DOCUMENT RESUME

ED 076 081 FL 004 106

AUTHOR Beebe, Ralph D.
TITLE The Determination of the Frequency of Syntactical Patterns in Present-Day Written Australiar English.
INSTITUTION Monash Univ., Clayton, Victoria (Australia).
PUB DATE 71
NOTE 16p.; In Linguistic Communications, 3, 1971

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Computer Programs; Descriptive Linguistics; *English; Language Patterns; *Language Research; Linguistic Theory; *Research Methodology; *Sentence Structure; Statistical Analysis; Surface Structure; *Syntax; Verbs; Written Language
IDENTIFIERS Australia

ABSTRACT

Confronted with the problem of determining the frequency of syntactical patterns in present-day written Australian English, the author employs a method of analysis which produces an output in the form of a two-dimensional line diagram showing all the syntagms comprising the sentence under analysis. For the remaining problem of sorting the diagrams into divisions and sub-divisions of syntagms, the author advocates the use of a method of linearization used for sorting structural diagrams of chemical compounds. A description of the methodology is provided along with an explanation of its adaptation to language analysis. (VM)

In: Linguistic Communications; 3, 1971.

THE DETERMINATION OF THE FREQUENCY OF SYNTACTICAL PATTERNS IN PRESENT-DAY WRITTEN AUSTRALIAN ENGLISH. Report dated 15th May, 1970.

Ralph D. Beebe
Monash University

In advising the writer on this project, Professor U. G. E. Hammarström had suggested that the frequency of English syntagms could be determined by examining a corpus of English sentences, dividing them first into sentence types, then sub-dividing the sentence types further, according to his system of syntactic terminology (Hammarström 1967). A manual sorting of sentences in that way would have been a process of great magnitude.

In searching for a more elegant method, the writer first aimed at a computing program which would have automatically analysed sentences into their syntagms. He hoped to be able then to modify the program to sort the sentences and their syntagms. Although such an analysis program had been developed by Bratley, Thorne and Dewar (1967), it proved to be incapable of being run on any computer in Australia due to computer-language incompatibilities. An alternative FAP program (Sager 1967) evolved at New York University did not provide an output in adequate form for the purposes of the project. No other programs were currently available.

As a manual analysis seemed therefore, inevitable, the writer turned his attention to other large-scale manual analysis work done previously. A fruitful area appeared to be in studies of the writing of children. Notable examples were those of La Brant (1933), Strickland (1962), Loban (1963), and Hunt (1965). These studies showed a

growing tendency towards a more formal delineation of sentence structure, but all indicated that a more complete study could not be made until some more detailed system of analysis had been devised.

The writer then turned his attention to using a method of analysis which he had himself developed primarily for teaching purposes. This method gave an output in the form of a two-dimensional line diagram showing all the syntagms comprising the sentence analysed. It was essentially a surface-structure analysis using a form of dependency grammar.

The problem still remained, however, of how to sort such diagrams into divisions and sub-divisions of syntagms.

The writer had observed that a somewhat similar problem of sorting chemical compounds expressed in the form of molecular-structure diagrams had been solved in various ways throughout the world. He selected one way devised by the U.S. Army Biological Laboratories (Wiswesser 1954) and currently popular with many U.S. drug companies.

The selected method first reduced the two-dimensional diagrams of molecular structure to linear strings of symbols, and then sorted the strings by conventional computer methods.

From the principles employed by Wiswesser, the writer succeeded in learning how to linearize his own two-dimensional diagrams of sentence structure, and the remainder of the project can now be completed by writing a suitable computer program for sorting the linear strings of symbols.

Further aid may be obtained in this phase of the project by studies of the programs used in organic chemistry and of new languages for the computer such as PL/1 and SNOBOL devised especially for sorting strings of symbols. Compatibility with the Monash University computer complex will be an overriding consideration.

A statistical analysis of the results will determine the required syntagm frequencies, and the syntagms might then be allotted hierarchical distinctions using Hammarström's proposed terminology.

By examining several different genres of present-day written Australian English, the syntagm frequencies among the genres can be compared, thus reducing the influence of errors in the syntactical analysis.

BRIEF DESCRIPTION OF THE WISWESSER SYSTEM

The method of linearization used for sorting structural diagrams of chemical compounds in the United States, devised by Wiswesser (1954), and revised by Smith (1968). first translates all conventional two-letter atomic symbols into single letters, and also provides single-letter identification symbols for groups of atoms forming commonly-occurring radicals. For example the halogens, bromine and chlorine, normally expressed by the symbols Br and Cl, become E and G, so that the following list emerges:

E bromine atom
F fluorine atom
G chlorine atom
H hydrogen atom (although H is mostly unexpressed)
I iodine atom

Added to the list are the following symbols for various groups:

Q hydroxyl group, -OH.

