

2. FREQUENCY DISTRIBUTION OF WORD OCCURRENCES

Introduction. To characterise the text extracts in Chapter 3, Section 5, the 3-surrounding graphic words of each language were sorted by frequency of occurrence; then a small number of word stems was selected from the most frequently occurring semantic words in each list. That the selection was made after establishing synonymies across the languages with a 'Chinese menu' approach has already been judged as unsatisfactory. Another aspect of the method which is unsatisfactory, and would be so even if selections were made separately within each language list, is that no general rule was given for the number of semantic words to select. Even if this decision was arbitrary, it is reasonable to require that the selection be consistent across the languages. Yet the languages differ appreciably in two respects which make this requirement more difficult to meet. First, they differ in the sizes of their lists; this is shown in Table 3-3, where it can be seen, for example, that English, Spanish and Polish have 26184, 1160, and 36 word occurrences, respectively. Second, as discussed in Section 1 of this Chapter (*Language Differences*, p.274), the languages differ in their ratio of analyticity/syntheticity; that is, the upper ranks of the language lists of some languages contain many more grammatical function words and many fewer inflectional variants of semantic words than those of other languages.

In this Section, the issue of how these differences between languages may be removed is addressed. In the first case, the problem is essentially one of re-scaling the frequency distributions of 3-surrounding words. The full type-token distributions, and not only tabulations of upper ranking words, must now be considered -- though our interest remains with the upper ranking words. This matter is considered in Subsection A. In the second case, as already discussed in Section 1, the problem is one of removing the effects of grammar which remain in the word lists after word isolation, viz. by grammatical function word removal and by inflectional stemming. If, after these operations, the frequency distributions of the 3-surrounding strings in the more-analytic languages and the more-synthetic languages become more alike, then a further, purely 'statistical' justification, is provided for them. This matter is considered in Subsection C.

Two other related issues are also considered. One concerns the adequacy of the 3-extracts as a text window for identifying the best diagnostic strings. This will be more fully discussed in later Sections, but some remarks seem most appropriate in the present context. In Subsection B, a comparison of the frequency distributions of words in the 3-

surrounds of two languages is made with those for larger extracts from the same languages, that is, with large language 'corpora'. This comparison may reveal faults in the use of such short extracts. For example, for present purposes more concentrated distributions would be more suitable; if the distributions from the 3-extracts prove to be less concentrated than those from larger extracts, then larger extracts would be recommended for the present study. The other matter follows from comments made in Section 1 about constructing lists of grammatical function words. In Subsection D, a less onerous and less 'grammatical' method for the identification of frequent grammatical function words is investigated. This uses the large language corpora already mentioned.

A. THE NUMBER OF WORDS TO BE SELECTED FROM LANGUAGES

Since the pooled 3-surrounds for each language differ greatly both in the number of graphic words (types) and in their number of occurrences (tokens), a problem exists for the method used in Chapter 3, Section 5 -- if language bias is to be avoided -- in the specification of the number of words that must be selected from each language. To use a fixed proportion of the total numbers of words (types) in each language, when they are ranked by decreasing numbers of occurrences, is to ignore the disproportionate increase in the total number of word occurrences with the total number of words (types). For example, as shown in Tables 3-7(a) and (b), English has 2680 words with 26184 occurrences, whereas Russian has 1128 words with only 2685 occurrences; that is, Russian has 42% of the number of distinct graphic words as English, but only 10% of the number of occurrences.⁸⁰ More suitable methods, which compensate for the total number of occurrences, might be the selection of all words with more than a certain proportion of occurrences each, for example 0.5%, or the selection of all words in rank order, till their cumulative number of occurrences exceeds a certain proportion of total occurrences, for example 50%. The former method guided the actual selection made, though rather loosely. For example, in Russian, selection was considered effectively to rank 16 ('MODEL'), with 0.86% of occurrences per word, but in Portuguese effectively to rank 24 ('1934'), with 0.5% occurrences per word. To clarify this matter, the exact form(s) of the frequency distributions for the languages must be studied.

⁸⁰If the top 1% of ranks was marked for consideration in each language, 27 ranks would be chosen for English -- which is approximately what was considered -- representing, cumulatively, about 50% of all word occurrences in English. But for Spanish and German, only 3 ranks would be considered -- that is, a three- and five-fold reduction in the number of words actually considered -- and representing only 32% and 12%, respectively, of all occurrences in those languages.

A Note on Data Presentation and Analysis. Frequency distribution data may be plotted in any of three ways. First, they may be presented in *Zipf form*, which, for the present case, places the number of occurrences per word (or other string) on the ordinate, and the rank number of words (ordered by decreasing ordinate values) on the abscissa. Unless qualified, the 'Zipf form' is also understood to mean that both axes have logarithmic scales, usually to base 10. In the main, such plots are only qualitatively compared.⁸¹ Second, where required, they may be presented in *Bradford form*; for the present case, the cumulative number of occurrences, by increasing rank number, is placed on the ordinate, and the rank numbers of words, logarithmically-scaled, are placed on the abscissa. To display data sets greatly different in size in Bradford form, the linearly-scaled ordinate must be either multiply-scaled, or be normalised by the total number of occurrences in each set (for example) and the values expressed as percentages. Third, there is usually no value in plotting frequency distributions for small data sets in Zipf form, for they are 'noisy' and invariably appear as a shallow and erratic stair. In Bradford form, the cumulation may mask some of this noise. However, something can be learnt of these distributions by plotting 'bounding' characteristics *allometrically*, that is for the present case, with the total number of unique words (ranks) on the ordinate, and the total number of occurrences on the abscissa, for each language and with both axes scaled logarithmically. Basic quantitative analyses are usually required in this case.⁸²

⁸¹(1) In the Zipf plots and Bradford plots presented here, every rank-frequency datum is plotted. The printing of points is suppressed and only the line directly connecting each point is shown. Since the axes have only discrete values, it is more common in Zipf plots for each successive data pair, where the ordinates have different values, to be connected by a step, that is, a vertical segment to the next value and a horizontal segment to the next rank. Steps are most evident in the uppermost ranks, where values change rapidly, and in the lowest ranks wherever there are long runs of tied ranks. It is difficult to compare a number of similar distributions when plotted in such a 'street-block' manner, leading to the present method of directly connecting data points. (2) It is felt that comparisons of shape without recourse to precisely-evaluated parameters and statistical tests are sufficient for most purposes here. Questions of the correct method of evaluating and comparing Zipfian parameters (see Chapter 3, Section 3C, *Consequences: Data Handling...*, p.123) are avoided, though -- as the data are not so portable without evaluations of parameters -- a disservice may be done thereby. In such comparisons, the *overall-shape* of each plot may be noted first: is it predominantly a straight-line, or predominantly a concave-down curve, or is it too disrupted by steps (the result of tied ranks) to interpret, etc? *Straight lines* can be fixed by any two of the following parameters: total number of occurrences (bounded area), total number of ranks (x-intercept), maximum number of occurrences/rank at rank 1 (y-intercept), and the slope (or exponent in the untransformed relation). Steeper lines (higher absolute values of exponents) represent more *concentrated* distributions; flatter lines *more-uniform* (or equitable) distributions. More parameters are needed to fix more complex curves, though the range of allowable curve types is limited by the rank ordering: the plotted curve must either remain level (ties) or fall with increasing rank. For example, the first three of the listed parameters should suffice for typical curves.

