BUTTERWORTHS PUBLICATIONS LTD.
88 KINGSWAY, LONDON, W.C.2

AFRICA: BUTTERWORTH & CO. (AFRICA) LTD.
DURBAN: 33/35 Beach Grove

AUSTRALIA: BUTTERWORTH & CO. (AUSTRALIA) LTD.
SYDNEY: 8 O'Connell Street
MELBOURNE: 430 Bourke Street
BRISBANE: 240 Queen Street

CANADA: BUTTERWORTH & CO. (CANADA) LTD.
TORONTO: 1367 Danforth Avenue

NEW ZEALAND: BUTTERWORTH & CO. (AUSTRALIA) LTD.
WELLINGTON: 49/51 Ballance Street
AUCKLAND: 35 High Street

U.S.A. Edition published by
ACADEMIC PRESS INC., PUBLISHERS
111 FIFTH AVENUE
NEW YORK 3, NEW YORK

Made and printed in Great Britain by
William Clowes and Sons, Limited, London and Beccles

PREFACE

THE present book contains an account of some of the results which have been obtained at Birkbeck College Computational Laboratory on the application of digital calculators to linguistic problems. Almost all of the material which has been published in the field is of a very general, not to say nebulous, character and it seemed worth recording the actual details of the techniques which we have used. There is one obvious exception to this—the detailed programmes for machine translation on APEXC—these are deliberately omitted for two reasons: firstly because many readers will not be professional mathematicians and would therefore not find them interesting, and secondly because the more important ones appear in the book *Programming for an Automatic Digital Calculator* by Kathleen H. V. Booth (Butterworths Scientific Publications, 1958).

Although the idea of machine translation originated at Birkbeck College in 1947, it was not until 1955 that extensive effort could be devoted to it, and this became possible because of the generosity of the Nuffield Foundation whose support the authors most gratefully acknowledge.

Thanks are also due to Miss Jill Metherell for performing so well the onerous task of preparing the typescript for the press.

The senior author wishes to add a note of personal thanks to Dr. J. F. Lockwood, Master of Birkbeck College, for his unfailing encouragement and help in solving the problems which confront a scientist when he ventures, like Daniel, into the den of his colleagues in the Faculty of Arts.

Birkbeck College,
October 1957.

A. D. B.
L. B.
J.P.C.

CONTENTS

1

HISTORICAL INTRODUCTION

It is impossible to trace the date, from the point of view of history, when man first considered whether machines could be applied to the resolution of various problems of language. From the time of the ancient Greeks various schemes[1] have been debated for setting up ideal languages to replace those which have grown through the history of intelligent man, and many of these imply, in their structure, the possibility that a mechanical means might be found for the treatment of all problems of linguistic analysis.

Much of the most important work on statistical analysis of language[2] resulted from the desire of telegraph engineers to utilize in the best possible manner the various transmitting media available. In particular, the problem of band width restriction for the transmission of spoken sounds over telegraph lines and cables formed an invaluable spur which culminated in the invention by Shannon of what is now called Information Theory[3]. Whilst these applications were directed primarily to the improvement of technological systems involving the transmission of information, nevertheless, the results obtained were also important for more literary spheres of linguistic endeavour.

The use of a machine to translate one language to another was first suggested by Booth in the course of conversations with Warren Weaver, in New York in 1946. At that time the problem was simply an intellectual exercise directed at finding yet another use for the new high-speed digital calculators[4], which were just coming into existence. Little support was given to machine translation and this, coupled with the lack of calculating machines themselves, limited progress in the field of translation to the development of ideas instead of the actual production of a translating machine or to the practical demonstration of translation. In 1947 Booth and Britten, working at the Institute for Advanced Studies, Princeton, produced a programme[5] enabling an automatic digital calculator to make use of a dictionary stored in its memory-organ to facilitate the looking-up of words presented to the machine on a standard teletype input. At a later point in this book the functions of

dictionaries by themselves will be considered in some detail, but it is appropriate to mention here that a dictionary of the ordinary sort implies the knowledge by a user of certain basic syntactical features of the language with which the dictionary is concerned. For this reason the early dictionary programmes would have been of little use in real language translation, even had they been put on a computing machine, supposing such to have been then available. In 1948 Booth and Richens considered further how a dictionary might be constructed in a form more suitable than ordinary dictionaries for the unintelligent processes available on a computing machine. This collaboration resulted in a pioneer paper[6] which suggested that the appropriate method for translation was to construct dictionaries involving not whole words, as is normally the case, but stems and endings. The validity of these ideas was tested very extensively on ten languages, ranging from the Romance to languages of the Near and Far East as well as to Russian and Arabic.

