Ralph D. Beebe

THE FREQUENCY DISTRIBUTION OF ENGLISH SYNTAGMS

1. INTRODUCTION

People use language principally as a means of communication, each language community determining the scope of the particular language it uses. Within the language, various sets and subsets may exist depending upon regional and social differences. Most studies of such subsets have been concerned with lexical variations. In this paper, a methodology is outlined by which the frequency usage of syntactic structures can be measured for subsets of a language, and for a language as a whole.

In order to establish homogeneity features within a subset, reliance is made on calculations of chisquare on batches of sentences within the subset. Before considering whether or not a more sensitive test should be applied to detect significant differences within and among subsets of a language, it seemed advisable to determine the frequency distribution of syntagms throughout particular genres. An extension of the methodology to accomplish this objective is described and the results for one type of syntactic structure within one genre are tabulated. These results indicate that the frequency distribution of syntagms within these limitations approximates to a Poisson distribution.

2. METHODOLOGY

There are five main steps in the methodology: the isolation of each and every syntagm in the corpus of sentences; the assembly of the isolated syntagms into a matrix, each row of which contains syn-

Acknowledgements. The author wishes to thank his colleagues in the Departments of Information Science and Mathematical Statistics at Monash University, and the staff of the Monash University Computer Centre for assistance and advice.

tagms having the same head-character; the comparison, for particular syntagms, of the frequencies with which the syntagms occur in different genres; the record-keeping of the quantities of sentences in which particular syntagms occur once, twice,...*n* times; and the comparison of those quantities with standard statistical distributions.

The first step in the isolation of the syntagms is the transcription of the syntactic structure of the sentences into linearized form, preparatory to their being used in the computer.

A criterion used for the selection of a grammar on which to base the syntactic structure description was that, for purposes additional to those required for this investigation, namely the support of remedial English-teaching courses, a manual system of parsing had to be provided for use by other users reasonably familiar with traditionally-accepted terminology. The methodology was designed to provide a means whereby comparisons of genres for frequency-usage of syntactic structures could be made, so that teaching courses for particular genres could emphasize the structures most frequently used in those genres.

The methodology had to be such that users quite unfamiliar with computer techniques could prepare the data in suitable form for feeding into the computer, and use prepared programs to obtain the necessary results.

For the above reasons, a dependency grammar was selected, details of which are shown in Figs. 1 and 2.

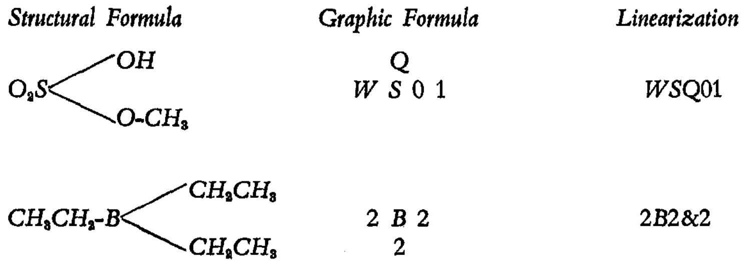
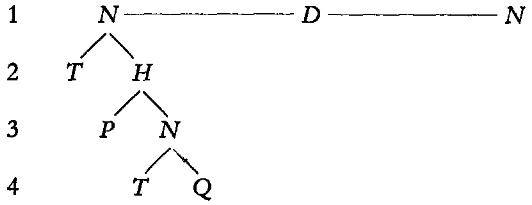


Fig. 1. *Wiswesser Line Notation*

The method for linearizing syntactic trees, such as those shown typically in Fig. 2, was based on the line-formula chemical notation developed by W. J. Wiswesser (W. J. WISWESSER, 1954; E. G. SMITH, 1968). The simple rules applying to the Beebe line notation are that

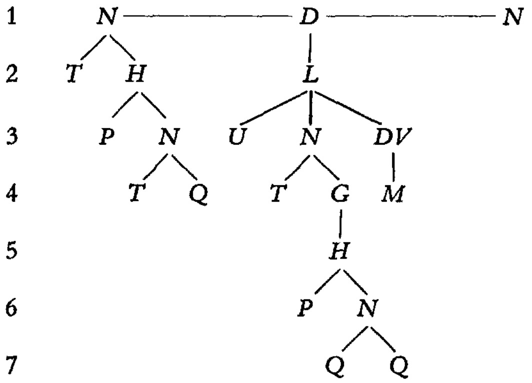
Sentence: *The policies of the new government supported isolationism.*



