MODELS FOR SYNTACTIC STRUCTURES: COMPUTATIONAL CHARACTERIZATIONS OF SOME SYNTACTIC CONSTRUCTIONS IN ENGLISH

by

William T. Clay

A thesis
submitted in partial
fulfillment of the requirements for the degree of
Master of Arts in the Department of Linguistics
Fresno State College

April, 1969

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Grammatical analysis involves the resolution of sentences into constructions consisting of related words or groups of words. Traditionally the relationships bear such names as modification, apposition, and predication. These grammatical relationships have been defined primarily in semantic terms. See, for example, Curme's <u>Syntax</u>.[1]

Consider the relationship of modification as exemplified by the construction <u>foolish</u> <u>man</u> in the sentence

The foolish man laughed.

The word <u>foolish</u> refers to an insubstantial quality (foolishness) which a man may possess; thus it makes the substantial and generic word <u>man</u> more specific. As modification is traditionally understood, the modifier limits or restricts the meaning of the word which is modified.

The function of <u>foolish</u> may indeed be to specify which man laughed. If contrastive stress is used, one says

The fóolish man laughed.

to distinguish that it was not the case that

[1] George O. Curme, Syntax (New York, 1931).

The clever man laughed.

But modification is, however, more than specification. There is also another interpretation. If contrastive stress is not used, _foolish_ may be thought to add information or explain in the manner of an appositive.

The man, a fool, laughed.

In order to elucidate the difficult concepts of grammatical relationship, one may examine the co-occurrence properties of words and phrases when uttered or written. Consider how the words _foolish_ and _man_ might be distributed in some English text. One would expect that each of these words would be more frequent than the phrase _foolish_ _man_ itself. That is, generally speaking, various words other than _man_ would follow _foolish_, and _man_ would not always be preceded by _foolish_. It follows generally that among the constructions consisting of an adjective followed by a noun, the phrase _foolish_ _man_ might be relatively infrequent. However, the words themselves — _foolish_ and _man_ — might be relatively frequent or relatively infrequent within their respective word classes — adjective and noun. Thus, in order to formulate a co-occurrence model for constructions like _foolish_ _man_, one is led to examine the co-occurrence properties of adjectives and nouns as classes of words.

An adjective-noun construction and its constituents have certain distributional relationships. Consider _foolish_ _man_ once again. Generally one may replace _foolish_ _man_, wherever it occurs, by _man_ and conversely, and not change the grammaticality of the sentence. However, usually one may not replace _foolish man_ wherever it occurs by _foolish_ nor _foolish_ by _foolish man_. In other words, in relationships of modification, _foolish_ is dependent, if not on _man_, on some noun; _man_, however, is independent of any adjective.

In purely distributional terms, there are three fundamentally different construction types. A construction of the first type, as described above, has a grammatical distribution similar to that of one and only one of its constituents. Such a construction generally exhibits the relationship referred to as modification.

Another type is exemplified by a construction consisting of a phrase followed by an appositive. The phrase _a fool_ stands in apposition to _the man_ in the sentence

The man, a fool, laughed.

Clearly, in any sentence where one finds _the man_ followed by the appositive _a fool_, he is free to delete the appositive without changing the grammaticality of the sentence. Conversely, one is free to add the appositive _a fool_ after every occurrence of _the man_ in a text. In fact, if the phrase _the man_ is not followed by the appositive _a fool_, one is free to delete _the man_ and substitute the phrase _a fool_. The grammaticality of the sentence would remain unchanged. In short, the distribution of the entire construction is similar to that of either of its constituents. It follows that, in relationships of apposition, the constituents are independent of one another.

The third distributional type is encountered when one considers the relationship of predication. The subject _the foolish man_ and the predicate _laughed_ are conjoined to make "a complete sentence."

The foolish man laughed.

Neither _the foolish man_ nor _laughed_ is a substitute for the sentence itself. The grammatical distribution of the sentence is quite different from that of either of its constituents. Subject and predicate are, in fact, dependent upon one another to form a sentence. This third distributional type is also found in such constructions as prepositional phrases. For example, one such construction is the phrase _at himself_ in the sentence

The foolish man laughed at himself.

Various hypotheses have been made regarding the distributional characteristics of the constituents of syntactic constructions in English. See, for example, the well known works of Nida[2] and Harris.[3] These hypotheses are rooted in the concepts of grammatical independence and dependence. In this paper, the co-occurrence properties of word classes are used to make these concepts explicit. In this way, a computational

---

[2] Eugene A. Nida, _A Synopsis of English Syntax_ (Norman, Okla., 1960), esp. pp. iii-lxviii for a discussion of construction types.

[3] Zellig S. Harris, _Structural Linguistics_ (Chicago, 1951), esp. pp. 325-348 for a discussion of distribution.

characterization of some English syntactic constructions is achieved. The characterization is verified with respect to a particular text and a particular set of word classes. For this text and word class system, the characterization is numerically adequate but by no means perfect. Whether a more complicated characterization or a more elaborate system of word classes, or both, would have led to better results remains an open question.

CHAPTER II

CONSTRUCTION TYPES

Traditionally grammarians have classified clauses as either "dependent" or "independent." Independent clauses may stand alone, as acceptable sentences; dependent clauses may not. Usually dependent clauses are joined in a sentence by using a subordinate conjunction such as _if ... then_, _if ..._, _since ..._, _that_, _which_, or _who_.

If he laughs, she will cry. (Subordination)

Clauses are made independent by the use of a coordinate conjunction such as _and_, _but_, _for_, _or_, _nor_, or _yet_.

He will laugh, and she will cry. (Coordination)

A sentence in which a relationship of subordination or of coordination is marked, as with a conjunction, is said to be _hypotactic_.[4] If however, the clauses are placed one after the other, with no grammatical relationship indicated, the sentence is _paratactic_.[5]

He laughed; she cried. (Parataxis)

---

[4] _Webster's New International Dictionary of the English Language_, 2nd ed. (Springfield, Mass., 1951).

[5] _Ibid_.

These traditional concepts, long used to describe relationships between clauses as constituents of sentences, are applicable, in part, to a description of relationships between words as constituents of phrases. In describing relationships between words, it is convenient to further subcategorize hypotactic constructions into *endocentric* and *exocentric* types.

