[*Eighth Annual Round Table Meeting on Linguistics and Language Studies*,
Georgetown University, 1957]

The Use Of SEAC In Syntactic Analysis

Richard B. Thomas
National Bureau of Standards

The purpose of this structure search is to determine whether the syntactical patterns of English sentences, expressed symbolically, show that a relatively small number of such patterns represents a significantly large number of sentences.

The routines written for SEAC (National Bureau of Standards Electronic Automatic Computer) examine the structure of sentences within a corpus of expository prose chosen from scientific and technical writings. SEAC performs several functions in the search: (a) accurate high-speed tabulation; (b) precise comparison of data; (c) compression of coded data in terms of syntactical equivalence relationships. Preparation of the material includes coding the elements of the source sentences functionally according to the scheme shown in Figure One. This is but one of many feasible analytical schemes, and it represents a rather gross cut made according to traditional grammar. Each functional element to be coded is given a notation comprising one number and one letter, as shown in Figure Two. The notations (maximum 9) of each sentence then become SEAC input.

Primary Search

Except for tabulation, the program rejects the letters in the codes. (The letters are included in the raw data so that they may be available for other programs.) The computer reduces the first incoming sentence code to its numerical pattern (e.g., 56416500000) and stores the pattern. Each subsequent candidate sentence is reduced in the same way and its numerical pattern is compared with all stored patterns. If a candidate pattern is identical with a stored pattern, a recurrence tally of 1 is added into the least-significant place of the stored pattern and the candidate pattern is rejected. But if the candidate pattern proves to be unique, it is stored along with the others. Check routines are included in the programming of this preliminary phase to reject data erroneously prepared.

When all sentences have been processed, the resultant unique primary patterns, with their tallies, emerge from SEAC via high-speed magnetic wire, along with tallies showing (a) the number of sentences processed, (b) the number of patterns stored as unique, (c) the number of sentences rejected because of errors in preparation or inscription, and (d) the number of patterns having 1 digit, 2 digits, ...9 digits.

Each time 50 sentences have been processed, SEAC prints out the number of unique patterns being held in storage.

Compressed Search

The unique primary patterns are then fed back into the computer for compression and a 1 is added to every pattern tally so that each tally will show the actual number of occurrences. The routine first rejects all-but-one of any digit that is contiguously repeated within a pattern. For example, 444166655 becomes 4165, but 414656 remains unchanged. The basis for compression in this manner is the assertion of equivalence relationships whereby "The little red hen clucks" (44416) is here considered to function syntactically as "The hen clucks" (416) or "The hen will cluck" (4166), etc. A compressed pattern is thus construed as a basic form from which all corresponding primary patterns could be developed by regular structural transformation.

Since a number of patterns whose primary forms were different would probably be identical in their compressed form, the compressed patterns are then compared, each with all others. When SEAC finds that two compressed patterns are identical, their respective tallies are added together; the sum is stored in the least-significant places of the first pattern in question and the other pattern is cleared to zeros. At the end of this operation, the unique compressed patterns with (some new) tallies are printed out, together with (a) the number of compressed patterns stored as unique and (b) the number of unique compressed patterns having 1 digit, 2 digits, … 9 digits. Separate routines are then employed to list the patterns in numerical order, as shown in Figure Three.

Results of the Search

The original corpus of 1002 sentences is a very small sample. The curve of cumulative occurrences (Figure Seven) shows little tendency to reach zero slope, as unique primary patterns were still occurring at a nearly fixed rate. The distribution of patterns according to the number of digits (notations) per pattern is shown in Figure Four. The components of the raw and processed data are shown in Figure Five. Figures Six and Seven express the rate of occurrence of new patterns.

The 1002 sentences yielded 541 unique primary patterns. Of these, the five most common are listed here in sequence, followed by the number of occurrences (in parentheses) and a sample sentence for each:

41665	(19)	The dog has run across the street.
414665	(17)	The dog with floppy ears has run across the street.
162	(16)	Dogs eat bones.
16434	(15)	Fido is the dog with floppy ears.
165	(14)	Fido ran across the street.
1665	(14)	Fido has run across the street.

The total number of sentences (95) shared by these patterns is less than 10 percent of the corpus.

