

## 5. GRAPHICAL ANALYSIS OF TEXT

*Introduction.* The informal criterion developed to select documents on the topic of Bradford's Law of Scattering had two components: (a) a precise graphical component, viz. that documents contain in the body of their text at least one occurrence of the word 'BRADFORD', or a variant of it, or equivalent forms in non-English languages; and (b) an inexact semantic component, viz. that at least one occurrence of '-BRADFORD-' is cited in a passage of text which the author has interpreted as dealing with a type of natural regularity. To obtain a satisfactorily precise selection criterion for the topic literature, the second component must be replaced by a simple graphical component. As noted, there was no small number of text pieces which obviously characterised all suitable 'BRADFORD'~containing passages, so a more formal analysis is required. The aim of the present Section is to present this formal analysis, and so complete the graphical selection criterion. A supplementary task of the analysis is to identify the various forms of 'BRADFORD' which occur in this literature.

### A. PREVIEW OF ANALYSIS

The methods employed in this section are now introduced and explained, and important terms to be used are defined. Very broadly, the analysis consists of the formation of a large list of what may be termed 'candidate text pieces' or 'candidate strings', followed by a progressive reduction of this list by various criteria to achieve the desired aim.

*What are Suitable Unit Strings?* Since the body of the text of documents is primarily a linear sequence of characters, it is customary to refer to the required text pieces most generally as character *strings*. The aim of the analysis is to identify strings which convey the notion of the particular regularity, which, with 'BRADFORD', are considered to define the topic literature. As noted in Chapter 2, suitable graphical correlates should be found in specific words, specific parts of words or in several specific words. The primary choice of string unit, then, is the *graphic word*, which, in the main, is a contiguous sequence of alphanumeric characters surrounded by spaces (or blank characters). Graphic words differ when their explicit form --that is the order and the identity of their characters -- differs, for example 'BRADFORD' and 'BRADFORD'S' are two distinct graphic words; in fact, the term *word-form* is often used synonymously with 'graphic word'. The example shows the convention to be adopted in referring to graphic words. A second string unit to be used later in this analysis is the *graphic stem*, a contiguous substring of a

graphic word, most often obtained by its right truncation; for example '-BRADFORD-' is a graphic stem of the graphic words 'BRADFORD', 'BRADFORD'S', 'BRADFORDIZE', and 'ZIPF-BRADFORD'. This example shows the convention to be adopted in referring to graphic stems. (A third string unit is the *graphic phrase*, that is several graphic words in a fixed sequence, but this will not be considered till the next Chapter). It should be noted that in the following discussion, it is usually convenient to drop the qualifier 'graphic' unless emphasis is required and ambiguity results. Repetitions of a specific string in text are referred to as *occurrences* of that string, or more customarily, as *tokens* of that *type*.

*Obtaining Specific Graphic Words: Extraction from Text.* The words to be identified occur in the text which surrounds any word with the stem '-BRADFORD-', or its equivalent forms in non-English languages, in the body of the text of all documents on the topic. To identify these words, we first extract from each document a passage of text, centred on each occurrence of such a word, and then isolate each appropriate word of the extract. In the present analysis, technical considerations limited the extract size to seven word places. These places were allocated to the 'BRADFORD' variant, and to the three words both preceding and following it, with the proviso that these words were taken from the same sentence, or the same sentence-like string, for example, a heading or a caption. Several conventions were adopted with regards to these extracts: the extracts are termed *3-extracts*, consisting as they do of a *key word* containing the *key stem* '-BRADFORD-', and up to 3 words on either side of it; the non-key words are termed the *3-surrounds*, or *3-surrounding words*. More generally, we may refer to *n-extracts*, with at most  $(2n+1)$  words, viz. a key word and the *n-surrounds*.

*Treatment of Extracts.* In the present analysis, 3-extracts were copied into a machine-readable database, as near as possibly verbatim. But the transcription is not so much a copying as an *encoding*, from the orthographies of the various languages into one standard machine code, ASCII. Since ASCII is most compatible with the orthography of English, the extracts undergoing the greatest distortion are from languages with orthographies most different from that of English. For languages with Roman scripts, graphic vandalism is most noticeable in the loss of diacritical marks and digraphs. For languages in Cyrillic and Hebrew scripts, transliterations into the Roman script of English, by protocols described elsewhere, were carried out prior to machine entry. Regrettably, the author was unable to so transliterate the scripts of Chinese and Japanese, and documents in these languages have not been analysed here. It is also problematic as to exactly what should count as a graphic word in the wholly or partially logographic scripts of these languages -- as it is with the logographic script of

mathematics. Some of the difficulties in encoding extracts, and of recognising graphic words therein, are discussed below.

Another issue to address here is the number of documents that must be analysed to obtain an adequate diagnosis of the topic literature. If only a small randomly-selected sample of documents from the major languages were used, there would be considerable reduction in the labour of preparing extracts. But part of the initial motivation for this analysis, the observation that no small number of obvious text pieces seems to characterise all suitable 'BRADFORD'-containing passages, suggests that at least a very large sample should be used. To err to caution, the greater part of the collection was used in the preparation of extracts, and additional procedures were devised to test all suitable documents.

*Selecting Characteristic Words: Word Ranking.* The initial observation -- that a great number of strings could be needed to characterise the whole collection on the topic -- suggests that in selecting these strings, every unique string of the same kind, in all extracts, should be treated as a candidate. But it is also likely -- if the topic has any integrity, and the extracts are sufficiently wide -- that a moderate number of strings in the extracts should characterise a large part of the collection. These strings would be the most common strings, that is the upper ranking strings when all strings in the 3-surrounds are ordered by their individual frequency of occurrence, in a sense to be defined. It would be less time-consuming to select these strings first, even leaving those documents or extracts not possessing them to be individually scrutinised for additional choices.

By what frequency measure should the strings in the 3-surrounds be ranked? There are several obvious measures of how strongly a string is characteristic of a subject literature. One measure is the number of occurrences of the string in the total literature of the topic, or, in the present case, the *number of occurrences* of the string in the set of all 3-surrounds from this literature. This measure could be normalised, for example, for the literature of each language, by the total number of comparable strings in the 3-surrounds. Another measure is the *number of documents* on the topic which contain, either exactly or at least, a fixed number of occurrences of the string in the whole literature, or, in the present case, in its 3-surrounds. Again, this measure could be normalised by the total number of documents. For the present, the frequency measure referred to, unless stated otherwise, is the first mentioned measure.<sup>113</sup>

---

<sup>113</sup>A third measure is the *number of n-extracts* in the literature of the topic which contain, either exactly or at least, a fixed number of occurrences of the string, i.e. how frequently the string occurs within *n*

*Characteristic Words and Diagnostic Words.* Words with high values of these measures may be strongly *characteristic* of the 3-extracts, but they are not necessarily *diagnostic* of them.<sup>114</sup> To be diagnostic, these highly-occurring words must be absent, or at least be relatively rare, in all other comparable passages in the literature which contain the word stem '-BRADFORD-', that is in the appropriate background literature for this analysis. (If there were no other documents containing the word-stem '-BRADFORD-', then the only diagnostic string needed would this stem). Since all the documents to be analysed have been judged to be on the topic by interpretation of these passages, one or more passage in each document should contain the required diagnostic strings, and, one hopes, the 3-extracts should contain the required diagnostic words. But the 3-extracts also contain frequently occurring words which are patently not diagnostic: these occur either equally frequently in a wider literature, including the appropriate background, or only in the part of the literature on the topic, that of some sub-topic. How are these words to be removed from consideration?

*Removing Non-Diagnostic Words.* The most efficient and general solution to the first part of this problem is to compare the characterisation of the passages on the topic with that of passages from the background against which the diagnosis is intended.<sup>115</sup> A far less efficient solution is to ignore the problem here and correct for it later: that is to pick only characteristic words in this analysis, and to test their diagnostic ability against a background literature in a subsequent analysis. This is inefficient in that it may well require repetitions of the initial analysis. But this is the procedure adopted here -- with two exceptions, which are now discussed.