Consider a typical English construction in which one word is subordinate to another: *foolish man*, in the sentence below. Assume that *foolish man* is, in fact, a "construction" in this sentence. Whether or not a particular pair of adjacent words in a sentence is a construction is determined by a structural analysis of the sentence, such as might be found in Nida.[6]

The foolish man laughed loudly.

Of the two words in the construction, *foolish* and *man*, only *man* is independent. That is, if one deletes *foolish* from the sentence, leaving *man* to stand alone in the place previously occupied by the construction, the sentence will remain grammatically acceptable, though a little less informative.

The ~~foolish~~ man laughed loudly.

---

[6] Nida, *op. cit.*

On the other hand, if *man* is deleted, the resulting sentence is meaningless. Clearly, *foolish* is dependent, if not on *man*, on some "understood" substantive.

The foolish (ones) laughed loudly.

Likewise, given that *laughed loudly* is a construction, *loudly* is dependent on *laughed*. If *laughed* is deleted, the sentence becomes meaningless.

The foolish man ~~laughed~~ loudly.

Clearly, *laughed* is independent.

The foolish man laughed ~~loudly~~.

A dependent word is an *attribute* in Hockett's terms.[7] In *foolish man*, *foolish* is an attribute; the independent *man* is the *head* of the construction. Such a construction, consisting of attribute and head, is *endocentric*. If the head is to the *right* of its attribute, as it is in *foolish man*, the construction is *right* endocentric. If the head is to the *left* of its attribute, as it is in *laughed loudly*, the construction is *left* endocentric. The structural symbols

---

[7] Charles F. Hockett, *A Course in Modern Linguistics* (New York, 1958), pp. 183 ff.

are from Nida.[8]

The foolish man laughed loudly.

Consider now the construction *was laughing* in the sentence below.

The man was laughing.

Only with difficulty could one argue that it is endocentric, right or left. He might contrive to say that *was* is a "helping verb," and thus, an attribute of the "main verb" *laughing*. On the other hand, he might argue that *laughing* is "descriptive of a state of being"; and thus it is the attribute in the construction. It is argued here however, that because neither *was* nor *laughing* may be deleted, neither is subordinate to the other. In a sense, they are coordinate; but the traditional concepts break down at this point. Coordinate clauses, it will be remembered, are independent. However, the constituents of the construction in question are in no sense independent; they are mutually dependent. Each is subordinate to the verb phrase. Such a construction, consisting of equally necessary and mutually dependent constituents, is *exocentric*.

The man was laughing.

[8] Nida, *op. cit.*

The rather vague concepts of "head" and "attribute" lead the traditional grammarian to err in the classification of at least one construction. Commonly, he believes that *the* is somehow an attribute of *man* in the construction *the man*, as in the above sentence; and thus that the construction is right endocentric. Because neither *the* nor *man* may be deleted, the construction is properly exocentric, rather than endocentric.

The man was laughing.

Relationships of subordination and coordination between adjacent words are not marked with conjunctions, as between clauses, but simply by the fixed grammatical ordering of the constituent words. Now consider some constructions in which the order of the constituent words is not fixed. In such constructions, either word may be on the left or on the right and either one may be deleted, without affecting the grammatical acceptability of the sentence. The constituents are coordinate, yet the coordination is unmarked. It will be remembered that unmarked coordination is called *parataxis*. One may consider parataxis to be a result of the deletion of a conjunction such as *and*, as in the sentences below.

The foolish insensitive man laughed.

or The insensitive foolish man laughed.

i.e. The man, who is foolish *and* insensitive, laughed.

or The man, who is insensitive *and* foolish, laughed.

The cold damp air is bad.

or The damp cold air is bad.

i.e. The air, which is cold *and* damp, is bad.

or The air, which is damp *and* cold, is bad.

These paratactic constructions are not constructions at all according to Nida, who would diagram the sentences differently.

The foolish insensitive man laughed.

The cold damp air is bad.

For a different analysis of grammatical structure, which tends to support the contention that the structures in question are paratactic, see Chomsky's "On the Notion 'Rule of Grammar'."[9] Chomsky would diagram these structures in still another way.

The foolish (and) insensitive man laughed.

The cold (and) damp air is bad.

The types of two word constructions discussed in this paper are summarized in Table 1.

---

9 Noam Chomsky, "On the Notion 'Rule of Grammar'," *The Structure of Language: Readings in the Philosophy of Language*, eds. J. A. Fodor and J. J. Katz (Englewood Cliffs, N. J.), pp. 119-136.

<u>Hypotactic</u>:	Subordination or coordination is marked by order or concord.		
	<u>Right Endocentric</u>:	The subordinate word is on the left.	
	<u>Left Endocentric</u>:	The subordinate word is on the right.	
	<u>Exocentric</u>:	Neither word is subordinate to the other; however, each is subordinate to the construction as a whole.	
<u>Paratactic</u>:	Coordination is unmarked. Each word is coordinate with the other and with the construction as a whole.		

Table 1 Construction Types

CHAPTER III

PROBABILISTIC MODELS

Suppose one is given some English text, containing possibly several hundred words as they might be uttered or written, of which the sentence below is a part.

The foolish man was laughing.

Consider again the right endocentric construction *foolish man*. Suppose that *foolish* is a relatively frequent attribute in the text. One finds several constructions like *foolish men*, *foolish people*, *foolish ideas*, and so on. However, constructions of which *man* is the head are few. For the given text, one would conclude that there is a small chance that given *foolish*, *man* will follow. There is, however, a relatively good chance that given *man*, *foolish* will precede.

Suppose the situation is the converse: *foolish* is an infrequent attribute and *man* is a frequent head. In this case, the conditional probability that *man* follows *foolish* is high. The conditional probability that *foolish* precedes *man* is low, however.

There are two other possible situations. The conditional probabilities of co-occurrence might be low, or they both might be high. In a text, in fact, a pair of words found in a right endocentric construction might stand in any of these general probabilistic relationships.