Although no machine was available at this time, the ideas of Booth and Richens were extensively tested by the construction of limited dictionaries of the type envisaged. These were used by a human untutored in the languages concerned, who applied only those rules which could eventually be performed by a machine. The results of these early 'translations' were extremely odd, a typical example being:

'De ziekte treedt dus zeer hevig op en heeft in vele gevallen een totale misoogst ten gevolge.'

'disease come thus very rapidly up and has in many case a (one) total amiss crop then follow (P).'

where the symbols in brackets give grammatical information derived by the machine from a consideration of the endings. Two things were realized: firstly, that when they became available calculating machines would be rather limited in the scope of their linguistic comprehension, and secondly, because of the great expense of storage organs, the vocabularies which could be held in the machine would be rather limited. This in turn led to the realization that it would be necessary to store in the machine, not a dictionary of the normal size, but one consisting of two parts: the first, a microglossary of those words peculiar to the subject being translated, and the second, a set of words of general utility. Some very rough examinations of scientific texts suggested that in each case about one thousand words would be adequate, this number to include both stems and endings. It was proposed that when a foreign language word was not to be found in the dictionary, the machine should produce it in

unmodified form at the output, so that it could receive the detailed attention of the human translator.

In July 1949 a historic paper entitled 'Translation'[7] was circulated by Warren Weaver to various potentially-interested workers in the United States, and this led to an upsurge of interest in the subject in that country. Despite the unfavourable reception of this paper by a number of distinguished linguists, a nucleus of younger and more energetic workers was immediately forthcoming to take up the development of the new subject, and in 1950 Reifler produced the first of his 'Studies in M.T.'[8], a fundamental paper in which was postulated the pre- and the post-editor. The pre-editor was to remove known ambiguities and difficulties from foreign language texts for which function he needed to know nothing of the language into which translation was to be effected. The post-editor, on the other hand, would render the output of the machine into a respectable grammatical form. Reifler has himself abandoned the idea of a pre-editor, and it is perhaps not unfair to say that at the present time there is a general agreement that neither of these two posts will be necessary for the translating machine of the future.

In 1951 a second fundamental paper appeared, written by Oswald and Fletcher[9], which discussed the analysis of German syntax in a form suitable for use on a computing machine.

In 1952 the general interest in the subject led to the sponsorship by the Rockefeller Foundation of a conference devoted entirely to machine translation and held at the Massachusetts Institute of Technology. At this conference all the workers currently interested in M.T. were able to meet together and hold discussions over a period of about a week. In the course of these several members seemed to reveal a fixed opinion that the subject was too difficult for any practical experiments. Others, however, were less sceptical, and agreed with the British view that experiments were the only way in which the subject would progress at all. Although it was not the business of the conference to reach any formal conclusions, there was a general feeling at the end that two major lines of attack should be pursued. The first of these was for analyses of word frequency and word meaning to be conducted on a large scale, both for a number of subjects and for all the languages for which workers could be found, and the second was that operational analysis and syntax should be developed, so that, even supposing machines were not already in a state to receive them, these rules would be available as and when required.

After the conference work proceeded on a considerable scale both in England and in the United States. Harper[10] and Oettinger[11],

for example, were occupied with various aspects of the Russian language, Oswald and Lawson[12] made word frequency analyses of the literature of neurosurgery, Yngve worked on the utility of partial translations[13], and Booth[14] devoted a considerable amount of attention to means whereby the time required for the examination of the contents of a dictionary could be reduced. Aspects of dictionary search will be considered in detail in Chapter 6, but it may be mentioned here that up to this time the only proposals for searching a dictionary were that each word should be stored in a storage location, whose code number was simply that of the letters of the word taken in aggregate. This it may be said had been simply demonstrated to be so inefficient as to be outside the range of any conceivable machine either at the time or in the foreseeable future. Alternatively, the foreign language word could be compared with the dictionary entries, starting at the beginning. This of course means that, on the average, half of the dictionary must be examined to determine the sense of any word.

