8. OPERATIONAL CONSIDERATIONS

Whatever the verdict of evaluation of one or more automatic indexing techniques, whether of the derivative, modified derivative, or assignment type, there are certain operational considerations and problems that typically affect any attempt to apply such techniques in actual production operations. These considerations, which also affect linguistic data processing operations in general, include input considerations, availability of methods or devices for converting text to machine-usable form, programming considerations, questions of format and content of output, and problems of customer acceptance of the machine products.

8. 1 Questions of input

Input considerations include, first, questions of the extent and availability of material which can be handled directly by the machine. This may be limited to title only, to title plus abstract, title plus other material, 1/ preselected text or automatically generated extracts; or it may in a few cases extend to full running text. Possible future requirements may extend to the processing not only of full text but of interspersed graphic material (equations, charts, diagrams, drawings, photographs) as well.

We have considered typical arguments for and against the limitation of input to titles only, to augmented titles, and to abstracts in other sections of this report. The points to be emphasized here are requirements for pre-editing or post-editing, provisions for error detection and error correction, the time and cost requirements of conversion equipment if material is not already available in machine-usable form, and the like. As Cornelius suggests:

> "Present day computers, if used for machine indexing, will be generally input limited and will require excessive data preparation. Causes of these limitations are: time required for translation to machine language, verification of this machine language, and the capability or lack of capability of correction in the input media." 2/

Examples of pre-editing requirements, even for the simple case of keyword-in-title indexing, include the spelling out of chemical symbols, the encoding or the omission of subscripts and superscripts, insertions of hyphens to prevent indexing of a word, and substitutions of blanks for hyphens in compound words to assure indexing of each component. 3/ For full text, a far more extensive and elaborate set of rules and conventions must be developed and applied. 4/ Other editing may be required for format standard-

1/ This may specifically include cited titles, as suggested variously by Bohnert, 1962 [69], p. 19; Giuliano and Jones, 1962 [229], p. 10; Swanson, 1963 [580], p. 1; Gallagher and Toomey, 1963 [205], p. 53; and as used in the SADSACT method, see pp. 98-99 of this report.

2/ Cornelius, 1962 [140], p. 42.

3/ See, for example, Kennedy, 1961 [311], p. 120.

4/ See, for example the sophisticated proposals of Nugent, 1959 [441], and Newman et al, 1960 [439].

ization, especially in the case of citation indexes compiled by machine. 1/ O'Connor notes, however, that "the provision of pre-editing information can slow down the keypuncher or typist, increase the chance of mistakes, and require more intelligence or training on the typist's part." 2/

Questions of error detection and error correction apply both to the original text and to transcribed versions if these are necessary. That is, the basic documents themselves may contain typographical errors, misspellings, and the like, and additional errors are bound to occur at all subsequent stages requiring human processing. Wyllys discusses the need for the correction of spelling errors, mentions suggested computer programs for detection, and cites a private communication from Stiles suggesting that the criteria for accepting words as valid be either that they are identified as already being in the system vocabulary or that they occur at least twice in the input item. 3/

Swanson's analysis of the reasons for retrieving irrelevant, and failing to retrieve relevant, material in the case of text searching on the nuclear physics abstracts includes typical data on the effect of errors. 4/ He found, for example, that failures to record hyphenated words, subscripts, superscripts and other special symbols accounted for about 5 percent of failures to retrieve relevant items, and errors in transcription of either text or search instructions accounted for another 3 percent of these failures. Errors in keypunching of the search requests alone accounted for 4 percent of the cases of irrelevant retrievals. By contrast, in the newspaper clippings experiments where the input material was already in machine-usable form transcription errors were not a factor but the input tape itself had many errors. In this special case, however, Swanson reports: "Garbles are not important simply because messages are sufficiently redundant to insure that even if one or two keywords for a given category are garbled, almost invariably others are present." 5/

The news clippings material used by Swanson represents one class of materials that are today initially available in machine-usable form, because the original recording of the message or text resulted in a machine-usable medium, such as punched paper tape. A punched paper tape is produced as the product of many typesetting operations, especially for newspaper and magazine publication, and this will be increasingly true in the future, together with computer-prepared tapes for input to automatic typographic composing equipment. To date, however, equipment to convert from these tapes to the particular machine language of a given computer processing system is largely non-available, is costly, and is highly subject to error. 6/

1/ See, for example, Atherton, 1962 [25], p. 4; Marthaler, 1963 [399], p. 22. However, at least one computer program has been developed to assist in this process. See Thompson, 1963 [600], p. II-1: "The present program takes bibliographic citations and automatically arranges then into a standard format in such a way that the various parts of the citation are unambiguously identified. These standardized citations can later be processed by sorting and matching procedures to identify similar citations and to effect various rearrangements."