⁸²In allometric plots, some property of an entity is related to a more generic property -- for example, the size of a part in relation to the whole -- by an expression of the form stated below; see, e.g., Gould (1966), and innumerable more recent citing papers in many disciplines. The data may be synchronic, as in the number of different words used in corpora of different sizes, or diachronic, as in the number of words used in a particular corpora at stages in its growth. We may express the relationship between the total number of distinct words used (y) and the total number of word occurrences of a corpus (the size, x), noting that $y \leq x$, as $y = a.x^b$, where $b \leq 1$ is a constant. For $b=1$, growth is *proportional*, that is the

Results: Frequency Distributions for the 3-Surrounds by Language. In Figure 4-1, the frequency distribution of occurrences of the 3-surrounding word-forms, in each of the five principal languages, are presented in Zipf form. Individual words are not identified, but may be obtained for the upper ranks of English, Russian and Portuguese, from Tables 3-7(a-c), and for the upper ranks of German from Table 4-4. Two features of the data are notable. First, and ignoring the topmost ranks (especially in the cases of Russian and German), all plots are, broadly, parallel straight lines, with slopes of c. -1. Seemingly, the dominant process in the distribution of word occurrence over word-form is similar for each of the languages. As the collection increases in size, there is a diagonal displacement of the plots, with proportional increases both in the number of occurrences for each of the established word-forms, and in the number of new word-forms.⁸³ Second, the number of occurrences for the topmost ranking words in Russian and German fall well below what may be expected from this description, while those for

proportion of y to x , $(y/x)=a$, is constant. For $b < 1$, growth is *allometric*, that is $(y/x)=a \cdot x^{b-1}$, with the negative exponent meaning falling proportions as x increases. Three additional points of reference may be made. First, the allometric plot of word-form vs. word occurrence may be converted into a plot of the *log type/log token ratio* vs. number of tokens: $(\log y/\log x)=b+(\log a/\log x)$. Second, the curve through the terminating points of a coherent set of Bradford plots, for example that for a collection growing in time, and which may be called the 'bound' or 'envelope', may be related directly to an allometric plot. The *Bradford 'envelope'* is a plot of, for example, the total number of occurrences vs. log total number of ranks for a set of collections, that is x vs. $\log y$. For $y = a \cdot x^b$ to hold, it can be shown that the envelope must be an exponential curve, as it invariably appears to be in practice. Third, in recent times, allometric plots have been related to self-similarity, with the exponent interpreted in terms of a fractal dimension (see Egghe & Rousseau, 1990, pp.308-312).

⁸³It is useful to coin a term, *Zipfian-proportional growth (ZP growth ?)*, for this type of growth, and to briefly discuss its features, for it is a reference point for all such comparisons of distributions which can be construed in dynamic terms. In the following discussion, we consider the frequency distribution in Zipf form. Two components are necessary for this growth: occurrences must add to established ranks and create new ranks. (i) Established ranks must increment multiplicatively for their portion of the line to increment linearly, that is they must display *proportional growth*, or constant relative growth. (For example, if rank 10 has 5 occurrences when rank 1 has 50 occurrences, then it must have five times as many or 25 occurrences when rank 1 has five times as many or 250 occurrences. The top 10 ranks would then gain 600 occurrences, from 150 to 750, approximately). (ii) But not all occurrences are available for concentration in the old ranks, otherwise the plot will truncate at the last old rank. New members must be *recruited* into the system, and at a specified rate (more accurately, at a specified declining rate) if the distribution shape is to be preserved. (In the example used, 40 new ranks must be created and stacked with occurrences to preserve the distribution shape; this might require 400 occurrences in addition to the 600 of proportional growth). (iii). Zipfian-proportional growth is growth where the appropriate recruitment rate is maintained (*proportional recruitment*), and old rank growth is proportional growth (*proportional concentration*). In the example used, the occupant of rank 1 has 50/150 or 33% of occurrences; after Zipfian-proportional growth, it does not have 250/(150+600) or 33% of occurrences, as it would with proportional growth, but 250/(150+600+400) or 22% of occurrences. For larger collections these differences are less dramatic. (iv). It should be noted that a distinction between the recruitment of sources and additions to (established) sources might be eliminated by viewing recruitment as addition to what has so-far only been a 'potential source'. (v). Two common trends in growth which modify the basic Zipfian-proportional pattern are decline in 'concentrating ability' in the upper ranks, and loss of recruitment of new members ranks. Either or both will lead to the development of a convex-upwards curve in middle ranks in the Zipf plot, a common distribution shape.

Spanish and Portuguese do not, and English is intermediate but closer to Spanish and Portuguese in form. In Russian and German, the top three or four words share occurrences more equitably than is the case in other languages, and in their own lower ranks. On closer scrutiny, it may be seen that the complete distributions for these languages are, in fact, slightly more equitable, that is, have slightly lower slopes.⁸⁴

The importance of the differences between languages in the uppermost ranks of the distributions, for the goal of selecting a consistent number of words in each language, is most apparent when they are transformed to Bradford plots. This has been done in Figure 4-2, where the ordinate has also been normalised by the total number of occurrences in each language. It may be seen, for example, that 50% of all word occurrences can be met, for both Portuguese and Spanish, by only the top 10 words (approximately); for English, the top 25 words are needed; for German, the top 46 words are needed; and for Russian as many as 107 of the top words are needed. The number of ranks considered for selection in each language departs considerably from these values. While the top 25 words were considered for selection for English. For Portuguese and Spanish more words, to ranks 24 and 17, respectively, were considered, and these cover more than 50% of all word occurrences in those languages. For German and Russian, many fewer words, only to ranks 15 and 16, respectively, were considered, and these cover much under 50% of all word occurrences in those languages. With this scaling of all 3-surrounding words, the former two languages are over-represented with respect to English, and the latter two are under-represented.