V carbonyl connective, $-\overset{O}{\overset{\|}{C}}-$

(carbon connected to three other atoms)

W nonlinear (branching) dioxo group as in $-NO_2$, $-SO_2-$. Not used for linear (unbranched) structures such as CO_2, SiO_2, NO_2, SO_2.

M imino group, $-\overset{H}{\overset{|}{N}}-$.

Z amino group $-NH_2$.

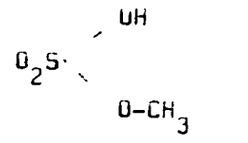
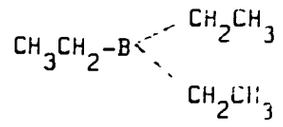
Numerals are used to show the number of carbon atoms in unbranched alkyl chains or segments.

Thus the following unbranched compounds are expressed in linear notation as shown:

(1)	$CH_3-\overset{O}{\overset{\|}{C}}-CH_3$	1V1
(2)	$CH_3CH_2-O-CH_2CH_3$	202
(3)	$HO-CH_2CH_2-OH$	Q2Q
(4)	$O_2N-CH_2-O-CH_2-NO_2$	WW1O1WW
(5)	$H_2N-CH_2CH_2CH_2-NH_2$	Z3Z

For branched compounds, a graphic formula is first interposed between the structural formula and the eventual linearization, rules being laid down for linearizing the graphic formula. In the following simplified description, these rules are abbreviated to the point of inadequacy, but they serve to demonstrate the basis for the eventual set of rules devised by the writer for his sentence diagrams.

Thus observe the following linearizations:

Structural Formula	Graphic Formula	Linearization
		WSQ01
		2B2&2

The rules state that the linearization of a graphic formula is performed by citing the symbols along a main chain until a branching point is reached, digressing along the branch, then returning, after the end of the branch is reached, to the main chain, inserting an extra symbol (&) before resuming the symbols of the main chain. If the branch terminates in a symbol which cannot be followed in any case along that branch by other symbols, then it is a 'terminating' symbol, and there is no need to insert the resumption symbol (&) when continuing along the main chain.

In the first example above, Q is a 'terminating' symbol known to be such by an organic chemist, so there is no need to use the resumption symbol when continuing along the main chain after dealing with the branch chain. In the second example, however, the branch symbols are not 'terminating' symbols, as they can each be followed along their branches by other symbols, information which again is known by the organic chemist who encodes the diagram.

Thus the inherent technical knowledge of the encoder enables him to encode correctly.

The Wiswesser system covers not only unbranched and branched chains, but also cyclic compounds, utilizing in all some 250 rules. In the encoding of sentence diagrams, however, only a few of the rules of the Wiswesser system are needed. These selected rules have been drastically simplified in the brief description given above. Their application to sentence-diagram encoding will now be described in detail.

APPLICATION OF THE WISWESSER SYSTEM TO SENTENCE DIAGRAMS

The appendix gives some examples of the encoding of sentence diagrams. The four basic types of English sentences, distinguished by their verb types, are encoded as follows:

(1)	John shuddered	N+D
(2)	John injured Jim	N+D+N
(3)	John was sick	N+B+Q
(4)	They elected John captain	R+FN+N

The D in the graphic formula of sentence (4) above has been omitted from the linearization. This has been done because D is an essential element of a factitive predicator F and can therefore be assumed to be present without being specifically mentioned. Its omission is similar to the omission of the hydrogen symbol from the alkyl group in the Wiswesser system.

A similar omission of the symbol for the preposition can be made in every prepositional phrase since every such phrase must commence with a preposition. It is only necessary to insert the symbol H for the phrase and go straight on to consider the other elements apart from the preposition. The normal element accompanying the preposition in the phrase is the noun, but that element can be replaced by various substitutes such as the pronoun, or non-finite verb. If the noun is present, it can be omitted from the linearization; only the symbol for its substitute need be included when such a substitute is present. On the other hand, any dependencies of the noun must be shown, as in sentences (5) and (6).

(5)	John struck Jim in anger	N+DH+N
(6)	John struck Jim in great anger	N+DHQ+N

There can be no ambiguity concerning the Q in sentence (6) since an adjective cannot be used to describe a preposition. The Q must be a dependency of the N in the phrase H.