2/ O'Connor, 1960 [444], p. 8.

3/ Wyllys, 1963 [653], p. 15.

4/ Swanson, 1961 [586], Appendix.

5/ Swanson, 1963 [580], p. 5.

6/ Compare, for example, Savage, 1958 [521], p. 11: "The use of tape as the original input to the process has offered a number of problems which have yet to be solved. One is the occurrence of typographical errors."

Moreover, to date, very little material in the scientific and technical literature is available in this form. As of 1961, it was reported that a survey by McGraw-Hill indicated that only about 2 or 3 percent of the publications in the United States were then prepared by typesetting tape, that most of this was in the form of Monotype tape which because of its 30-column width and special format is not generally compatible with tape reading equipment, and that tapes had many errors in them which would require considerable effort to correct. 1/ As of late 1963, Bennett reports:

> "Computer processing of natural language text material requires that a body of data be available in machine-readable form. At present such a body of data results only from a direct human copying process. An inquiry into existing transcriptions of text which were machine-readable showed that they were abbreviated both in terms of completeness and in number of symbols represented. As an alternative text produced as a by-product of typesetting operations is clearly an eventual possibility, but present practices make the detection of unit delimiters such as ends-of-sentences difficult." 2/

In the future, both machine-usable text from publishers and printers and the similarly machine-usable paper tape produced as a byproduct from the original keystroking of manuscript on such equipment as Flexowriters and Justowriters may alleviate this problem for new items. Nevertheless, the wealth of the world's present literature, the informal and unpublished technical reports of high current interest but limited initial distribution, and material acquired from foreign sources, will continue to pose for the foreseeable future major problems either of automatic reading of the printed page or of human re-transcription at high cost.

While there have been many promising developments in automatic character recognition techniques, the devices that are now available for production use are limited to small character sets, such as a single alphabet in a single font, often of special design. The multi-font page reader is not only not yet commercially available but may not become so for some years to come. Even if it were, there are many unresolved and as yet incompletely specified problems involved in the development of suitable rules for the machine so that it can distinguish between title or page number and text, figure caption and text, author's name in a cited reference and the title of the paper cited, and the like. A case in point, not only for automatic reading equipment of the future but for machine processing of machine-usable material available today, is the difficulty of machine recognition of punctuation marks as used for different purposes. 3/

In the absence, then, both of scientific and technical documents already in machine language form and of character recognition equipment capable of reading the printed page, we are left with the unsatisfactory situation of re-transcribing input material either by use of a tape typewriter or by keypunching to punched cards. That this situation is unsatisfactory and is a major bottleneck in machine processing of text in excess of the bibliographic citation data only is evidenced by such typical statements as these:

1/ Cornelius, 1962 [140], p. 47.

2/ Bennett, 1963 [50], p. 141.

3/ See Bennett quotation above; Luhn, 1959 [384], p. 22, and Coyaud, 1963 [143].

> "The expense of transcribing such documents in their entirety will be justifiable to a limited extent only and it may, therefore, be assumed that automatic processing will be mainly applied to future literature." 1/
>
> "As long as we are limited to using the equipment that is available now, the preparation of data for input will be an expensive procedure and a major cost factor in automatic processing of natural language." 2/
>
> "... In a discussion of indexing by machine, we must recognize the preparation of input to the system as the major item of cost of operation." 3/
>
> "Present inability to read documents automatically would make it necessary to punch cards or tapes, an operation likely to be even more expensive than reading by humans." 4/

In addition to the high costs of manual retranscription, it is also noted that keypunching "tends to undermine the purpose of natural text retrieval by requiring human effort at the input end of the process." 5/

In particular, keypunching or keystroking requirements undermine the purposes of rapid indexing as well as filing for retrieval by virtue of the time required to transcribe text. Horty and Walsh report, for example:

> "Flexowriter operators can produce between 1400 and 1800 lines per day of statutory text. Keypunch operators used in previous experiments could punch approximately 100 lines per hour of alphabetic materials, but could not maintain this rate for a sustained period of time." 6/

Thus, until such time as more versatile character recognition equipment is available, even some of the most ardent advocates of full text processing are forced to the use of considerably less than full text for other than research purposes. Swanson comments, for example:

> "... One must note that the manual recording of text may be exorbitantly expensive. If so, a judicious selection process may permit a reasonable compromise between the expense of input and the depth of indexing which results. For example, it is reasonable to select the title, abstract, table of contents (if any), sub-headings, and key sentences or paragraphs." 7/

1/ Luhn, 1959 [384], p. 2.

