only a moderate correlation, indicating that the three-point scoring system for frequency and seriousness of error may require intensive calibration sessions before different raters can use it as an accurate index for error tabulation.

In spite of this inconsistency, the Correctness Score showed a surprisingly high correlation with the holistic ratings. Several researchers (Scott and Tucker 1974, Larsen-Freeman and Strom 1977, Larsen-Freeman 1978, Gaies 1980, Perkins 1980, Brodkey and Young 1981, Homburg 1984) have emphasized the importance of tabulating error in the objective assessment of writing. The results of this study, showing a high correlation between a measure of frequency and seriousness of error and the holistic ratings, support this conclusion. The ESL teachers who evaluated the thirty essays analyzed in this study appear to have been strongly influenced by the factor of error in their judgments.
Unfortunately, the Correctness Score cannot be viewed as a wholly objective measure, a factor reflected in the disappointingly low correlations of the reliability check (see Chapter III, page 36). Brodkey and Young's (1981) contention that the Correctness Score is a "powerful discriminator" among different levels of proficiency cannot, therefore, be supported by the results of this study. Only a broader analysis, using a larger essay corpus with two or three essays from each student, and with raters skilled in the recognition of three levels of error, can determine the eligibility of the Correctness Score as a useful index of ESL student writing quality.

SUMMARY

The results indicate that there were no remarkable differences in syntactic complexity among thirty college-level ESL student essays analyzed in this study. Figures for the three measures of the clause analysis technique did not show significant correlation with the holistic ratings.

The Correctness Score, a measure of frequency and seriousness of error, was found, however, to show a significant positive correlation with the holistic ratings. This finding indicates that evaluators took the occurrence of error as a major consideration when grading the sample.

The results must be qualified by questions about the dependability of the holistic rating scheme. Variation
among the two experienced raters used for this study was surprisingly high. The results for the measures of syntactic complexity may also have been influenced by the fact that two-thirds of the sample were written by students from one proficiency level, this, perhaps, being the cause of the rather narrow distribution of figures for information block length and embedding depth.

The only significant finding of this study was, then, that the Correctness Score reflected evaluators' perceptions of ESL writing quality. Questions remain, however, about the dependability of the Correctness Score as a find discriminator.
CHAPTER V

CONCLUSION AND IMPLICATIONS

CONCLUSION

The results of this study cannot support any strong conclusions about the relationship between syntactic complexity, as measured by the clause analysis technique, and evaluators' perceptions of ESL writing quality. On the other hand, the results do indicate a definite relationship between the evaluators' ratings of the student essays and the frequency and seriousness of error, as measured by the Correctness Score.

Since two-thirds, or twenty out of thirty, of the students who contributed essays for the research corpus came from the highest proficiency level, identified with TOEFL scores of 490+, the sample cannot be said to be representative. Therefore, the nature of the sample itself may have had some influence on the results indicating an apparent lack of correlation between the two measures of subordination, information block length and embedding depth, and the holistic ratings. The data do reveal that the five essays which were rated 2.5 or below on the holistic scale all ranked below the mean in embedding depth. This finding points to a possible relationship between embedding depth
and ratings of essays from the lower levels of proficiency. However, this group of five essays cannot be used to extrapolate any definite conclusions, particularly when the reliability of the holistic ratings themselves is open to debate.

The reliability of the clause analysis technique measures was surprisingly high, considering the fact that these measures were developed for first language research. However, it is also clear that consistent and reliable analysis is a problem when readers are faced with the task of subjectively judging the ESL writer’s intentions in prose which is characterized by a significant number of syntactic errors. The error-free T-unit was designed to address the problem of measuring syntactic complexity in ESL writing, but this method would only appear to be effective in analyzing essays which are relatively error-free. A highly reliable measure of syntactic complexity for ESL writing at lower levels of proficiency has yet to be developed. The clause analysis technique, designed as it was for the analysis of adult first language writing, cannot be advocated for detailed, discriminating, and accurate analysis of second language essays.