Compression of the 541 primaries yielded 247 unique compressed patterns. The five most common are listed here, as above; the parenthetical number is the number of original sentences represented by the compressed form:

4165	(80)	The dog ran across the street.
165	(62)	Fido ran across the street.
41465	(56)	The dog with floppy ears ran across the street.
54165	(35)	Finally the dog ran across the street.
416424	(28)	The dog ate the bone which he had dug up.

The total number of sentences (261) shared by these compressed patterns represents only 26 percent of the corpus. The first four compressed patterns listed above contain at least one adverbial element in every case and lack objects or predicative nominatives; these patterns represent 233 sentences or 23 percent of the corpus.

The results must be viewed as specifically inconclusive because the corpus is small. But the technique of compression appears valid and useful for examining possible “base” or “kernel” forms of syntax.

The computer is admirably suited to this type of search. To duplicate the search on other types of equipment would require considerably more time and more complex operations, especially in the comparison and compression phases. The total SEAC running time for this program was about 40 minutes.

FIGURE ONE. Coding Scheme for Syntactic Analysis

LEXICAL UNITS	PHRASES	DEPENDENT CLAUSES
(A)	(B)	(C)
1, 2, 3 - Noun, Pronoun	Noun Infinitive Gerundive Prepositional	Noun
1 A - Subject	1 B - Subject	1 C - Subject
2 A - Object (d.&i.)	2 B - Object (d.&i.)	2 C - Object (d.&i.)
3 A - Pred. Nom.	3 B - Pred. Nom.	3 C - Pred. Nom.
4 A - Adjective	4 B - Adj. Infinitive Participial Prepositional	4 C - Adjective
5 A - Adverb	5 B - Adverb Infin. Prep.	5 C - Adverb
6 - Verb of Independent Clause		
6 A - Main verb	6 B - Auxiliary	6 C - Modal Auxil.

NOT CODED: Connectives (relative pronouns, coordinating conjunctions, subordinating conjunctions, conjunctive adverbs); absolutes; appositives; interjections; non-functional expletives; internal structure of phrases and dependent clauses; and elements which modify portions of "B" or "C" structures (except verbs).

Such structures have been omitted from the coding because they do not affect the basic structure of the independent clause.

FIGURE TWO. SAMPLES OF CODES AND PATTERNS

```
                                                    (SEAC Input)
   5-A   6-B  4-A   1-A   6-A 5-A
1. When does the balloon go up?                     5A6B4A1A6AØ
                                                    5AØØ ØØ ØØ ØØ Ø-
 ——— 2-C ———         1-A  6-A
2. Whatever Lola wants, Lola gets.                  2C1A6AØØØØØ-

  1-A 6-C  6-B  6-B   6-A   5-A   5-A
3. It could have been solved more simply.           1A6C6B6B6AØ
                                                    5A5AØØ ØØ ØØ Ø-
          6-A 4-A  2-A  5-A
4. Polly, put the kettle on.                        6A4A2A5AØØØ-

   4-A     4-A       4-A      4-A    1-A
5. The electronic automatic digital computer
       6-A   4-A 2-A                                4A4A4A4A1AØ
     dropped a bit.                                 6A4A2AØØ ØØ Ø-
```

Primary patterns formed from the foregoing sentences:

1. 5 6 4 1 6 5 Ø Ø Ø Ø Ø
2. 2 1 6 Ø Ø Ø Ø Ø Ø Ø Ø
3. 1 6 6 6 6 5 5 Ø Ø Ø Ø
4. 6 4 2 5 Ø Ø Ø Ø Ø Ø Ø
5. 4 4 4 4 1 6 4 2 Ø Ø Ø

Compressed patterns formed from the primary patterns:

1. 5 6 4 1 6 5 Ø Ø Ø Ø Ø
2. 2 1 6 Ø Ø Ø Ø Ø Ø Ø Ø
3. 1 6 5 Ø Ø Ø Ø Ø Ø Ø Ø
4. 6 4 2 5 Ø Ø Ø Ø Ø Ø Ø
5. 4 1 6 4 2 Ø Ø Ø Ø Ø Ø

FIGURE THREE

Unique Compressed Patterns - Listed - Last Two Digit Places
Show Number of Occurrences