2/ Ray, 1961 [496], p. 51.

3/ Howerton, 1961 [282], p. 327.

4/ Levery, 1963 [359], p. 235.

5/ Doyle, 1959 [168], p. 2.

6/ Horty and Walsh, 1963 [280], p. 259.

7/ Swanson, 1963 [580], p. 1.

> "Costs come much more into line if we make available to the machine something on the order of one per cent of the full text. Then, of course, the problem of selecting that one per cent presents itself." 1/

8.2 Examples of Processing Considerations

A second major area of operational considerations involves the machine processing problems, given a specified input. For most of the automatic derivative, and modified or normalized derivative, schemes, this is primarily a question of the limitations of machine language to a vocabulary of, typically, no more than 64 distinct characters for input, internal manipulation, and output. In addition, the limited number of characters that can be packed into a single machine-word complicates internal processing, storage, file look-up (i.e., against exclusion or inclusion lists), and sorting operations.

Arbitrary truncation of text words to, say, 6 characters per word, leads to certain computer processing or storage economics. However, it leads also to complications in the selection of words either to be included (clue word lists) or excluded (stop lists) in many of the proposed methods both for derivative and for assignment indexing. Additional problems of artificial homography are created. Obvious examples are "Probab-le, -ility"; "Condit-ion, -ional," "Freque-nt, -ntly, -ncy," "Commun-ity, -ication;-al", and the like. Barnes and Resnick include in their studies of the effectiveness of an SDI System 2/ the use of 6 different truncation levels (from 4 to 9 characters). No significant differences were found in terms of the number of hits (matches of a new item to a user's profile which he considered to be of definite interest to him) but there were significant differences in the number of notifications sent him, as presumably matching his interest, and the amount of "trash" (irrelevant items) among these notifications.

The importance of the selection criteria in derivative indexing, operationally considered, is largely a matter of the length and the contents of the stop lists. Variability in practice among the various producers of KWIC indexes has previously been noted, 3/ but there are some interrelated and interlocking factors which affect the quality, the costs, and the customer acceptance of this type of machine-generated index. First, the number of pages in a printed index is directly related to the total costs of producing that index. 4/ The amount of material covered on a single page can be increased by photographic or other type of reduction (e.g., the 96 lines per page of the Bell Laboratories KWIC program output are reduced by xerography to 62 percent of the machine output page size), (Kennedy, 1961 [311]) but the reduction must not be such as to exceed reasonable limits of legibility.

This, in turn, means that the number of entries generated for each title (obviously, a function of the words that survive stop list purging) needs to be held to a reasonable minimum. Thus:

> "One of the major limitations of the published index stems from the conflict between the quantity of text that must be placed between the covers and the capacity of the printed page to handle it. The size of the page and the legibility of the printing determines the maximum density of characters which can be read without special aids." 5/

1/ Swanson, 1962 [584], pp. 470-471.

2/ Barnes and Resnick, 1963 [36]. See also p. 148 of this report.

3/ See discussion, pp. 65-66.

4/ See Markus, 1963 [394], p. 16.

5/ Taine, 1961 [592], p. 153.

The question of stop list effectiveness therefore becomes an operational factor as well as one that may affect the quality and acceptability of the product. On the other hand, too generous a purging of the input titles may of course reduce the utility of the title index by the elimination of too many potential access points and, in particular, many that users may be most tempted to look for.