Make the strong assumption that one can place any English word in one and only one word class. Consider then two English texts; the first is a sequence of words, and the second is a sequence of word classes obtained by marking the class of each word in the first. The words which belong to any two adjacent word classes, which are constituents of the same construction, might stand in any of the previously described word level co-occurrence relationships; however, the word classes themselves will necessarily exhibit, for the whole text, one and only one of the four possible co-occurrence relationships.

Consider a right endocentric construction. The word class on the left may be deleted and the sentence will remain grammatically acceptable. This is not the case for the word class on the right. It is hypothesized that the conditional probability of the right constituent following the left is greater than the conditional probability of the left constituent preceding the right. For example in English, nouns may stand alone or be preceded by adjectives, but adjectives will very likely be followed by nouns. In symbolic terms, if the ordered pair A B of the word class text represents a right endocentric construction, then the conditional probability of B given A $P(B|A)$ is greater than the conditional probability of A given B $P(A|B)$. The hypothesis so far is that the grammatical construction right endocentric may be characterized by the inequality

$$P(B|A) > P(A|B)$$

In graphical terms, the pairs of conditional probabilities for right endocentric constructions are expected to fall in a region bounded by the axis of abscissas $P(A|B) = 0$ and a line determined by the equation

$$P(A|B)/P(B|A) = k_1 \quad , \text{ where } 0 < k_1 < 1$$

The value of k_1 that expresses just how large $P(B|A)$ may be in relation to $P(A|B)$ may be estimated experimentally.

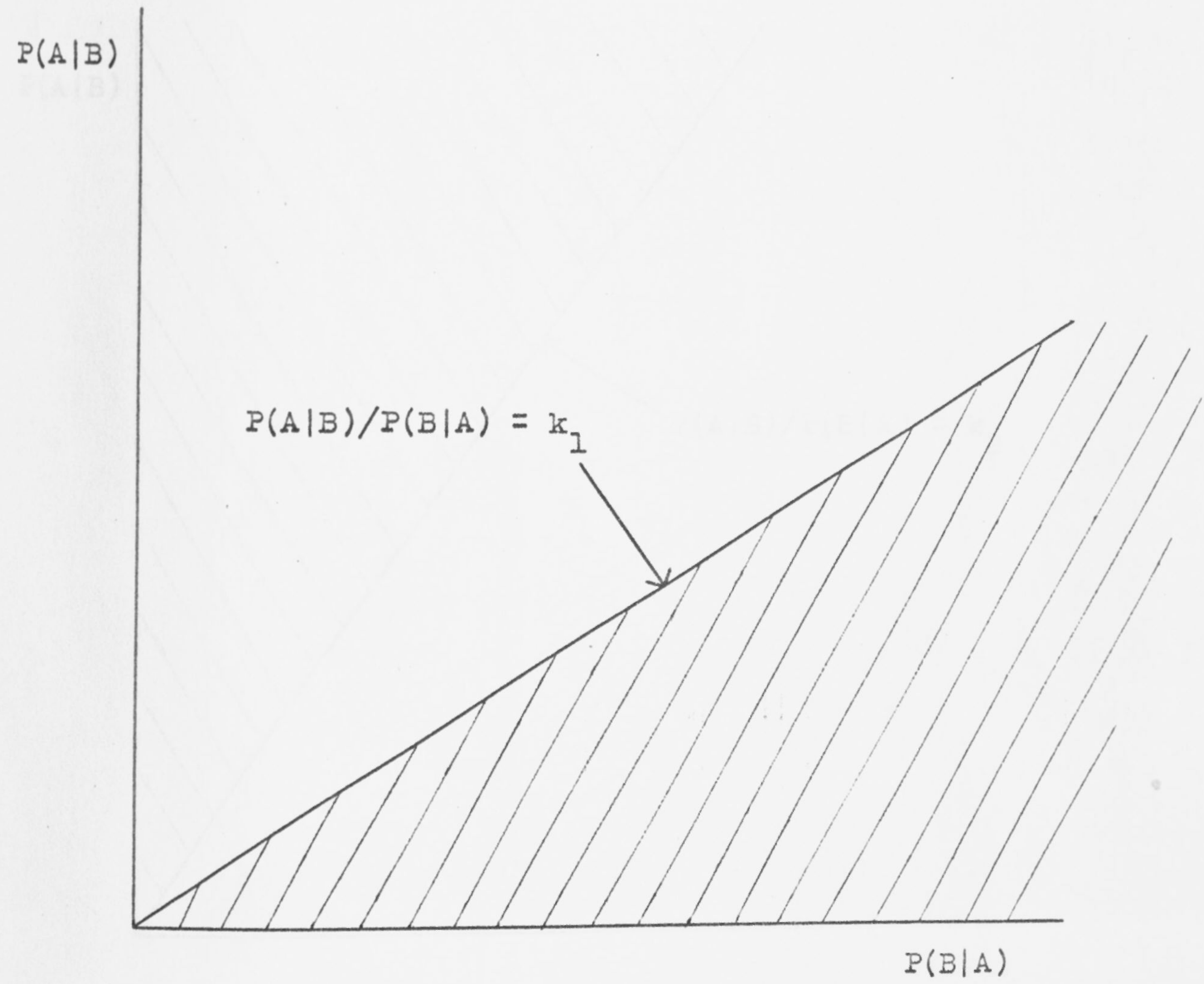

Figure 1 Hypothetical Distribution of Right Endocentric Constructions

By arguments similar to those made regarding right endocentric constructions, it may be hypothesized that a region of left endocentricity, as shown in Figure 2, is bounded by the axis of ordinates $P(B|A) = 0$ and a line determined by the equation

$$P(A|B)/P(B|A) = k_2 \quad , \text{ where } 1 < k_2 < \infty$$

The value of k_2 that expresses just how large $P(B|A)$ may be in relation to $P(A|B)$ may be estimated experimentally.