16000000 14
62000000 09
65000000 01

14600000 02
16200000 23
16300000 07
16400000 11
16500000 62
21600000 02
41600000 03
51600000 06
56200000 01
61400000 02
62500000 06
64100000 02
64200000 01
65100000 01

14250000 01
14630000 02
14650000 09
15620000 02
15650000 03
16240000 04
16250000 05
16340000 02
16350000 01
16420000 07
16430000 07
16450000 06
16520000 03
16530000 01
16540000 05
16560000 04
21620000 01
41450000 01
41460000 03
41620000 05

41630000 03
41640000 07
41650000 80
51460000 01
51620000 08
51640000 11
51650000 18
54160000 05
56160000 01
56410000 07
56420000 02
61650000 01
62520000 01
64140000 05
64150000 01
64240000 04
64250000 06
65410000 01
65420000 01

14625000 02
14642000 01
15642000 02
16245000 01
16414000 03
16415000 01
16424000 22
16425000 15
16434000 20
16435000 02
16462000 01
16545000 03
16562000 03
16565000 09
41416000 01
41426000 01
41456000 01
41462000 02
41463000 13

41464000 04
41465000 56
41565000 03
41624000 04
41625000 02
41642000 12
41643000 05
41645000 09
41654000 07
41656000 03
46165000 01
51464000 01
51465000 02
51565000 01
51625000 02
51642000 05
51645000 01
51652000 01
51654000 04
51656000 01
54146000 06
54156000 01
54161000 01
54162000 02
54164000 07
54165000 35
54641000 01
56161000 01
56165000 01
56414000 03
56416000 01
56425000 01
56434000 01
56561000 01
61434000 01
61625000 02
61642000 01
64145000 02
64163000 01

64165000 01
64245000 02
65414000 01
65415000 01

14642100 01
14642400 03
14643400 03
14656400 01
14656500 01
15642400 03
16242500 02
16424500 06
16425200 01
16425400 03
16454200 01
16542400 02
16562400 02
16564200 02
16564500 02
41456300 01
41456400 02
41456500 03
41462400 02
41462500 03
41463500 01
41464200 01
41464300 05
41464500 07
41465200 01
41465400 08
41465600 01
41562400 01
41562500 01
41565600 01
41624500 01
41642400 28
41642500 11
41643400 12

41643500 01
41646500 01
41652400 01
41654200 01
41654300 01
41654500 09
41656200 02
41656400 01
41656500 13
45416500 01
51464500 02
51562500 01
51614300 01
51642400 08
51643400 02
51652400 01
51656400 01
51656500 03
54146400 04
54146500 13
54162500 02
54164100 01
54164200 03
54164300 01
54164500 01
54165400 04
54165600 02
54641500 01
54643500 01
61642400 01
65642400 01
14564240 01
14642450 02
14656450 01
15642450 01
16424240 01
16454250 01
16564150 01
16564240 01

FIGURE THREE (Continued)

