Whole Text Information in the Legal Office *

M.A. Heather


INTRODUCTION

If the machine is made for man and not man for the machine, the machine must master the modes of man. But man communicates mainly by natural language and stores his information mostly in documents composed of full text. The powers of the machine in these two fields are still quite puny despite their extreme importance and relevance to everyday human affairs. Machines at present still find natural language unpalatable and it is partly due to this difficulty in getting at the meaning expressed in natural language that the amorphous structure of documents of full text cannot be too easily handled by automatic methods. It may yet be some time before the problems of developing satisfactory methods to deal properly with natural language are resolved and in the meanwhile it is necessary to make progress in improving machine capabilities for manipulating documents in full text form. The prime purpose of this paper is to concentrate only on one very pertinent aspect of the latter - machine aids to comprehension. This is of most relevance to one category of full text, namely the whole text, but it also has some significance for other full text forms.

The principles seem to be generally applicable but they will be discussed here in one particular subject area, that of the Law where active work is in progress in the study of the processing of whole texts. The Law relies heavily on the full text and legal information systems have been forced to meet the problems early.

WHOLE TEXTS

It is convenient to distinguish the different categories of full text that the machine is called upon to deal with and to consider them in three classes: "loose text", "short text" and "whole text".

* M.A. Heather is a professor at the School of Law, Newcastle Polytechnic, Newcastle-upon-Tyne, NE1 8ST, UK.

* Based on a presentation at the International Conference on Man/Machine Systems organized by the Computing and Control Division of the Institution of Electrical Engineers of London and held at the University of Manchester Institute of Science and Technology, Manchester, U.K. 6th-9th July 1982.
Loose text is composed of fragments of full text often to be found as extra words of explanation in a database that is not primarily in full text form. Thus a bibliographic database in addition to the text fields of Author, title, etc., may also have notes in phrases or clauses but which are not in full grammatical sentences. Explanatory captions for figures and diagrams are a second example.

Short text consists of an abbreviated form of full text. It can only be achieved by the relaxation of the strict rules of natural language. There is the immediate result that there may be a loss in precision in the communication of information. The amount of licence taken in writing short text can vary from the familiar abrupt staccato style of telegraphy and the telegram to one closer to a normal use of language but with the occasional use of elliptic constructions such as the dropping of the definite article.

Whole text on the other hand is that which is expressed in proper full free text form. It is important to distinguish the form of the text from the nature of its contents. A precis or an abstract may be in either short or whole text form. Brevity and conciseness are not necessarily synonymous. Often a short text form may be used because there is not the time or intellectual energy available to produce an equivalent succinct whole text.

Where electronic mail is in regular use the trend appears to be to use short text rather than whole text. The art of letter writing was developed in an age when the pace of life was slower. The standard of using the whole text for correspondence once established has been maintained. It appears that those who readily use the short text for telematic communications would still hesitate to do so on headed notepaper or when using their own personal stationery. There are important implications for business and commerce if short text teleconferencing and electronic mail replace interactive oral negotiations face to face and by telephone when there is not the attendant “to be confirmed in writing” in whole text. Where commercial contracts and arrangements are expressed in short text many complicated legal problems can arise in the construction and interpretation. It has been shown by Picot et al. in an attitudinal survey of 477 office managers that the three most important criteria for a communication channel are (in order) “unambiguous understanding of content”, “speediness” and “certainty of exact wording”. Yet it would appear that the second mitigates against the other two criteria.

It may also be noted that teletex and viewdata systems in general can only display a maximum of about 150 words in one frame because of the constraints of the domestic television receiver. Such systems therefore also show a preference for short text. In disciplines such as the Law it is often necessary to construct long structured whole text sentences to achieve certainty of meaning. It may often not be possible to display at one time a complete sentence from an Act of Parliament on a TV screen.

It is for these reasons that in considering the extraction of information from
full text documents processed by machine, we are mostly concerned with
whole texts. However documents are often composite and may contain a
mixture of whole text, short text and loose text where the two latter act as
adjuncts and aids to sorting and understanding the whole text where there is
embedded the "end-information" sought by the user.