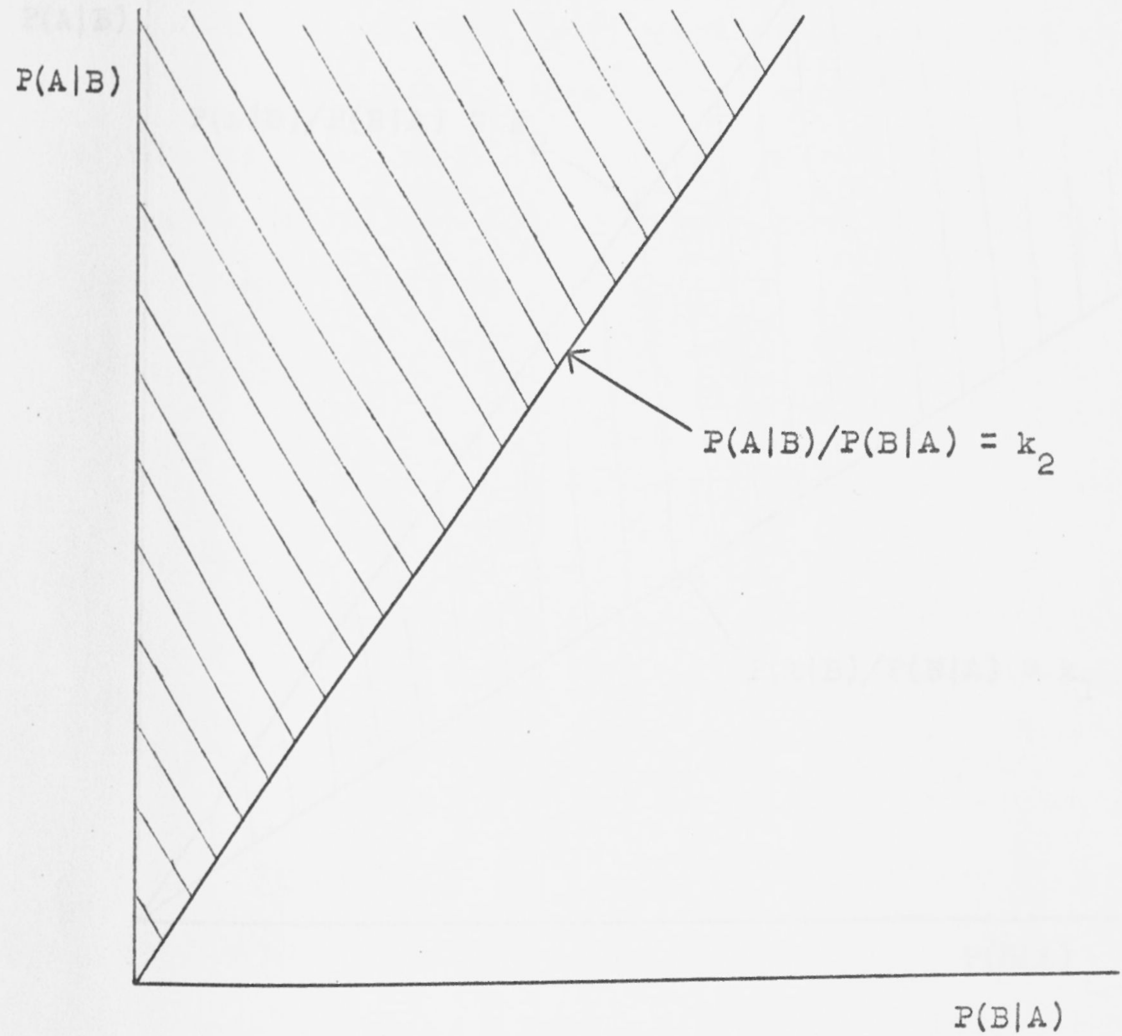

Figure 2 Hypothetical Distribution of Left Endocentric Constructions

It will be remembered that neither constituent may be deleted from an exocentric construction. It is hypothesized that a region of exocentricity, a region of nearly equal conditional probabilities, lies between the regions of endocentricity. The hypothesis is depicted in Figure 3.