It must also be noted that the scores for embedding depth and information block length in the essay sample show a high correlation with each other: +.927 in a Spearman rank order correlation. Although the measure of embedding depth
was designed to measure the "quality" of writing complexity, while information block length is a measure of "quantity," the very high correlation indicates that these two measures have a very reliable relationship. Information block length, measuring the number of clauses per main clause, was shown to be a more reliable measure, although embedding depth showed a slightly higher correlation with the holistic ratings. In any case, it would appear that only one of these measures need be applied as a measure of subordination.

An interesting finding of the analysis was that scores for embedding depth and information block length did not show any remarkable spread, or distribution. This would seem to indicate that all of the students were subordinating structures at roughly the same level of complexity. Since this was the case, it is perhaps understandable that raters did not appear to take into account the rather minor differences in syntactic complexity when evaluating different essays.

Finally, while a number of studies have established a relationship between complexity of syntax and ratings of college-level ESL essays, all of these studies have used measures borrowed from, or adapted from, first language research. Measures devised to analyze growth in first language acquisition over a number of years may serve little
purpose in measuring levels of syntactic complexity in the intensive environment of college-level ESL programs.

The guidelines for the TWE (Test of Written English) examination stress variety of syntax, rather than syntactic complexity. The goal of the test appears to be fluidity of expression, rather than the manipulating of complex and weighty structures learned in the ESL writing class. Current trends in language teaching indicate a departure from prescriptive notions that "good" writing may be characterized by conscious manipulation of structures in subordinate relationships. The student is left to develop a personal style arising from his or her own knowledge of the language. Thus, the idea put forward by Hunt (1965) and his followers, that more syntactic complexity may reflect higher competence, could be invalid.

While the results of this study do not justify any definite conclusion about the relationship of syntactic complexity to evaluators' perceptions of student writing quality, the high correlation between the Correctness Score and the holistic ratings indicates that the factor of error played a major part in readers' judgments.

This result contradicts current theory that error should not, in fact, be a deciding criterion in the rating of ESL writing. One of the hypotheses put forward in this study was that error would only be a significant factor in the rating of the essays when combined with a measure of
subordination. However, the findings demonstrate clearly that frequency and seriousness of error is a major factor in evaluators' perceptions of ESL writing quality.

Since the Correctness Score did show a significant correlation with the holistic ratings, it could be promoted as a useful index of ESL writing quality. However, the Correctness Score itself demonstrated reliability problems. The inter-rater reliability check showed that there was significant disagreement among raters about the constitution of first-, second-, and third-degree errors. The measure is, in fact, quite subjective and cannot be recommended as a fine-toothed discriminator among many different levels of proficiency.

Any conclusions in this study must be qualified by the reliability analysis of the holistic ratings, posited as the criterion reflecting the evaluators' perceptions of the quality of the student essays. Taking into account that the TWE guidelines acknowledge a one-point difference on the six-point rating scale as acceptable, the two raters used in this study showed an agreement of .90. However, the raters gave identical scores to an essay only eight times. If exact agreement were established as the criterion for determining reliability, then the raters, both of whom underwent the same training in evaluating according to the TWE guidelines, showed agreement only 26 percent of the time. From this evidence, it must be concluded that the
holistic ratings were not a dependable criterion in this study and may not be suitable for further research in ESL writing.

In conclusion, the results do not support the hypotheses that two measures of subordination, information block length and embedding depth, would be valid indices of ESL writing quality as determined by holistic ratings. However, the nature of the sample precludes any definite conclusions that these two measures are not indicative of evaluators' perceptions of ESL writing quality. The Correctness Score, a measure of frequency and seriousness of error, did show a high correlation with the holistic ratings, leading to the conclusion that the factor of error played a significant role in the rating of the essay sample. Reliability problems, however, indicate that the Correctness Score has limited value as a fine-toothed discriminator at different levels of proficiency.