*Removing Non-Diagnostic Words: The Grammatical Function Words.* A preliminary screening-out of one group of potentially bad diagnostic words can be made within the chosen procedure. These are the *grammatical function words*, words which have a

---

words of the stem '-BRADFORD-'. This measure could be normalised by the number of *n*-extracts in the total literature. More complex measures could be of use, e.g. the number of documents which have at least a fixed proportion, say 50%, of their *n*-extracts with at least one occurrence of a string -- i.e. using '-BRADFORD-' in conjunction with the string on at least half of the occasions that they use '-BRADFORD-' at all.

<sup>114</sup>As used here, the term *characteristic* means 'typical' or, more strongly, 'always noticeably present', where 'noticeably' is interpreted as 'having a high frequency measure'. It does not necessarily imply 'distinguishing', which effectively requires the addition of '... and noticeably conspicuously absent elsewhere'; this is reserved for the term *diagnostic*. Negative characteristics or diagnostics, defined by interchanging 'present' and 'absent' etc., are not considered here.

<sup>115</sup>As in conventional studies of recall and precision, when some of the foreground (a sample of a topic's literature) and some of the background (samples of related literatures) are present in the analysis. Suitable diagnostic strings are selected directly. Using the terms of conventional studies, the present analysis deals only with the enhancement of *recall*; testing the *precision* of the selection is delayed.

primarily syntactic function. Critically, these words tend to be amongst the most frequently occurring of all words, in all categories of discourse -- hence their poor diagnostic ability. Just as critically, they belong to several grammatical word classes ('parts of speech') which have relatively small and closed memberships in each language. Therefore, they may be identified in dictionaries by their grammatical class (for example: articles, conjunctions, prepositions, pronouns, auxiliary verbs), and fully and explicitly listed as graphic words. In indexing practice, grammatical function words -- or their most frequent members -- belong to those words proscribed as non-indexing terms, and which are explicitly tabulated in *stopword lists*. Thus, the *stopword lists* of A&I Services provide a ready source for many grammatical function words.

The grammatical function words may be automatically screened-out with safety, leaving what may be termed notional or *semantic* words. (These terms are not intended to imply that grammatical function words convey no meaning at all, though by comparison with nouns in particular, they convey little when removed from their association with these words). The screening of semantic words for bad diagnostics, if it is to be done with complete safety, requires the characterisation of a particular background. In the creation of practical *stopword lists*, though, a number of such words may be obvious without a full frequency analysis once this background is specified; for example the word 'LIBRARY' seems unlikely to distinguish the literature on library management from the wider literature of Library and Information Science.

*Removing Non-Diagnostic Words: Words Characteristic of Sub-topics Only.*

Assistance with the second part of the problem -- the identification and removal of high-frequency semantic words that are used only in a sub-topic literature -- is provided by a comparison of the two measures described above, for each word in the list. A word with a high number of occurrences in the text extracts disproportionate to the number of contributing documents, is likely to be an important word only for some specialist theme within the topic. It is less likely to reflect the style of expression of an author or school, without regard to theme. (Of course, if all high-frequency semantic words of a topic were so distributed, the integrity of the whole topic would be called into question). Rejection of such words should not be automatic, and they should probably be first assessed semantically.

*Selecting Stems from Characteristic Words.* Having assembled in principle a list of semantic words which are characteristic, and possibly diagnostic, of the 3-extracts in each language, we turn our attention to converting these words to stems. The motivation for using word stems, rather than full graphic words, is as follows. It is

common for different graphic words, with suitably-placed common substrings, to convey sufficiently similar meanings, for present purposes, within one language. For example, it may be asserted that the aforementioned graphic words 'BRADFORD', 'BRADFORD'S', 'BRADFORDIZE', and 'ZIPF-BRADFORD', are sufficiently synonymous to their common substring '-BRADFORD-', wherever they may occur in the literature, for the substring to be substituted for them without appreciable change of meaning. This being the case, it is economical of some commodity such as page space or computer memory, or at least it is more elegant, to specify only this one string instead of four strings. With this step, the initial assertion is now taken to be generally true, that is that *all* graphic words containing this stem are sufficiently synonymous. If this wider assertion is correct, and unsuspected graphic words with this stem occur where the four initial graphic words do not, then the stem will characterise more text (extracts, documents) than these four initial words. In summary: stemming gives greater return (characterisation, retrieval) for smaller input (specification of strings), without loss of precision. If however, the wider assertion is false, there will be a loss in precision; the possibilities for error in stemming are obvious.

So rather than using, as candidates for diagnostic strings, a large number of the more frequently-occurring graphic words from the 3-surrounds of each language, it was decided to use a smaller number of frequently-occurring stems. The hyperbolic shape of the word frequency distributions (in Zipf form), at least for the large languages with long lists of 3-surrounds, ensures that the frequently-occurring stems can be identified from the frequently-occurring graphic words; that is, words not in the upper ranks can not produce stems in the upper ranks. (For smaller languages with shorter word lists, where the distributions are likely more uniform, this may not be the case, and lower ranking words must also be considered). The relatively small number of words to be treated means that this process can be performed manually, and perhaps more accurately than might be possible with existing stemming algorithms. Manual stemming was required, if only from the fact that suitable algorithms for the 18 languages other than English were not known to the author.

*The Problem of Texts in Many Languages.* So far, the graphic word lists obtained, in principle, from the 3-surrounds of each language have been treated separately. Needless to say, there are relatively few graphic words common to different languages which have common meaning -- even borrowed words acquire inflections characteristic of the borrowing language. Eventually, however, if the literature on the topic is to be treated as a semantic whole, the separate language lists must be combined. There are two broad approaches to this. One is to give priority to the graphic or textual aspect, wherein the separate lists remain 'incommensurable entities'. The only useful allowable operation is a

concatenation of lists, with the set of diagnostic strings from each language preserved in the large total list. The other approach is to give priority to the semantic aspect, and establish approximate synonymies between words and stems across all languages. This approach can be best followed where, as here, the domain of discourse is common to all languages and rather narrow. Effectively, this approach involves the translation of all words into one language, and the construction of a smaller common set. All strings in other languages must be subsequently derived from this set. In the present study, English would fulfil this role of a semantic metalanguage, as well as remaining one of the object languages.<sup>116</sup>

*Selecting Characteristic Words and Stems Across Languages.* It was decided to adopt an intermediate approach which recognised the topic as a semantic whole but also acknowledged the graphic isolation of each language. First, to select candidate words, the top words of each language should be selected separately, but with the selection of each such word, the approximately synonymous words in the other languages -- if such exist -- should also be selected for these languages. (Synonymies should not be established within one language, however, other than for the narrow purpose of stemming). The question of exactly how many ranks should be considered in each language would remain open until the actual distributions could be studied, though clearly more ranks would have to be considered from the larger languages. Second, only after the separate lists of candidate words were established for each language, and there had been compression of the separate lists to form lists of candidate stems, should the second approach above be completely implemented: that is, subsuming semantically-equivalent stems across the language lists under the stem of the semantic metalanguage. Since this language is, in fact, English in a different guise, the convention to be adopted for such expressions will be to bold the normal English representation, for example **'-BRADFORD-'** is the semantic metalanguage stem, '-BRADFORD-' is its representation in English and Portuguese, and '-BREDFORD-' is its representation in transliterated Russian.

*Final Choice of Characteristic Stems: Evaluating the Candidates.* At an earlier stage, to remove words which are not diagnostic against background text from a list of words that are characteristic of the text of a topic, it was convenient to order words by their number of occurrences in the whole text. However, in characterising a collection of documents on a topic, it is of more value to use the second measure introduced above,

---

<sup>116</sup>It should be noted that to form the set of documents under analysis, text passages around 'BRADFORD' or its equivalents had to be interpreted. Except for Hebrew, Japanese and Chinese documents -- and those documents not in English for which English translations could be obtained -- the present author had to translate these passages word-by-word into English. That is, the semantic aspect is already assumed in the formation of the collection, and the word lists are not incommensurate.

viz. the *number of documents* in which the word or stem occurs, at least some chosen number of times, in (for example) the combined 3-extracts of each document. The most generous value of the parameter is *one occurrence per document*, that is a document need contain the stem in only one 3-extract. Using a different parlance, we may say that the document is *retrieved* from the collection *by* a search for *that stem* in proximity to the stem '-BRADFORD-'. The individual ability of each candidate stem to characterise or retrieve the documents of the collection may be evaluated by this measure, and all candidate stems may be ranked accordingly.