This is an example of the inherent technical knowledge of the encoder enabling him to encode correctly, a parallel operation to that of the organic chemist encoding chemical compounds by the Wiswesser system.

The advantages of the linearization system become more evident when more complicated sentences are considered. See Appendix, sentences (7) and (8).

It is clear that the sorting of the strings is, comparatively speaking, the least problematical part of the project.

BIBLIOGRAPHY

1. Bratley, P., Dewar, H. and Thorne, J. P. 1967

Recognition of Syntactic Structure by Computer. Nature, Vol.216,December 9

See also: Hamish Dewar, Paul Bratley, and James Peter Thorne, 1969

A Program for the Syntactic Analysis of English Sentences, CACM,Vol.12,No.8,Aug.

2. Hammarström, G. 1967

On Linguistic Terminology, Actes du Xe Congrès International des Linguistes, Bucarest, 28 août-2 septembre, Vol.I,pp.321-325

3. Hunt, Kellog W. 1965

Grammatical Structures Written at Three Grade Levels, NCTE Research Report No. 3, National Council of Teachers of English, 508 South Sixth Street, Champaign, Illinois, 61822

4. LaBrant, Lou L. 1933

Study of Certain Language Developments of Children in Grades Four to Twelve inclusive, Genetic Psychology Monographs Vol.XIV,No.5,Nov.

5. Loban, Walter 1963

The Language of Elementary School Children NCTE Research Report No. 1 National Council of Teachers of English, 508 South Sixth Street, Champaign, Illinois

6. Sager, N. 1967

Syntactic Analysis of Natural Language, Advances in Computers, Vol.8

7. Smith, Elbert G. 1968

The Wiswesser Line-Formula Chemical Notation
McGraw-Hill Book Company.

8. Strickland, Ruth G. 1962

The Language of Elementary School Children: Its Relationship to the Language of Reading Textbooks and the Quality of Reading of Selected Children. Bulletin of the School of Education, Indiana University, Vol.38.No.4,July

9. Wiswesser, W. J. 1954

A Line-Formula Chemical Notation Thomas Y. Cromwell Company:New York

APPENDIX

1. SYMBOL CODE FOR STRUCTURAL DIAGRAM

Adj	-	Adjective
Adv	-	Adverb
AG	-	Appositive Group
C	-	Conjunction
CG	-	Coordinate Group
Cl	-	Clause
Comp	-	Complement
D	-	Degree
Exp	-	Non-finite Expression
F	-	Frequency
FV	-	Finite Verb
Fac Pred	-	Factitive Predicator
M	-	Manner
N	-	Noun
Neg	-	Negation
NFV	-	Non-finite Verb
O	-	Object
P	-	Place
Phr	-	Phrase
Pn	-	Pronoun
Prep	-	Preposition
S	-	Subject
Sup	-	Supplement
T	-	Time

2. SYMBOL CODE FOR GRAPHIC FORMULA AND LINEARIZATION

A	Appositive
B	Being verb
C	Coordinator
D	Doing verb
E	past participlE
F	Factitive predicator
G	inG verb-form
H	prepositional pHrase
I	Intensifier
J	reJector
K	infinitive
L	cLause
M	Modifier
N	Noun
O	cOmpound verb
P	Preposition
Q	Qualifier
R	pRonoun
T	deTerminer
U	sUbordinator
V	passiVe verb-form
W	having, costing, or Weighing verb
X	non-finite eXpression
Y	numeralitY
Z	possessive
&	return to main chain
+	governing relationship

3. EXAMPLES OF THE ENCODING OF SENTENCE DIAGRAMS

(1) Sentence: John shuddered

Structural Diagram:

S(N) (John) —— FV (shuddered)

Graphic Formula:

N —— D

Linearization: N+D

(2) Sentence: John injured Jim

Structural Diagram:

S(N) (John) —— FV (injured) —— O(N) (Jim)

Graphic Formula:

N —— D —— N

Linearization: N+D+N

(3) Sentence: John was sick

Structural Diagram:

S(N) (John) —— FV (was) —— Comp (Adj) (sick)

Graphic Formula:

N —— B —— Q

Linearization: N+B+Q

(4) Sentence: They elected John captain

Structural Diagram:

S(Pn) (They) —— Fac Pred —— O(N) (John)

[FV (elected) + Comp(N) (captain)]

Graphic Formula:

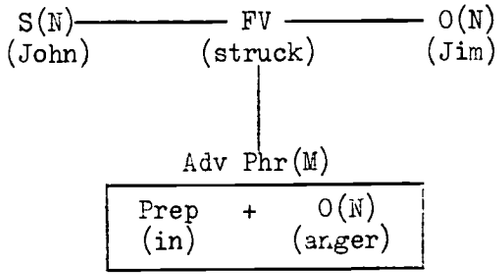
R —— F —— N

[D + N]