We may digress briefly to consider the smaller languages. In Figure 4-3, allometric plots are employed for all 19 languages. In Figure 4-3(b), the number of unique word-forms used in the language 3-surrounds is plotted against the total number of word-form occurrences; that is, the number of types is plotted against the number of tokens. In Figure 4-3(a), the number of equal-occurrence or equal-productivity classes in the separate distributions is plotted against the same variable. In either figure, the data for both the smaller languages and the five larger languages lie approximately on single straight lines, though deviations for several of the languages with fewer than 50 word-

⁸⁴A quantitative analysis is clearly needed at this level of detail; it may then be seen, for example, that Portuguese has an overall steeper slope than Spanish, and so on. The overall lower slopes for Russian and German, lead to them having longer tails in the lower ranks, that is a larger membership in the lower productivity classes. Thus, Russian has 1.59 times the number of word forms which occur only once, and 1.30 times the number of word-forms which occur twice, as has Portuguese (777 vs. 490, and 348 vs. 267, word-forms respectively). To study this region of the distributions, where the interest is in the size of equal-productivity classes rather than in the identity of its members (see Chapter 3, Section 3C, *Exact Form of ...*, p.118, and footnote 67), data are better treated in *Lotka form*.

form occurrences are marked in the plot of equal-occurrence classes.⁸⁵ Substantial deviations between distributions are revealed even in these bounding characteristics, but they lie within the range of variation of the larger languages, remarked on earlier. These plots *suggest* that the same general distribution processes of word occurrence over word-form apply in each language, except possibly where the samples are very small indeed. There is nothing to suggest that the smaller languages bring additional problems to the present analysis.

Further Analysis of Scaling Frequency Distributions. The problem of scaling the frequency distributions of the five major languages may better illustrated by considering the transformation of the simplest ideal case in the Zipf form into the Bradford form. We start with a language whose frequency distribution in Zipf form can be closely approximated by a straight line with a slope of -1, and which has its the first ranking word with p_1 occurrences -- that is, not dissimilar to English, if p_1 is assigned 2474. If we let $p(r)$ stand for the number of occurrences of a word of rank r , then the distribution, over r_{\max} ranks, can be described by the rectangular hyperbola:

$$p(r) = p_1 r^{-1} \quad \text{for } r = 1, 2, \dots, r_{\max} \text{ (and so } r_{\max} = p_1).$$

The Bradford form of this distribution is the number of occurrences summed from rank 1 to rank r -- here written as $P(r)$ -- and plotted against the logarithm of rank r , over the range of the ranks, that is:

$$\{ P(r) = \sum_1^r p(r) \} \quad \text{vs.} \quad \log(r) \quad \text{for } r = 1, 2, \dots, r_{\max}$$

There are difficulties in approximating this summation of discrete values with an integration over continuous values, particularly in the tails of the distribution, where critical parameters p_1 and r_{\max} are estimated.⁸⁶ One approximation, which is not

⁸⁵(1) The lines plotted were fitted by least-squares regression. In Figure 4-3(a), the line fitted to all but the five lowest points is: $\log y = -0.259 + 0.513 \log x$ ($n=14, r=0.996$, 95% confidence limits for slope are 0.484 and 0.542). In Figure 4-3(b), the line fitted through all data is: $\log y = 0.412 + 0.711 \log x$ ($n=19, r=0.988$, 95% confidence limits for slope are 0.654 and 0.769). It should be noted that the uppermost point (for English) is far-removed from the other points and is strongly influential on the regression. (2) The scales are logarithmic, which tends to 'suppress variation' in the data for the eye. For example, in Figure 4-3(a), as remarked, the difference between the number of word-forms for Russian and for Portuguese, viz. $(1128-631) = 497$, is of the order of the total number of word forms for Portuguese (631). This feature, and a consequence, the effects of included outliers (such as the data for English here), underlies Fairthorne's remark about such plots, viz. "... a straight line law connecting any empirical data always can be achieved with the aid of suitably scaled logarithmic paper and a robust conscience ..." (Fairthorne, 1969, p.331)!

⁸⁶Starting with the Zipf form, we have a sequence of vertical line segments or 'spikes', ideally $\{(1, p_m), (2, p_m/2), \dots, (r, p_m/r), \dots\}$, which has been overlain by the continuous curve, $p(r) = p_m r^{-1}$, to be read only at $r = 1, r = 2, \dots$ etc. The Bradford form is the successive summation of the length of these discrete spikes, plotted against discrete r , and the problem is to find its counterpart in the integration of this approximating curve. (Integration measures the area bounded by the curve and the abscissa, so to represent a finite sum for each r , the area must be closed). The Zipf form needs to be recast. A solution is to associate with each spike a finite area under the curve with certain properties -- effectively to replace each spike by a vertical rectangle, or 'tile', with the same height and with unit width. The tiles are laid down to best cover the area under the curve. A good method is to place them with their *vertical*

particularly good, but is sufficient for present purposes, is, for the same range of values of r :

$$P(r) = \sum_1^r p(r) \approx \int_1^r p(r) dr + p_1 = p_1 + p_1 \ln r$$

This plots as a straight line in Bradford form with a slope of $0.434p_1$ and an intercept of p_1 . The total number of occurrences in the collection, which may be written P_{tot} , is thus:

$$P_{\text{tot}} = P(r_{\text{max}}) = p_1 + p_1 \ln r_{\text{max}}$$

And so, by substitution and a rearrangement of terms:

$$[P(r) - p_1] / [P_{\text{tot}} - p_1] = p_1 \ln r / p_1 \ln r_{\text{max}} = \log r / \log r_{\text{max}}$$

Therefore, instead of using a Bradford plot with a normalised ordinate, as we have in Figure 4-2, which plots:

$$\{ 100 P(r) / P_{\text{tot}} \} \text{ vs. } \log (r)$$

so we may, following Mitsevich (1975b), plot instead:

$$\{ 100 [P(r) - p_1] / [P_{\text{tot}} - p_1] \} \text{ vs. } \{ 100 \log r / \log r_{\text{max}} \}$$

In this plot, *any* 'ideal' distribution in Zipf form, with a slope of -1, will thus be transformed into *one* straight line with slope +1, running from the (0,0) to (100,100) -- both axes using percentage scales in this case. Distributions with forms different from this will not be so transformed. The data in Figure 4-2 has been replotted in this manner in Figure 4-4. The distributions most like the simple ideal distribution, those for English, Portuguese and Spanish, are transformed more closely into the diagonal line, while the distributions for Russian and German transform into a separate concave-upward curves.

This Figure demonstrates graphically one way of scaling out of size differences between collections, and exposing fundamental differences in the form of the distributions. It also demonstrates that for distributions of the same form, we select ranks more equitably when the logarithms of chosen ranks are in proportion to the logarithms of the total number of ranks.⁸⁷ Of course, for reasonably regular curves such as these, further transformations, involving slopes other than -1, and some additional parameter to describe the 'flattening' in the initial ranks, could be introduced to align all curves; for

midline on the spike at r , and their width spanning from $(r-1/2)$ to $(r+1/2)$; the summation from $r = 1$ to $r = r^*$ is then replaced by the integration from $(1/2)$ to $(r^*+1/2)$, so that $P(r) = p_1 \ln(r^*+1/2) - p_1 \ln(1/2)$, instead of the equation used. This method was used by Leimkuhler (1967) -- and, in fact, the scaling required here could be accomplished directly by fitting Leimkuhler's equation to the data. The poor method used here places the *right vertical edge on the spike*, the tiling consistently falling below the curve, rather than overlapping it in the better method. Accordingly, in the top ranks, the relative cumulative error can be shown to be about three times higher than with the better method; but as the rank increases both errors fall away, until the long tail of tied lower ranks is approached, which introduces more complications. In practice, the scatter in the actual data -- real data don't necessarily lie on a regular hyperbola -- will probably be of more consequence in determining the error.