Linearization: *NT-HP-NT-Q/D$N*
Level Numbers: 12123234340101

Fig. 2 (*a*). *Typical Linearization of a Sentence*

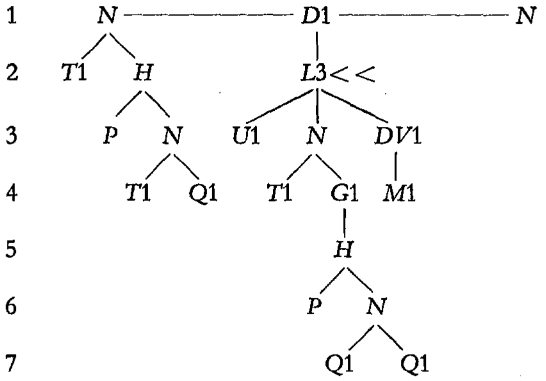
Sentence: *After the problems arising from previous international agreements had been happily resolved, the policies of the new government supported isolationism.*



Linearization: *NT-HP-NT-Q/DLU-NT-GHP-NQ-Q——DVM$N*
Level Numbers: 121232343401232343456567676543233401

Fig. 2 (*b*). *Typical Linearization of a Sentence*

starting from the first character of a sentence group such as the subject group, verb group, or object group, and proceeding from the leftmost path of the tree, characters are recorded in sequential order until a termination of a branch of the tree is reached. At that point, a return to the nearest branching point in the tree is made, recording a dash for each step retraced from the terminating point to the branching point. The next leftmost path is then followed with the same retracing process until all the paths in the group have been followed to their terminations. A boundary-marker is then inserted in the linearization to mark the boundary between the subject group and the verb group, and the procedure repeated for the verb group until all the paths in that group have similarly been followed. Another boundary marker is then inserted in the linearization, and the next group is then processed until the end of the sentence has been reached.



Linearization: *NT*1-*HP*-*NT*1-Q1/*D*1*L*3<<*U*1-*NT*1-G1*HP*-*N*Q1-Q1——*DV*1 *M*1$*N*

Level Numbers: 12212323443440112222332344344565677677654323334401

Fig. 2 (*c*). *Typical Linearization of a Sentence*

The data is then key-punched in the following sequence: a header card for the number and title of the batch of sentences; a card for the natural text of the sentence; a card for the linearization.

This data is then fed into the computer, where it is processed by a very simple program which contains the following instructions: read the characters in the linearization one by one, and if the first character is alphabetical, prepare a record string commencing with the numeral 1. If the next character in the linearization is also alphabetical, increment this numeral by one and enter the result in the next position in the record string. Continue so doing until a dash is met. For a dash, reduce the numeral by one, and for a group boundary marker, record a zero. The effect of this program is to produce for each sentence a record string of numerals, as shown in Fig. 2, corresponding to the level numbers of the nodes in the syntactic tree.

The header card, the natural text card, and the linearization card are then transferred to magnetic tape records together with the numerical record string representing the assignment of the levels. These records are then available for calling upon for processing for the isolation of the syntagms.

The isolation procedure commences by first drawing the records of the linearizations, one at a time, into an array, while at the same time drawing the corresponding record of level numbers into another array. Both arrays are then scanned simultaneously under the following conditions: each character of the linearization is transferred as it is scanned, to a further array, while the level numbers corresponding to the characters are checked until a number is encountered less in value than the number of the first character in the linearization. The transferred string of characters then represents the syntagm which has been isolated.

The scanning process is then repeated starting from the next character in the linearization until the next syntagm has been isolated. The procedure is continued starting from successive characters in the linearization until all the characters in the linearization have been processed.