41456420 01

41456540 01

41456560 01

41462450 01

41464240 11

41464250 02

41464340 07

41465450 02

41465640 01

41465650 07

41562540 01

41564240 01

41565650 01

41642450 01

41645420 01

41653450 01

41654240 01

41654340 02

41656420 02

41656430 01

45436410 01

51456540 01

51465450 01

51654240 01

51656420 02

54146420 02

54146450 03

54146540 02

54146560 02

54164240 04

54164250 02

54164340 01

54165250 01

54165450 01

54165650 03

64541450 01

14656424 01

16464245 01

16565424 01

41456425 01

41464245 01

41465643 01

41545625 01

41645424 02

41645434 01

45414642 01

54146565 02

54154642 01

54165424 01

165 454345 01

414645424 01

414656435 01

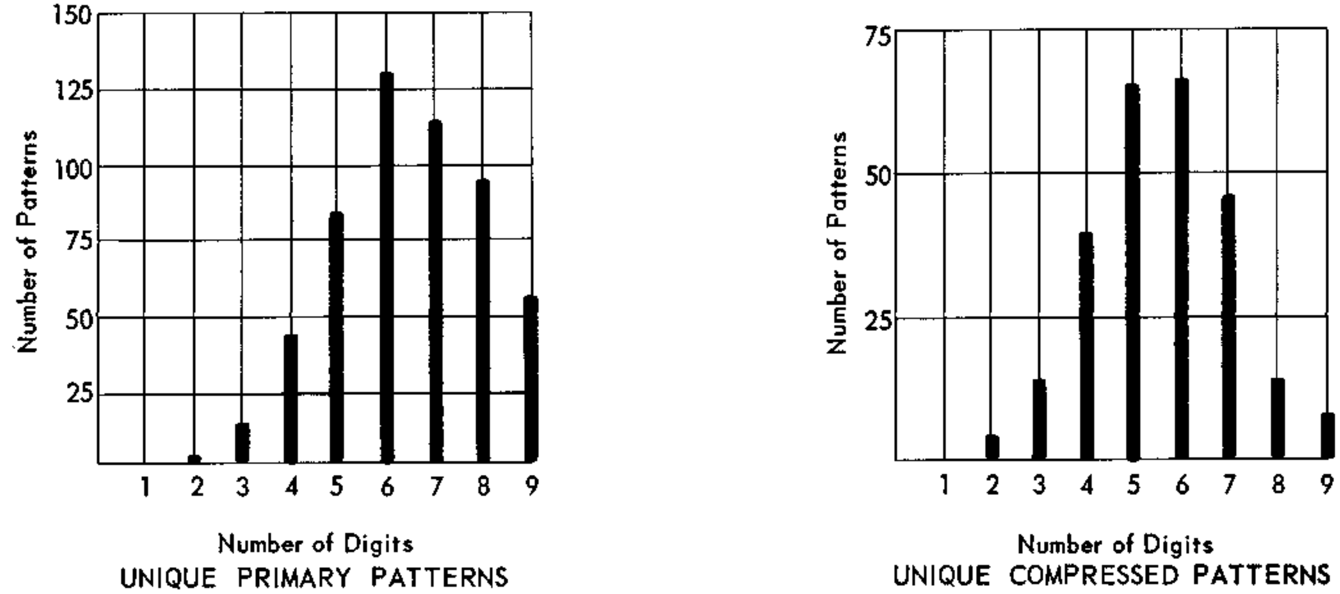
454162424 01

541464543 01

541465643 01

564156424 01

FIGURE FOUR. Distribution of Patterns according to Number of Digits



UNIQUE PRIMARY PATTERNS

UNIQUE COMPRESSED PATTERNS

FIGURE FIVE. Components of Raw and Processed Data

	(A)	(B)	(C)	Total
(1)	970	10	15	995
(2)	279	23	48	350
(3)	98	9	12	119
(4)	1368	435	54	1857
(5)	305	645	147	1097
(6)	1002	432	154	1588
Total	4022	1554	430	
COMPONENTS OF RAW DATA				

Components of Unique Primary Patterns:	(1)	529	(4)	1162
	(2)	211	(5)	676
	(3)	69	(6)	844
Components of Unique Compressed Patterns:	(1)	243	(4)	414
	(2)	112	(5)	281
	(3)	39	(6)	300

FIGURE SIX. Percentage of Unique Primary Patterns in Each Increment of 100 Sentences Processed