⁸⁷Rather loosely, $P(r)/P_{\text{tot}} \approx \ln r / \ln r_{\text{max}}$, so for two distributions, $\ln r_1 / \ln r_2 \approx \ln r_{\text{max}1} / \ln r_{\text{max}2}$.

which see the General Zipf-Mandelbrot Function, Chapter 3, Section 3C, p.121). But this avenue has little merit here, as will be discussed below.

Conclusion and Comment. The manner used in Chapter 3, Section 5, for selecting as candidates for diagnostics different numbers of words from the different languages, bears no systematic relation to the sizes of the different language collections, and must be considered deficient as a method. Some suggestions for improving the method, by scaling out gross differences of size, have been considered, but the problem of distributions with different forms remains. Further study may now be directed to other channels. We might more profitably inquire as to what factors account for the differences in the form of the distributions between languages, most notably between Russian and German on the one hand, and Portuguese and Spanish, and perhaps English on the other? If these are identified, can they be removed by other than mathematical manipulation? This is considered next.

B. COMPARISONS OF 3-SURROUNDS AND LANGUAGE CORPORA

Perhaps a first question should be: Are the differences found between languages in the frequency distributions presented a product of the method of word extraction, or a characteristic of the particular literature, or do they reflect more fundamental differences between the languages themselves? In the first case, it may be possible to alter the method of taking text extracts and so eliminate the differences, while the second case, the least likely explanation, would be particularly interesting to study. We might answer this question by inspecting the frequency distributions of graphic word-forms for extensive and general corpora of different languages, and comparing them with those obtained here for the 3-surrounds. This will be done for English and for German, languages which show somewhat different distributions in the upper ranks of the 3-surrounds. For English, the general corpus used was the Brown University Corpus of over one million word-occurrences in a variety of text, for example press reportage, fiction, and scholarly writing, and written mainly by American authors (Kucera & Francis, 1967); the *Brown Corpus* has already been introduced in Section 1E of this Chapter, on p. 282. For German, the general corpus used was Meier's 1967 revision of the corpus of Kaeding, based on some ten million word-occurrences in mainly literary text by German authors in the 19th and early 20th centuries; hereafter, this is termed the *K-M Corpus* (Meier, 1967).⁸⁸

⁸⁸The Brown Corpus and the K-M Corpus are comparable to the 3-surrounds in that: (i) their data are simple tabulations of *exact graphic word-forms* rather than of 'dictionary words', lexical words, or 'lemmas', as occurs with (for example) with Juilland & Chang-Rodriguez (1964) for Spanish; Steinfeldt

Results. Kucera & Francis (1967, Graph B1, p.358) plot the frequency-rank data of the full Brown Corpus in Zipf form, and obtain a straight line with a slope of c. -1 to a good order of approximation; the ten upper ranks and the lowermost ranks show some deviation. The upper 10,000 ranks of this distribution are copied into Figure 4-5, where the full data for the English 3-surrounding words are repeated.⁸⁹ The Brown Corpus plot lies parallel to that of the English 3-surrounds, though diagonally displaced. As discussed earlier, the difference between the two distributions is primarily one of size. There is a greater 'flattening' in the first three ranks of the 3-extracts plot than in the Brown Corpus plot, though both plots show similar deviations in their initial ranks.

A similar plot of frequency-rank data for the full K-M Corpus was not available, but approximately the upper 8,000 ranks of this distribution are also plotted in Figure 4-5, where the full data for the German 3-surrounding words are also repeated.⁹⁰ Over the greater part of its range, the distribution of the K-M Corpus data is broadly parallel to the other distributions presented, though further diagonally displaced by its greater size; more accurately, the slope of this plot is somewhat steeper than those of the English language plots. Of particular interest is the 'flattening' in the upper 20 ranks, and the extreme 'flattening' in the upper three ranks, of the K-M Corpus plot. The latter resembles the extreme 'flattening' in the upper four ranks of the plot of the German 3-surrounds.

It is apparent that there is greater similarity between the two plots of one language (German: K-M Corpus and 3-surrounds; English: Brown Corpus and 3-surrounds), than there is between the plots for different languages, despite size differences of orders of magnitude.⁹¹

(1973) for Russian; or Francis & Kucera (1982); and (ii) they sample text written for adults, rather than for children (with pedagogical intent) as (for example) with Carroll et al. (1971) for English, and to a lesser degree, Steinfeldt (1973). Suitable corpora of graphical words for Russian and Czech have apparently been prepared at Brown University but these were not available for comparison here; see Kucera (1968).

⁸⁹All data to rank 100 are plotted, and thereafter points are selected at (approximately) exponentially-increasing intervals of rank. Thus, after rank 100, the line loses detail: the beginnings of the steps from larger sets of tied ranks are suppressed. The data are taken from Table B1 of Kucera & Francis (1967, pp.300-307); their full plot is Graph B1, p.358.

⁹⁰All data to rank 100 are plotted, and thereafter points are selected at (approximately) exponentially-increasing intervals of rank. Thus, after rank 100, the line loses detail: the beginnings of the steps from larger sets of tied ranks are suppressed. Also, detailed data after rank 7994 (frequency = 101) are not provided. The data are taken from Part B (Rank Lists) of Meier (1967, Vol.2, pp.111-137).

⁹¹The identity of the top words in the Brown and K-M corpora may be read indirectly from Tables 4-3 and from 4-4. The large corpora are, of course, based on selected 'extracts', though these are many orders of magnitude wider than the extracts used here. For example, the Brown Corpus uses 500

The appropriate total values for the Brown and K-M Corpora were also plotted allometrically in Figure 4-6, which repeats Figure 4-3(b) for the 3-surrounds of the 19 languages.⁹² Both data lie close to the straight line fitted to the data for the 3-surrounds, though they are respectively nearly two and three orders of magnitude larger. Again, it appears that the same very broad distribution process of word occurrence over word-form applies in each language, to 3-surrounds and larger corpora alike.⁹³

Conclusion and Comment. It is concluded that the differences between languages in the 3-surrounds, most apparent in the upper ranks of the distributions, do reflect differences between languages and do not result from a language-specific bias in the method of extraction, or in some feature of this particular literature. A different procedure for taking extracts would not remove these differences. In fact, the principal differences between the distributions of the 3-surrounds and the larger Corpora are due to extract size alone. So we might next inquire as to what could account for the observed differences between the form of the frequency distributions for languages, more particularly the differences between Russian and German, on the one hand, and Portuguese, Spanish, and (to a lesser degree) English, on the other? Zipf (1935, Plate IV, Chapters II and V) reported a very marked flattening in the upper ranks of the frequency distribution of Latin vis-à-vis English. He suggested that this may be due to Latin being a more inflectional (that is, synthetic) language, and English being a more isolating (that is, analytic) language, though no detailed supporting arguments were provided.⁹⁴ If this is the basis for the observed language differences, that is they arise from grammatical differences, then the correct route to adjusting for these differences would lie in the removal of grammatical function words and grammatical inflections.