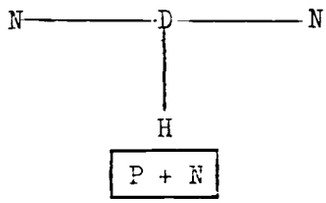
Linearization: R+FN+N

(5) Sentence: John struck Jim in anger

Structural Diagram:



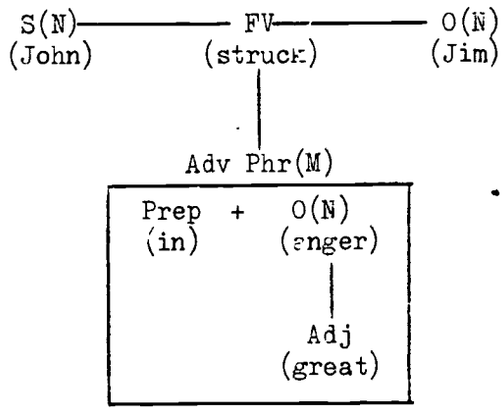
Graphic Formula:



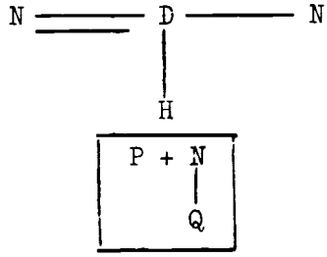
Linearization: N+DH+N

(6) Sentence: John struck Jim in great anger.

Structural Diagram:



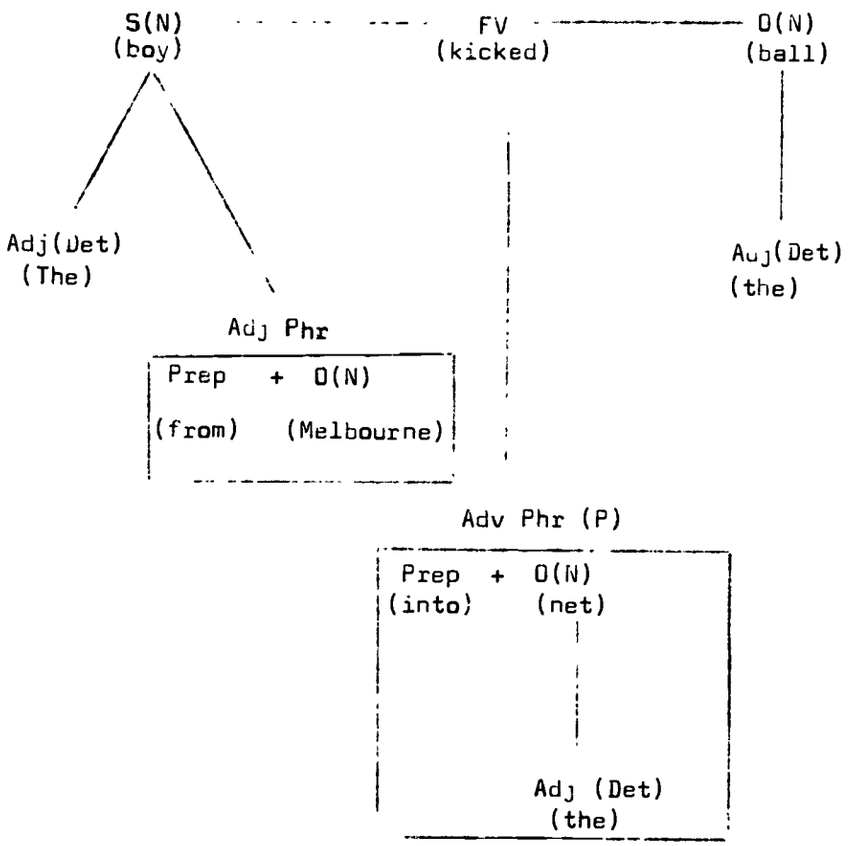
Graphic Formula:



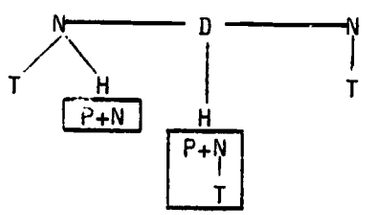
Linearization: N+DHQ+N

(7) Sentence: The boy from Melbourne kicked the ball into the net.