The final choice of strings (for example, stems) can be made from the cumulative value of the measure as a function of the cumulative number of strings selected -- what may be termed the *retrieval profile* of the candidate strings. Correctly, this profile should be termed the *maximum retrieval profile*, where the strings are so ordered as to ensure the greatest cumulative value of the measure, for any cumulative number of strings. However, retrieval profiles are not immediately obtained with the chosen measure -- the number or percentage of documents retrieved -- since the value of the measure for two strings is not a simple calculable function of the individual values; rather, it involves Boolean set operators. If by the documents retrieved or characterised by two strings X and Y, we mean the documents retrieved by X and by Y separately, not excluding those with both strings X and Y, then the value of the measure for both strings is the number of documents in the *union* of the sets of documents retrieved by the individual strings X and Y.<sup>117</sup> To obtain the desired retrieval profile, and the final choice of strings, it is necessary in principle to check the unions of the sets of documents retrieved for all individual candidate strings, in all possible combinations. In practice, it is likely that the initial strings will be amongst those in the top ranks, when candidate strings are ordered by their individual values of the measure.

*Selecting Strings: An Extra Procedure.* Finally, it may be noted that it is not necessary, but only convenient, to evaluate strings with respect to the established (in principle) 3-extracts. With a manageably small list of good candidate strings, it would be possible to scrutinise longer extracts containing the stem '-BRADFORD-' *directly* in the documents themselves. This technique will be usefully employed below in Subsection C, §3, p.211.

---

<sup>117</sup>The *intersection* of the respective sets of documents provides the number of documents which have an occurrence of both strings together in their text passages. Measures could be developed involving the joint presence of two (or more) strings in each text extract. As the text extracts used in the present analysis are quite short, this approach will not be developed. In fact, as used to diagnose the documents on the topic, the measures used are already of a *joint occurrence*, that of the string and of '-BRADFORD-'; both components are required in close proximity. In Boolean terms, the collection is to be characterised by each of its documents possessing at least one occurrence of: ('-BRADFORD-' and string X, in close proximity) or ('-BRADFORD-' and string Y, in close proximity) or ...etc.

*Selecting Strings: A Semantic Correlation.* A strictly unnecessary but sensible constraint on the strings finally selected is that -- when containing or used in association with the word 'BRADFORD' -- they must be unproblematically indicative of the topic. In the present case, this means that the strings should readily relate to the inexact semantic component (b) of the informal selection criterion, which was stated in the introduction to this Section on p.191.

## **B. DETAILS OF METHODS**

*Records Analysed.* At the closure of the collection in mid-1993, 1310 analytical-level scholarly documents which meet the two components of the informal selection criterion had been located. With the removal from this group of 94 documents in Chinese and Japanese, 1216 documents were suitable for the text word analysis needed to find a graphical replacement for the semantic component of this criterion. However, a full analysis was restricted to those 1040 documents which were published prior to 1987 and in a single language each. The 176 documents published either at a later date or in two or more languages were analysed subsequently, but by a different method (see Subsection C, §3, p.211).

*Data Preparation.* In each of the 1040 documents selected for text analysis, every occurrence of a word containing the stem '-BRADFORD-' was noted. For each occurrence, a piece of text containing this word was copied, in upper case and without diacritical marks, as a single item into a new vector field, (116) BRADTEXT, in the database of documents, the BRAD File. Each such text extract consisted of the three words before the 'BRADFORD' variant, the 'BRADFORD' variant itself (that is the keyword), and the three words after it, all from the same sentence or sentence-like string. Where the keyword was closer to the beginning or the end of its sentence than three words, fewer surrounding words were entered, or, alternatively, the appropriate 3-surrounds could be considered as blanks. Where the key was misspelt or abbreviated, the extract was suitably labelled. Likewise, the nature of the text from which the extract was taken, for example whether it was from a heading, or from a footnote, and so on, was also recorded. Full details of the protocol used in the preparation of these 3-extracts, and the structure of the vector field created, are provided in Section 2 of the Appendix.

Details of the 3-extracts from the 1040 documents are given in Tables 3-3 and 3-4. 7067 extracts were obtained for analysis in 19 separate languages, providing a total of 42,734 words of which 7067 were keys and 35,667 were surrounds. 6293, or 89.0%, of the

extracts come from the main body of the text, 450 (6.4%) from captions to figures and tables, and 237 (3.4%) from headings in the body of the text. The mean number of extracts per record is 6.8, and the mean number of surrounds per record is 34.3. English accounts for over two-thirds of the documents (703 or 67.6%) and nearly three-quarters of the keys and surrounds (5196 and 26,184, respectively, or 73.5% and 73.4%) that are to be analysed. It should be noted that this analysis of 1040 documents is not biased for language: the proportions of documents in the major languages are similar to those in the full analysable collection of 1338 documents -- that is, to the 1310 documents in the informally-defined collection supplemented with the 28 'PRE', 'PAR', and 'DIF' documents - - when the Chinese and the Japanese documents are likewise removed. For example, the proportions here for English (with 703 documents), Russian (144) and Portuguese (46) are 67.6%, 13.8%, and 4.4%, whereas in the full collection they are, respectively, 67.4% (838 documents), 12.9% (160), and 4.3% (54).

*Data File Analysis.* Two programs, *ContextA* and *ContextB*, were written to convert the vector field of 3-extracts in the BRAD File into lines of a DFsortfile. Each item (that is, each 3-extract) in the field was transferred onto a line of the DFsortfile, and initialised with the document number, extract number, language, and extract type. Within the extract, each 3-surround, or a blank string, was placed in one of six fixed positions with respect to the centrally-placed key.<sup>118</sup> By appropriate operations, the DFsortfile could be separated into two DFsortfiles, one consisting only of keys and the other only of 3-surrounds, which could then be separately analysed. An additional program, *Dfriqs*, was written to convert each of these files back into two new vector fields in the BRAD File, one with the keys as items, and the other with the pooled 3-surrounds (without regard to position) as items. Thus the text extracts could be variously analysed with a range of programs operating either on RIQS-type or DFsort-type files on the IBM3090. In addition, separate databases were constructed for the 3-extracts, for the pooled 3-surrounds, and for the keys, on a 486 PC using the BRS/SEARCH software. Further information on File Manipulation Tools is given in Section 3 of the Appendix.

## C. RESULTS

---

<sup>118</sup>The programs *ContextA* and *ContextB* broke each vector item into an ordered set of words, and, where present, the initial one-letter code for the source of extracts other than the text. They then identified the keyword and its position in the ordered set. An analysis of keywords showed that keys could be identified by their inclusion of the substrings '-BRADF-', 'BREDF-', or 'BRAEDF-', except in the cases of errors or abbreviations which were noted appropriately in the initial one-letter code. These cases required special treatment. Special treatment was also needed for the few cases where a 'BRADFORD' variant appeared more than once in an extract, i.e. both as a key and as a 3-surround; such 'BRADFORD' variants would be keys in the previous or the subsequent extract(s).

Results are presented in three Subsections:

- §1. the characterisation of the keywords in the 3-extracts;
- §2. an analysis of the 3-surrounding words to establish -- with the keys -- a likely set of diagnostic strings; and
- §3. an extension of the analysis to documents from which 3-extracts were not prepared.

## §1. CHARACTERISATION OF KEYWORDS

*Key Word-forms.* The extract keys were separated by language and sorted into graphic word-forms or types. A synopsis of the results is presented in Table 3-5. Noteworthy is the great variety, even within one language, of the exact form of the key, which results primarily from inflectional endings and from hyphenated compounds. Nevertheless, there is generally a strong concentration of occurrences in the principal form found in each language. In English-language documents, for example, there are 55 distinct graphical word-forms in the keys for 5196 occurrences or word tokens; but half of all occurrences (2609, 50.2%) are provided by the one-word form 'BRADFORD', and this word-form is found at least once in two-thirds (467, 66.4%) of the documents.