IMPLICATIONS

The results of this study and the questions raised by the analysis suggest several implications for further research and the ESL writing class.

Given that 67 percent of the sample came from one proficiency level, no definite conclusion can be drawn about the relationship of degrees of subordination with evaluators' perceptions of ESL writing quality. A larger
sample, drawing from several different proficiency levels, would lend credence to any conclusions established. In addition, the measures of information block length and embedding depth would have to be refined in order to produce more accurate analysis of subordination levels in ESL writing. Clear guidelines regarding the treatment of erroneous syntax must be established. The problem of reliability may be answered in part by using the most reliable measure, information block length, as an index of subordination.

However, some studies (Perron 1977, Crowhurst and Piché 1979) have found that syntactic complexity in student writing varies in different modes of discourse and when different audiences are identified. In addition, the TWE guidelines specify variety of syntax, not complexity, as an important criterion for evaluation. These factors indicate that any measure of syntactic complexity in college-level ESL writing may, in fact, prove of little value for assessment purposes. Students may adopt different strategies when writing for an examiner, and test evaluators, in any case, appear to be interested in more than subordination per se.

The high correlation established between the Correctness Score and the holistic ratings implies that a measure of frequency and seriousness of error may be a useful assessment tool. However, the discriminatory powers
of the Correctness Score are open to question and the measure cannot be recommended as a reliable and fine discriminator at many different levels of proficiency.

The finding that the factor of error did play a major role in the rating of the essay sample suggests two significant implications for ESL instruction. First, it is not altogether encouraging that raters appeared to take error as such an important criterion in their evaluations. Error is only one of the many criteria listed by the TWE Scoring Guidelines. Others include organization, development, attention to detail, focus on the writing task, unity, coherence, progression, syntactic variety, and range of vocabulary. While this study did not objectively analyze the relationship of these factors to the holistic ratings, it would appear from the results that the raters placed an inordinate amount of emphasis on error as a factor in evaluation. The implications are that raters need to be retrained to take into account other criteria, such as those described by the TWE, and give them equal weight.

Second, the Correctness Score results appear to reflect teacher concerns about error in ESL student writing. These concerns could be addressed by devoting more time in the ESL class to the analysis of student errors, emphasis of student correction of error, and instruction in those areas which may allow for rapid progress, such as spelling. Care should be taken, however, not to over-emphasize error to the
exclusion of organization, coherence, range of vocabulary, and all the various criteria established by the TWE as indicative of good writing. The essay must be evaluated as a complete act of discourse and not as the sum of its errors.

Finally, the use of holistic ratings as a reliable measure of writing quality for research purposes must be carefully examined. Accurate and reliable results can be expected only with the most professional training sessions, involving experienced raters and a scoring guide. An analytical scoring scheme, such as that suggested by Jacobs et al. (1981), may be found to be a more reliable method of grading. Using this scheme, scores are allotted to each essay in five areas: content, organization, vocabulary, language usage, and mechanics. Use of this kind of scoring format may also help to reduce the emphasis on error evident in the rater evaluations of the essay sample used in this study.

Many variables may affect the quality of an ESL student essay. Only further research can establish reliable, empirically-based conclusions on the relationship of syntactic complexity, and frequency and seriousness of error, to evaluators' perceptions of ESL writing quality.
REFERENCES


San Jose, C.P.M. 1972. "Grammatical Structures in Four Modes of Writing at Four Grade Levels." Dissertation Abstracts International 33: 5411A. Syracuse University.


APPENDIX A

TEST OF WRITTEN ENGLISH (TWE)
SCORING GUIDELINES *

Readers will assign scores based on the following scoring guide. Though examinees are asked to write on a specific topic, parts of the topic may be treated by implication. Readers should focus on what the examinee does well.