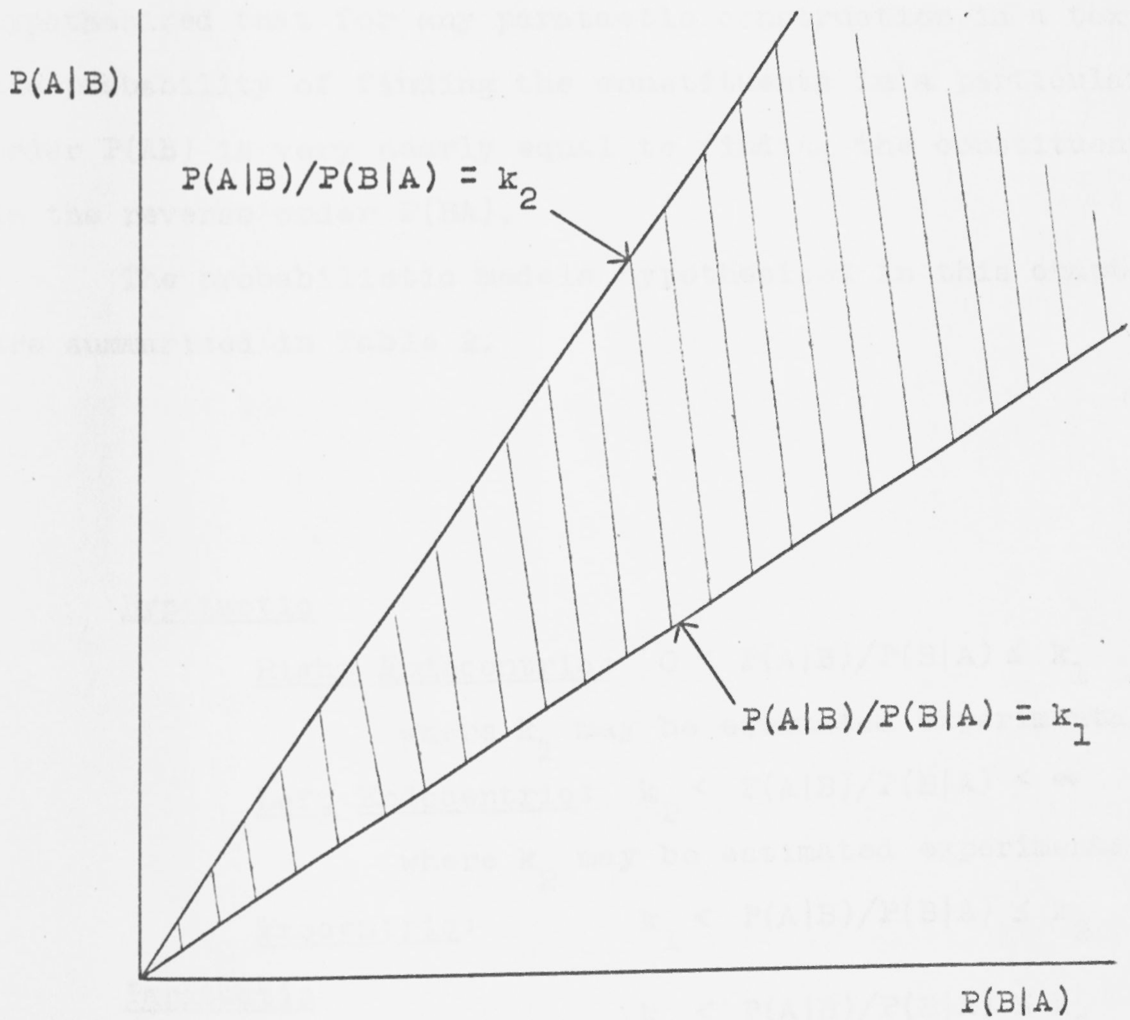

Figure 3 Hypothetical Distribution of Exocentric Constructions

The constituents of an exocentric construction are coordinate with one another. The coordination is marked by a specific and irreversible ordering of the constituents. The constituents of paratactic constructions are coordinate also. Graphically, paratactic constructions are expected to fall in the exocentric region. However, in paratactic constructions, the order of constituents is not specified grammatically and is freely reversible. Therefore, it is hypothesized that for any paratactic construction in a text, the probability of finding the constituents in a particular order P(AB) is very nearly equal to finding the constituents in the reverse order P(BA).

The probabilistic models hypothesized in this chapter are summarized in Table 2.

<u>Hypotactic</u>

<u>Right Endocentric</u>: $0 < P(A|B)/P(B|A) \leq k_1$,
where k_1 may be estimated experimentally.

<u>Left Endocentric</u>: $k_2 < P(A|B)/P(B|A) < \infty$
where k_2 may be estimated experimentally.

<u>Exocentric</u>: $k_1 < P(A|B)/P(B|A) \leq k_2$

<u>Paratactic</u>

$k_1 < P(A|B)/P(B|A) \leq k_2$

and $P(AB) = P(BA)$

Table 2 Probabilistic Models of Construction Types

CHAPTER IV

PREPARATION OF DATA

In order to make a statistical analysis of an English text, the sequence of words was first converted to a sequence of symbols representing word classes. For this study, the criteria for word class membership were mainly derivational. Criteria have been selected which facilitate the use of these same word classes in comparative studies. The number of word classes was limited to eight, in view of the small size of the available computer (an IBM 1620.) Each word class and sentence boundary was identified by one of nine digits, and it was upon sequences of these digits that the computer performed calculations.

The word classes as used are:

(1) Nouns: mass, count, common, proper, nominalized adjectives; in general, whatever a good dictionary calls a noun.

(2) Determiners: articles, possessive pronouns adjacent to nouns, and closed class adjectivals.

(3) Pronouns: if not otherwise, as above, generally whatever the dictionary calls a pronoun.

(4) Qualifiers: open class adjectives and adverbs, derivationally related.

(5) Verbs: inflected forms, with exceptions as below.

(6) Auxiliaries: A class of words with similar distributional characteristics: helping verbs *do*, *is*, *has*, et cetera, before participles; closed class adverbials *fast*, *quick*, *slow*, et cetera; and the modals.

(7) Participles: present and past.

(8) Particles: conjunctions and prepositions; again, generally whatever the dictionary calls a conjunction or a preposition.

(9) Sentence boundary.

The English text analyzed consists of about 1500 words, the first 54 sentences of "The Changing Year," from Rachel Carson's *The Sea Around Us*.[10] The manner in which word classes were marked is shown below.

```
 For the sea as a whole, the alternation of
9 8  2   1  8 2   4     2          1      8

 day and night, the passage of the seasons,
  1   8    1     2     1     8  2     1

 the procession of the years, are lost in its
  2      1      8  2    1      5    7  8  2

 vastness, obliterated in its own changeless
     1          7      8   2   2       4

 eternity.  But the surface ...
     1     9 8   2     1
```

The word class sequence for the entire text appears as a sequence of digits on page 36 in the Appendix. The conditional probabilities of co-occurrence for the various word classes,

---

[10] Rachel L. Carson, "The Changing Year," *The Sea Around Us* (New York, 1961), pp. 41-47.

as calculated for the text, are shown on pages 37 and 38.

For the 54 sentences of the English text, structural diagrams were prepared, generally following the analysis of Nida.[11] See Figure 4. As further examples, full diagrams appear in the Appendix; see Figure 8.