*Key Word-stems.* Despite the great variety of form, nearly all keys in Roman script contain the stem '-BRADFORD-', and nearly all keys in appropriately-transliterated Hebrew and Cyrillic scripts contain the stems '-BRADFORD-' and '-BREFDORD-', respectively. The number of documents in each language which contain at least one occurrence of either of these keystem is summarised in Table 3-6. It can be seen that for English, for example, all 703 documents analysed contain at least one occurrence each of the keystem '-BRADFORD-', although six records also contain some abbreviated form of the key (for example 'B-LAW') and four records also contain erroneous keys (for example 'BRANDFORD-ZIPF'). For Portuguese, of the 46 documents analysed, 45 contain at least one occurrence of the keystem '-BRADFORD-', but the one document with an erroneous key word ('BRADFORT') does not contain any occurrence of the correct keystem. In total, 891 documents (85.7%) in 17 languages contain at least one occurrence of the keystem '-BRADFORD-', and 146 documents (14.0%) in 4 languages contain at least one occurrence of the keystem '-BREFDORD-', after appropriate transliteration to Roman script if necessary. Three documents, however, do not contain either keystem in their text. This is clearly an error for two of the documents, which contain only one keyword each, but does not appear to be the case for the third document, which is in Polish and contains two uses of the keystem '-BRAEDFORD-'.

## §2. ANALYSIS OF THE 3-SURROUNDING WORDS

*Ranking Word-forms by Numbers of Occurrences and of Documents.* The words surrounding the keys were separated by language and sorted by exact graphic form. For each word-form, the number of occurrences, and the number of documents containing at least one occurrence of the word-form, were determined. A table was prepared for each language with these data and in which the word-forms are ranked in order of decreasing number of occurrences. Tables 3-7(a-c) present, inter alia, the upper portions of these tables for the top three languages in the analysis, English, Russian and Portuguese, in that order. Noteworthy is the high prevalence of grammatical function words in the upper ranks by either measure, and most especially in English and Portuguese. It may be seen, for example, that out of 2680 word-forms in the English 3-surrounds, that with the greatest number of occurrences is 'THE', which functions grammatically as an article; it provides nearly one tenth (2674, 9.45%) of all 26,184 occurrences or tokens. The word-form found in the 3-surrounds of the greatest number of documents in English (501 of 703, or 71.3%) is 'OF', which functions grammatically as a preposition.

*Removing Poor Diagnostics.* Plausible grammatical function words were identified and removed from each language table, leaving a ranked list of semantic words forms.<sup>119</sup> In Tables 3-7(a-c), for English, Russian and Portuguese, the resulting semantic words have been bolded; in Tables 3-8(a), 3-9(a), and 3-10 the grammatical function words that were removed for all 19 languages are separately listed. It may be noted in Tables 3-7(a-c)

---

<sup>119</sup>(1) While it is not difficult on a casual basis to recognise many grammatical function words in a language one knows quite well, it is not so easy a task to explicitly define and assemble a full list of such words, and a progressively more difficult task as unfamiliar languages are included. Here, grammatical function words were taken to be words other than nouns, verbs (but not auxiliary verbs), and (with less precision) adjectives and adverbs. Since the words under consideration are isolated from their context, it is not always possible to be sure what grammatical class (or lexical word) a graphic word belongs to; e.g. does a particular occurrence of 'HAS' in English text represent the auxiliary verb or the verb of possession, and of 'ROUND' the preposition or the adjective. The crude procedure used was to assume all representations of the graphical word were in fact a grammatical function word, if a majority, or even a large minority, of them appeared to be so in the text. Also, the grammatical function sought may reside not in an isolated word but in two or more adjacent words, a more likely situation when many languages are considered and exact word-word synonymies do not exist. Some mistakes have been made, e.g. the English 'ACCORDING' (chiefly, 'ACCORDING TO') and the Russian 'SOGLASNO' (as well as the Hungarian 'SZERINT', Czech 'PODLE' and Slovak 'PODLA') should have been treated as non-semantic words. However, it is doubtful that any proper semantic word likely to have significant diagnostic value has been unjustly expelled from these lists. (2) In addition to grammatical function words, a class of odd-looking words, in the main quite rare, have been removed. These words appear as a consequence of policies adopted in the encoding of text, especially with regard to the encoding of in-text bibliographic surrogates and mathematical text (see Appendix). Thus mathematical 'words' such as 'B' and 'LOG(N)', and the bibliographic '<R>', have been treated as stopwords; however, 'words' representing the operators, e.g. '=' and '>', have been treated as semantic words. A (perhaps idiosyncratic) justification for this is that stopped terms are rather like pronouns, acquiring more than limited meaning only when their reference is explicated, whereas retained operators are verb-like.

that the top semantic word-form in English by either measure is 'LAW', accounting for 7.1% (1846) of occurrences and appearing in 68.8% (484) of documents. Likewise, the most frequent semantic word-form in Russian is 'ZAKON' (2.8% occurrences, in 34.7% documents), and in Portuguese is 'LEI' (9.5% occurrences, in 87.0% documents).

To identify highly-occurring semantic words that are perhaps more characteristic of sub-topics than of the whole topic, the ranks of the top word-forms by number of occurrences and by numbers of containing documents were compared by two methods: by the simple subtraction of respective rank numbers, and by dividing the occurrence frequency by the respective document frequency. For any upper-ranking word-form, a high absolute difference between rank numbers -- values which are bolded in Tables 3-7(a-c) -- indicates that it occurs in extracts disproportionately limited to few documents for its rank. It has been suggested that such word-forms may be poor diagnostics for the whole topic. An example is provided by the word-form 'MULTIPLIER' in Table 3-7(a). It is ranked 30th by frequency of occurrences but only 57th by frequency in documents, appearing in only half the documents (39 as compared with 65-80) that one might expect from the former ranking. In general, however, there is a good matching between the two rankings, especially in the uppermost ranks, which is only in part due to the prevalence of grammatical function words.<sup>120</sup> So from this aspect at least, the majority of the semantic word-forms listed would appear to be suitable candidates for diagnostics.

*Selecting Top Strings Across Languages.* The top ranking semantic word-forms in the 3-surrounds of each of the 19 languages are compared in one chart, Tables 3-8(a), 3-9(a) and 3-10; word-forms are aligned by decreasing frequency of occurrence. It can be seen, for example, that the top word-forms of the top five languages are: 'LAW' (for English, with >5% of all occurrences -- in fact, as already noted, 7.1% of all occurrences); 'ZAKON' (for Russian, with between 2.5% and 5% of occurrences); 'LEI' (for Portuguese, with >5% occurrences); 'LEY' (for Spanish, with >5% of occurrences); and 'GESETZ' (for German, with between 2.5% and 5% of occurrences).<sup>121</sup>

---

<sup>120</sup>(1) In comparing the two rankings, the magnitude of both the difference and the quotient have to be judged relative to that of word-forms with similar rankings (by one of the measures). Local rank differences become progressively more distorted as more and larger rank dislocations occur, and the maximum possible quotient of frequencies declines as rank (and maximum number of occurrences) increases. This procedure becomes useless at lower ranks, but word forms with low frequencies are not of interest anyway. (2) Good matching between the two rankings is shown in the *recurrence of zero values when individual rank differences are summed downwards*, e.g. as far as (occurrence) rank 19 in English (to 'WHICH'), rank 24 in Russian (to 'LOTKA'), and rank 15 in Portuguese (to 'UM').

<sup>121</sup>No precise solution was found to the problem of how to scale out the vastly different sizes of the different language literatures (as shown in Table 3-3). Unless a disproportionate weighting is afforded the smaller literatures, there would be no point in extending the analysis beyond English and Russian. As a case in point, the word form 'MULTIPLIER', considered above to be somewhat restricted in its use in English for a diagnostic, occurs nearly as frequently in the combined extracts as all Danish words (94 vs.

To establish synonymies between words across languages, English equivalents of the top-ranking semantic word-forms of the non-English languages were obtained; these are presented in Tables 3-8(b) and 3-9(b), which preserve the form of Tables 3-8(a) and 3-9(a), and in the lower portion of Table 3-10. By reference to both sets of tables, it can be seen that word-forms listed for one language quite frequently have 'exact' semantic counterparts in at least some of the other lists -- not surprisingly, as there must be some common coinage in the discourse of a common topic. For example, the top English semantic word-form 'LAW' has a counterpart in the top rank of the languages Russian ('ZAKON'), Portuguese ('LEI'), Spanish ('LEY'), German ('GESETZ'), Czech ('ZAKON'), Serbocroatian ('ZAKON'), French ('LOI'), Romanian ('LEGEA'), Polish ('PRAWEM'), Bulgarian ('ZAKONA'), Hebrew ('CHOK'), Swedish ('LAG'), and Ukrainian ('ZAKONOM'); and near the top rank in Hungarian ('TORVENY'), Slovak ('ZAKON'), and Italian ('LEGGE').