LEGAL TEXTS

The excessive storage capacity and computing power requirements for han-
dling documents in full text have so far prevented the widespread use of
computers for that application. Technological capabilities however are now
such that we can expect to see a large scale move to the electronic medium
for storing and retrieving full text documents. Law has been a pioneer in
this field. This has been of necessity because in the solution of legal pro-
blems fine detail may be critical. This fine detail is only to be found in the
whole text. Not only may any given word be significant but even, as Niblett
has pointed out, for Sir Roger Casement, the very positioning of a comma
proved literally fatal. Legal data bases have therefore been established in
whole text form. Because of the size of whole texts, the largest electronic
data bank in Europe is probably that in Italian law set up by the Camera of
Deputies (Pagano) and possibly the largest in the world is that of the Ameri-
can legal information retrieval system LEXIS which has recently been esti-
imated by Harrison to hold 35 x 10^9 characters. Part only of the law is lit-
tle better than no law and it is very expensive to establish a comprehensive
data base. Bing and Harvold describe very many attempts in Europe and
North America to establish legal information retrieval systems which have
failed and some of which have even led the promoters into bankruptcy. The
accounts given in reference may provide cautionary tales for those who
wish to establish whole text data bases in other subject areas.

Indeed the Law is a full text information science which has been in existen-
ce for many centuries and it is interesting to note the impact of technologi-
cal developments. It is noteworthy that changes have not always been for
the better. Two centuries after the invention of printing it was found that
the promulgation of English statutes was less effective than in the days
when Acts of Parliament were copied by hand and dispatched to each shire.

An interesting example relating to the extracting of information from whole
texts can be found in the conveyancing of real property. Unless land has a
title that has been registered at HM Land Registry, the right of ownership
and the nature of the ownership is proved by production of title deeds. The-
se deeds are invariably in whole text. It has been normal practice for centu-
ries to provide a purchaser or a mortgagee with a summary of the relevant
title deeds in a document known as an abstract of title. An Abstract was
produced on brief paper (406 x 330 mm) usually in short text form. At one
time it was fashionable to compress the text further by using abbreviated
spellings for most words. Nevertheless it was in a form of continuous prose but the structure of the document was presented to the eye by tabulating the starting position of each fresh item of information. Thus dates and the value of revenue stamps were given in the left-handmost margin, then details of covenants and certificates. At the third position began the recitals, the habendum, redendum, provisos and conditions. The fourth set out any trusts to which, the property was subject. The fifth was about halfway across the page and gave the details of the owner, the parcels clause, receipts, registrations and enrolments. The result of this was that is was very easy for the eye to skim through many pages of abstract and find for instance all mentions of trusts. A solicitor acting for a purchaser could then quickly and confidently investigate the title of the property he was buying for his client without having the title deeds, although it was still necessary to compare the original deeds with the abstracted version, and a clerk could be dispatched to the vendor's solicitor to carry out that task. However the significant point is the effect of the introduction of the photocopying machine. It immediately became easier to copy than to abstract. The result was that the purchaser's solicitor only received a copy of the whole text version and was denied the convenience of the short text structured format. The whole text form obviously contained more information but there was a deterioration in the quality of communication of information. It is very important to safeguard against this. Labour saving machines have tended to reduce the standard of craftsmanship and we cannot let brain-aiding devices bring about a comparable deterioration in intellectual achievements. When facsimile transmission becomes commonplace there may be a temptation to send large quantities of textual material without regard to how the receiver is to sort all the information to extract only what he needs to know.

Whole texts supply the fine detail of information. The whole text primary source materials for English Law amount to about a thousand million words. As the Law has to be viewed in toto it means that this requirement to be able to focus down to the level of a single word demands a resolution of $1 \times 10^9$. This is an exacting standard for any machine but it is the specification for an ideal automatic legal information retrieval service. Attention will now be directed to the more immediate and quite urgent need to provide better facilities to assist human apprehension of full text information stored in the machine.

TEXTUAL GRAPHICS

Over the centuries many languages have borrowed the Greek root γραφαί to mean either writing or drawing. Sometimes one meaning dominated, sometimes the other. Whole civilizations can be characterised by their choice between a letter based and a pictorial form of written communication. This semiotic competition exists today and at the present time it is the picture gra-
phics which appear to be in the ascendancy. Current developments in both hardware and software combine to make computer aided design very attractive and to produce some very pleasing results in computer simulation and visualization. Unfortunately the same cannot be said for textual graphics. Even upper and lower case is not always available. Some networks achieve economy and some line printers achieve high speeds by not supporting lower case working. Three examples might be given where the machine is still very much inferior to methods used in manual paper based documentary systems. The first is in browsing. There appears to be some form of combined motor/cognitive skill that enables man to flick, skim and scan through a hand-held book. The judicious use of the contents page, indexes, page headings or just random sightings of parts of the text allow information to be located in a remarkably short time. The best that a machine can do seems to be to switch on some pseudo-random number generator to select records to display to the user haphazardly. The only control that the user has is usually of the rate at which documents are displayed to him. What is lacking is some form of mental servo-mechanism to guide the direction of progression for seeking records at random.