A related problem has to do with the number of pages required because of the length of the title line allowed in the listings. A suggestion advanced by Brandenberg (1963 [80]) is the assignment of numeric codes to the machine stop words used and the insertion of these codes into the listed title line in the place of these presumably insignificant words. Thus one of the KWIC entries for the title, "Determining Aspects of the Russian Verb from Context in Machine Translation" might go from:

```
RMINING ASPECT OF THE   CONTEXT IN   MACHINE TRANSLATION.  /DETE to:
ERMINING 032 416 712 RUS   CONTEXT 308   MACHINE TRANSLATION.  /DET
```

This particular example was picked at random from a KWIC index utilizing a 103-106 character title line, 1/ but it was deliberately shortened to the 60-character line length found in many such indexes in order to illustrate effects of chopping and wrap-around. Coincidentally, it also illustrates some of the difficulties of designing a well-balanced exclusion list since in this case the purged word "aspect" is apparently being used in a technical sense rather than in the common one of "Various aspects of...". By accident, this case does show rather severe "aspects" of the chopping problem in the loss also, for this entry, of "Russian" and "verb" although they would of course be picked up in the entry blocks for these words. Certainly, however, the claimed advantages of context checking are not striking, even without the introduction of the numeric codes. It is true that for excluded words longer in length than those in our example the possible conservation of the character-space to reduce the chopping effects for the same length line may result in improvements. However, the replacement of, for example, "Preliminary investigations of..." by numeric codes would hardly assist the user in determining quickly from the many possible entries under "..." which he should select for further personal perusal.

Turning to the case of automatic assignment indexing, the processing considerations likely to be involved in operational factors affecting the evaluation of a system are much less easily exemplified. Obviously, conditions that hold for research experiments on small (and usually, especially selected) samples do not necessarily relate to requirements in potential productive applications. Exceptions are the problems of the sizes of term-term and term-document co-occurrence correlation matrices that can be readily manipulated, previously mentioned, 2/ and the concurrent problems of the size, and hence the representativeness, of inclusion lists or clue-word vocabularies that can be accommodated.

Both Maron and Borko found, even in their limited test samples, a certain proportion of new items that could not be indexed or categorized at all because these new items did not contain any of the clue words recognizable by the system. 3/ Due perhaps to longer selective clue word lists, as well as to the special nature of his items, Swanson found no instances, for 775 test items, of failure to assign because of lack of indicative clues in the input material. In the case of 60 tests against the SADSACT model, which uses approximately 1, 600 words drawn from a "teaching sample" of items previously indexed to descriptors, (related by frequency of co-occurrence to any of 70-odd descriptors with whose

1/ Walkowicz, 1963 [629], pp. 136 and 137.

2/ See pp. 108 and 160 of this report.

3/ See Maron, 1961 [395]; also Borko and Bernick, 1963 [78].

assignment they had co-occurred), the machine had a sufficient basis in the input material for the derivation of a selection-score for at least 12 descriptors for each new item. The items were closely similar to, though not identical with, the source items from which the word associations with descriptors assigned had been drawn. The sample is obviously critically small. Nevertheless, the possibility that extensive clue word lists, notwithstanding the incorporation of trivial and even erroneous associations, can be used as effectively as smaller, more precise, and more carefully tailored lists, but with significant gains in memory space or computational requirements, is suggestive. A somewhat related conclusion, again reflecting the effect of processing requirements, is stated by Needham as follows:

> "The main point to be made is that theoretical elegance must be sacrificed to computational possibility: there is no merit in a classification program which can only be applied to a couple of hundred objects." 1/

In KWIC type derivative indexing by machine, except in terms of allowable character sets and word-lengths conveniently processed, the problem of appropriate programming languages does not arise to any serious extent. For the processing of material in research on natural language text, however, the choice of interpretative and compiler types of automatic programming languages may involve computational requirements which, while being inappropriate in a production situation, offer considerable flexibility and versatility for experimental purposes. Examples of special programs of this type include the use of Yngve's COMIT by Baxendale and Knowlton, the development and use of FEAT by Olney, Doyle, and others at SDC, and the use of list-processing techniques in the General Inquirer system. 2/ Yngve describes the use of his program as follows:

> "COMIT has also been used in the experimental work in information retrieval of Baxendale and Knowlton at IBM. The purpose of their COMIT program was to accept as input the title of a document and to produce as output, not only descriptors, but pairs of descriptors which are roughly of the form adjective-noun. The purpose of the work is to automatically generate, from document titles, retrieval words of a more specific nature than simply Boolean functions of the existence of certain words in a title." 3/

The FEAT program was designed originally for word and significant-word-pair frequency counts. Olney describes the program in part, as follows:

> "FEAT is designed to perform frequency and summary counts of words and word pairs occurring in its natural text input; i.e., text written in ordinary English and transcribed into Hollerith code according to some set of keypunching rules. To focus attention on the semantic aspects of word pairs rather than on their syntactic aspect, pairs of which one member is a function word, such as 'the', 'is', 'by', etc., are excluded."
>
> "Using a bucket list structure of the type proposed by C. J. Sheen in FN-1634, the program sorts each incoming word serially, constructing a list within each of 256 buckets for good words of a given alphabetic range ... and another list within each good word entry for the Doubles and Reverses which will be ordered alphabetically

1/ Needham, 1963 [433], p. 8.

