STATISTICAL PARSING BY COMPUTER

K. H. V. BOOTH
University of Saskatchewan
Saskatoon, Saskatchewan
Canada

Several projects are at present under way in Canada, under the auspices of the National Research Council, with the object of producing English to French translation on a computer.

One of these is at the University of Saskatchewan, where we are attempting to produce a translation which although not perfect, can be processed by a post-editor with less effort, and less knowledge of English, than would be required to do the complete translation.

To do this, a text of 20 000 words, taken from the Canada Year Book, 1962, is being used as a basis for the dictionary, idiom list, etc., and also as a test sample.

This paper gives results on a statistical parsing method which seems promising both from the point of view of accuracy and of

GRAMMATICAL CODING

N	Noun
V	Verb
A	Adjective
B	Adverb
R	Pronoun
D	Determiner (the, a, etc.)
L	Linking Verb
X	Auxiliary Verb (to be, to have)
I	Intensifier (very, much, etc.)
P	Preposition
E	Relative Pronoun (which, who, etc.)
S	Subordinator (because, than, when, etc.)
C	Connector (and, but, etc.)
G	Negative (not)
Z	Preverb (to, by)
Y	Exclamation

Fig. 1

speed. Parsing is defined here as coding each word of a sentence into one of the 16 grammatical categories in Fig. 1. These are based on the classification used by W. S. Stolz. [1]

Each word of the text was first coded by hand. In doing this, proper names and most hyphenated words were treated as groups and coded as a single word. The coded text was then analysed to give the number of occurrences of groups of codes. Group sizes were taken from 1 word up to 5 words, and the analysis was made within sentences. Thus 5 tables were obtained, of which Fig. 2 taken from the quadrigram list, is typical. More extensive examples of these tables will be found in "Machine aided translation with a post-editor".[2]

QUADRIGRAM GLOSSARY	
CAN,	35
CANA	7
CANC	15
CANE	3
CANI	2
CANL	8
CANP	62
CANV	14
CANX	26

Fig. 2

A dictionary of the text was then prepared, giving all possible grammatical classes to which each word could belong even if it were not used in that sense in the text. As an example, "in" is entered as "BPVZ", that is "adverb, preposition, verb, preverb", although its use as a verb is very rare.

The text was also processed for idioms, an idiom being defined as any group of words which cannot be translated separately to make sense. Proper names were included in the idiom list although it is proposed to treat these by a separate routine eventually.

The complete text was then re-coded automatically, idioms being removed first, and the statistical tables used as a basis for coding. In this experiment, quadrigrams were used and Fig. 3 shows a typical output after coding. The entries under the various columns need some explanation: col. 2 indicates where idioms have been located, thus in the example, "coastal waters" and "such as" have been found on the idiom list. W and P signify word and punctuation. Col. 3 gives the dic-

[1] Stolz, W. S., *Syntactic constraint in spoken and written English*, Ph. D. thesis, Wisconsin 1964.

[2] Booth K. H. V., "Machine aided translation with a post-editor", in *Machine Translation*, Ed. A. D. Booth, North-Holland Publishing Co., Amsterdam 1967.

tionary information on grammatical codes. "Of", for example, may be either preposition or pre-verb (PZ). Col. 4 gives the code finally selected by the program. Col. 5 gives the four choices for each word, one for each quadrigram in which it figured. Col. 6 gives the "probability" of the quadrigram starting with that word being correct. The latter figure was calculated by dividing the number of occurrences of the quadrigram by the total number of possible quadrigrams found in the tables and selecting that with the highest probability. Thus the quadrigram "marine biology and meteorology" was given the coding "ANCN", as can be seen by following the diagonal line down as shown, and this group of codes actually occurred 56 times in the tables, as against 116 occurrences of other possible graphs (AACA etc.) giving a probability of 0.48 in col. 6.

The final selection in col. 4 was made firstly on a "majority vote" basis from the four choices in col. 5, and, if these were not decisive, on the probability figures in col. 6.

1.	2.	3.	4.	5.	6.
A	W	AD	D	D	0.78
COMPREHENSIVE	W	A	A	AA	1.0
DESCRIPTION	W	N	N	NNN	1.0
OF	W	PZ	P	PPPP	0.98
THE	W	D	D	DDDD	0.65
COASTAL	I	A			
WATERS	I	N	N	NNNN	1.00
OF	W	PZ	P	PPPP	1.00
CANADA	W	AN	N	NNNN	0.50
WOULD	W	X	X	XXXX	1.00
REQUIRE	W	V	V	VVVV	0.56
INFORMATION	W	AN	N	NNNN	3.84
FROM	W	P	P	PPPP	0.66
SCIENCES	W	AN	N	NNNN	0.91
SUCH	I	P	P	PPPP	0.66
AS	I	BEIPSZ			
OCEANOGRAPHY	W	AN	N	ANNN	0.73
,	P	,	,	, , , ,	1.00
MARINE	W	AN	A	AAAN	0.48
BIOLOGY	W	AN	N	NNNN	0.95
AND	W	C	C	CCC	0.00
METEOROLOGY	W	AN	N	NN	0.00
.	P	.	.	.	0.00

Fig. 3

In an analysis of the complete text of 20 000 words, 910 errors ere made, giving an error rate of 4.6 per cent. It is hoped that me of these can be eliminated by more sophisticated use of the tatistics, and also by using word probabilities. Thus the word "but" as incorrectly identified 34 times out of the 62 times it occurred, nd much better identification would have been obtained using the tatistics on this word which give a probability of 0.97 of it being connector.

As a check, a similar analysis was made on 5000 words of ext, also from the Canada Year Book, for which the dictionary en- ries were listed, but which had not been included in the original tatistics, and which had not been checked for idioms. The error rate was 5.2 per cent.

The analysis was then repeated for groups of 2, 3, and 5 words, and the error rates against group size are shown in Fig. 4. It will be seen that there is a steady improvement with increasing group size, and it would be interesting to see how far this could be taken. A difficulty arises, however, and the statistics become progressively more "sparse" with larger groups, and it would be necessary to hand-code further text in order to obtain satisfactory data for the tables. This is now being done.

NO. OF WORDS IN GROUP	ERROR RATE FOR 20,000 WORD SAMPLE
2	6.8%
3	6.3%
4	4.6%
5	3.5%

Fig. 4

This program was run on an IBM 7040 computer with disc store, and the total time taken for input, dictionary look up, idiom removal and parsing was on average 0.35 secs. per word. Of this time, the parsing occupied 90 m.s./word.

The author wishes to acknowledge gratefully the work of Mrs. J. Andrews, Miss C. Brown and Mr. C. Stock in coding the text, preparing the dictionary and idiom list, and writing the computer programs respectively.

The work has been financed by grants from the National Research Council of Canada whose assistance is gratefully acknowledged.