In 1954 an experiment in the machine translation of a language was conducted at Georgetown University with the assistance of the I.B.M. Corporation[15]. In this, a computing machine was programmed to translate selected sentences from Russian using a restricted vocabulary of about 250 words. Also of historic importance in 1954 was the first appearance of a periodical devoted entirely to the machine translation of language. This was called 'M.T.' and is produced by the M.I.T. under the editorship of Yngve and Locke.

Work on translation in England had, however, been severely limited by the fact that the active workers could devote only a very small part of their time to this subject. But in 1955, the Nuffield Foundation made an extremely generous grant to Birkbeck College, University of London, for the furtherance of the work on machine translation. This grant could not have been better timed. The machine APEXC[16] was now working in a satisfactory manner. The enormous interest which had accompanied the inception of these machines, to see to what extent they affected the problems of numerical analysis, mathematics, physics and engineering, had largely been satisfied, so that a certain amount of machine-time was available. The art of programming, too, had risen to heights which enabled the problems of linguistic analysis to be contemplated without undue trepidation.

With the coming of the Nuffield grant, the group at Birkbeck College expanded, so that not only were programming aspects considered, but also the detailed analysis of languages themselves. It was in 1955 that Booth suggested the first practical logarithmic

method for dictionary search. This will not be discussed further here, except to say that for a dictionary of one million words, which under the second of the schemes mentioned above would take about half a million examinations to locate any given word, it was now possible by the logarithmic technique to define the position of the word in about twenty look-up operations.

At the same time Cleave commenced work on an examination of the possibilities for transcribing standard English into Braille. , This at first seemed to be a project of little interest from the linguistic point of view, but soon showed itself to contain many of those elements peculiar to real language, for example, the existence of context[17]. Brandwood was similarly examining the French language in considerable detail, so that a realistic programme for the translation of French into English could be constructed[18]. All of this work proceeded with great rapidity, so that, by the middle of 1955, machine programmes had been devised and applied to the machine to translate from the French language on the basis of restricted dictionaries. A further highlight of 1955 was the appearance of the first book on machine translation[19] of language, edited by Booth and Locke, and containing essays by all the active workers in the field.

The following year saw continued activity in the field both in England and in the United States, and particularly from a newcomer in Moscow. Work in England and in the United States was devoted not only to the general problems outlined above, but also to the examination of the possibilities of a meta-language, which might facilitate the problems of translation between various languages of a group. The work in Moscow was chiefly concerned with translation from English into Russian, and led to a paper by Mukhin[20], first read at the April, 1956 conference of the I.E.E. in London. A number of examples which are claimed to be actual translations are given in this paper, and some of these will be considered later in this book.

To sum up the situation at this time, it is perhaps fair to say that the group at Birkbeck College had produced valid methods of translation, and tested these on an actual computing machine, that the group in the Soviet Union had done the same thing, but that apart from the I.B.M. Georgetown experiment of 1954, the United States had been particularly backward in the testing of translation routines on any actual machine. On the other hand it should be noted that the American workers had started on the construction of a dictionary machine specially designed for translation, and this, when it is available, will be of the greatest help to the work which is going on.

So much, then, for the problems of translation as such. In

parallel with these problems are various others, sometimes of a higher, sometimes of a lower degree of sophistry. There is, for example, the problem of the analysis of the frequency of occurrence of words in a given text. Workers such as Shannon had made use of pencil and paper methods of analysis which were laborious and could result in accurate estimates only when applied to short segments of text. The availability of machines had suggested that the machines themselves might be used to construct the frequency analyses required in machine translation, and considerable work in this field was carried out by the Nuffield research group at Birkbeck College. Another problem of the same generic type is that of constructing concordances for given texts, that is, lists, usually in alphabetic order, of the words in those texts, each word being accompanied by a set of page and line references to the place of its occurrence. Such concordances are of course familiar to linguists, having been available for many years for such classics as the Bible and Shakespeare. The interest at Birkbeck College in this field was chiefly engendered by some earlier research work[21] on the Dialogues of Plato, which had been considerably hampered in the past by the absence of such concordances. Parallel work in this field has been carried out by the I.B.M. Corporation, and it appears that some of this work is now being put to practical use in the preparation of a concordance for the works of Thomas Aquinas.