Scores

6 Clearly demonstrates competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.

A paper in this category
-- is well organized and well developed
-- effectively addresses the writing task
-- uses appropriate details to support a thesis or illustrate ideas
-- shows unity, coherence, and progression
-- displays consistent facility in the use of language
-- demonstrates syntactic variety and appropriate word choice

5 Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will have occasional errors.

A paper in this category
-- is generally well organized and well developed, though it may have fewer details than does a 6 paper
-- may address some parts of the task more effectively than others
-- shows unity, coherence, and progression
-- demonstrates some syntactic variety and range of vocabulary
-- displays facility in language, though it may have more errors than does a 6 paper

4 Demonstrates minimal competence in writing on both the rhetorical and syntactic levels
A paper in this category
-- is adequately organized
-- addresses the writing topic adequately but may slight parts of the task
-- uses some details to support a thesis or illustrate ideas
-- demonstrates adequate but undistinguished or inconsistent facility with syntax and usage
-- may contain some serious errors that occasionally obscure meaning

3 Demonstrates some developing competence in writing, but it remains flawed on either the rhetorical level or syntactic level, or both.

A paper in this category may reveal one or more of the following weaknesses:
-- inadequate organization or development
-- failure to support or illustrate generalizations with appropriate or sufficient detail
-- an accumulation of errors in sentence structure and/or usage
-- a noticeably inappropriate choice of words or word forms

2 Suggests incompetence in writing.

A paper in this category is seriously flawed by one or more of the following weaknesses:
-- failure to organize or develop
-- little or no detail, or irrelevant specifics
-- serious and frequent errors in usage or sentence structure
-- serious problems with focus

1 Demonstrates incompetence in writing.

A paper in this category will contain serious and persistent writing errors, may be illogical or incoherent, or may reveal the writer’s inability to comprehend the question. A paper that is severely underdeveloped also falls into this category.
Papers that reject the assignment or fail to address the question in any way must be given to the Table Leader. Papers that exhibit absolutely no response at all must be given to the Table Leader.

### APPENDIX B

**SYMBOLS USED IN TRANSCRIPTION OF WRITTEN ESSAYS ON A WORD PROCESSOR**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>X . . . X</td>
<td>Words deleted by student, e.g. 'We went to XscoolX school.'</td>
<td>'We went to XscoolX school.'</td>
</tr>
<tr>
<td>&lt; . . . &gt;</td>
<td>Insertions by student, e.g. 'We went to &lt;our&gt; school.'</td>
<td>'We went to &lt;our&gt; school.'</td>
</tr>
<tr>
<td>XX</td>
<td>Heavy deletion by student, e.g. 'We ate XX some apples.'</td>
<td>'We ate XX some apples.'</td>
</tr>
<tr>
<td># . . . #</td>
<td>Directions by transcriber, e.g. #skips line#</td>
<td>#skips line#</td>
</tr>
</tbody>
</table>
A "5" Essay

People's knowledge about life is acquired in many ways. One is learning from one's parents, relatives and friends. The other is learn through one's own experiences. In my opinion, most of the things we learned about life are from other people like parents and friends.

When I was in high school, I had a fight with a boy in my class. After the fight I didn't dare go to school, because his big brother was also in my school. My father advised me to go and talk to that guy. I took my father's advice and the result was really much better than I expected. I said sorry for beating him the other day and I expressed my hope that I hadn't hurt him. After the talk, we became friends again. What I learned from this is that people are not so mean as they appear to be sometimes. Also, words are more powerful than fists.

After working in a factory for three years, I forgot most of things I learned in school. I hesitated at the opportunity of entering for university exam. My friends strongly persuaded me to go ahead because I wouldn't do well in society without at least college education. I tried and
I succeeded. From this I learned that life is in some degree conquerable, as long as one is determined. Finally, I believe that people learn most of the things about life from others before they step into society. But once they are grown up, they acquire more knowledge through their own experience. By saying "step into society," I mean when they are really independent of their family.