As they are produced, the isolated syntagms are fed into the appropriate rows of a matrix, so that each row contains only syntagms having the same head character. Before each syntagm is entered into a row of the matrix, the previous entries in that row are scanned to see if a similar syntagm has already been entered. Only when there has been no previous entry of a similar syntagm, is a new syntagm entered in that row. If there has been a similar entry, a frequency count adjacent to the entry is incremented by one.

Optionally, each new entry can have an additional entry inserted adjacent to it of the dominant node in the syntactic tree. The depth of the syntagm, calculated by subtracting the level of the head character from the maximum level of the syntagm, is also entered, as is the length, calculated by adding the number of characters, including dashes, in the line notation for the syntagm.

A control feature is included, which permits the matrix either to be cleared at the end of a batch, or to stay filled so that syntagms from successive batches of the same genre can be added, to produce cumulative results.

Rows of the matrix can be selectively printed.

At the end of a run, a summary of the entries in each row of the matrix is made. The type/token ratio is calculated by dividing the number of discrete entries by the total number of syntagms in the row.

Graphic plotting programs have been written to display scatter diagrams of length against depth, of quantities of discrete and of total syntagms against quantity of sentences, and of type/token ratios against quantity of sentences. See Figs. 3 and 4.

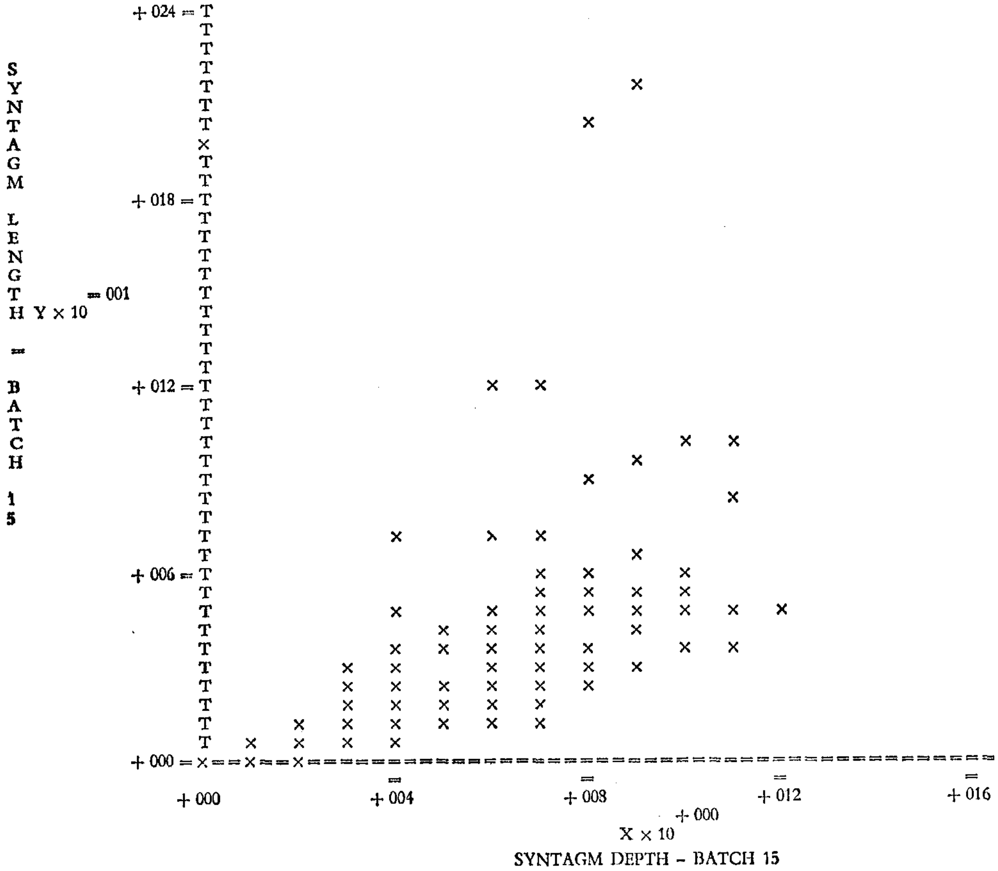


Fig. 3. *Syntagm Length/Depth (Newspaper Editorial)*

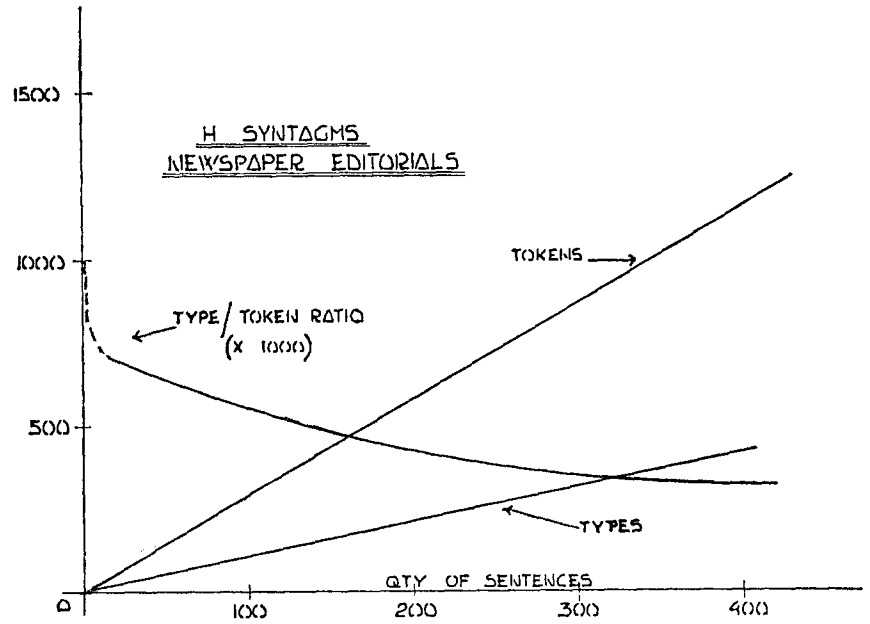


Fig. 4. *Type/Token Ratios*

The system of parsing allows for the inclusion of any degree of sub-categorization of syntactic categories. See Tables 1 and 2.

TABLE 1 – *Symbol Code for Sentence Structure Linearization*

A	Appositive	K	infinitive	V	passiVe verb-form