C. REMOVAL OF FUNCTION WORDS AND STEMMING OF SEMANTIC WORDS

extracts, each of c. 2000 words of (largely) uninterrupted text; the English 3-surrounds consist of 5196 extracts, each of c. 6 words, separated into two parts by the word 'BRADFORD' or a variant of it

⁹²The Brown Corpus has 1,014,232 graphic word-types and 50,406 word-tokens (Kucera & Francis 1967); the K-M Corpus has 10,910,777 graphic word-types and 258,173 word-tokens.

⁹³But mindful of Fairthorne's remark in footnote 85, p.309, perhaps a 'robust conscience' is required to draw this conclusion; for example, more data might reveal that a family of convex-upward curves best describes the situation.

⁹⁴To quote Zipf (1935, p.47) " ... Thus the deviation of the most frequent words of Latin below the standard curve of English may well be connected with the greater degree of inflection of Latin (see Chapter V), or with the facts that Eldridge's count [i.e. English] was based on written prose while the Plautine count [i.e. Latin] was based on verse to be declaimed. These suggested problems merit future investigation ..." I have added the language names in square brackets.

The effects of removing grammatical function words, and of semantic word stemming, on the frequency distributions of the 3-surrounding words will next be investigated. In the first case, the investigation will be confined to the English, Russian and Portuguese word lists, and in the second case, to English and Russian only. If the frequency distributions for the different languages are more alike in form after these operations, then the prospect of a language-invariant selection of candidate strings is improved, and the use of these operations in the present analysis will receive additional justification. This result would also implicate differences in 'grammatical strategy' in establishing the exact form of the frequency distributions of languages, as suggested by Zipf (1935).

§1. THE REMOVAL OF GRAMMATICAL FUNCTION WORDS

In the analysis below, grammatical function words were identified by direct recognition, as used in Chapter 3, Section 5. While it is most unlikely that the results obtained here would be altered by greater exactitude, it must be noted that this is an unsatisfactory methodology. A better technique would be that described in Section 1E of this Chapter, p.281. In addition to grammatical function words, a few mathematical terms have also been removed, as discussed earlier; however, the removed words may still be justifiably referred to in what follows by the shorter term *function words*.

A Note on Rank Plots and Decomposition of Distributions. In the analysis below, the frequency distribution for all 3-surrounding words in each language will be presented with those of its components, the function words and the semantic words; the distributions will be plotted principally in Zipf form. In comparing these rank-frequency plots, it must be remembered that the two component distributions are plotted against their own rankings, so that the abscissa effectively has three rank scales. The decomposition is not like that in a size-frequency plot, where the ordinate value of the combined distribution, at any abscissa value, is the sum of its two component ordinate values. To clarify how various decompositions of a simple 'rectangular hyperbolic' distribution might appear in Zipf plots, several theoretical curves were generated. In Figures 4-7(a,b), the distribution has been decomposed with a *constant proportion* between the ranks of the two components, in two ways. First, in (a), alternate members are placed into separate distributions, that is words of total rank 1,3,5, ... are assigned to ranks 1,2,3, ... for one component, and words of total rank 2,4,6, ... are assigned to ranks 1,2,3, ... for the other component; this is a 50% decomposition. Second, in (b), every fifth member is assigned to one component, and the remainder to the other; this may be termed a 20% decomposition. It may be seen that the component plots quickly come to parallel the total distribution as rank increases. In contrast -- and of more

interest here -- in Figure 4-7(c), the distribution has been decomposed with a *decreasing proportion* for one component, from 70-80% to 10-20% of ranks over the first 200 ranks. That is, in the first 10 ranks of the total distribution, only two go to component one, and eight to component two, whereas towards rank 200 the situation is reversed. It may be seen in the Figure that the plot for the initially-favoured component one follows the whole distribution closely initially, but then drops rapidly away, while that for later-favoured component two begins at low values but converges to the total plot. Figure 4-7(c) serves as a reference in the following discussion.

Results. In Figure 4-8, the frequency distributions of occurrences for all words, for function words, and for semantic words, are plotted in Zipf form for the English 3-surrounds. The component distributions have only been plotted as far as an ordinate value of five occurrences per word, that is, to rank 570 overall, to rank 148 for function words, and to rank 422 for semantic words. Thus, 74% (422/570) of all words above rank 570 are semantic words. It is apparent from the Figure that the function words are concentrated in the upper ranks of the full distribution; after rank 20 in function words, the plot falls away with a slope of c.-1.5, and after rank 100, with a slope of c.-2. From rank 4 in semantic words, the semantic words show a more equitable distribution, the plot falling with a slope of c. -0.5 initially but converging to the overall slope of c.-1; the three top semantic words, viz. 'LAW', 'DISTRIBUTION', and (to a lesser extent) 'SCATTERING', follow a separate steep plot with a slope of c.-2. With reference to the theoretical analysis above, the function words occur in decreasing proportion of all words, when ranked by occurrence.

This pattern for the English 3-surrounds may be presented more directly as the percentage of semantic words, in successive groups of all ranked words, plotted against overall word rank. In Figure 4-9, successive groups of 20 and 60 words have been used, the latter smoothing the variation of the former. The initially rapid, and thereafter gradual, rise in the proportion of semantic words, to c. 85% of adjacent ranks at rank 500, is apparent. It is matched by a 'hyperbolic' decline in the proportion of function words. It might be noted that the pattern to rank 100 is apparent in data in an earlier table, Table 4-1.

In Figure 4-10(a,b), the frequency distributions of occurrences for all words, for function words, and for semantic words, are plotted in Zipf form for the 3-surrounds of Portuguese, and of Russian, respectively. In both cases, the component distributions have only been plotted as far as overall rank 150; for Portuguese, this is to rank 55 for function words, and to rank 95 for semantic words; for Russian, this is to rank 40 for

function words, and to rank 110 for semantic words. (Thus, above this rank, 63% of words are semantic for Portuguese and 73% are semantic for Russian). If a comparison is made with English to, say, overall rank 100, we find that c. 55% of 3-surrounding words are semantic words in Portuguese and English, but c. 70% are semantic words in Russian; alternatively stated, more of the 100 most frequently-occurring words are function words in Portuguese and in English, than they are in Russian -- at least in this literature and in these extracts. For Russian, the function words fall out from rank 2, whereas for Portuguese they fall out from about rank 20, as was the case with English; the resulting distribution of semantic words more closely follows that for all words with Russian, than with Portuguese or with English.