Structural Diagram:



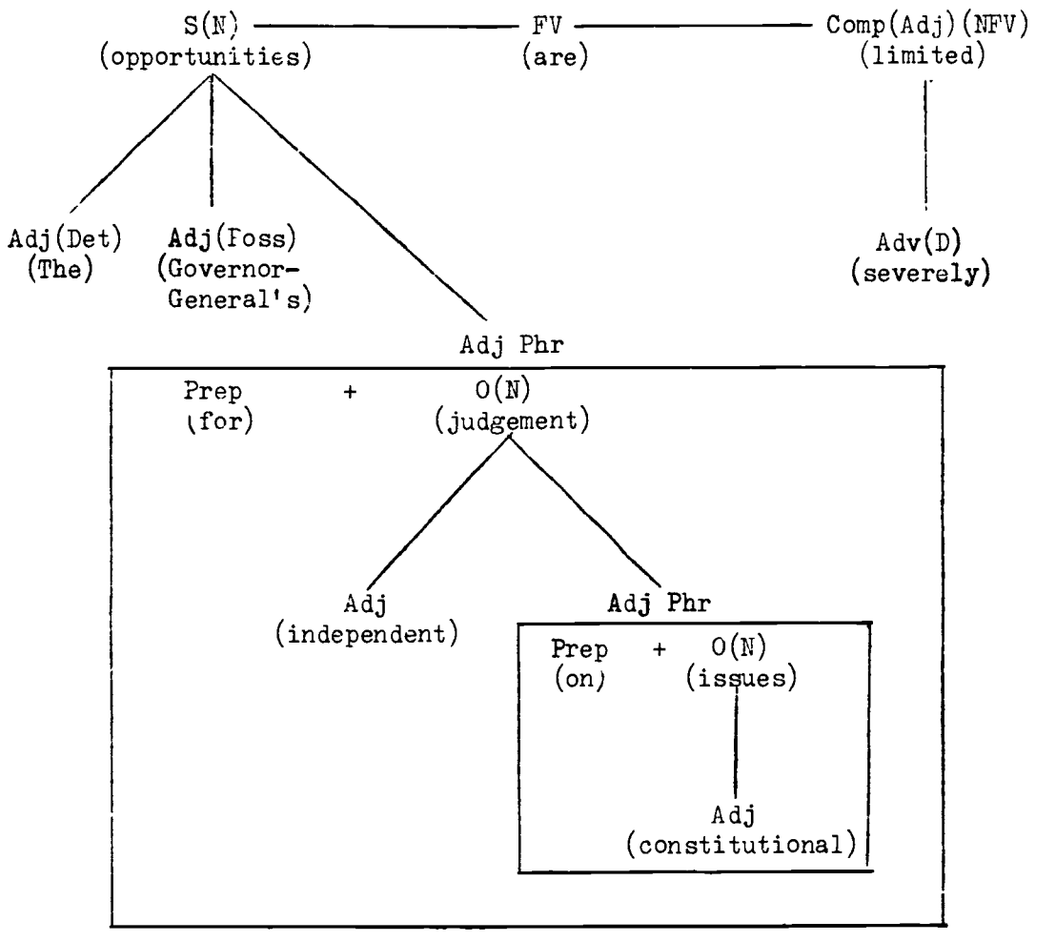
Graphic Formula:



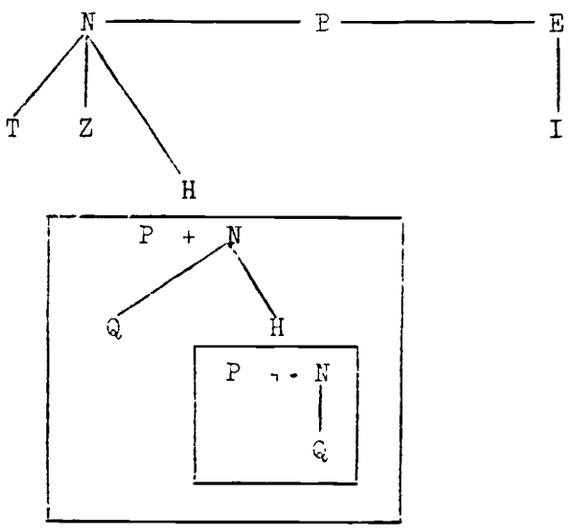
Linearization: NTH+DHT+NT

(8) Sentence: The Governor-General's opportunities for independent judgement on constitutional issues are severely limited.

Structural Diagram:



Graphic Formula:



Linearization: NTZ&HQ&HQ+B+EI