Computers in Linguistics

计算机在语言学中的应用［英］

Christopher Butler

Basil Blackwell
World Publishing Corp

Computers in Linguistics

Christopher Butler

Basil Blackwell
World Publishing Corp

First published 1985

Basil Blackwell Ltd
108 Cowley Road, Oxford OX4 1JF, UK

Basil Blackwell Inc.
432 Park Avenue South, Suite 1505,
New York, NY 10016, USA

British Library Cataloguing in Publication Data

Butler, Christopher, 1945–
Computers in linguistics.
1. Linguistics — Data processing
I. Title
410'.28'54 P98
ISBN 0–631–14266–5
ISBN 0–631–14267–3 Pbk

Library of Congress Cataloging in Publication Data

Butler, Christopher.
Computers in linguistics.

Bibliography: p.
Includes index.
1. Linguistics — Data processing. I. Title.
P98.B84 1985 410'.028'5 85–7540
ISBN 0–631–14266–5
ISBN 0–631–14267–3 (pbk.)

Reprint authorized oy Basil Blackwell
Reprinted by World Publishing Corporation,Beijing,1991
Licensed for sale in China only
(excluding Hong Kong, Macao and Taiwan)
ISBN 7–5062–1047–9

Computers in Linguistics

Acknowledgements

I should like to express my gratitude to the copyright holder for permission to reproduce Sylvia Plath's poem 'Apprehensions' (From Sylvia Plath's *Crossing the Water*, London: Faber and Faber 1971, copyright Ted Hughes; and from *Winter Trees*, New York: Harper & Row, publishers) in a form suitable for input to the computer (chapter 14). My thanks are also due to Michael Stubbs, not only for permission to use a modified version of his lexical density program in chapter 14, but also for the valuable comments he made on early drafts of the book. I am also grateful to Peter Ford for his comments on chapter 1. Finally, my main debt is to the many students who, during the course of their efforts to assimilate the material presented here, have taught me a great deal about how to teach arts computing. For the many inadequacies which remain in the book I am, of course, solely responsible.

General introduction

It is probably not too much of an exaggeration to claim that the 'micro-processors' which are at the heart of computer systems are the single most important influence on our lives today, and will assume even greater importance in tomorrow's world. The sphere of influence of computers extends from space exploration to our everyday experience of computerised sales points in supermarkets, from the navigation of an aircraft to the booking of tickets by its passengers, from the control of electricity generation to the preparation of the bills which drop through our letter-boxes. The present generation of children are learning to accept such things as ordinary and are already, in their schools, gaining familiarity with the principles of computing. For many of us, however, the computer remains strange, powerful, even rather sinister. Yet, despite the slow emergence of so-called 'expert systems' mimicking human decision-making processes, computers are still very stupid beasts, incapable of action unless programmed by their human masters.

If properly understood, the computer can bring enormous benefit, not least in areas of academic interest which have hitherto been the traditional preserve of dedicated researchers spending long years on the often tedious manual analysis of their materials. The purpose of this book is to show how the computer can aid those who have a professional interest in the study of language. Such interests range over a wide field: one researcher may be engaged in isolating characteristics of the literary style of a particular author; another may wish to obtain information on the frequency and range of vocabulary items or syntactic constructions in technical texts in a foreign language; a third may wish to construct and test a model of the frequency distributions of words in a text; a fourth may have amassed a large body of data which he wishes to analyse for correlations between some linguistic parameter and factors such as age, social class or sex. In all these cases and many more, it is likely that computational methods will be of considerable value.

If computers are really rather stupid machines, needing much human effort before they can be made to do any job, then why, the reader may wonder, are they claimed to be of such value in such a rich variety of tasks? It is true that the machine is capable of only a few basic sorts of operation: as we shall see a little later, it can move items of information from one storage location to another, compare one item with another, and carry out arithmetical operations. The working section of a computer can be thought of as simply a multitude of tiny switches, each of which can be on or off, so that even such apparently simple operations as the comparison of two symbols, or the addition of two numbers, must in fact be broken down into a pattern of binary, on/off events. However, the operation of these 'switches' is electronic rather than mechanical, and herein lies the source of the computer's phenomenal powers. Electrical pulses travel at extremely high speeds, so that the computer works at rates many millions of times greater than those of a human operator. In order to appreciate the order of magnitude involved, let us consider one or two examples. Time yourself roughly in doing the following addition sum:

$$\begin{array}{rl} 15826498164801498\,2 & \\ \underline{642184330776515998} & + \\ \hline \end{array}$$

How long did you take? 15 seconds? 30 seconds? A powerful computer could carry out several million such additions in a single second. Consider a further example, this time from the area of language study. It took Ione Dodson Young about 25 years to produce a concordance (that is, a list of all the occurrences of each word, with context for each) for the complete poetic works of Byron, using traditional manual methods. In 1967, a computer generated a concordance of the complete works of Blake in about three hours (though this does not include the time taken to get the text into computer-readable form, a problem which we shall discuss briefly later). With the latest generation of computers, the time taken would be even less.

