

Appendix IV

C.L.R.U. MECHANICAL ABSTRACTING- TEST I

by
C. Wordley

June, 1959

This paper describes a test carried out in the Cambridge Language Research Unit's punched-card laboratory on Mechanical Abstracting, The text that is analysed may be found in "Fictitious Sentences in Language", by M. Masterman.

The procedure of analysis is as follows*:

The words in the text are firstly numbered (a text position indicator). Then the text is transferred onto punched-cards. Each card is equivalent to one word of text. The first sixteen columns on the card represent the alphabetic coding of the word; if a word has more than eight letters, only the first eight are coded. The code used is of a very simple nature; each letter of the alphabet is assigned a number. Thus A = 01, B = 02, etc. The next three columns on the card register the position of the word in the text; e.g. if the tenth word in the text is considered, holes would be punched in the 0,1,0 positions of the seventeenth, eighteenth and nineteenth columns respectively.

The pack of cards produced has the following operations carried out on it:

- 1) It is sorted into alphabetic order. This groups all the same words together.
- 2) The pack of cards is then transferred to the collator which is instructed to insert a marker card in between each of the groups, thus separating the word groups.
- 3) Each group is then taken separately and passed through the sorter. The number of cards (words) in this group is registered on the Vidor Counter.

The frequencies of the words in text are listed below. No word which occurred less than three times is listed; nor were the articles "a" and "the" and the conjunction "and" considered.

* This procedure is very similar to that invented by H.P. Luhn of I.B.M. Research. This procedure has been extensively tested by him and his associates.

<u>Occurrences.</u>	<u>Word.</u>	<u>Occs.</u>	<u>Word.</u>	<u>Occs.</u>	<u>Word.</u>
3	actually	3	area	3	between
3	for	3	her	3	I
3	is	3	kitchen	3	living
3	my	3	or	3	other
3	play	3	put	3	things
3	time	3	what	3	which
4	be	4	box	4	by
4	mother	4	with	4	would
5	he	5	this	5	up
6	she	7	it	8	John
8	to	9	pen	10	his
16	of	16	was	22	in

The following collator operations were then carried out on the pack of cards.

- 1) The first card in each card group was extracted.
- 2) This pack of cards is then matched with the rest of the pack. Each of the cards which occurs in both of the packs is extracted. From this operation there are four packs: those cards which occur only once; two identical packs of those cards which occur 2 or more times; and the remaining pack, the cards which occur three or more times.
- 3) One of the identical packs is taken and matched with the pack which gives the cards of three or more occurrences. This generates four more packs. Those cards which occur 2 times, two identical packs of cards which occur 3 or more times, and the remaining pack in which the cards are of 4 or more occurrences.
- 4) The pack which gives the cards with 2 occurrences is then matched with the remaining pack of 2+ occurrences obtained in (2). This enables those cards which have an occurrence of more than two to be obtained for this pack.
- 5) All those packs in which are cards that have more than 2 occurrences are gathered together. These are then sorted into the order of the text.

From the sorted cards, it is possible to go through the text and underline those words which have a frequency of 3 or more. Each sentence is then looked at individually. If 50% or more of the words in the sentence are amongst the most frequent, this sentence is included in the abstract.

The following abstract was obtained by this method:

This young mother was preparing lunch in (the) kitchen in (a) house in which kitchen (and) living-area were contiguous. In order to attract his mother's attention, every time she passed, carrying dishes, between kitchen (and) living-area, John would put, just between her feet or actually under them, (the) largest of (a) nest of cardboard square boxes, pasted over with coloured alphabetic symbols, which happened to be at (the) time his favourite toy. Little John was looking for his toy box. Finally he found it. (The) box was in (the) pen. "Things go in pen", he said, "Things go in pen". His mother, warned by this sign, hastened in to see what he was up to. (The) idea of putting things in play-pens had rooted itself stereotypically in (the) mind of John. (The) handlamp was in (the) pen. (The) inkstand was in (the) pen.

A word underlined is one of the frequent words. A word that is in brackets is one which was not considered.

From this abstract, it is easy to see that the idea of "things being in pens" is predominate.