*Word Stems.* It can also be seen in Tables 3-8, 3-9 and 3-10 that there are often several word-forms in the upper ranks of each language list which share common stems and common meaning. The word-form differences prove to arise from (terminal) grammatical inflections and from compounding. For example, in the former case, the English word-forms 'LAW' and 'LAWS' prove to be singular and plural forms of one lexical word which is grammatically a noun; the Russian word-forms 'ZAKON', 'ZAKONA', 'ZAKONU', 'ZAKONOM', and 'ZAKONE' prove to be singular forms for various cases (nominative, genitive, dative, and so on) of one lexical word which is grammatically a noun. In the latter case, in German, a compound word containing 'GESETZ' is 'STREUUNGSGESETZ' [scattering law]. As suspected, the stemming of word-forms seems advisable, that is the diagnostic strings should be word-stems.

*Selecting Specific Word-Forms.* The following procedure was adopted for picking a small set of diagnostic strings suitable for all languages. First, several of the uppermost ranking semantic word-forms were selected in each language, with more ranks considered for the languages with the greater numbers of documents in the collection; these may be referred to hereafter as the 'larger' or 'major' languages. In this selection, preference was given to word-forms known to be distributed widely across documents within the language, and withheld from word-forms suspected of being too widely distributed in the language to qualify as diagnostics. Second, the semantic equivalents of each selected word-form were identified in the lists for other languages, and also

---

108) and appears in more English documents than there are German documents (39 vs. 37). The procedure adopted, as shown in Tables 3-8(a), 3-9(a) and 3-10, was to choose more of the upper ranking word-forms from the 3-extracts of the larger languages (with have more word-forms), but each cut-off point was chosen somewhat arbitrarily. This matter is reconsidered in Chapter 4, Section 2, p.307.

selected. Third, the various inflectional and compound forms of this wider class of selected word-forms -- which will be conflated to stems -- were identified within each language, and also selected. For convenience, all word-forms selected in this manner may be termed *significant* words for the topic. The significant words chosen are bolded in Tables 3-8, 3-9 and 3-10.

The rationale for the choice of specific significant word-forms is as follows:

- The three top-ranking English semantic word-forms, viz. 'LAW', 'DISTRIBUTION', and 'SCATTERING', were selected first. As mentioned, equivalent forms of 'LAW' dominate the top few ranks of all languages. Equivalent forms of 'DISTRIBUTION' and 'SCATTERING' are found in the top five ranks of semantic word-forms in all languages where they occur at all, except in German and Serbocroatian. The equivalents of 'SCATTERING' in the Romance languages suggest that 'DISPERSION' in English (which ranks 44 in the semantic words, and 83 in all surrounds) should also be included. The next five ranking English word-forms were also considered: two are inflectional variants of prior forms, and 'DATA' is considered to be of too general use to be a diagnostic, but 'FORMULATION' and 'ZIPF' seem suitable. Equivalents of 'ZIPF' occur at similar rank in the next three top languages, and at rank 2 in French. Equivalents of 'FORMULATION' occur intermittently in the upper ranks across the tables (including rank 1 in Dutch and rank 3 in Danish); 'FORMULATION' ('FORMULATE', etc.) and its equivalents have been distinguished from 'FORMULA' and 'FORM' and their equivalents, an operation requiring some care.<sup>122</sup>
- The uppermost ranks of Russian introduce 'INFORMATSIYA', discounted as too general, and 'MODEL', equivalents of which occur at rank 29 in English, and at rank 23 in Portuguese. In addition, word-forms containing '-ZAKONOMERNOST-' have been selected, in part because of affinity with 'ZAKON'.<sup>123</sup>

---

<sup>122</sup>Several 'word equivalents' which appear in the uppermost ranks in the tables were passed over because they were judged to occur too frequently in the wider literature of information science, and so likely to be poor diagnostics for this topic -- with respect to that background. These include 'DATA' (English 'DATA', Portuguese 'DADOS', French 'DONNEES'); 'INFORMATION' (Russian 'INFORMATSIYA', Czech 'INFORMACI', and English 'INFORMATION' at semantic-word-rank 103 and all-surrounds-rank 178), and 'APPLICATION' (German 'ANWENDUNG'). This is now considered to have been an ill-advised procedure. The only background relevant to the analysis is the set of 3-surrounds of all other 'BRADFORD'~containing passages in the literature. The best that can be said for the particular choices is that it is likely that many of these passages do deal with themes in Library and Information Science, and that later analyses show the oversights to be of no consequence.

<sup>123</sup>The selection of '-ZAKONOMERNOST-' is questionable. The most frequent form, 'ZAKONOMERNOST', has rather restricted use -- see Table 3-7(b) -- although its other forms do not. It is translated into English as 'law (of nature)' but also as 'natural regularity or pattern', so it has no exact counterpart in English or the other major languages (but cf. Slovak 'ZAKONITOSTI'). Correctly, other language equivalents such as the English 'REGULARITY' (semantic-word-rank 176, surrounds-rank 279) and 'PATTERN' (semantic-word-rank 95, surrounds-rank 167) should have been included as well, unleashing a wave of additional

- The uppermost ranks of Portuguese introduce 'MULTIPLICADOR' and '1934', which have English equivalents at ranks 9 and 21 -- the former, a questionable diagnostic in English, being thereby reinstated.
- The uppermost ranks of Spanish introduce 'NUCLEO' and 'ZONAS', which have English equivalents at ranks 54 and 25. Since the English word-form 'CORE' (at semantic word rank 87, and all surrounds rank 155) is invariably given as a translation of 'NUCLEO' and its equivalents ('YADRO', 'KERN', etc.), it was allowed to accompany 'NUCLEUS' as a courtesy.
- The uppermost ranks of German introduce 'REGEL', which has an English equivalent 'RULE' at rank 272 (in semantic words, and rank 393 in all surrounds), and reinforce a prior choice with 'FORMULIERUNG' and 'FORMULIERTE'. 'ANWENDUNG' [application] is discounted as too general.
- The word-form 'LOTKA' was also selected, in part because it is consistently distributed in the upper ranks of the major languages (and French) though avoiding selection in any one. No additional word-forms need to be introduced for the smaller languages.

*Stemming Selected Word-Forms.* Fourteen sets of significant word-forms which have common stems were selected in English, and 14 or fewer similar sets were selected for each of the other languages. Within each language, these sets of word-forms were then reduced to their stems. Table 3-11 presents details of the stemming of sets of significant word-forms, for the three top languages, as follows: for English, the four sets of word-forms which reduce respectively to the stems '-LAW-', '-DISTRIBUT-', '-SCATTER-', and '-DISPERS-'; for Russian, the four sets of word-forms which reduce respectively to the stems '-ZAKON-', '-RASPREDELEN-', '-RASSEYAN-', and '-ZAKONOMERNOST-' (the last being further reducible to the first in an alternate treatment); and for Portuguese, the three sets of word-forms which reduce respectively to the stems '-LEI-', '-DISTRIBU-', and '-DISPERS-'. In this regard, a contrast between Russian and Portuguese is most apparent; for example there are eight distinct word-forms containing '-ZAKON-' in the Russian 3-surrounds (including, for example, 'ZAKON', 'ZAKONOM', and 'ZAKONOV'), but only two distinct word-forms containing the semantically-equivalent stem '-LEI-' in the Portuguese 3-surrounds (viz. 'LEI' and 'LEIS'). English is intermediate, supplementing its few inflectional forms (in the present case, 'LAW', 'LAW'S' and 'LAWS') with imaginative compounds and derived words (in the present case, 'LAW-CORE' and, compounded with the key, 'BRADFORD-LAW' and 'B-

---

synonymising. In the grouping of stems adopted below, it is allowed to accompany 'ZAKON', as 'DISPERSION' is allowed to accompany 'SCATTER'.

LAW'). Subsequently, the (significant) word-stem will be used to name the particular set of significant word-forms used to create it, and to stand for any of its members; for example, '-ZAKON-' will stand for all or any of the eight forms listed, and '-LEI-' for either or both of the two forms listed. Also, it now proves convenient to explicitly recognise the approximate synonymy of stems across all 19 languages, and to order stems in each language in the same (semantic) way.