Not only is the computer incredibly fast, it is also extremely accurate. Mistakes can sometimes occur because of breakdown of the machine's internal mechanisms, but such cases are much rarer than mistakes due to errors by human programmers. Furthermore, the machine will be just as accurate when performing the millionth calculation or comparison as when performing the first: unlike human beings, it does not suffer from fatigue, boredom or lapses of memory. The mention of memory brings us to a further important point: a computer has, associated with its central processor, a certain amount of internal storage capacity or 'memory'. A small computer, such as the simplest micro-computers selling for around £40, may have only about 1 000 individual storage locations (in the jargon of computing, this would be referred to as 1K, where a K unit represents 1 024 storage locations), while a powerful computer may have several orders of magnitude more. However, this is not the limit of the computer's ability to store data since, as we shall see, additional ('secondary') storage capacity is usually available in the form of magnetic discs or tapes which can hold large amounts of data. This, then, is a further important factor: the computer and its secondary storage devices can hold a very large amount of data indeed (the Bible, or the complete works of Shakespeare, would present no problem whatever for a large machine), whereas the minds of human beings have much more humble capacities.

To sum up: the modern computer gains over the human researcher in its extremely high speed of operation, its relentless accuracy and its huge capacity for storing information. It is, however, capable of only a very few basic operations, so that considerable human ingenuity is needed to reduce complex problems to the machine's level of coding in terms of electrical impulses which are either 'on' or 'off'. It is with ways of inducing the computer to tackle problems in the study of language that the present book is concerned. Part I gives a brief introduction to the modern digital computer and surveys its applications in linguistic and literary studies. This part of the book includes an introduction to some 'pre-packaged' tools available for text analysis, for those who do not wish to learn a programming language. Part II offers a thorough introduction to the programming language SNOBOL4, which was specifically designed for the analysis of natural language texts.

Contents

Part I

Computers in linguistic and literary research

1 An introduction to the computer

1.1 Types of computer

When we use the term 'computer', we normally mean a 'digital' computer, which uses counting as its fundamental mechanism. There is another type, known as an 'analogue' computer, which works by measuring various quantities related to the phenomenon under investigation. Analogue machines are normally used only for rather specialised purposes, and we shall assume throughout this book that we are dealing with a digital computer.

Computers can also be classified according to their size and power. Nowadays, a wide range of 'micro-computers' is available for home use, as well as for restricted use in business and scientific applications. These are compact, desk-top machines with limited capacities, though the increasing availability of high-capacity 'secondary storage' devices (see p. 6) for such machines is making it possible for quite serious applications to be implemented on them. Rather larger and more powerful are the 'mini-computers' which are now to be found in even quite small businesses, as well as in industry and academic institutions. At the top end of the scale we have multi-user 'mainframe' computers, which may cost millions of pounds and are capable of handling complex tasks involving very large amounts of data. Because of its complexity and the large quantities of data often involved, linguistic and literary computing is usually carried out on the more powerful machines, and it is this level of work with which the present book is concerned. Some applications, however, can be implemented on less powerful computers, and with the very rapid pace of development in the micro-computer area it is likely that we shall soon be able to perform even quite sophisticated, large-scale analyses in our own homes.

In the following section the structure of a multi-user computer system will be described briefly; for further details readers should consult an elementary textbook on general computing, such as Hunt and Shelley (1983).

1.2 The fundamental components of a computer system

1.2.1 Computer hardware

The overall system The actual physical components of a computer system are often called the 'hardware'. The basic hardware components are shown diagrammatically in figure 1.1.