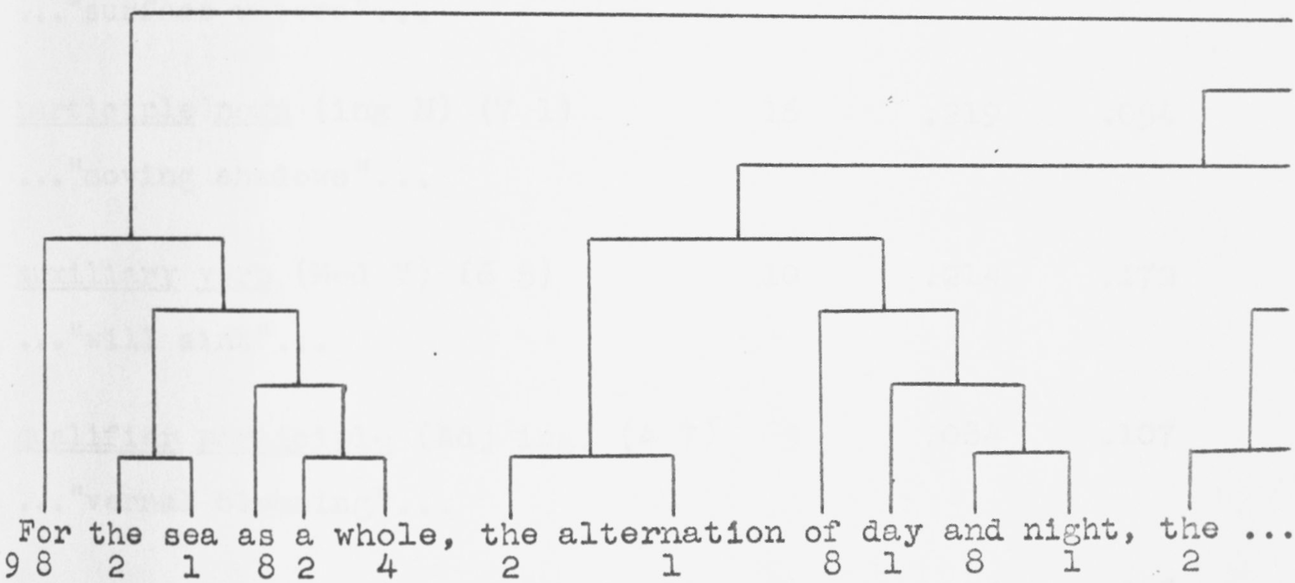

Figure 4 Partial Sentence Diagram — Word Classes Marked

Each construction consisting of two adjacent words was marked as to type, in the manner discussed in Chapter II. Then the constructions were grouped according to type; the results are presented in Table 3. Pairs of word classes of each type found in the text are listed with their frequencies and their conditional probabilities of co-occurrence. All examples of pairs of word classes are from the English text.

[11] Nida, op. cit.

Right Endocentric Constructions

A B	FREQ	P(B|A)	P(A|B)
qualifier noun (Adj N) (4 1)	81	.677	.216
..."changeless eternity"...			
noun noun (N N) (1 1)	21	.089	.089
..."surface waters"...			
participle noun (ing N) (7 1)	16	.219	.054
..."moving shadows"...			
auxiliary verb (Mod V) (6 5)	10	.214	.172
..."will sink"...			
qualifier participle (Adj ing) (4 7)	3	.084	.107
..."vernal blooming"...			
qualifier particle (Adv Prep) (4 8)	2	.118	.046
..."far below"...			
noun participle (N ing) (1 7)	1	.054	.215
..."spring flowering"...			

Left Endocentric Constructions

A B	FREQ	P(B|A)	P(A|B)
participle adverb (ing Adv) (7 6)	1	.076	.101
..."slipping down"...			

A B	FREQ	P(B|A)	P(A|B)
verb pronoun (V Pron) (5 3)	1	.035	.075
..."lose them"...			
verb adverb (V Adv) (5 6)	1	.035	.043
..."rises again"...			
particle particle (Conj Conj) (8 8)	1	.042	.043
..."as though"...			

Exocentric Constructions

A B	FREQ	P(B|A)	P(A|B)
determiner noun (Det N) (2 1)	158	.602	.451
..."the sea"...			
particle noun (Prep N) (8 1)	41	.194	.159
..."at sea"...			
auxiliary participle (Aux ing) (6 7)	11	.185	.139
..."is changing"...			
particle verb (Prep V) (8 5)	8	.039	.137
..."to claim"...			
verb participle (V -ed) (5 7)	7	.142	.129
..."are stirred"...			
particle qualifier (Conj Adj) (8 4)	7	.052	.135
..."or green"...			

A B	FREQ	P(B|A)	P(A|B)
particle pronoun (Prep Pron) (8 3)	5	.039	.300
..."of them"...			
verb qualifier (V Adj) (5 4)	3	.119	.084
..."are different"...			
particle participle (Prep ing) (8 7)	3	.056	.182
..."of existing"...			
noun verb (N V) (1 5)	1	.117	.494
..."spring progresses"...			
verb particle (V Prep) (5 8)	1	.369	.103
..."come up"...			
verb noun (V N) (5 1)	1	.035	.008
..."find shelter"...			

Paratactic Constructions

A B	FREQ	P(B|A)	P(A|B)
qualifier participle (Adj ing) (4 7)	2	.084	.107
..."endless, hurrying"...			
participle qualifier (ing Adj) (7 4)	1	.087	.067
..."pushing green"...			

Table 3 Construction Types and Pairs of Word Classes

CHAPTER V

RESULTS

In order to estimate the values of k_1 and k_2 as defined in Chapter III, the graphs shown on pages 31-33 were prepared. Using the data given in Chapter IV, the pairs of word classes were marked along with their frequencies of occurrence. Figure 5 shows a region of right endocentricity determined by the expression

$$0 < P(A|B)/P(B|A) \leq k_1 = 2/3$$

Within this area, one finds 99 or 74% of the 134 right endocentric constructions. The region of exocentricity, as shown in Figure 6, is determined by the expression

$$2/3 < P(A|B)/P(B|A) \leq k_2 = 1$$

Of the 246 exocentric constructions, 220 or 90% are found within this region. Finally, as shown in Figure 7, a region of left endocentricity is determined by the expression

$$1 < P(A|B)/P(B|A) < \infty$$

All four of the left endocentric constructions fall within this area.

Consider the distribution of all the construction types throughout the three regions. These data are summarized in

Table 4 — a three by three table. The frequency for each construction type in each region is given.

Hypothesized Regions	Construction Types		
	Right Endocentric	Exocentric	Left Endocentric
$0 < P(A\|B)/P(B\|A) \leq 2/3$	99	2	0
$2/3 < P(A\|B)/P(B\|A) \leq 1$	31	220	0
$1 < P(A\|B)/P(B\|A) < \infty$	4	24	4

Table 4 Observed Distribution

If the hypotheses were perfectly confirmed, Table 4 would show frequencies greater than zero only along the diagonal from the upper left to the lower right. That is, the distribution of construction types would be as shown in Table 5.