Tables 3-12(a-d) lists each significant word-stem found in the 3-extracts for each language, in this common order. For each stem, the number of subsumed word-forms, both in the 3-surrounds *and* compounded with the keystem, are noted. It should be remarked that, at this stage of the analysis, the keyword is now reinstated in its 3-surrounds, that is the full extract is now analysed.

*Evaluating Stems.* In Tables 3-12(a-d) the values of the 'characteristic' measures introduced earlier are given for the word-stems selected in each language. One measure is the number of occurrences of the stems in the 3-extracts; this has also been normalised both by the total number of word-forms in the extracts, and by the total number of extracts, for the language. Of more value here is the other measure, the *number of documents* that contain at least one occurrence of the stem in the 3-extracts of the document; this has also been normalised by the total number of documents for the language. Thus, it can be seen that for English, as many as 70.6% (496) of the 703 documents contain at least one appropriately-sited word with the stem '-LAW-', resulting from 1916 occurrences of this stem in the 3-surrounds in four different word-forms (viz. 'LAW', 'LAWS', 'LAW'S' and 'LAW-CORE'), and from three occurrences in the keys in two different word-forms (viz. 'BRADFORD-LAW' and 'B-LAW'). Likewise, 42.8% (301) of documents contain the stem '-DISTRIBUT-', 29.6% (208) of documents contain the stem '-SCATTER-', and so on. Also noteworthy for English is the importance of compounds of '-ZIPF-' with the keystem. Altogether, this stem occurs in the 3-extracts of 26.3% (185) of documents; in the 3-surrounds alone, it occurs in less than half this number of documents (12.6% or 89 documents) in seven different forms, but in compounds with the keystem '-BRADFORD-' only, it occurs in two-thirds of these documents (17.5% or 123 documents) in 12 different forms.

Other languages show broadly similar patterns to English. To compare languages, it is now convenient to refer to the stems which are approximately synonymous across all 19 languages under common semantic rubrics, and, as established earlier, the stems of the semantic metalanguage will take the form of the English stems, but in bold font. Thus, it can be seen that the stem '**-LAW-**' occurs in a moderate to high percentage of documents in

most languages; in English, '-LAW-' occurs in 70.6% (496) of documents; in Russian, '-ZAKON-' occurs in 76.4% (110) of documents; in Portuguese, '-LEI-' occurs in 89.1% (41) documents; in Spanish, '-LEY-' occurs in 55.3% (21) of documents; and so on. Also, the importance of the stems '-ZIPF-', and (to a lesser degree) '-LOTKA-', in compounds with the keystem, is apparent in many languages. However, as may be expected from the earlier cross-language comparison of word-forms, equivalent stems do not retrieve equivalent percentages of documents (or occur equally frequently in the 3-extracts) in all languages. For example, whereas the stems '-DISTRIBUT-' and '-DISPERS-' occur in Portuguese in 21.7% (10) and 45.7% (21) of documents, respectively (as '-DISTRIBU-' and '-DISPERS-'), they occur in English in 42.8% (301) and 3.7% (26) of documents, respectively (as '-DISTRIBUT-' and '-DISPERS-').

*Problems with Cross-Language Comparisons.* The comparison of languages in the previous paragraph raises several problems of method. First, complete isomorphic synonymies have not been established. For example, the only stem matching the Russian stem '-ZAKONOMERNOST-' is the Slovak stem '-ZAKONITOST-'; it could be argued that either they should both be subsumed under '-ZAKON-', or that a more assiduous search for semantic equivalents in other languages be undertaken. Again, it could be argued that the English stem '-SCATTER-', which occurs in 29.6% (208) of documents in that language, is a more credible derivative of '-DISPERS-' than is the scarce English stem '-DISPERS-', or that both these stems should be combined under the same semantic rubric. Second, it is unlikely that there will be one common order of best retrieval across all languages, and any useful single ordering of semantic stems. The 'solution' adopted for this problem is to judiciously group stems so as to obtain a single ordering of stem groups. Clearly, the approach used in this analysis for treating multiple languages, involving a mix of graphical and semantic operations, is an unwise choice.

*Stem Groups, and Stem Group Ordering.* Stems were allocated into three groups, somewhat arbitrarily, but directed to obtaining as far as possible in each language, the greatest total retrieval in the first group, and the least total retrieval in the third group. The allocation of stems is as follows:

- Into *Group A* were placed the stems '-LAW-', '-DISTRIBUT-', '-SCATTER-' (and '-DISPERS-'), that is all word-forms stemmed by '-LAW-' in English, '-ZAKON-' in Russian, and so on through to '-DISPERS-' in Italian and '-ROZPODIL-' in Ukrainian. In addition, the Russian and Slovak stems, '-ZAKONOMERNOST-' and '-ZAKONITOST-', respectively, were included, as semantic relatives of '-LAW-'. Word-forms from this Group clearly dominate the top ranks

by occurrence in the extracts of all 19 languages, and several stems individually show high retrieval of documents.

- Into *Group B* were placed '-ZIPF-' and '-LOTKA-', which occur (with one exception) in the higher ranks of the five major languages, and have intermediate individual retrieval levels. As remarked, they are notable for frequently forming compounds with the keystem, enhancing their retrieval value. They are, of course, more conspicuous as personal names, with completely unambiguous synonymies in this literature.

- Into *Group C* were placed the remaining stems: '-FORMULAT-', '-MULTIPLI-', '-ZON-', '-1934-', '-MODEL-', '-NUCLE-', '-CORE-', and '-RULE-'. Examples of words containing these stems are, respectively: 'FORMULATION' (English) and 'FORMULIERUNG' (German); 'MULTIPLIER' (English) and 'MULTIPLICADOR' (Portuguese); 'ZONES' (English) and 'ZONAS' (Spanish); 'MODEL' (English) and 'MODEL' (Russian); 'NUCLEUS' (English) and 'YADRO' (Russian); 'CORE' (English); and 'RULE' (English) and 'REGEL' (German). With the exception of the first, these stems occur less frequently in the upper ranks of all top five languages than do prior stems, and, in fact, each is more the contribution of one language only. Group C is intended as a sort of 'safety net' for peculiarities of individual languages.

*Evaluation of Stem Groups.* Profiles for documents retrieved for each language are then obtained in stem-group order, and in stem-group increments; that is for Group A with a total of 4 stems, then for (Group A or Group B) with a total of 6 stems, and finally for (Group A or Group B or Group C) with a total of 14 stems. This may not be the maximum retrieval profile over all 14 stems, but it is most likely to be so, or nearly so, for the first half dozen or more stems. The retrieval measure for each Group of stems is, of course, found first, and in the same manner, that is as the number of stems in the union of the sets of documents retrieved separately by each stem in the Group. The results, for each stem group, and for each language and all languages combined, are listed in Table 3-13.<sup>124</sup> The number and percentage of documents which are not retrieved by these word-stems are also listed. As well, the retrieval profiles are displayed graphically for four languages (English, Russian, Portuguese, and German) in Figure 3-5. The values

---

<sup>124</sup>Retrieval profiles for the number-of-occurrences-measure, displaying the total proportion of words occurring in the 3-extracts (e.g.) for a choice of stems, may be obtained by directly summing the percentage of occurrences for each stem listed in Tables 3-12(a-d). Summing in the group order used here, it can be shown that (e.g.) for English: (i) forms of Group A stems represent 10.8% (3377) of word occurrences in the English extracts, whereas forms of '-LAW-' alone represent 6.1% (1919) occurrences; (ii) with Group B stems added, 13.0% (4076) of word occurrences are represented; (iii) with Group C stems added, 14.8% (4640) of word occurrences are represented; thus (iv) 85.2% of all word occurrences in the English extracts do not contain one of the 14 selected stems. Given the high frequency of occurrence of many grammatical function words, it would perhaps be more meaningful to express these word occurrences in terms of semantic words, rather than in terms of all words.

for English are approximately those of the whole literature. The predominant stem '-LAW-' has also been granted separate entry in these profiles. The following results may be noted:

- (i) Word-forms from the four *Group A* stems occur at least once in the 3-extracts of 86.3% (898) of the 1040 analysed documents. Amongst the seven languages with 10 or more documents, this value ranges from 75.7% for German to 95.7% for Portuguese. For comparison, word-forms of the single stem '-LAW-' occurred in 71.1% (739) of the analysed documents, or, in the top seven languages, from 55.3% for Spanish to 89.1% for Portuguese.
- (ii) The addition of word-forms of the two stems of *Group B* (based on the two personal names) enhances retrieval to 89.5% (931) of the analysed documents. Amongst the seven languages with 10 or more documents, this value ranges from 78.4% for German to 100% for Czech.
- (iii) The addition of word-forms of as many as eight more stems from *Group C* enhances retrieval to 93.0% (967) of the analysed documents. Amongst the seven languages with 10 or more documents, this value ranges from 84.6% for Hungarian to 100% for Czech and Portuguese.
- (iv) 7.0% (73) of documents do *not* contain even one word-form of any of the 14 chosen word-stems in their 3-extracts. Amongst the seven top languages, five have unretrieved documents; for these languages, this proportion ranges from 5.3% for Spanish to 15.4% for Hungarian. The greatest numbers of unretrieved documents are in English and Russian (respectively 53 and 10), but the respective percentages (7.5% and 6.9%) are unexceptional.