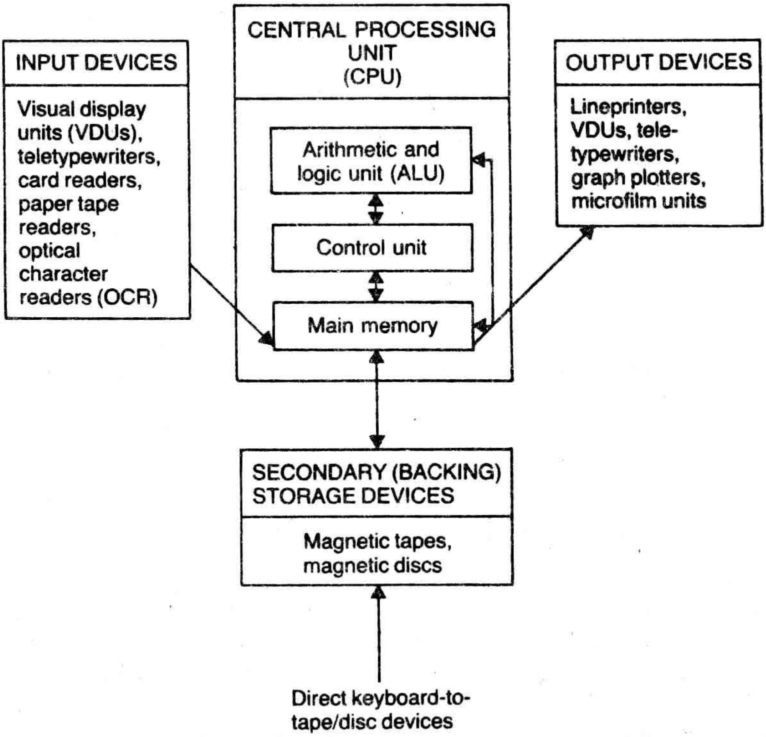
CENTRAL PROCESSING UNIT (CPU)
INPUT DEVICES
Visual display units (VDUs), teletypewriters, card readers, paper tape readers, optical character readers (OCR)
OUTPUT DEVICES
Lineprinters, VDUs, teletypewriters, graph plotters, microfilm units
Arithmetic and logic unit (ALU)
Control unit
Main memory
SECONDARY (BACKING) STORAGE DEVICES
Magnetic tapes, magnetic discs
Direct keyboard-to-tape/disc devices

Figure 1.1 The basic hardware components of a computer system

Clearly, we need not only the central processing unit itself, but also some means of getting information into the computer, and some way of getting information out again after processing. Further, we may need secondary storage devices to supplement the limited storage capacity of the central processor itself. Input and output devices and secondary storage are often termed 'peripherals', since they are peripheral to the central processor, and may in fact be situated at some considerable distance from it. We shall now consider each of these principal hardware components in turn.

The central processing unit As shown in figure 1.1, the CPU consists of three parts. The heart of the machine, where the operations involved in comparison or calculation are carried out, is the arithmetic and logic unit (ALU). It will be remembered from the introductory chapter that the computer carries out all its arithmetical and logical operations in terms of a binary code representing 'on' and 'off' states of electrical circuits. In order to perform these operations, the ALU needs to know what tasks it is to perform, and on what data. This information is obtained from the main memory of the computer. Although the main memory may not be large enough to hold all the information relevant to the complete solution of a complex problem, it holds whatever information is relevant at a given point in the solving process. Conceptually, the main memory consists of a number of storage locations, each numbered to provide an 'address' so that the information in each location can be retrieved. Physically, the memory unit usually consists of a large number of microscopic electrical circuits embedded in a silicon chip. The third fundamental part of the CPU is the control unit, which supervises the flow of information between the main memory and the ALU, thus co-ordinating the various stages of processing.

Input devices Methods of getting information (data and instructions) into the computer have undergone important changes over the last few years. Until relatively recently, the usual method was to encode the information on punched cards. A typical card is divided into 12 rows, each with 80 columns. Each symbol which can be recognised by the computer (letters, numbers, punctuation marks, and so on) is represented by a unique combination of punched holes in any given column. The cards are prepared using a card punch, which has a keyboard very like a normal typewriter keyboard. Information punched on cards is usually checked, or 'verified', before it is passed to the computer: a second operator keys the information in again, but on a machine which checks for identity between the new input and the characters already punched. After verification, the cards can be fed into a card reader, which senses the pattern of holes and translates it into a representation of the characters in the computer. Cards can be read at an average rate of approximately 1 000 per minute, that is about 1 350 characters per second. Input by means of cards has a number of disadvantages: mistakes are slow to correct, since a mispunched card must be rejected and repunched; cards may become damaged; large card 'decks' require considerable storage space. Card input has now been largely superseded by other methods.

A second input method is to use paper tape. Like cards, paper tape has a number of rows, or 'channels' across it (often eight), and a pattern of punched holes represents a character. Tape is cheaper than cards, and requires less storage space, but is susceptible to tearing. Furthermore, the continuous nature of paper tape makes it rather difficult to insert or delete material. Paper tape readers work at a maximum rate of about 1 500 characters per second.