In Figure 4-11, the distributions of the upper ranks of semantic words in the 3-surrounds are directly compared for the three languages in Zipf form.⁹⁵ Once again, the distribution for Portuguese is most dissimilar in form to that for Russian, with the distribution for English intermediate but more like Portuguese. The removal of function words has elevated the very top ranks of the plots, and somewhat depleted the following ranks, for both Portuguese and English. In the former, the elevation is confined to only one rank, to the word-form 'LEI', whereas in the latter, it extends to three ranks, to the word-forms, 'LAW', 'DISTRIBUTION', and 'SCATTERING'. The plot for Russian retains the 'flattening' seen in the upper ranks for all Russian words, but this is reduced from five ranks to three ranks, that is -- with the removal of the function words 'v' and 'l' -- to the word-forms, 'ZAKON', 'ZAKONA', and 'RASSEYANIYA'. It is clear that the removal of grammatical function words has not aligned the frequency distribution of Russian with that of Portuguese and of English.

Conclusion and Comment. The contrast already noted between Russian, on the one hand, and Portuguese and English on the other, for all 3-surrounding words, has not been reduced by the removal of grammatical function words. However, their removal has 'unmasked' disproportionately high occurrences of the top semantic words in the latter two languages. As grammatical function words are most certainly poor diagnostic words, their higher proportion in the upper ranks of words in Portuguese and English than in Russian cautions against the blind usage of threshold ranks for all words, as

⁹⁵It is not possible to present these three distributions in modified Bradford form, for comparison with the distributions for all words in Figure 4-2, since semantic words have not been determined beyond an all-word-rank of 150 for Russian and Portuguese and 570 for English. If, however, percentages on the ordinate are calculated with respect to the cumulative number of semantic word occurrences to rank 150, as an imperfect substitute for the total number of semantic word occurrences, it can be shown that: 50% of such occurrences are provided by the three top semantic word-forms in Portuguese, by the five top semantic word-forms in English, and by the 15 top semantic word-forms in Russian. The extent of the differences in the upper ranks of the semantic words, between English and Portuguese on the one hand, and Russian on the other hand, is clear.

developed in Subsection A above. If the scaling discussed was implemented, then the relatively few upper ranks recommended for Portuguese and English, vis-à-vis Russian, could be further proportionally depleted by the removal of function words, biasing in favour of Russian. The ranks actually chosen now seem a better choice.⁹⁶

The differences observed between the distributions of function words and semantic words in each language assuredly results from the fact that, in the composition of text, function words are selected from a *small closed class* of words, while semantic words are selected from a *vast open class*. This difference is further magnified as graphic 'semantic words' are much more likely to be inflected than graphic 'function words'. Thus, as more and more passages of text, in a more analytical language, are brought into the analysis, the possibility of recruitment of new function words rapidly falls, and word occurrences concentrate in established members. In contrast, the possibility of recruitment of semantic words is high and does not diminish greatly.⁹⁷ High recruitment and low concentration produce more equitable distributions, but the pattern for semantic words would be complicated by semantic matters, viz. the subject nature of the text. Passages from a specialist discourse could well have semantic words selected disproportionately from a *small restricted subclass* of specialist semantic words. As more and more passages of such text enter the analysis, these words would behave like the more common, but not the most common, function words. Of course, widening the subject scope of the text analysed would dilute any such effect, allowing the function words to dominate more and more of the upper ranks. (This is put to good use in Subsection D below). It may not even be possible to expand a corpus that is so resolutely 'subject-narrow' for this effect to persist. The very high concentration of some semantic words in the above analyses ('LAW', 'DISTRIBUTION', and 'SCATTERING' in English; 'LEI' in Portuguese) is less likely to arise in this way. It is more likely to arise from the manner in which the extracts were taken. These are words which associate or collocate very strongly with the keystone of the extract. Widening the extracts within the topic literature should see the slow dilution of these words, and a 'straightening' of the plot for the upper ranks of semantic words, as it moves diagonally with increasing corpus size.

⁹⁶It may even be more appropriate to extract only semantic words in a string from the text, and to overlook function words, from the very outset; extract size would then be measured directly in semantic words. This policy was not followed, in part to allow for *exact* phrase analysis in Section 5. An intermediate policy, would be to measure extract size in semantic words, though (if necessary) extracting all words to the chosen semantic length. It must be admitted that, in the preparation of the 3-extracts for Portuguese (especially), it was occasionally regretted that one more word could not be included in an extract populated with short prepositions and articles.

⁹⁷In the terminology of footnote 83, p.310, we could say that function word growth rapidly shifts from Zipfian-proportional to proportional since recruitment virtually ceases, while semantic word growth remains Zipfian-proportional.

§2. SEMANTIC WORD STEMMING

For a pure analytic language, the grammatical function words play the same role, of composing semantic words into a sentence, as the grammatical inflections do in a pure synthetic language. In the simplest case, a number of function words and one semantic word-form, in some fixed order, in an analytic language, have as their exact semantic counterpart in a synthetic language, a single inflected semantic word. The removal of grammatical function words from analytic languages must be complemented by the removal of grammatical inflections from synthetic languages, if any balanced comparison of their text-semantic link is to be made. Since actual languages are a mix of both pure types, both procedures must be carried out on them, producing a comparable set of semantic stems. Unless, as is clearly not the case, Russian were to have the same analytic/synthetic ratio as English (or Portuguese), the previous comparison of semantic words would still be 'grammar-biased'. It may be that the continuing differences observed in the form of their frequency distributions reflect this residual grammar-bias, and that it will be removed if semantic stems are compared.