*Unretrieved Documents.* An examination of the 3-extracts of the unretrieved English documents did not suggest any small number of word-forms which would lead to all 53 being retrieved. Rather, the pattern is that already observed: of diminishing gain with word-stem addition. Some word-forms in the English semantic word list of a slightly lower rank than were considered in the selection procedure, viz. 'CURVE', 'ANALYSIS', and 'SET', were each noticed in two or three documents, but most documents had 3-extracts with no common semantic surrounds. It should be noted that 62.3% (33) of these documents had only one extract, which is proportionally more than twice that for the 650 English documents retrieved (177 or 27.2%). It should also be remarked that an appreciable number of these documents *did* contain word-forms from Group A stems,

but at greater distances from a keyword -- from four to 20 words -- than were allowed in the present extracts; the use of larger extracts is suggested.

*Conclusion.* It appears that document characterisation or retrieval is a matter of diminishing returns, and that many stems of more limited individual occurrence may be necessary to capture all 1040 documents. Expressed in more topical terms, the distribution is Bradfordian. This bears out the initial qualitative observation; however, that so great a part of the collection can be characterised by a limited number of strings, suggests that the initial observation was more of 'the trees' than of 'the forest' -- by exceptions in occurrences rather than by commonality in documents. It may be that the procedures used can be improved to produce higher document retrieval, but at least they seem free of any language-specific defect. Exactly how many stems one chooses to characterise the literature depends on the acceptable level of retrieval and on the acceptable size of the selection criterion. A moderate choice could be to dispense with Group C stems and accept the a loss of 10.5% of this literature, that is of 109 documents. Some languages would lose disproportionately, for example about 20% of documents in Spanish, German and Slovak would not be recovered. But nearly 90% of all documents could be characterised with only the six stems of Groups A and B, suitably-sited. This matter will be returned to in Subsection D (p.213).

### **§3. EXTENSION OF ANALYSIS.**

The above analysis has not included a number of suitable documents, viz. 147 documents published in one language after 1986, and 29 documents published simultaneously in two or more languages. Using results from this analysis, it is comparatively easy to incorporate these documents in the manner now described. In each of the 147 single-language documents, every occurrence of the key stems '-BRADFORD-' or '-BREDFORD-' was noted. Then, *any* occurrence of a word-form based on Group A stems of the appropriate language was sought within 3 words, and within the same sentence, of the key stem. Where none was found, the procedure was repeated with Group B stems, and then subsequently with Group C stems. It was thus determined if each document could be retrieved by Group A stems alone, by (Group A or Group B) stems, by (Group A or Group B or Group C) stems, or not at all. A similar analysis was carried out on the 29 'bi-' or 'multi-lingual' documents, using -- as far as they were in hand -- both or all language versions with the appropriate stems. It was decided that satisfaction of the criteria in any one language was sufficient to qualify the document.

Results of this procedure are presented in Table 3-14 for the 147 recent single-language documents, and in Table 3-15 for the 29 multi-language documents. No document failed

to be retrieved from a critical defect in the key stems; the key stems '-BRADFORD-' or '-BREDFORD-' sufficed. The pattern of results for the recent single-language documents matches closely that of the earlier analysis (Table 3-13) in all important respects -- at least with respect to the major languages. Overall, for example, 91.8% of the 147 documents are retrieved by Groups (A+B) stems, while 5.4% of documents are not retrieved even with the inclusion of Group C stems; in comparison, 89.5% of the 1040 older single-language documents are retrieved by Groups (A+B) stems, while 7.0% of documents are not retrieved even with the inclusion of Group C stems. Likewise, the pattern of results for the few multi-language documents conforms sufficiently closely to that already established. Here 93.1% of the documents are retrieved by Groups (A+B) stems, with no improvement from the inclusion of Group C stems; that is 6.9% of documents remain unretrieved with all three stem Groups.

As these results are compatible with those of the previous analysis, they may be added to them, and a better description obtained of the primary collection. The combined results are presented in Tables 3-16 and 3-17. They pertain to 1216 documents of the 1310 'paper-like' documents meeting the primary selection criterion, the remainder of the documents being in Chinese or Japanese. There are only minute variations in percentages from Table 3-13 with respect to the major languages, so that prior conclusions may be extended to these 1216 documents. Thus, for these documents, the four stems of Group A retrieve 1053 (86.6%) documents, the six stems of Groups A and B retrieve 1093 (89.8%) documents, and the 14 stems of the three Groups retrieve 1133 (93.2%) documents; but some 83 (6.8%) documents still fail to be included. A benefit of extending the analysis, at least to the most recent documents, is to affirm that description obtained is quite stable, and most likely will carry forward to beyond the present.

#### **D. THE COMPLETED SELECTION CRITERION**

With completion of the formal analysis of the text around occurrences of the key stems '-BRADFORD-' or '-BREDFORD-', it should now be possible to give a precise definition to the collection of documents interpreted as on the topic of Bradford's Law of Scattering -- at least, for that part not written in Chinese or Japanese. Unfortunately, no small set of words could be found in the extracts which would completely characterise all these documents. Rather, the results present a diminishing-returns profile, with many documents retrieved with only a few word-stems, but with progressively fewer additional documents retrieved for each additional word stem used. Complete recall appears quite impracticable. So a problem remains of how to complete the required criterion. One possible solution is to cast the selection criterion in the form of a retrieval profile, either

in tabular form or as a function of the cumulative number of stems, thereby allowing any number of sub-collections to be extracted. This may be useful for some purposes, but for most bibliometric analyses of a subject literature which require only a single collection, it seems unduly complicated.

A better solution for present purposes would be to choose one selection criterion only, striking a balance between fewer word stems and higher retrieval. In terms of collection size, there is little to be gained in amplifying the selection criterion beyond the limited set of the most-common word-stems -- beyond this, it increases disproportionately slowly for the added specification. A moderate choice appears to be to use the six stems of Groups A and B -- suitably sited near the seventh stem '-BRADFORD-'. This choice retrieves 1093 (89.9%) of the semantically-defined documents analysed above, but would not admit 123 (10.1%) of them. Adding in the eight word stems of Group C would only admit 40 (3.3%) more documents and still reject 83 (6.8%). (It may be noted that the four word stems of Group A admit fully 1053 (86.6%) documents, and the two Group B word stems also only admit a further 40 documents, but whereas six word stems hardly strain the utility of the selection criterion, fourteen certainly do). Correctly, one should probably explore the relative stability of other properties of interest in making the choice. If there were no appreciable changes in, for example, proportions or distributions of these properties from that of the full collection by excluding documents not in Groups A and B, but there were when documents in Group B were also excluded, then the suggested choice of criterion would be better grounded. In this regard, it must be noted that some languages do lose disproportionately with this choice, for example as compared with 89.9% retrieval over all languages, only 84.6% of documents in Hungarian and Slovak would be retrieved; while others would gain, for example over 98% of documents in Portuguese and Czech would be retrieved.