A second problem is how to make comparisons between documents and within a document. In order to ascertain the meaning of a passage in a legal document such as a section in an Act of Parliament it may be necessary to refer constantly to other parts of the same Act where terms are defined, to case reports on the point and even different reports of the same case. These are the primary sources; there are then also the secondary sources such as commentaries and practitioners' text books. Thus a solicitor or a member of counsel may need a dozen works of reference open before him to advise on some very narrow but precise point. With full concentration on the conceptual threads between the documents the human brain can spare little more diversion that is required to direct the head towards the right document. Superficially the eye can only concentrate on a small amount of text at one time but in practice it requires some very sophisticated hardware and software to provide the equivalence of a dozen open books plus several bookmarks and the odd finger and thumb.

The third problem relates to the display format of whole texts and can be demonstrated by reference to figure 1 where there is reproduced a two page extract from the version published by Her Majesty's Stationery Office of the Supreme Court Act 1981 Part III. It illustrates only some of the components of whole text to be found in the published Statutes in Force. To provide the reader with the fullest aid to comprehension the Queen's Printer finds it necessary to use a hierarchical structure of headings, cross notes, shoulder notes, margin notes and footnotes, to break the text into numbered sections and sub-sections, schedules and paragraphs and to provide the title in various forms (the long title, short title, preamble, etc.). Section 72 of the Supreme Court Act 1981 is an example of a section where there are local definitions which control how that section has to be read. It is therefo-
rc foolish to try to read sub-sections (1) or (2) without also having sub-section (5) close to the field of view. Furthermore as subsection (5) itself contains sub-paragraphs (that is at the level of the sub-sub-sub-section) it is essential to have the proper indented display. The qualified definition of "related offence" in sub-section (5) is then quite easy to apply. However this is a fairly straightforward example. Often references must be made to some more global definition to be found elsewhere in the statute, in other statutes on similar subject material or even from interpretation arising in case reports.

Clearly the typographical layout, styles of type, etc carry information and there will be a reduction in the quantity of information conveyed unless these are retained or transformed in the conversion to the electronic medium. If the aim is only to use the computer as a centralized continuously updated databank of documents in whole text form which are to be routed to a local print terminal as required, then the problem is related to computer typesetting, procedures for which are now well established. However information retrieval from whole texts is an interactive process and this principal advantage of using the machine will be greatly reduced if a large quantity of printed material has to be produced in hard copy form at each stage.

Her Majesty's Stationery Office make available for the purpose of electronic information retrieval a version of the data tapes which are used to print Statutes in Force in the form illustrated in figure 1. It is therefore possible to compare and contrast the quality of information communicated by the printed page with that displayed at the computer terminal. HMSO publish a specification of their format styles but an analysis of the HMSO data tapes on the Northumbrian Universities Multiple Access Computer has identified forty different logical kinds of text required to provide the various physical formats. As display may be required on a variety of different terminal devices it is not sufficient just to use control codes. Each kind of text has to be stored as a structured entity. Each entity can then be transformed to an appropriate representation using features such as indentations, rotations, highlighting, foregrounds, flagging, flashing characters, reverse video, colour and what one will or better still according to standards which are or should be commonly agreed.

**Current developments**

Techniques for catalectic graphics are at present to be found in word processing systems and also as a secondary role in pictorial graphics where text material is often included to supplement diagrams. These applications have encouraged the introduction of larger display screens. The A4 size screen can display twice or three times the number of lines that can appear on the traditional VDU although the width does not allow many more characters than the familiar 80 columns. Word processors being orientated towards data in-
put and editing may provide facilities for larger clearer text characters but often without any increase in the total number to be exhibited at a time. Some however provide for two documents to be shown side by side. Software is available for the split screen and for partitioning into virtual screens which can be manipulated with the dexterity of an animated cartoonist but the dimensions are usually too small for any realistic quantity of whole text. Likewise the 132 character option which is now being added to many standard 80 character VDUs provides too small a size text unless the screen size is of at least 15″. The large demonstration type display units can usually only provide large size character fonts. What is needed is one about the size of a draftsman’s light table. It could be composed of integrated segments but must support ordinary size text.