A more involved application of the same sort is to the stylistic analysis of a work by purely mechanical means. Here again there is a considerable literature, but previous workers have considered the subject, not within the framework of a machine, but for analysis by a human linguist acting in a mechanical way. Roughly speaking, a problem of this type is the following:

Given a set of books, for example the Dialogues of Plato, and given also that the actual dates of production of some of these are known. Find, by an analysis of word frequencies, structure and so on, the position of unknown works in chronological order within the date framework known unequivocally.

This problem will be discussed at a later point in the book, and it will be shown that problems of this sort are indeed quite amenable to purely mechanical resolution on a computing machine.

The examples of linguistic analysis just mentioned are only a few of those which are now coming into the ambit of computing machines. Problems of the same sort, although not so conspicuous for their linguistic content, concern the cross-referencing of libraries, the indexing of scientific and other information, telephone directories, and certain problems of mathematical logic. Not all of

these problems will be examined in this book, but it is hoped that in the following chapters sufficient information is given to enable people interested in the fringe applications of the subjects to work out methods for using an automatic computing machine to resolve their problems in the way already used for language.

REFERENCES

[1] ELLIOT, W. V., 'Isaac Newton's "Of an Universall Language"' *Mod. Lang. Rev.*, LII (1), (1957), 1–18

[2] SHANNON, C. and WEAVER, W., 'The Mathematical Theory of Communication', Illinois Univ. Press, Urbana, (1949)

[3] BRILLOUIN, L., 'Science and Information Theory', Academic Books, London, (1956)

[4] BOOTH, A. D. and BOOTH, K. H. V., 'Automatic Digital Calculators', (2nd Ed.) Butterworths, London, (1956)

[5] BOOTH, A. D. and BRITTEN, K. H. V., 'Coding for A.R.C.', (2nd Ed.) Inst. Adv. Study, Princeton, (1947)

[6] BOOTH, A. D. and RICHENS, R. H., 'Machine Translation of Languages', Wiley, New York, (1955), 24–26

[7] WEAVER, W., 'Translation', Mimeographed, New York, July, (1949)

[8] REIFLER, E., 'Studies in Mechanical Translation, No. 1, M.T.', Mimeographed, Washington, (1950)

[9] OSWALD, V. A. and FLETCHER, S. L., 'Proposals for the Mechanical Resolution of German Syntax Patterns', Mod. Language Forum, 36, (1951), 1–24

[10] HARPER, K. E., 'The Mechanical Translation of Russian: Preliminary Report', Mod. Language Forum, 38, (1953), 12–29

[11] OETTINGER, A. G., 'A Study for the Design of an Automatic Dictionary', Mimeographed, Harvard, (1953)

[12] OSWALD, V. A. and LAWSON, R. H., 'An Ideoglossary for Mechanical Translation', Mod. Language Forum, 38, (1953), 1–11

[13] YNGVE, V., 'Mechanical Translation of Languages', Wiley, New York, (1955), 208–226

[14] BOOTH, A. D., *Nature*, 176, (1955), 565

[15] DOSTERT, L., 'Mechanical Translation of Languages, Wiley, New York, (1955), 124–135

[16] BOOTH, A. D., M.T.A.C. VIII, (1954), 98

[17] CLEAVE, J., 'Information Theory', (Ed. C. Cherry) Butterworths, London, (1956), 184–194

[18] BRANDWOOD, L., 'The Translation of a Foreign Language by Machine', *Babel*, 11 (No. 3), (1956), 111–118

[19] BOOTH, A. D. and LOCKE, W. N., (Ed.) 'Machine Translation of Languages', Wiley, New York, (1955)

[20] MUKHIN, I. S., 'An Experiment on the Machine Translation of Languages', *Acad. Sci. U.S.S.R., Moscow*, (1956)

[21] BRANDWOOD, L., 'Analysing Plato's style with an Electronic Computer', Inst. Classical Studies, (Univ. London) Bulletin No. 3, (1956), 45–54

2

THE NATURE OF CALCULATING AND DATA PROCESSING MACHINES

BEFORE the actual application of a machine to problems of linguistics can be considered, it is first necessary to see how language data can be presented to a machine, and this involves a study of the way in which data can be recorded in permanent form. Language no doubt arose in the first instance merely as vocal or other form of sonic expression. The spoken word can be made permanent in many ways, for example, it can be recorded upon a gramophone record, upon a magnetic tape in a tape recorder, it can be recorded on the sound track of film or, and more fundamentally, it can be committed to paper in the form of handwriting, typewriting or print.