The most common method of input for a multi-user system now is the use of a terminal with a keyboard, like a typewriter keyboard, on which the required information can be typed. Increasingly often, this terminal is itself a micro-computer. Most terminals incorporate either a visual display unit (VDU), which displays the information on the screen as it is typed, or a teletypewriter (often abbreviated to 'teletype'), which prints the typed information on paper, so enabling the user to obtain hard copy. Such terminals are very flexible: they can be used for direct interaction with the central processor, or for storing information on secondary storage devices (see p. 6). Furthermore, the correction of errors and the alteration of previously stored information is generally quite simple, since most computer systems have powerful editing facilities for use with terminals. Terminals can be situated at some considerable distance from the central processor, so that an increasing number of users can interact with their local computer, or indeed with a network of computers in more remote locations, from terminals in their own offices.

All the methods of input outlined so far have the disadvantage of being very time-consuming and costly. This is an especially important problem for the linguistic and literary researcher, who often needs to use a large data base for his work. A relatively recent but increasingly available input technique, optical character recognition (OCR), overcomes this problem to a large extent. For some time now, computers have been used to sense highly standardised marks on, for example, multiple choice examination papers. Now, a refinement of this technique can be used to 'read' a wide variety of printed and typed materials, and to convert the data into a form which can be stored on magnetic tape or disc (see pp. 6–7). Typical of such devices is the Kurzweil Data Entry Machine (KDEM), which is able to recognise a number of standard type founts, and can be trained to read others. The machine matches the pattern of a given character, analysed as a matrix of dots, against a store of known character patterns, and accepts the character if it can be matched to within a certain degree of tolerance. If no such match can be made, the character is displayed on a screen for the user to check and enter the correct character. Given a cleanly printed copy on good paper, the KDEM can read at a rate of some 250–300 words per minute. There can be no doubt that as these machines become cheaper, OCR will be a standard input mechanism for linguistic and literary computing in the future.

Output devices The most common way of obtaining printed output from the computer is to use a lineprinter, consisting of rows of characters attached to a chain or drum. As the correct character is selected, a magnetically controlled hammer prints it in much the same way as a typewriter hammer. Usually, lineprinter output is printed on perforated and folded sheets of paper with a width of 120–36 (often 132) characters, though it is possible to print on smaller slips of paper, specially prepared forms, and so on. By the use of different character sets, output in non-Roman alphabets can be produced if required. Lineprinters can print at the rate of about 1 500 lines per minute. Other printers becoming commonly available include laser printers and smaller daisy wheel and dot matrix printers, the latter two often being attached to the terminal, so that the need to wait for centralised output is avoided.

In scientific applications, much use is made of plotters which can produce graphical output. Graph plotters are also of interest to the linguistic and literary researcher, since they can be adapted to plot the characters of non-Roman alphabets without the need for special lineprinter character sets.

Information can also be output to VDU and teletype terminals. VDUs are most suitable for cases where the output is simply to be inspected, but no hard copy is needed; if a printed copy is required, a teletype can be used.

A final output method of considerable interest to linguistic and literary researchers is computer output microfilm (COM), a device which takes information on magnetic tape (see below) and converts it, at high speeds, into a highly reduced photographic image on microfilm, which may be in roll form or as individual 'microfiches'. Since the output from a linguistic or literary analysis can often be voluminous, microfilm offers a cheap and compact way of storing results. The only disadvantage is that a special reader is needed to magnify the image to normal size for inspection. Hard copy is easily generated by means of a printing attachment.

Secondary ('backing') storage We have seen that the central memory capacity of a computer is limited (although large mainframe machines can contain very large internal memories), so that it may not be able to hold, at any one time, all the information necessary for the solving of a problem. This is often the case, for example, where the processing of large amounts of text is being undertaken. Thus secondary, or 'backing', storage is often needed to supplement the main memory. Information stored on secondary devices can be transferred to and from the CPU as required. Secondary storage can be used not only as temporary holding devices, but also as permanent stores for information which will be required repeatedly for processing. There are two main types of secondary storage, magnetic tapes and magnetic discs, each with their own advantages and limitations.