Hypothesized Regions	Construction Types		
	Right Endocentric	Exocentric	Left Endocentric
$0 < P(A\|B)/P(B\|A) \leq 2/3$	134	0	0
$2/3 < P(A\|B)/P(B\|A) \leq 1$	0	246	0
$1 < P(A\|B)/P(B\|A) < \infty$	0	0	4

Table 5 Hypothesized Distribution

On the other hand, if the hypotheses were totally unconfirmed, one might find as an extreme case the artificial distribution shown in Table 6.

Hypothesized Regions	Construction Types: Right Endocentric	Exocentric	Left Endocentric
$0 < P(A\|B)/P(B\|A) \leq 2/3$	44.7	82	1.33
$2/3 < P(A\|B)/P(B\|A) \leq 1$	44.7	82	1.33
$1 < P(A\|B)/P(B\|A) < \infty$	44.7	82	1.33

Table 6 Artificial Distribution

The values $k_1 = 2/3$ and $k_2 = 1$ were found by trial and error. That is, the regions were hypothesized in such a way as to achieve maximum observed frequencies along a diagonal, and minimum frequencies elsewhere, in a three by three table. In other words, the attempt was made to make Table 4 appear as much as possible like Table 5 and least like Table 6.

One may obtain a measure of the degree to which the hypotheses are confirmed by first calculating chi-square.[12]

$$\chi^2 = \sum_{i,j} \frac{\left[n_{ij} - (n_{i.}\, n_{.j}/n)\right]^2}{n_{i.}\, n_{.j}/n}$$

[12] Alexander M. Mood, Introduction to the Theory of Statistics (New York, 1950), pp. 277-280.

It is the case that $n_{i.} = \sum_j n_{ij}$ and $n_{.j} = \sum_i n_{ij}$, where n_{ij} is the number of observations in the i^{th} row and the j^{th} column of the table.[13] Given χ^2, a measure of the degree of confirmation is provided by the Tschuprow coefficient

$$T = \sqrt{\frac{\chi^2}{N\sqrt{(p-1)(q-1)}}}$$

where N is the total number of observations, p is the number of columns in the table, and q is the number of rows in the table.[14] For a three by three table

$$T = \sqrt{\frac{\chi^2}{2(N)}}$$

For the artificial distribution of construction types shown in Table 6, the computed values are $\chi^2 = 0$ and $T = 0$. In this case, the value of T is at its lower limit.

The upper limit of T is one. This value is obtained if $\chi^2 = 2(N)$. There are 384 observations; thus $2(N) = 768$. For the hypothesized distribution shown in Table 5, $\chi^2 = 506$. The inequality of 2(N) and χ^2 in this case is the result of the unequal numbers of different construction types. If the three values along the diagonal are each changed to 128, then $\chi^2 = 768$. To obtain a T value of one for the perfect confirmation of the hypotheses which is depicted in Table 5, one may use the equation

$$T^* = \sqrt{\frac{(768/506)\chi^2}{2(N)}}$$

---

[13] Ibid.

[14] Maurice G. Kendall, The Advanced Theory of Statistics (London, 1947), p. 320.

Given 384 observations, it follows that

$$T^* = \sqrt{\frac{\chi^2}{506}}$$

For the observed distribution shown in Table 4, $\chi^2 = 284$ and $T^* = .749$, which indicates a relatively high degree of confirmation.

Though generally large, the exact value of T* clearly depends, for one thing, upon the designation of word classes. The data for the statistical analysis presented here were obtained by first placing each word of an English text in one and only one word class. A word which was derivationally a noun, for example, was marked as a noun wherever it occurred. If one, however, allows a word to be a member of more than one word class, the value of T* may be much larger. Consider the 21 occurrences of noun noun. The noun on the left is an attribute of the other; it functions in the same manner as an adjective. If one forms a new word class called adjectival, including these attributive nouns and adjectives, then the number of right endocentric constructions found in the hypothetical area is increased from 99 to 120. The number of exocentric constructions in the hypothetical right endocentric area is thereby reduced to 10. Given these changes $\chi^2 = 361$ and $T^* = .845$.

The paratactic constructions present a problem similar to the one encountered with noun noun as a right endocentric construction. It will be remembered that it was found useful to consider nouns which function as adjectives to be adjectivals. Similarly, it will be useful to consider the participles found in each of the paratactic constructions, qualifier participle and <u>participle qualifier</u>, to be adjectivals. Each participle is coordinate with the other constituent, <u>qualifier</u>, and the function of each is certainly attributive. Participles so considered, the paratactic constructions become one: <u>qualifier qualifier</u>, for which $P(B|A)$ is 0.033 and $P(A|B)$ is 0.033. The order of the constituents is reversible. Clearly, $P(B|A)$ is equal to $P(A|B)$ for either order. It follows that $P(AB)=P(BA)$. At least one pair of word classes may be said to occur as a paratactic construction, and thus confirm the hypothesis made in Chapter III regarding paratactic constructions.

The ratio of conditional probabilities between adjacent word classes in a text provides a minimally simple description of co-occurrence. Yet this function seems to provide a numerically adequate characterization of English binary syntactic constructions. However, better and possibly more interesting results would have been obtained if the procedure used in this study had been extended to take into account three or more word classes in sequence, rather than just two. This extension would require a larger computer than the one which performed the calculations for the study presented here. It has become a truism that the diversity of English sentences is "infinite." Any calculation concerning sequences of linguistic units quickly overtaxes the storage capacity of a limited storage computer such as the IBM 1620. Finally, one could wish for better techniques for the analysis and interpretation of multiple classifications in $n \times n$ tables. For this, one must await advances in the science of statistics.

LIST OF WORKS CITED

Carson, Rachel L. "The Changing Year," The Sea Around Us (New York: Signet, 1961).

Chomsky, Noam. "On the Notion 'Rule of Grammar'," The Structure of Language: Readings in the Philosophy of Language, eds. J.A. Fodor and J.J. Katz (Englewood Cliffs, N.J.: Prentice-Hall, 1964).

Curme, George O. Syntax (New York: D.C. Heath and Co., 1931).

Harris, Zellig S. Structural Linguistics (Chicago: University of Chicago Press, 1951).

Hockett, Charles F. A Course in Modern Linguistics (New York: Macmillan, 1958).

Kendall, Maurice G. The Advanced Theory of Statistics (London: C. Griffin, 1947).