Procedure and Results. Sufficient numbers of semantic words in English and Russian were stemmed manually to produce only the top 40-50 most frequent stems in the appropriate 3-surrounds. A word-frequency table was made for each language, and words were sorted alphabetically. Words differing only in inflectional suffixes, as determined from personal knowledge (English), or from dictionaries (Russian), were conflated and their frequencies added, when it appeared that the combined frequencies would exceed c.50 occurrences in English, or c. 10 occurrences in Russian.⁹⁸

The frequency distributions of the top 40-50 semantic stems, for both the English and the Russian 3-surrounds, are plotted in Figure 4-12, along with the distributions of the semantic words for each language repeated from Figure 4-11. For English, the distribution of the stems lies above that for word-forms. From ranks 2 to 8, the stems distribution is straighter (or less concave-upwards) than the words distribution, and from rank 10, slightly steeper and convergent towards it. Nevertheless, the stems distribution matches the general form of the semantic word distribution. The top two, or possibly three, stems, viz. '-LAW-', '-DISTRIBUT-', and possibly '-SCATTER-', have values above what

⁹⁸A most useful Russian dictionary was Katzner (1984). In scanning these lists, it was also necessary to remain alert to critical prefixes, for example in Russian, the inflectional prefix 's' indicating the perfective aspect of certain verbs, as shown in 'FORMULIROVAT' imperfective, 'SFORMULIROVAT' perfective, [formulate].

may be expected from lower ranks. These stems are those of the top three word-forms, 'LAW', 'DISTRIBUTION', and 'SCATTERING', which also show proportionally elevated values over other semantic words.⁹⁹ For Russian, the distribution of stems generally lies above that of word-forms, and does not appear to converge in the range shown. The top five stems, viz. '-ZAKON-', '-RASPREDELEN-', '-RASSEYAN-', '-MODEL-' and 'INFORMATS-', are those of the top ten word-forms.¹⁰⁰ Most importantly, the stems distribution has 'straightened out' somewhat in the upper ranks, and so has a form much closer to that of English than that of the semantic words distribution. The top one, and possibly four, stems even have values above what may be expected from lower ranks, in the manner of English. It must also be noted that the top three stems for Russian are synonyms of the top three stems for English, that is, respectively, '-LAW-' and '-ZAKON-', '-DISTRIBUT-' and '-RASPREDELEN-', and '-SCATTER-' and '-RASSEYAN-').¹⁰¹

Conclusion and Comment. The limited data *support* the suggestion that where the frequency distributions for all words, and semantic words, are of different form, the frequency distributions of semantic stems will be much more alike in form. In other words, the differences of form reflect differences in the grammatical strategies of languages, which interpose, as it were, between the semantic similarity of the discourse and the graphical appearance of the text. The removal of grammatical functional words *and* grammatical inflections from languages must be recommended as a prerequisite to scaling out gross size differences between language collections on the same subject. However, it must be stated that the data presented here are slight; also, the methodologies for function word and stem identification are hardly rigorous.

It should be reiterated that the 3-surrounding words are not from randomly selected text extracts, but are words chosen by close proximity to one specific word-form, or variants of it. The frequency distributions of stems for English and for Russian appear to be themselves influenced by two factors, a general distribution of semantic stems associated

⁹⁹The top four word-forms are 'LAW', 'DISTRIBUTION', 'SCATTERING' and 'DATA', and the top six stems are '-LAW-', '-DISTRIBUT-', '-SCATTER-', '-APPL(Y)-', '-FORMULAT-' and '-DATA-'. The hollower word-form curve results in part from a strong dispersion of the stems '-APPL(Y)-' and '-FORMULAT-' into word-forms.

¹⁰⁰The ordinate value of stems for a rank may lie below that of word-forms for the rank since the ranking scales for stems and word-forms are different. This is obvious at ranks 10 and 11 in Russian, where it seems to result from a high number of word-forms in the upper word-form-ranks mapping into a small number of upper stem-ranks; e.g. '-ZAKON-' at stem-rank 1 is derived from words at word-form-ranks 1,2,4, and 6. The stems plot must eventually cross the word-forms plot, since there are word-forms with one occurrence only which map into stems with one occurrence only.

¹⁰¹Unfortunately, it has not been possible to form a modified Bradford plot, for stemming has not been extended sufficiently far to obtain a sensible denominator. This plot would determine the number of semantic ranks needed in each language to achieve a certain proportion of all semantic stem occurrences. Since the distributions now have a more similar form, these rank values should be similar, once size differences are scaled out.

with the topic to various degrees, for example, '-APPL(Y/I)-', '-FORMULAT-', '-ZIPF-', '-DATA-', and '-LITERATUR-' in English; and '-LOTK-', '-INFORMATS-', and '-OSNOV-' in Russian, and the collocational 'magnetism' of the specific keystem, which promotes the occurrence of some stems, for example, '-LAW-' and '-DISTRIBUT-' in English, and '-ZAKON-' and '-RASPREDELEN-' in Russian.

D. THE IDENTIFICATION OF GRAMMATICAL FUNCTION WORDS

The process of specifically removing grammatical function words would not have been necessary here if sufficient numbers of extracts, containing the word-stem '-BRADFORD-' (or '-BREDFORD-') in a context judged *not* to be on the topic, had been prepared for each language. Since many grammatical function words occur very frequently in all text, they would appear with approximately-equal high frequency both in the topic extracts and in the background extracts, and so be identified and removed, along with common and frequent semantic words, by comparison. With the present procedure, where the background is not considered until late in the analysis, it is important that these obviously bad diagnostics be removed as early as possible in the analysis to minimize labour. Other good reasons for their early removal have been suggested above. As discussed in Chapter 3, Section 5A, p.195, and in Section 1E of this Chapter, p.276, the problem is to find some simple technique to quickly identify at least the most frequent grammatical function words in different languages.

A suitable technique is suggested from observations made in §1: The Removal of Grammatical Function Words (p.317) of Subsection 2C of this Chapter. There it was surmised that function words would strongly dominate the upper ranks of large and general word occurrence lists. That this is the case for English, at least, can be seen in the Brown Corpus of written English (Kucera & Francis 1967). As a preliminary exercise, the upper portion of the rank listing of this Corpus was scanned, and graphic words were 'individually recognised' as belonging to either functional or semantic classes. The most likely highest ranking semantic word ('SAID'), occurs at rank 53; in the top 100 ranks there are only seven likely semantic words, and in the top 200 ranks there are only 29. Thus, by simply deleting from the English 3-surrounds an appropriate number of graphic words listed in the top ranks of the Brown Corpus, the most frequent and general function words of written English could be removed, albeit (perhaps) with a small number of very common semantic words. This procedure will now be considered more carefully for English, German and Russian.

Procedure and Results. In Table 4-3, the 100 most-frequently occurring words in the English surrounds are tabulated inter alia with both their rank by frequency of occurrence in the Brown Corpus, and a measure (percentage of samples) of how generally they occur throughout that corpus. As examples: the article 'THE' has rank 1 both in the 3-surrounds and in the Brown Corpus, and occurs in 100% of samples of the latter; the noun 'LAW' has rank 2 in the 3-surrounds, but in the Brown Corpus it only has rank 309, and occurs in only 19.8% of samples. Some words occur rarely in the Corpus, for example 'SCATTERING', or not at all, for example 'ZIPF'. More germane, all but three of the 45 likely function words previously identified in this list occur in the upper 100 ranks of the Brown Corpus, while no semantic word in the list does.¹⁰² Alternatively stated, if the top 100 words in the Brown Corpus were deleted from the 3-surrounds, leaving the italicised words in the Table, the resulting list would be nearly identical to the list of English semantic words presented in Table 3-8(a).¹⁰³

Similarly, in Table 4-4, the 54 most-frequently occurring words in the German 3-surrounds are tabulated with their rank by frequency of occurrence in the K-M Corpus of written German (Meier, 1967). Likewise, the articles 'DER' and 'DIE' take the first two places both in the 3-surrounds and in the K-M Corpus, while the noun 'GESETZ' [law] has rank 4 in the 3-surrounds but only rank 333 in the Corpus. And likewise, all but two of the 32 function words already identified in this list, occur in the upper 100 ranks of the K-M Corpus, but no semantic word in the list does so. The two defaulting words are actually function words in English passages and identifiable by the procedure in the previous paragraph. Alternatively stated, if the top 100 words of the K-M Corpus were deleted from the surrounds, and English passages appropriately treated, the resulting list, the italicised words, would be identical to the list of German semantic words presented in Table 3-9(a).