2/ Stone, et al, various references, p. 137 of this report.

3/ Yngve, 1962 [655], p. 26.

on that word . . . If there are four different Double types of which the first word is 'external' the addresses of the four different second words form a new list which is linked to the entry for 'external'. Each word type occurs only once in core, and all word pairs of which it is a member refer to it by means of its core addresses."

"The program could process millions of words, automatically generating frequency counts far larger than the Thorndike and Lange counts, which cost many man-years, and in addition, FEAT would provide complete lists of word pairs (Doubles and Reverses), which, so far as we know, have never been counted in a sample of appreciable size, despite their importance for semantic analysis of text."

FEAT is used, together with a modified version of the Proto-Synthex program, and special output formatting routines, for another SDC program, the Descriptor Word Index Program, which produces a content-word-concordance for natural language text as well as statistics reflecting the type of words that occur, frequencies of occurrence, and positional data, (Olney, 1960 [457], 1961 [456]; Stone, 1962 [574].

The IPL-V list-processing language is used by Kochen in some of his work on simulated concept processing by machine. Programs for accepting sentences written in a formal language which was constructed of names and logical predicates (inserted either from a console or in the form of punched cards), for updating and re-organizing a file of such sentences, for storing and manipulating metalinguistic sentences such as "If X is author of Y and Y pertains to topic Z, then X has worked on Topic Z", for interrogating the file, and for tracing associations between names linked through various predicates, have been written in this language. [1]

8.3 Output Considerations

Turning to operational problems of output, the question of limitations of computer printout language to, in most cases, a single set of upper case alphabetic characters, numerals, and a few special symbols, [2] is a serious factor in customer acceptance with respect to appearance -- format, legibility, readability. Involved here are questions previously mentioned. Where, in the only presently available outputs of machine-generated indexes, the KWIC type permuted title indexes, should the indexing access point "slot" be on the page? Should all or only part of the title be displayed? Should 60- or 106-character lines be used? More detailed discussion of these and related points are provided by, for example, Youden (1963 [658]) Kennedy (1962 [311]) and Brandenberg (1963 [80]).

A separate, but related question, is how much identification, and in what form, should be provided for the item itself either directly as a part of the index entry or by cross-reference to the address of more detailed information. There seems to be quite general agreement that the typical user needs something more than author's name and title

[1] Kochen, et al, 1962 [328], p. 34.

[2] See, for example, Lipetz, 1960 [365], p. 252: "A disadvantage of keypunched cards, however, is the lack of capacity to record or to print other symbols than a one-case alphabet, one case of arabic numerals, and about a dozen punctuation marks and miscellaneous symbols. Citations in the scientific literature generally make use of a much larger number of significant symbols: multiple cases, multiple fonts, italics, boldface, Greek letters, mathematical symbols, etc." Note, however, that Chemical-Biological Activities, a digest produced by Chemical Abstracts Service, uses printouts of the modified IBM 1403 chain printer, using 120 characters (see Fig. 5).

alone to guide him. 1/ However, if the full bibliographic citation, perhaps the abstract as well, is to be printed out by machine, the problems of limited character set are even more severe. This problem is today being solved, in some cases, by separate operations involving sorting and assembly of the full citations and abstracts of the items indexed, separately prepared, for photographic reproduction or typesetting. Hopefully, this partial solution will become obsolete as automatic type-composition equipment and computer-prepared typesetting techniques become more generally available.

Operational considerations thus involve the costs, the availability, and the limitations of equipment now usable for machine-generated index production. Schultz and Schwartz report, as of October, 1962.