B	Being verb	L	cLause	W	Weighing, costing or having verb
C	Coordinator	M	Modifier	X	non-finite eXpression
D	Doing verb	N	Noun	Y	numeralitY
E	En verb-form	O	cOmpound verb	Z	possessive
F	Factitive	P	Preposition	/	subject/verb separator
G	InG verb-form	Q	Qualifier	$	verb$object separator
H	prepositional pHrase	R	pRonoun	=	verb=complement separator
I	Intensifier	T	deTerminer	%	verb%supplement separator
J	reJector	U	sUbordinator	+	truncation indicator

Table 2 – *Examples of Sub-Categorization*

L1	Restrictive Relative Clause	R1	Personal Pronoun
L2	Unrestrictive Relative Clause	R2	Relative Pronoun
L3	Adverbial Clause	R3	Reflexive Pronoun
L4	Noun Clause	R4	Emphatic Pronoun
L5	Noun Clause in Apposition	R5	Indefinite Pronoun
M1	Normal Form of Adverb	Q1	Normal Form of Adjective
M2	Comparative Form of Adverb	Q2	Comparative Form of Adjective
M3	Superlative Form of Adverb	Q3	Superlative Form of Adjective

A refinement of the parsing allows the inclusion of left and right arrowheads to display departures from arbitrarily-selected norms of the linear order of the words of the natural text. Comparisons of quantities of syntagms with and without such arrowheads provide measures of the actual departures from the norms, and thereby provide an evaluation of the validity of the norms. The levels-assignment program contains a provision for not changing the level when scanning a subcategorization number or an arrowhead.

As each syntagm is isolated and fed into the matrix, a corresponding entry is made in a duplicate matrix, this second entry containing the number of times that the syntagm has been found in the one sentence. When the end of the sentence has been reached, there is also entered at that point in the second matrix an adjacent entry of an increment in an array at a position corresponding to the number of times that the syntagm was found in that sentence. The second matrix therefore contains a growing record of the number of sentences in which the syntagm appears once, twice, ...*n* times.