*Analysis of Documents Excluded by Proposed Selection Criterion.* An extension of the notion of the stability of properties beyond the chosen boundary, with the widest import, is whether any documents of special 'significance' or 'worth' with regards to the topic are excluded by the choice of boundary made. Ideally, all such documents should lie well inside the chosen boundary. As raised in Chapter 2, and more fully discussed in Chapter 5, readily measurable properties could be used to capture some of this notion; these will be used here with little elaboration. We might consider documents as having special 'significance' with regards to the topic if: first, they were strongly 'on' the topic; and second, even if they were only moderately 'on' the topic, they were also quite 'influential' in the literature on the topic. An approximate measure of the strength of document 'aboutness' in the present case is the number of times the word stem '-BRADFORD-' appears

in the body of the text of documents. An approximate measure of the influence or relevance of a document on a topic is the number of times that document is cited in the literature on the topic, possibly adjusted for the effect of time. The 123 documents excluded by the proposed selection criterion were studied with respect to these properties.

First, we consider excluded documents which are strongly on the topic. The frequency distribution of our measure, the number of occurrences of '-BRADFORD-' per document, is strongly skewed: while the minimum and mode value of one occurrence per document is found in 70 (56.9%) documents, unfortunately 19 (15.4%) documents each have five or more occurrences. With respect to topic aboutness, the omission of these documents with five or more occurrences of '-BRADFORD-' is clearly unsatisfactory.<sup>125</sup> Of the 19 documents, 13 are in English, while two each are in Russian, Spanish, and German -- the problem is proportionally spread over the different languages. It should be noted that if Group C stems were to be added to the selection criterion, 10 of these documents would be included but nine would still defy inclusion. Further, it may be shown that if the extracts were widened some ten-fold, 15 of these documents would be included using only Group A word stems; but, even then, and with all three Groups of word stem used, three of the documents strongly on the topic still stubbornly refuse to be included.

Second, we consider excluded documents which are at least moderately on the topic and which have been considered relevant by many authors writing on the topic. As a measure of the former we will take documents with three or more occurrences of '-BRADFORD-', of which there are 36 (29.2%) in this group. As a measure of the latter we will take documents cited at least seven times by the literature in the semantically-defined collection of 1310 documents, but will not make any adjustment for time effects. Unfortunately, seven documents of the 123 meet both criteria; five are in English and one each in Russian and German. Two are most serious omissions: one document with four occurrences of '-BRADFORD-' has 68 citations, and one document with five occurrences of '-BRADFORD-' has 19 citations. If Group C stems were to be added to the selection criterion, three of these documents would be included but three would still defy inclusion. If the extracts were widened some ten-fold, four of these documents would be

---

<sup>125</sup>(1) For two of the five documents with the most occurrences of '-BRADFORD-', the number of these occurrences proves to be a poor measure of topic aboutness, since a preponderance of usages are unrelated to the Law of Scattering per se; however, for the remaining 17 documents, the converse is the case, and the problem remains. (2) In the full semantic collection (including Chinese and Japanese) this distribution is likewise skewed but more strongly. The minimum and mode value of 1 occurrence per document is found in 28.9% (vs. 56.9%) documents, and 39.5% (vs. 15.4%) documents each have five or more occurrences -- i.e. there are proportionally more 'strongly on' documents in the full collection than in those excluded here.

included using only Group A word stems; but, even then, and with all three stem Groups used, two of these documents would still not be included.

A third method of assessing the importance of the exclusion of 123 documents from the informally-defined collection of 1310 documents by the adoption of the completed selection criterion is in terms of the number of documents indexed by the chosen A&I Services which are also excluded. There are 153 such documents in the informally-defined collection. Five (3.3%) of these are excluded with the completed graphical selection criterion, which is a smaller percentage ( $123/1310 = 9.4\%$ ) than for all documents; that is the use of the completed criterion does not shift the focus of the collection away from the interpretation of the A&I Services. If Group C stems were allowed into the graphical selection criterion, then only three A&I indexed documents would be excluded; if extracts were widened ten-fold, one of these documents would still be excluded.

In summary, a not-inappreciable number of these 'topic-significant' documents are excluded by a selection criterion using only six word stems from Groups A and B. Relative to the size of the literature included, the omissions are not appreciable, but they are regrettable. However, even more generous selection criteria, with 14 word stems or with extracts of 60 words, still fail to include some 'topic-significant' documents, and one A&I Services~indexed document. It appears, then, that any such precise criterion that is practical to use will fail to include a portion of the semantically-defined collection, and more seriously, a small portion of the possibly 'topic-significant' documents in that collection. Unless one is prepared to abandon the idea of a precise collection definition - or at least a simple one based on word stems -- there is no option but to reclassify these documents as marginal documents. The policy adopted is to retain the selection criterion based on the six word stems of Groups A and B, and to create a new supplementary sub-collection for the 123 excluded documents of the informally-defined collection; this will have the status code (18) SEM (for 'semantic') in the BRAD File. Hereafter, this graphical selection criterion alone is considered to define the topic literature. However, for other purposes, it may be permissible to still append these 123 documents to the precisely defined collection, as the 'PRE', 'PAR', and 'DIF' sub-collections of marginal documents may also be appended. A major drawback with the 'SEM' sub-collection is that, unlike the other includable marginals, it does not have a precise definition, so its membership is open to interpretation.<sup>126</sup>

---

<sup>126</sup>Of course, this 'collection' could be precisely characterised by the remainder of the retrieval profile, i.e. 40 documents (33% of this collection) are recoverable with '-BRADFORD-' adjacent to the eight Group C word stems, and additional word stems such as '-RANK-', '-SET-' and '-CURVE-' could be brought in to improve recovery. But the same problem arises, and soon this collection will shed an imprecisely

*Additional Remarks.* Several problems with the proposed selection criterion remain to be addressed. First, the criterion has not yet been shown to diagnose the literature interpreted as on the topic, but only to characterise it, or a large part of it. How adequately the criterion diagnoses, that is distinguishes from the background, that which it characterises, remains to be tested in the next section. It may be that it will have to be rejected, and a new criterion constructed. Second, the connections of the proposed selection criterion to the earlier interpretation of the topic should be explicated, although such formalities seem disingenuous with this choice of stems in A and B. The obvious connections bode well for the criterion providing a clear diagnosis. Third, to have general applicability the criterion must be extended to the substantial literatures in Chinese and to Japanese; in other words, 'symbol patterns' semantically equivalent to the word stems used must be found for these languages. The present author's assumption is that these equivalents will likewise be found among the most frequent comparable units in comparable extracts -- the message remains the same, as it were. Hereafter, we proceed as if this assumption were true, and that the graphical criterion applies to all these Chinese and Japanese documents. Thus the final graphically defined collection on the topic of Bradford's Law of Scattering has 1187 analytical-level scholarly documents.<sup>127</sup>

*Conclusion.* The following precise graphical selection criterion has been proposed for the literature on the topic of Bradford's Law of Literature Scattering: *A document is selected as 'on the topic' if it contains in the body of its text at least one occurrence of a word-form that is derived from the word-stems '-LAW-', '-DISTRIBUT-', '-SCATTER-', '-DISPERS-', '-ZIPF-' or '-LOTKA-', placed within three words of, and in the same sentence as, the word-stem '-BRADFORD-'.* This criterion recalls some 90% of the literature in the semantic collection or 1187 documents; the 123 documents in this collection which are excluded by the proposed selection criterion are retained in a special 'SEM' sub-collection.

---

defined marginal collection of its own; and so on. There may be more promise in defining the smaller marginal collection of only the 'significant' omissions, but this will not be pursued at this time. More generally, this problem emphasises that an author who wishes his/her documents to be quickly associated with well-recognised research topics, should use the common (or standard) terminology associated with those topics somewhere in these documents.

<sup>127</sup>A slightly more accurate procedure would be to assume that 10.1% (123/1216) of the 94 Chinese and Japanese documents interpreted as on the topic -- i.e. with '-BRADFORD-' in the context of the particular regularity -- also fail to be retained by the graphical criterion. The final graphically defined collection would then have only 1178 documents, and there would be 123 + 9 = 132 excluded SEM documents. Of course, in the correct procedure, comparable extracts (of what width?) from Japanese and Chinese documents should have had their comparable components frequency-analysed prior to establishing semantic equivalency, and the selection of the 'word stems' made with these two languages also in mind.

The criterion recalls some 91% of documents indexed as on the topic by the three A&I Services or 148 documents.