Mood, Alexander M. Introduction to the Theory of Statistics (New York: McGraw-Hill, 1950).

Nida, Eugene A. A Synopsis of English Syntax (Norman, Okla.: Summer Institute of Linguistics, 1960).

Webster's New International Dictionary of the English Language, 2nd ed. (Springfield, Mass.: G. & C. Merriam Co., 1951).

APPENDIX

98218242181812182121821578217822419821154921821667978118717821482121821518192115
82158218218585824718219682358218219158241824182184187418712218178241821278411821
81568241241821352418216352419213582183548588166522787198215687871352241879821881
15218218198241818241211677249624158576872418941816778218241378218213781136782167
88213821367212718218241821915782121815768666821683986812154724115821241814858418
19681158182182124162454821181821191651218324157983822115241928215841881657621481
92116536216436935241818218521874218187441821812187824183216578213652822418196215
21821181818218181218321165936521824782121824127121821198247482124182158592158419
21158682188222411821982418213541818271821918181654848424178218241817828211921571
82182241968622181578241824182193527182182124182419418241821582178241837184196821
21152419821818211748882718121848282115821821963382165682418215221818471821968154
18158212124818181811782182418219824847218215479215282488322419865418382168241811
35854182182296821212165784418211587241872182767682166878247198821658241824217815
68216769821215787138378218413365852193524178824171852718212178821821821217224181
12172182182198181216762418219621182782215368218219247875782182191582441821121787
881811819152181858218191217166724182415824124717821824191455821824132184167216582
35382352198278218277828211828211821158241819827182124111581741818181821524174882
4198124411667821821838219241582184174186466724184181351821858399 18 CARDS

Table 7 Word Class Sequence for the Entire Text

A	B	P(B|A)	P(A|B)
1	1	.089	.089
1	2	.095	.126
1	3	.038	.350
1	4	.010	.033
1	5	.117	.494
1	6	.070	.376
1	7	.054	.215
1	8	.416	.511
1	9	.106	.735
2	1	.602	.451
2	2	.054	.054
2	3	.007	.050
2	4	.259	.610
2	5	0.000	0.000
2	6	0.000	0.000
2	7	.046	.139
2	8	.025	.023
2	9	.003	.018
3	1	0.000	0.000
3	2	.119	.018
3	3	.047	.050
3	4	0.000	0.000
3	5	.333	.160
3	6	.190	.115
3	7	.095	.043
3	8	.166	.023
3	9	.047	.037
4	1	.677	.216
4	2	.033	.014
4	3	.008	.025
4	4	.033	.033
4	5	.016	.022
4	6	.008	.014
4	7	.084	.107
4	8	.118	.046
4	9	.016	.037
5	1	.035	.008
5	2	.238	.072
5	3	.035	.075
5	4	.119	.084
5	5	0.000	0.000
5	6	.035	.043

A	B	P(B|A)	P(A|B)
5	7	.142	.129
5	8	.369	.103
5	9	.023	.037
6	1	0.000	0.000
6	2	.114	.028
6	3	.028	.050
6	4	.028	.016
6	5	.214	.172
6	6	.114	.115
6	7	.185	.139
6	8	.285	.066
6	9	.028	.037
7	1	.219	.054
7	2	.131	.043
7	3	0.000	0.000
7	4	.087	.067
7	5	.010	.011
7	6	.076	.101
7	7	.032	.032
7	8	.395	.120
7	9	.043	.075
8	1	.194	.159
8	2	.551	.605
8	3	.039	.300
8	4	.052	.135
8	5	.039	.137
8	6	.019	.086
8	7	.056	.182
8	8	.042	.043
8	9	.003	.018
9	1	.150	.021
9	2	.188	.036
9	3	.075	.100
9	4	.037	.016
9	5	0.000	0.000
9	6	.188	.144
9	7	.018	.010
9	8	.339	.060
9	9	0.000	0.000

Table 8 Conditional Probabilities of Cooccurrence for the Various Word Classes

GRAMMATICAL ANALYSIS BY STATISTICAL METHODS

BY

WILLIAM TAYLOR CLAY

SUBPROGRAM ONE

```
ZZJOB  Z                       SCHOOL OF ARTS AND SCIE
ZZFORX
      DIMENSION IC(80),FIP(9,9),PBA(9,9),PAB(9,9)
      DO 1 I=1,9
      DO 1 J=1,9
      FIP(I,J)=0
      PBA(I,J)=0
      PAB(I,J)=0
    1 CONTINUE
    2 READ 3, (IC(I), I=1,80)
    3 FORMAT (80I1)
      DO 6  K=1,79
      I=IC(K)
      J=IC(K+1)
      IF(9-I) 5,4,5
    4 IF(9-J) 5,7,5
    5 FIP(I,J)=FIP(I,J)+1.
    6 CONTINUE
      GO TO 2
    7 DO 11 I=1,9
      FIT=0
      DO 8 J=1,9
      FIT=FIT+FIP(I,J)
    8 CONTINUE
      DO 11 K=1,9
      IF(FIT) 10,9,10
    9 PBA(I,K)=2
      PAB(K,I)=2
      GO TO 11
   10 PBA(I,K)=FIP(I,K)/FIT
   11 CONTINUE
```

```
      DO 14  J=1,9
      FIT=0
      DO 12 I=1,9
      FIT=FIT+FIP(I,J)
   12 CONTINUE
      DO 14  K=1,9
      IF(FIT) 13,14,13
   13 PAB(K,J)=FIP(K,J)/FIT
   14 CONTINUE
      DO 20  I=1,9
      DO 20  K=1,9
      IF(PBA(I,K)-2.) 15,16,15
   15 IF(PAB(I,K)-2.) 18,16,18
   16 PUNCH 17, I,K
   17 FORMAT (10X,I2,3X,I2)
      GO TO 20
   18 PUNCH 19, I,K,PBA(I,K),PAB(I,K)
   19 FORMAT (10X,I2,3X,I2,2(10X,F6.3))
   20 CONTINUE
      CALL EXIT
      END
ZZZZ
```

Table 9   FORTRAN II Computer Program for the Calculations Performed for this Study