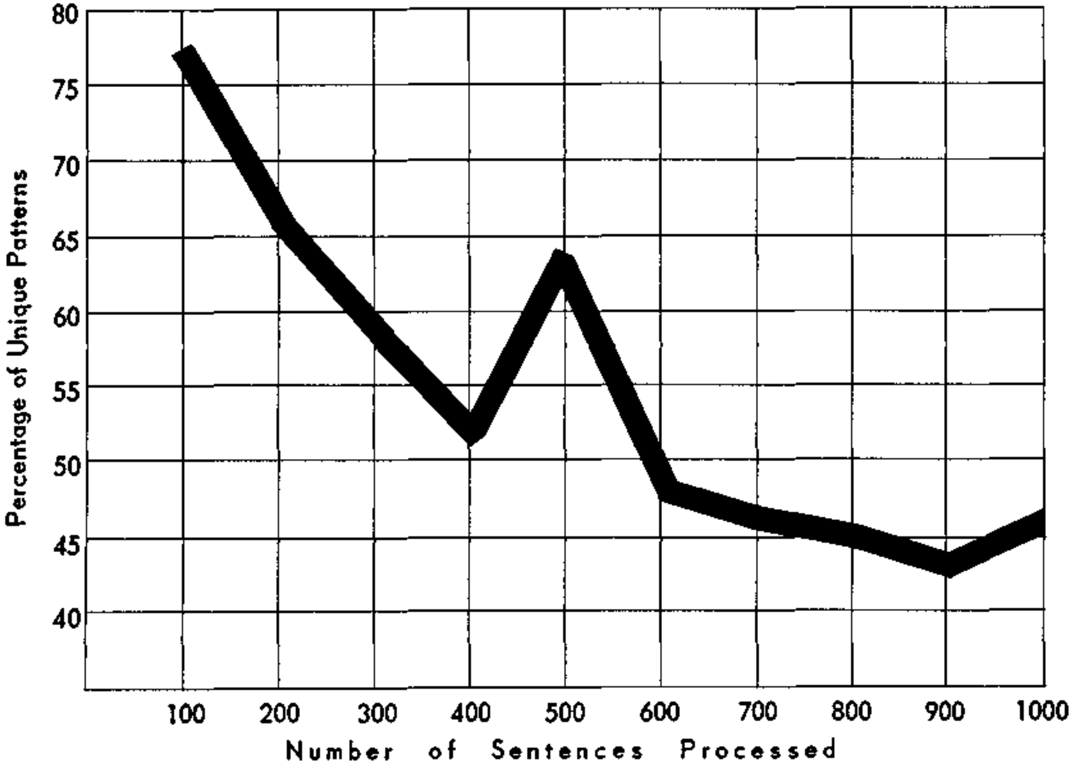
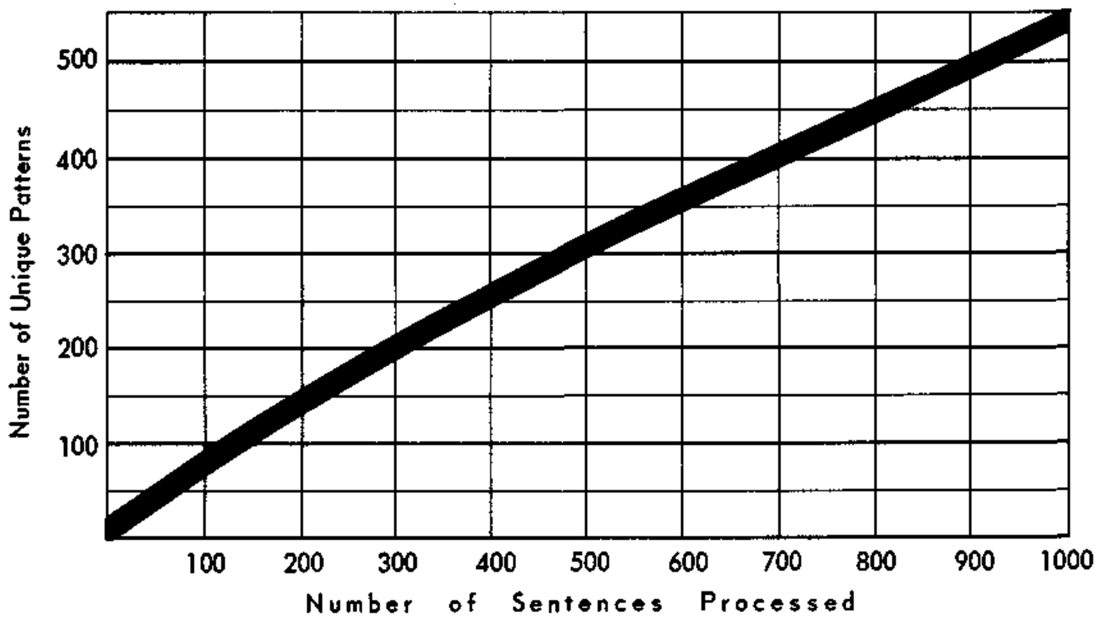


FIGURE SEVEN. Cumulative Number of Unique Primary Patterns.


Number of Unique Patterns
Number of Sentences Processed