Magnetic tape is available in reels up to 2 400 feet long, and is similar to that used on domestic tape recorders, except that it is rather stronger and wider. Across the width of the tape are a number of tracks (usually nine) containing spots which can be either magnetised or unmagnetised. The pattern of spots across the tape at any one point represents a character, just as does the pattern of holes punched on paper tape or cards. Information is written to the tape, and read from it, by means of read/write heads, one per track. Magnetic tape is a 'serial' storage device, that is, the information can be read only in the order in which it was encoded on tape. This means that the entire tape may need to be scanned in order to retrieve certain items, so that access to information is relatively slow. However, magnetic tape is cheap and easily stored, and a 2 400-foot tape can hold up to 30 million characters, which can be transferred to the main memory at a rate of up to 800 000 characters per second. The information on a tape can, of course, be erased when no longer required, and replaced by fresh information. For these reasons, magnetic tape is widely used, especially for applications where serial searching is not disadvantageous.

Magnetic discs are rather like gramophone records, arranged in packs mounted on a central rotating shaft. As on tapes, information is stored as patterns of magnetised and unmagnetised spots in tracks. In most disc units there is, for each surface, a head which moves across the tracks to write information to the disc or read from it; some disc units have a number of fixed read/write heads. Unlike the head of a magnetic tape unit, that of a disc unit can access any item of information without having to scan all the previous information; that is, it is a 'direct access' system. This greatly decreases the search time for information, as compared with a magnetic tape: the average search time for a large disc unit would be about 10 milliseconds (i.e. 0.01 sec.). Transfer rates to the CPU are also very high, up to about 1.2 million characters per second. The capacity of a large disc is extremely high, up to about 1 280 million characters. Thus discs offer considerable advantages over tapes; disc cartridges are, however, very much more expensive than magnetic tapes. As with tapes, information on discs can be erased and replaced as required. Mention should also be made of the small flexible 'floppy discs' which are becoming increasingly important, especially for micro- and mini-computer systems. These discs are loaded in cartridge form, and typically hold up to 500 000 characters.

Whatever the type of storage device used, information is organised into what are called 'files', each of which consists of a number of 'records', which may correspond to a punched card, a line on a VDU screen, or whatever. We might, for instance, set up a data file containing a particular poetic text, each record encoding one line of verse. The instructions to the computer to do a particular job may also be retained in files (see section 1.4).

1.2.2 Computer software

So far, we have concentrated on computer 'hardware', the actual machinery which constitutes the CPU and peripheral devices. However, this sophisticated machinery can do nothing unless it is given instructions, these instructions being referred to as the 'software'. The present book is concerned principally with the kinds of software which are of especial usefulness in solving problems connected with language study. It should be realised, however, that a large amount of software, written by computer manufacturers and by systems designers, is involved in the control and integration of the various components of a computer installation. When such an installation is running smoothly, the average user can, of course, remain blissfully unaware of the complex flow of information regulating not only the running of his own job, but also its interaction with jobs being run by other users.

1.3 What can the computer do?

This whole book is, of course, concerned with what computers can do in the field of linguistic and literary research. Here, we shall merely suggest a general classification of the kinds of task for which the computer is suited.

Firstly, the computer may be used as a *sorting* device. It can compare two items stored in its memory, on the basis of their numerical value or their position in an alphabetical sequence. This function of the computer is probably more important in the social and psychological sciences than in the physical sciences, and is the basis of many applications to language study, from the analysis of questionnaire data to the production, from textual data, of word lists in alphabetical or frequency ordering.

Secondly, the computer may be used as a very sophisticated *counting* device, able to perform complex calculations at a very high speed. This function is of paramount importance in the physical sciences, where the processing of large amounts of numerical experimental data (often referred to as 'number-crunching'!) is often required. The calculating function is also important, however, in many other areas, including the study of language. As we saw in the introductory chapter, many linguistic studies require the collection and statistical processing of quantitative data, and this processing is a task for which the computer is very well suited.

These two functions, of sorting and counting, are the basis of those linguistic applications of computing concerned with the analysis of textual data, data from tests, and the like. These are the applications on which we shall be focusing here. There are, however, other types of application, which we shall mention only in passing in chapter 2, and which involve the use of the computer as a *model-testing* device or as a *simulating* device. Both of these functions are ultimately derived from the sorting and counting functions, but perhaps require a few words of separate explanation.

Many of those interested in the study of language for various purposes spend much of their time constructing models of the phenomena they are studying. The purpose of a model is to provide a simplified representation of a restricted range of phenomena, which can be tested and modified. Models may be concrete (for example a hardware model of the human vocal tract) or abstract (for instance a mathematical model of the vocabulary distribution in a text, or a set of linguistic rules modelling part of the native speaker's grammar). Computers are useful in testing the extent to which the model is an accurate reflection of the phenomena it purports to represent. We might, for example, use a computer to generate all the sentences allowed by a certain set of linguistic rules, and then