> "There are two major bottlenecks in automated index production caused by inadequate equipment development at the present state-of-the-art:
>
> "1. There is no way of using automatic input of the printed page or the indexer's notes;
>
> "2. There is insufficient flexibility in the forms of output available for a computer-produced index.
>
> Both of these areas are being worked on by equipment manufacturers, and an early solution has been promised." 2/

In general, operational considerations of this type do not affect the appraisal of automatic assignment indexing techniques, because these have not yet been developed to the point of practical application on any realistic scale. Moreover, the difficulties of problem definition and basic understanding of language and meaning yet remaining to be resolved are such that radical new advances in computer technology, associative memories, character readers and pattern recognition devices may completely alter the picture before practical systems are ready for operational tests. Thus, for example, it is claimed:

> "It appears desirable to begin experimentation with automatic indexing so that solutions will become known by the time character recognition equipment will have passed the laboratory stage." 3/

Similarly, Doyle suggests that the "present rate of solution of the intellectual problems of IR is sufficiently slow that these advanced devices will be in common use long before IR will truly benefit from their presence", and he urges that researchers proceed as though such machines were already with us. 4/

1/ Compare, for example, Montgomery and Swanson, 1962 [421], p. 366: "This study suggests that indexing should be based on more than titles and that a bibliographic citation system should present to the requestor something more than titles"; See also, in addition to references cited, p. 61, footnote 1, IBM "ACSI-matic auto-abstracting project...", Vol 3, 1961 [290], p. 89: "The use of titles in document searching without any additional abstract seems to lead to a high number of ... errors, i.e., accepting documents which should be rejected, as not enough information is available to judge the pertinence of documents."

2/ Schultz and Schwartz, 1962 [531], p. 432.

3/ Levery, 1963 [359], p. 235.

4/ Doyle, 1961 [169], p. 3.

9. CONCLUSION: APPRAISAL OF THE STATE OF THE ART IN AUTOMATIC INDEXING

Notwithstanding the difficulties of evaluation we have discussed, we shall herewith attempt to evaluate the present state of the art in automatic indexing techniques, using such available criteria as seem most appropriate. First, we suggest that all of out initial questions except possibly the last, can today be answered affirmatively. "Is indexing by machine possible at all?" To this we can answer an unequivocal "yes" in view of the many examples of KWIC type indexes extant and in practical use. Secondly, "Is what can be done by machine properly termed 'abstracting', 'indexing', or 'classifying'?" If, by definition, word indexing of any kind is not "properly termed... indexing", then, as we have seen, automatic derivative indexing, such as KWIC, or the selection of words to serve as index tags based upon the frequencies of their occurrence in text, is not so either.

The fundamental Luhn concept for indexing based on word frequencies is, as we have seen, straightforward: namely that, after disregarding the most frequent "common words", especially those that are syntactic-function words -- articles, conjunctions, prepositions, and the like, together with those words that occur infrequently in a given text, the remaining high frequency words should give a reasonable indication of what the author was writing "about". Critiques of the Luhn position have been made on several-fold grounds:

(1) Information-theoretic - that, in fact, the most information is conveyed by the least frequent words.
(2) Absolute vs. relative frequencies of usage within specialized fields.
(3) Modifications of semantic purport by contextual and syntactic associations.
(4) Problems of synonymity and, conversely, of orthographically identical words. 1/
(5) Multi-aspect points of interest, and future need of access to material the author himself did not emphasize.

The last point raises again the criticisms that have been made against derivative, extractive or "word" indexing of all types. To repeat, although such procedures may index "as the author himself indexed best -- in his own language", the significant points are (1) there may be peripheral, minor, or unrecognized aspects of his topic and incidental information disclosed, of future interest to others, which the author himself is in no special position to recognize, and (2) notwithstanding the "author's own terminology" being current usage rather than the "fossilized" vocabulary of any previously established classification or indexing scheme, this very "currency" changes from field to field and, quite literally, from day to day. Nevertheless, it should be re-emphasized that the validity of these criticisms is not limited to automatic derivative indexing as such, but rather is applicable against any indexing system whatsoever, manual or machine, which is so strictly limited to author-terminology, author-emphases, and the consideration of the document at hand as a self-contained entity, without regard to other documents in a collection, in a particular field, and without respect to specific user needs. By contrast to this type of limitation, more promising approaches should stress both similarities and differences between a new document and previously received documents, between documents "belonging" to some definable category, or not, and even, as responsive to a particular user's profile-of-interest, or not.

1/ See Baxendale, 1962 [42], pp. 67-68: "... resolution of orthographic ambiguities is a non-trivial and over-riding prerequisite for the computer processing of text...", p. 67.