THE END

A "3.5" Essay

It is very useful for my life both listening to the advise of family or friends and learning about life is through personal experience. Above all, I think those factors are very important for me depending on each situation.

The advice of family and friends encourage me to work successfully when I have some depression. Also, their advices is very frankly because they are familiar with me. Sometimes they blame my faults and encourage me to do something. When I graduated high school, my father advised me to go to the military Academy because he also had graduated the military Academy. So I accepted his advice. Thus, I had lots of experience during the service periods after graduated the military Academy. So I believe that the
advice of family or friends is a great role to make the my future plan.

Another useful model is that it is to learn other personal experience. Former great people's experience lots of influence to me. Before I came to here, I saw some people who are very success in their field. They already studied in America when they are young age. Their experience moved my mind to came to America. Their experience is life knowledge to me. Also their experience became motive that I came to study here. Thus, their personal

A "2" Essay

Basicly there are tow ways of learning. First one some people believe that tht the best way is to listening. And other believe that the bast way is through personal experience. each ones of these has advantages and disadvantages, I will start with the advantages. First of all listining to familiy an friends advice is useful, because it is protect you or guid to the right way. Also it is sav for money and time, For example teacher guid student to the right way of studing If obey you will suceed and same time, If not you well fail and lose on the amount of money for a new course. In the work fields if worker listine to his supervisor, he will not have any trable with his work. Briefly flowing the advice from people are warchess that is useful. It is the other side (advantage of people who
like to learning by their experience they will learn more, because trying they is different from heart. #many lines later#

In my opinion th peopl should <learn> from the advice snad sometimes through ther persn exprnc
APPENDIX D

EXAMPLES OF THE ANALYSIS

Clause Analysis Technique

From my point of view, I would prefer the best way of learning about life is by listening to the advice of family and friends. I'm not saying that it's bad to learn only by one's personal experience.

There are some advantages for learning by personal experience. It brushes up one's own personality and oneself. Because you must learn everything by yourself. As the example, I have a friend, he has lost his relative by car accident when he was 16 years old. And he has decided to quit school and start to make his own living by himself without any other people's help or advice. He's now twenty-nine years old and has his own factory. He gave me many valuable advices and tips from his own experience. I think learning by oneself him very responsible and very concerned person.

Sometimes, people proceed their life to wrong way. Because people are not mature enough to make own right decision for life and other things.
If you'll <be> able to get a lot of advices or tips from others, you can evaluate and make better decisions. From my own experience, I learned so many things from my folks, school teachers and friends. I had arguments sometimes with them. I was right sometimes and wasn't sometimes. But this way of learning makes me that kind of person who made careful quick and right decision for my past life. This is why I would prefer to learn by advices of other people, not <only> by yourself.

SENTENCES (//) = 18

CLAUSES (//) = 3

MAIN CLAUSES (A) = 2

EMBEDDING VALUE = 54

"A" CLAUSES 22×1 = 22

"B" CLAUSES 10×2 = 20

"C" CLAUSES 4×3 = 12

\[ \text{EMBEDDING VALUE} = 54 \]
Correctness Score

Personal experience is more preferable than the listening to the advice of family and friends to learn about life.

There is a most important advantage in the personal experience. That is that he or she got something by himself or herself from his or her experience. It's hard to get new things but once he or she got, it really works.

On the other hand, listening to the advice of family and friends is a good way to understand life easily but it's just conception. If he or she try to do new thing by using this conception only, he or she will get some trouble because he or she has never done it before even if he or she has knowledge of it. Of course, it's better to have some knowledge then to have nothing.

Here is and example of these. Some of my friends told me that low class workers are not smart. But it was not true. I got part-time job which was a same as low-class job. After I spent a lot of time with them, I found new things about them. The things are that they knew a lot of things which I don't know and they handle it very easily even if it's hard for me.

I want say that only knowledge doesn't work but experience is really works even if there is a limit of it.

ERROR SCORE = 29