From preliminary investigations, it was found that *n* rarely exceeded ten. The provision of an array with fourteen elements was therefore ample for the keeping of such records.

Usage frequencies extracted from the first matrix can be used as input data for a chisquare calculation program to ascertain the homogeneity of usage within a batch of sentences, or to find if there are significant differences between genres. See Tables 3-6.

TABLE 3 – *Comparison of Editorials from Australian Newspapers*

Entity	1	2	3	4	5	6	7	8	9	10
Batch	1	2	3	4	5	6	7	8	9	10
HP-N	16	9	12	9	14	15	10	10	6	6
HP-NT	5	6	5	3	3	6	8	2	2	3
HP-NQ	5	3	2	5	6	5	3	2	4	7
HP-NZ	1	1	4	2	3	1	1	1	1	0
HP-NN	0	1	0	0	1	0	1	0	0	3
H[P]-NT	1	1	3	1	0	2	1	0	3	1
HP-NT-Q	1	1	1	2	0	4	1	2	3	4
HP-NT-N	3	0	0	2	1	0	1	1	1	0
HP-NT-HP-N	0	0	1	1	2	2	1	0	2	2
HP-R	1	5	3	2	1	3	2	1	0	1
HP-CN-N	0	0	2	0	1	1	5	1	0	0

Chisquare = 101.88 Degrees-of-freedom = 90 Significance N.S.

TABLE 4 – *Comparison of Editorials from Australian Newspapers*

Entity	1	2
Batch	1-10	11-20
HP-N	152	107
HP-NT	60	43
HP-NQ	37	42
HP-NZ	14	15
HP-NN	5	6
H[P]-NT	13	13
HP-NT-Q	18	19
HP-NT-N	10	9
HP-NT-HP-N	5	11
HP-R	20	19
HP-CN-N	9	10

Chisquare = 9.76 Degrees-of-freedom = 10 Significance N.S.

TABLE 5 – *Comparison of Higher School Certificate (HSC) Essays*

Entity	1	2	3	4
Batch	51	52	53	54
HP-N	10	10	10	18
HP-NT	7	13	10	5
HP-NQ	2	4	5	4
HP-NZ	1	7	5	12
HP-NN	3	1	0	1
H[P]-NT	3	0	0	1
HP-NT-Q	4	3	3	6
HP-NT-N	0	2	0	0
HP-NT-HP-N	0	1	2	0
HP-R	5	7	4	3
HP-CN-N	4	1	6	3

3 Entities, having total occurrences fewer than 5, have been ignored

Chisquare = 27.33 Degrees-of-freedom = 21 Significance N.S.

TABLE 6 – *Comparison of HSC Essays against Australian Newspaper Editorials*

Entity	1	2
Batch	$\frac{1\text{-}20}{3}$	51-54
HP-N	81	48
HP-NT	34	35
HP-NQ	26	15
HP-NZ	10	25
HP-NN	4	5
H[P[-NT	9	4
HP-NT-Q	12	16
HP-NT-N	6	2
HP-NT-HP-N	5	3
HP-R	13	19
HP-CN-N	6	14

Chisquare = 26.31 Degrees-of-freedom = 10 Significance **

3. RESULTS

No significant differences were found in the relative proportions of particular syntagms (prepositional phrases, in the cases displayed) among Batches 1-10 of one genre (newspaper editorials), although differences significant at the .05 level were noted among Batches 11-20 of the same genre. When the totals for each group of ten batches were compared, however, there were no significant differences.

Four batches of another genre, Higher School Certificate Examination (University Entrance) Essays, produced no significant differences within the genre, but did produce significant differences at the .01 level when compared with the same number of sentences from the previous genre.

It appears from the results obtained that the frequency usage of syntactic structures is a valid criterion of distinctiveness, and can be used as a means of comparing genres.

The entries in the second matrix give the frequency distribution of each syntagm within the corpus. See Table 7.