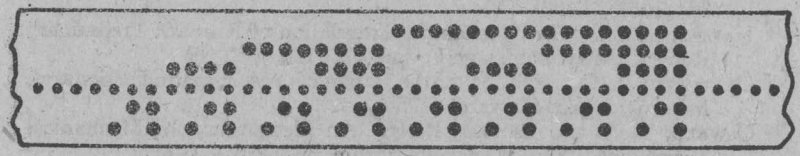

Figure 2.1. Teletype tape showing binary punching 0–31

From the point of view of machine handling, none of these forms is at present acceptable, and it turns out that, to feed data into a machine, it is first necessary to represent it in some form of numerical code. The encoding of letters has been studied very extensively in connection with the transmission of data over telephone and telegraph lines, as was mentioned in Chapter 1. Historically, one of the oldest forms of recording and encoding is in the form of a five-hole punching on what is known as teletype tape. In *Figure 2.1* is shown a small section of tape, and in Table 2.1 there is indicated the numerical code which is nowadays accepted by international convention for the encoding of alphabetic characters for transmission over telegraph lines. A cursory examination of the encoding system shown in the table will reveal that there is no particular logic behind

the code symbol associated with each alphabetic character. The reader who has a knowledge of mathematics will see that the numerical symbols constitute numbers in the so-called binary scale of notation, but it does not happen, for example, that A corresponds to the number 1, B to 2, C to 3, and so on. In fact, for mechanical

Table 2.1

		○Start	Code elements	●Stop			○Start	Code elements	●Stop
			1 2 3 4 5					1 2 3 4 5	
A			● ● ○ ○ ○		P	0		○ ● ● ○ ●	
B	?		● ○ ○ ● ●		Q	1		● ● ● ○ ●	
C	:		○ ● ● ● ○		R	4		○ ● ○ ● ○	
D	Who are you?		● ○ ○ ● ○		S	′		● ○ ● ○ ○	
E	3		● ○ ○ ○ ○		T	5		○ ○ ○ ○ ●	
F			● ○ ● ● ○		U	7		● ● ● ○ ○	
G	*Optional Characters*		○ ● ○ ● ●		V	=		○ ● ● ● ●	
H			○ ○ ● ○ ●		W	2		● ● ○ ○ ●	
I	8		○ ● ● ○ ○		X	/		● ○ ● ● ●	
J	Bell		● ● ○ ● ○		Y	6		● ○ ● ○ ●	
K	(		● ● ● ● ○		Z	+		● ○ ○ ○ ●	
L	)		○ ● ○ ○ ●		Carriage return			○ ○ ○ ● ○	
M	.		○ ○ ● ● ●		Figures			● ● ○ ● ●	
N	,		○ ○ ● ● ○		Letters			● ● ● ● ●	
O	9		○ ○ ○ ● ●		Line feed			○ ● ○ ○ ○	
					Space			○ ○ ● ○ ○	

Elements which cause the setting of teleprinter combination discs or perforation of reperforator tape are shown thus ●. They are often referred to as marking elements, whilst those of the opposite kind are known as spacing elements.

Each group of code elements is preceded by a start signal and succeeded by a stop signal. The latter may be of 1 or 1½ units duration.

handling, there is considerable reason to revise the alphabetic coding system to one more in accord with the needs of the subject.

Another example of a coding of alphabetic data into numerical form is the Morse code. Here, however, very much more thought has been given to the detailed form of the encodement. This will be seen by an examination of Table 2.2. In brief, the Morse coding

system is such that the number of numeric symbols required to encode an alphabetic character has the property that the more frequent characters have the simpler numerical code, whereas the less frequent characters are coded in a manner which required the transmission of more digits.