The same protocol has been used for the 50 most-frequently occurring words in the Russian surrounds, using the smaller Steinfeldt Corpus of written Russian (Steinfeldt, 1973). This corpus uses 'dictionary or lexical words' rather than 'graphic words or word-forms'. For comparison, the word-forms of the 3-surrounds have had to be inflectionally

¹⁰²The three missed function words are: 'BETWEEN', ranking 124 in the Brown Corpus; 'BOTH', ranking 125; and 'N' which has a very low rank. The last 'word' reflects differences between this study and the Brown University study in the interpretation of graphic words: here, 'N' is invariably a free term in an equation, and so would be considered as part of a mathematical formula, all cases of which are treated as one word, with rank 92, in the Brown Corpus.

¹⁰³The 42 removed stopwords occur in 93.2% of samples whereas the 54 retained semantic words occur in only 12.3% of samples; these are mean values for strongly skewed distributions so the median values of 98.2% and 8%, respectively, are more representative. In essence, the stopwords occur widely but the semantic words are quite restricted in use. In these calculations, the three missed stopwords and the 'word' '=' are excluded (and see previous footnote).

stemmed, producing 39 lexical words, and re-ranked. The data are presented in Table 4-5. The lexical word ZAKON [law] is ranked first in the 3-surrounds, followed by the function words *v* and *ı*; in contrast, in the Steinfeldt Corpus, these function words occupy the first two ranks and ZAKON takes a rank of only 843. In a similar manner to English and German, all but two of the 19 function words already identified in this list occur in the upper 100 ranks of the Steinfeldt Corpus, and no semantic word in the list does so.¹⁰⁴ Alternatively stated, if the top 100 'dictionary words' of the Steinfeldt Corpus were deleted in all their various inflectional forms from the 3-surrounds, the resulting list, the italicised words, would be nearly identical to the list of Russian semantic word-forms presented in Table 3-8(a).

Conclusion and Comment. The procedure of deleting from the 3-surrounds of a language, any graphic word listed in the top 100 ranks of a large and general word count for that language, appears to be a safe and speedy procedure for the elimination of the most frequent function words. It may be performed simply and automatically to produce results similar to that obtained in an earlier section by a poorly described process. Obviously, a suitable corpus must be available, and the language must be similar in type to English, German and Russian. More generally, this is simply the first step in the process of removing bad diagnostic words from the 3-surrounds by comparison with a specific background; in this case, the background is very general, and common to both the topic and any chosen specific background.

Two parallel matters need to be addressed. The first concerns the principal language of the topic, English. The Brown Corpus is of American English, while a large proportion of the English documents studied are written by British authors. To check if there are differences between American and British written English of relevance to the present procedure, a corpus comparable to that of the Brown Corpus, viz. the Lancaster-Oslo/Bergen or LOB Corpus (Hofland & Johansson, 1982), was consulted. In Table 4-3, words occurring in the LOB Corpus with a frequency that is significantly different (in a statistical sense) from that in the Brown Corpus are indicated (Hofland & Johansson, 1982, pp.471-544). The differences are slight, and of no concern to the present study.

Second, the selection of precisely the top 100 words has no special status. It is simply a natural unit in our counting system at which all, or nearly all, of the function words, as identified by other means, and none of the semantic words, which occur in the upper (50

¹⁰⁴The two missed function words are: *PRİ* at rank 179 in the Steinfeldt Corpus, and *BYLO* at a very low rank. This may reflect stylistic differences between the 3-surrounds and the Corpus. The Corpus is formed primarily from works of fiction and magazine and newspaper articles, and is slanted towards children's material.

or 100) ranks of the 3-surrounds, were thereby named and excluded. Were the selection widened, more function words would be excluded but so would more semantic words. These semantic words are words with general or multiple meanings, for example, in Table 4-3, they include 'WORK' (rank 120), 'FOUND' (rank 165), and 'SET' (rank 208); but with narrower interpretations, such words could be significant for some topics.¹⁰⁵ A guide to choosing the number of ranks selected should be the size of the corpus. Earlier it was suggested that function words occupy more and more of the upper ranks as the corpus size grows.¹⁰⁶ If this is correct, then more and more ranks could be selected with safety with larger corpora.

E. CONCLUDING REMARKS

This Section began with the problem of how to select from the ranked lists of 3-surrounding words in each language so as to prevent any bias resulting from differences between languages -- specifically, with respect to list size and in the analytic/synthetic ratio. The solution appears to be the removal of grammatical function words and inflectional affixes, followed by a suitable re-scaling of the rank-frequency distribution of semantic stems. This also allows for the identification of semantic stems with disproportionately high occurrences, most likely due to their collocational 'binding' to the keystem '-BRADFORD-'. Regarding these conclusions, several points must now be raised. Since we are ultimately interested in characterising documents, not text extracts per se, a more suitable measure to use appears to be document frequency rather than occurrence frequency (effectively extract frequency). This requires that a threshold be set for the number of occurrences per document needed to characterise a document, a matter to be considered in the next section, Section 3. However, it might be restated from Chapter 3, Section 5, that there are only minor ranking differences in the upper ranking 3-surrounding words between these distributions. The important point is that it is not the sum of the number of occurrences, as used here in Bradford plots and in the Mitsevic correction, but the union of the number of documents that is of interest. This complicates the problem of string selection, as will be discussed in Section 4. In fact, it is recommended there that no preselection of words or stems be made from the 3-

¹⁰⁵Even fixing the cut-off point at 100 words, the words 'TIME' and 'ZEIT' -- plausibly of significance in some cases -- would be deleted from the English and German using the Brown and K-M Corpora, respectively.

¹⁰⁶It is probably dangerous to compare corpora from different language types for this characteristic, but it is observed that: the first semantic word (SKAZAT') has rank 29 in the Steinfeldt Corpus of 387,211 'dictionary word' occurrences; the first semantic word ('SAID') has rank 56 in the Brown Corpus of c. one million 'graphic word' occurrences, and the first semantic word has rank 90 ('ZEIT') for the K-M Corpus of 10 million 'graphic word' occurrences.

surrounds at all, and that as far as possible all strings be tested for their contribution to document retrieval.