TABLE 7 – *Distribution of Syntagms in* 415 *Sentences*

Syntagm	Quantity of Syntagms per Sentence										
(in Beebe Line Notation)	0	0	1	2	3	4	5	6	7	8	9
	(Quantities of Sentences)										
HP-N	227	188	129	45	11	3	0	0	0	0	0
NP-NT	323	92	82	7	3	0	0	0	0	0	0
HP-NQ	343	72	66	5	0	0	0	0	0	0	0
HP-NT-Q	379	36	35	1	0	0	0	0	0	0	0
HP-NT-N	395	20	20	0	0	0	0	0	0	0	0
HP-NZ	388	27	25	2	0	0	0	0	0	0	0
HP-R	378	37	35	2	0	0	0	0	0	0	0
H[P]-NT	389	26	26	0	0	0	0	0	0	0	0
HP-NN	404	11	10	1	0	0	0	0	0	0	0
HP-NT-HP-N	397	18	18	0	0	0	0	0	0	0	0
HP-CN-N	396	19	19	0	0	0	0	0	0	0	0

Comparisons of specimen syntagms (prepositional phrases) are given in Table 8. It will be observed that the actual distributions are almost exactly in conformity with the distributions expected from a Poisson distribution.

This study must be considered as a pilot study to prove the methodology. Further processing of a much larger corpus will no doubt reveal more illuminating results.

TABLE 8 (*a*) – *Syntagms with More than One Occurrence in at Least one Sentence.*

	0	1	2	3	4
HP-N	227	129	45	11	3
HP-NT	323	82	7	3	0
HP-NQ	343	66	6	0	0
HP-NT-Q	379	35	1	0	0
HP-R	378	35	2	0	0

The above data gained from analysis of twenty newspaper editorials were tested for goodness of fit to the Poisson distribution, with the Chisquare test. In all (five) cases, the fit was acceptable at the 10 % level of significance.

TABLE 8 (*b*).

n_i is the number of sentences in which a given syntagm appears x_i times.

e_i is the expected value of n_i using the Poisson distribution.

TABLE 8 (*c*).

HP-N

x_i	0	1	2	3	4
n_i	227	129	45	11	3

$\Sigma n_i = 415$

$\Sigma x_i n_i = 264$

$$\text{Mean} = \frac{\Sigma x_i n_i}{\Sigma n_i} = \frac{264}{415} = 0.63613$$

TABLE 8 (*d*).

HP-N

x_i	0	1	2	3	4 and over
n_i	227	129	45	11	3
e_i	219.68	139.74	44.45	9.21	1.92
diff	7.32	10.74	0.55	1.79	1.08
$\frac{diff^2}{e_i}$	0.244	0.827	0.007	0.273	0.608

Chisquare = 1.959 Degrees of freedom = 3 Significance = N.S.

TABLE 8 (*e*).

HP-NT

x_i	0	1	2 and over
n_i	323	82	10
e_i	322.23	81.53	11.24
diff	0.77	0.47	1.24
$\frac{diff^2}{e_i}$	0.002	0.003	0.136

Chisquare = 0.141 Degrees of freedom = 1 Significance = N.S.

TABLE 8 (*f*).

HP-NQ

x_i	0	1	2 and over
n_i	343	66	6
e_i	343.90	64.64	6.46
diff	0.90	1.36	0.46
$\frac{diff^2}{e_i}$	0.002	0.058	0.033

Chisquare = 0.093 Degrees of freedom = 1 Significance = N.S.

TABLE 8 (*g*).

HP-NT-Q

x_i	0	1	2 and over
n_i	379	35	1
e_i	379.60	33.84	1.56
diff	0.60	1.16	0.56
$\frac{diff^2}{e_i}$	0.001	0.040	0.201

Chisquare = 0.242 Degrees of freedom = 1 Significance = N.S.

TABLE 8 (*h*).

HP-R

x_i	0	1	2 and over
n_i	378	35	2
e_i	377.78	35.50	1.72
diff	0.22	0.50	0.28
$\frac{diff^2}{e_i}$	0.001	0.070	0.046

Chisquare = 0.116 Degrees of freedom = 1 Significance = N.S.

REFERENCES

E. G. SMITH, *The Wiswesser Line-Formula Chemical Notation*, New York, 1968.

W. J. WISWESSER, *A Line-Formula Chemical Notation*, Thomas Y. Cromwell Company, 1954.