The other principal form of numerical coding for data in use at the present time is on punched cards. A typical example of one of these is shown in *Figure 2.2*, p. 17; punched cards were first devised by Herman Hollerith for application, in a purely numerical context,

Table 2.2

Letter	*Probability*	*Morse symbol*	*Binary code*	*Letter*	*Probability*	*Morse symbol*	*Binary code*
Space	0·2			F	0·022	··—·	0010
E	0·105	·	0	U	0·022	··—	001
T	0·072	—	1	M	0·021	— —	11
O	0·065	— — —	111	P	0·018	·— —·	0110
A	0·063	·—	01	Y	0·012	—·— —	1011
N	0·059	—·	10	W	0·012	·— —	011
I	0·055	··	00	G	0·011	— —·	110
R	0·054	·—·	010	B	0·010	—···	1000
S	0·052	···	000	V	0·008	···—	0001
H	0·047	····	0000	K	0·003	—·—	101
D	0·035	—··	100	X	0·002	—··—	1001
L	0·029	·—··	0100	J	0·001	·— — —	0111
C	0·023	—·—·	1010	Q	0·001	— —·—	1101
				Z	0·001	— —··	1100

to the problem of the analysis of census data. A typical card has eighty columns, each of which is capable of receiving a punching in any one of twelve positions. It follows that, since there are more than twelve letters to be encoded, it is either necessary to use more than one column to represent in numerical form a single letter, or alternatively that several punchings must be placed in a single column when alphabetic data is to be recorded on the card. Business machine manufacturers have, in fact, adopted the second of these alternatives. The normal method of designating the various positions on a card is shown in *Figure 2.2*, and when desired to represent alphabetic characters, a punching in the X or Y positions accompanies a second punching in one of the normal numerical positions. An alternative to this alphabetic code is given in *Figure 2.3*, p. 18.

The third means of data representation acceptable to a calculating machine is of far more recent origin, and is a development of the

magnetic tape recorder used for speech. In this case, however, magnetic tape is used in such a way that data is recorded on its surface in pulse form, that is, the number is first represented in a binary code. This binary code is then recorded on the tape surface, using a positive pulse of electrical current through the recording coil to represent 'one', and a negative pulse to represent 'zero'. It may be argued that, if a virgin tape were taken, it would be sufficient to record only ones *or* zeroes, the other being represented by the absence of any signal. The use of this system, however, is open to a number of objections. The first of these is that it is always dangerous to represent information—and zero is just as much information as one—by the absence of information, since this absence can in fact be only relative. All electrical apparatus produces a certain amount of what is known as 'noise', and it may only be said that the absence of information means noise less that some completely definite and non-zero upper limit. The second objection to presenting digits of one sort by electrical impulses and of the other by the absence of an electrical impulse lies in the details of the means used to recover the data at a later date. To recover data in which only ones are recorded on a virgin surface requires the use of a clock track, that is, a set of pre-recorded markers on some part of the tape, these markers indicating that certain time intervals have elapsed during which no electrical signal has been generated, when, in fact, it could have been generated. The alternative to the use of such time markers is very accurate synchronization of the recording medium with some external source of time measurement. It is certainly true that modern tape-recording machinery is capable of a high degree of synchronization, but it is nevertheless still extremely undesirable to assume such constancy of reproduction when the need for this accuracy can be eliminated by the simple device of using two forms of electro-magnetic impulse to represent the two types of binary digit concerned.

At the present time magnetic tape is just coming into use with computing machines, but up to now it has not been much used with programmes of linguistic interest. The principles which accompany its use are identical to those which would be used with paper tape, and it will not be necessary to consider it further in this book.

PUNCHED CARD MACHINES

Since it is only intended to relate the discussion of linguistic data to calculating machines, it is not necessary here to enter into a detailed description of telegraph systems for handling data, because, at

present, these systems are capable only of transmitting and reassembling data at two places, and not, in general, of making any transformation of the data during the course of its progress. The first machines which could absorb numerical data from a record were the punched card business machines. These machines are still of great importance, not only in office accounting, but also in certain aspects of linguistic work. They have the considerable advantage that a number of operators can be used to prepare cards either from different texts or from the same text, and that these cards can be brought together as a whole at the end of the operation and then presented to the computing machine. It was in fact on just such punched card accounting machines that the first Birkbeck College experiments in language translation by means of a simple dictionary were carried out.

Today the range of punched card machines is very great, but, from the linguistic point of view, it is necessary only to mention a few of these. The first machine to be considered is the card punch. This can take various forms from the extremely simple hand-operated type, shown in *Figure 2.4*, p. 19, to a complicated electrical system with storaged facilities for data and automatic handling for cards, shown in *Figure 2.5*, p. 19. The storage affords means for remembering data which has been punched on a single card, so that it may be repunched on successive cards.