Work is proceeding at Newcastle Polytechnic using the Stanford Public Information Retrieval System (SPIRES) generalized database management software to experiment on the effectiveness of various two-dimensional display formats for whole text. It has been found so far that although a 132 character VDU is essential on account of the 60% increase in the total number of words which can be viewed, nevertheless even with enhanced video the text is not too easy to read. Indeed a line length of 132 characters is too long for comfort; it is interesting to contrast this observation with the popularity in the past that the brief size paper has had among lawyers both for abstracts of title and for instructions to counsel. It seems then that the size of spacing between lines is also important. To have to resort to double spacing with 132 character lines of course immediately reduces the advantageous word density. So far the most effective format on a 132 character VDU has been found to be single spacing with double columns despite some wastage between the two columns of utilizable space and the unnecessary telecommunication cost of sending blanks. Double columns also make heavy use of the facilities of the host computer or else demand a very intelligent terminal on account of the necessary two-dimensional buffering. However narrow columns do not provide much space to use the layout for assisting the reader to gain a view of the logical structure of the text as can be seen by comparing a page of this publication with that reproduced in figure 1.

Conclusions

The capacity to handle larger data sets and its ability to perform logical operations on such data make the electronic digital computer a powerful intellectual tool for processing textual material but this is an area where, in comparison with numerical data, its potential has only just started to be exploited. No attempt has been made here to explore all the possibilities but rather in view of the theme of this international conference attention has been restricted in the main to one narrow but crucial point of how the machine must be able to communicate with man’s form of communication – whole
PART III

(4) Rules of court—
(a) may provide for securing such sittings of any Division of the High Court during vacation as the senior judge of that Division may with the concurrence of the Lord Chancellor determine; and
(b) without prejudice to paragraph (a), shall provide for the transaction during vacation by judges of the High Court of all such business in the High Court as may require to be immediately or promptly transacted.

(5) Different provision may be made in pursuance of subsection (3) for different parts of the country.

Other provisions:

75.—(1) In any proceedings to which this subsection applies, a person shall not be excused, by reason that to do so would tend to expose that person, or his or her spouse, to proceedings for a related offence or for the recovery of a related penalty—
(a) from answering any question put to that person in the proceedings; or
(b) from complying with any order made in those proceedings.

(2) Subsection (1) applies to the following civil proceedings in the High Court, namely—
(a) proceedings for infringement of rights pertaining to any intellectual property or for passing off; or
(b) proceedings brought to obtain disclosure of information relating to any infringement of such rights or to any passing off; and
(c) proceedings brought to prevent any apprehended infringement of such rights or any apprehended passing off.

(3) Subject to subsection (4), no statement or admission made by a person—
(a) in answering a question put to him in any proceedings to which subsection (1) applies; or
(b) in complying with any order made in those proceedings,
shall, in proceedings for any related offence or for the recovery of any related penalty, be admissible in evidence against that person (unless that person married after the making of the statement or admission against the spouse of that person.

(4) Nothing in subsection (3) shall render any statement or admission made by a person as there mentioned inadmissible in evidence against that person in proceedings for perjury or contempt of court.

(5) In this section—
"intellectual property" means any patent, trade mark, copyright, registered design, technical or commercial information or other intellectual property;
"related offence", in relation to any proceedings to which subsection (1) applies, means—
(a) in the case of proceedings within subsection (2)(a) or (b)—
(i) any offence committed by or in the course of the infringement of passing off to which those proceedings relate; or
(ii) any offence not within sub-paragraph (i) committed in connection with that infringement or passing off, being an offence involving fraud or dishonesty;
(b) in the case of proceedings within subsection (2)(c), an offence revealed by the facts on which the plaintiff relies in those proceedings;
"related penalty", in relation to any proceedings to which subsection (1) applies means—
(a) in the case of proceedings within subsection (2)(a) or (b), any penalty incurred in respect of anything done or omitted in connection with the infringement of passing off to which those proceedings relate;
(b) in the case of proceedings within subsection (2)(c), any penalty incurred in respect of any act or omission revealed by the facts on which the plaintiff relies in those proceedings.

(6) Any reference in this section to civil proceedings in the High Court of any description includes a reference to proceedings on appeal arising out of civil proceedings in the High Court of that description.

THE CROWN COURT
Composition of court

75.—(1) Subject to the provisions of section 8(1)(c), 74 and General 75(2) as respects courts comprising justices of the peace, all provisions, proceedings in the Crown Court shall be heard and disposed of before a single judge of that court.

(2) Crown Court Rules may authorise or require a judge of the High Court, Circuit judge or Recorder, in such circumstances as are specified by the rules, at any stage to continue with any proceedings with a court from which any one or more
texts. Until there are improvements at that interface much of the machine's potential will lie untapped. It is hoped that this discussion relating to just one area of application, the Law, will provide for those some background information on what the consumer needs.

References