After data has been processed through one of these punching machines, it is generally necessary to check it in some way. The checking operation is conveniently performed by means of a verifier shown in *Figure 2.6*, p. 20. Here a card, previously punched, is placed into the checking device and a different operator from the operator who prepared the card in the first place goes through the motions of preparing the card again. As the various keys are depressed, the electrical sensing mechanism examines them to see whether the depressed keys correspond with those keys which would have been depressed to produce the holes in the given position on the card. If, however, an error occurs, the device gives a signal in the form of a red light and refuses to feed the card. In this case the operator knows either that she has committed an error in operating the checking punch, or that there is an error in the card itself. This method of checking by reiterating the operation previously performed is open to certain well-known objections, but generally it enables a satisfactory check to be made, and is common in many forms of business accounting.

When the cards have been prepared and checked, various forms of analysis can be performed upon them by means of the punched

card calculating machinery. A frequent operation is that of sorting and this is performed in the device shown in *Figure 2.7*, p. 20. Here a stack of cards is passed through sensing apparatus, and cards bearing various pre-assigned combinations of punchings are distributed into one of twelve hoppers shown in the figure. In this way it is easy to take a pack of cards, for example, and to sort them into groups which have, say, zero in column one, one in column one, two in column one, and so on. It can also be shown that by a sequence of these operations it is possible to take a pack of cards arranged in random order and to reassemble it into numerical order, alphabetical order, etc. No actual operation of calculation is involved, although types of sorting apparatus exist in which the cards distributed between the various hoppers actuate counters, so that the number of cards in each hopper is available after the cards have passed through the mechanism. This type of device is very useful when it is desired, for example, to ascertain the number of words starting with A, B, C, D, and so on, in any given pack.

The principal calculating device in the punched card range is the tabulator, shown in *Figure 2.8*, p. 21. With this it is possible to take the data, in this case of course numerical, stored on a card and to combine it with the data which has been stored on other cards, usually by the processes of addition and subtraction. The tabulator is also provided with printing facilities whereby totals and sub-totals of data can be printed and in addition alphabetic information which is held on a card can be output at the same time. Thus the tabulator can, in principle, perform the functions of printing out dictionaries and statistical data.

Another extremely useful machine is the collator, which has various applications not only in linguistics, but also in the handling of accumulations of data in general. A typical collator is shown in *Figure 2.9*, p. 21. The apparatus is provided with two feeds, into which can be placed stacks of cards. The action of the collator is to take the cards from each of the stacks, examine two from one stack and one from the other to see if certain conditions of ordering are satisfied, and then to distribute them to one of four output hoppers according to various pre-assigned criteria, which are set up on the plug-board shown in the front of the machine. One use of the collating machine will be described in Chapter 5, and it was this which enabled a mechanical dictionary to be synthesized in the early stages of machine translation.

Finally there is the reproducer or gang-punch. This device simply prepares a new card in accord with data supplied to it. This data can either be taken from cards which it feeds from its own

hopper or alternatively from the output of a tabulating machine or collator. The use of the gang-punch will be described later, when we come to consider the realization of actual dictionary experiment.

At the present time electronics is extending the range of punched card machinery, and not only are more rapid devices available, but also such things as electronic multiplying punches are becoming commonplace. So far none of these more highly specialized arithmetical machines have found an application in linguistics, and, since it appears unlikely that they will find application, we shall not discuss them further here. Before we leave the subject of punched card machines, it is appropriate to give a short summary of the speeds of the operations of which these machines are capable. These are given in Table 2.3.

Table 2.3

Machine	*Speed, cards/min.*
Hand punch	2–3
Keystor punch	4–10
Verifier	2–3
Sorter	450–660
Tabulator	100–150
Collator	240
Reproducer	100

ALL-PURPOSE DIGITAL CALCULATORS

The revolution in arithmetical technique which made possible a detailed consideration of linguistic problems from the machine point of view was the invention of the all-purpose digital calculator. It is not our purpose here to enter into the history of these devices, except to say that the first conception of an all-purpose machine was that of Charles Babbage in 1833. In fact, no machine of a type which would be readily applicable to linguistic problems was produced until the early 1950s. For this reason it is not considered necessary to describe the various attempts, starting in the mid-1930s, which led to the eventual realization of such a machine. Nowadays numerous machines of the required type are available both in university laboratories and from manufacturers of business calculating equipment. The description which follows can be taken as typical of the facilities which are available on any of the better types of these machines.