Anthony A. Lyne

ON ROOTING OUT WORDS WITH INFLATED FREQUENCIES

IN WORD-COUNTS OF SPECIALIZED REGISTERS

This paper is concerned with the problem of ensuring that a corpus is properly representative of a target language variety for lexicometric purposes. More particularly it presents a device for remedying any residual bias which may reveal itself only when a word-frequency count has been completed.

When a word-count of a text or corpus of texts is carried out, the results can be displayed in the form of a list of all the word-types in the corpus, arranged in order of decreasing frequency, with the actual frequencies of the items placed alongside. Such a list constitutes a description of the vocabulary of the corpus. As a description it is a very partial one, since it gives only one kind of information, viz. frequency. Nevertheless it does have two considerable virtues: it is precise and it is exhaustive. We know that the first item in the list, for example, really is the most frequent item in the corpus and we know that the list contains every item in the corpus without exception (unless, of course, the list is deliberately abridged for publication purposes).

Often, however, the focus of the investigator's attention is not on the corpus for its own sake. Rather has he assembled the corpus with the intention of its being a representative sample of a whole language or, more reasonably, of a particular variety of that language, for example, a 'register' (a variety according to use). When an item's corpus frequency is divided by the number of running words in the corpus, we obtain an estimate of that item's probability of occurrence in the target population, i.e. in the register under investigation: $F/N \rightarrow p$. What it is vital to recognize is that with this shift of focus from the corpus to the underlying register the list ceases to be either a precise or an exhaustive description: what were precise frequencies are now only approximate probabilities (ever more approximate as we move down the list); what was an exhaustive list is no longer so, because we know that an increase in sample size will inevitably result in an influx of items not previously instantiated in the corpus.

I said a moment ago that a corpus is often assembled with a view to its being a representative sample of a target register. Achieving a representative sample is always fraught with difficulties, and registers are certainly no exception. In the first place, if a register is thought of as a language variety arising from a particular, recurring combination of circumstances, then the limits of the register will be no less hazy than the limits of this combination of circumstances. In the second place, even within a relatively narrow register it will perhaps always be possible to distinguish subregisters and sub-subregisters, which cut across each other in complex ways. From this it follows that

there is a danger of the corpus being biased in favour of one
or more of these subregisters at the expense of others. To take
the case of my own count of French business correspondence (FBC),
I found that there was no clear boundary between commercial letters
(my target register) and letters of other types, e.g. ones of
a personal or legal nature. Secondly, within the target register
itself, some letters originated in France, others in Belgium;
some were inquiries, others complaints, others apologies and so
on. Here we have two independent dimensions of variation already
within FBC and obviously we could go on (Lyne 1975).

The larger the sample the smaller the risk posed by accidental
bias. But practical considerations of time and money invariably
leave the investigator regretting his corpus was not larger than
it is. It took me a full year to persuade a large number of com-
panies to part with their jealously guarded, authentic letters.
Then, of the 1000 or so letters so laboriously collected, it
proved necessary to jettison some 300 in order to avoid having
the corpus swamped with letters emanating from just two sources
and dealing with a severely limited range of topics. Despite
these precautions it was clear, on completion of the count, that
the 80,000 word corpus was by no means free of accidental bias.
For example, common sense tells us that the word crème (cream)
does not have as high a probability of occurrence in FBC as does
remerciement(s) (thanks), yet both these lemmata occurred 26
times and were ranked approximately 350 out of 3500. The reason
for the inflated frequency of crème was that a batch of letters,
amounting to 4% of the total corpus, originated from a firm dealing
in liquid soap, referred to as crème, and all the occurrences
of this lemma were concentrated there. In contrast, the occurrences
of remerciement(s) were evenly spread throughout the corpus.
Nor was crème an isolated case. We know intuitively that certain
other items, like brique, tube and calendrier, which were ranked
even higher than crème, must be overrepresented for similar
reasons.

A corpus, then, even when assembled carefully, rarely turns
out to be a representative sample, certainly not a perfectly
representative sample, of the target register. What can be done
about this?

It is now generally agreed that it is very important to take
into account not just the frequency of each item but also its
dispersion, that is the degree to which its occurrences are dis-
tributed evenly throughout the corpus. In order to do this, we
must divide the corpus into a number of sections, preferably
of equal size, and for each item note the number of sections
in which it occurs. In my own count, for practical reasons, I
had to content myself for the most part with a rather rudimentary
dispersion measure known as 'Range'. This merely takes account
of the presence or absence of an item in each section. There
are however a number of alternative, more discriminating dis-
persion measures, which take account of each item's precise sub-
frequency in each section. It was while doing some test soundings
with one such measure, viz. Juilland's 'D', that I noticed the
phenomenon which is the main object of this paper.

The dispersion measure known as 'D' was devised by Alphonse

Juilland and his associates and used in the well-known series of
frequency dictionaries of the Romance languages. The details
of how D is calculated need not detain us here.[1] Suffice to say
that D takes a value ranging between one, for an item whose sub-
frequencies in all the sections are identical, and zero, for
an item whose occurrences are all in a single section.

The authors of the Juilland dictionaries operated with five
equal sections, which were deliberately heterogeneous, that is
one consisted of fragments of novels, another of plays, another
of technical prose and so on. For the present purpose I too used
five equal sections but an important point to bear in mind is
that my corpus was aimed at a single register - not just French
business correspondence but FBC of a routine nature and slanted
towards import-export transactions. In contradistinction to
Juilland's corpus, then, mine was intended to be homogeneous.
The 670 letters were in fact allocated to the five sections simply
on the basis of the order in which I had collected them. Batches
of letters from a single source will normally all be in one section
or occasionally spill over into a second, so the five sections
are in principle undifferentiated; they are not, a priori, composed
of letters of five distinct types.

Let us look now at my test soundings using Juilland's D. I
took a high-frequency item ranked 150/3500 in my list by decreasing
frequency, together with the items on either side of it - 30
items in all - thus forming a block sharing approximately the
same frequency, viz. between 72 and 57. I then established the
D values for all 30 items and listed them in order of decreasing
D value (see Table 1, first list). It will be seen that the
first item, cas, very nearly achieves the optimum D value of
one (actually $\overline{0.932}$), since its five subfrequencies all lie within
the narrow limits 14 $\pm$ 3. In contrast, the last item, dévouer, has
subfrequencies ranging from 41 to 3 and this relatively uneven
distribution is reflected in the lower D value of 0.442. I per-
formed the same operation on three further 30-item samples from
the frequency bands 41 - 38 (around rank 250), 18 - 17 (around rank
500) and 10 (around rank 750). Details of these are also shown
in Table 1.

My next step was to display the four sets of 30 D values
in the form of a graph (Figure 1), the items being arranged left
to right by decreasing D value. We may note in passing that the
location of the four traces in the vertical dimension relative
to one another confirms Charles Muller's observation that D is
positively correlated with frequency (Muller 1977:72). Even more
interesting for our present purpose is the extent to which the
D values in each frequency band are bunched in the upper half
of the scale. There seemed no reason a priori why all the D values
for a particular frequency band should not form an unbroken con-
tinuum from the most evenly distributed item at the top left
hand of the figure to the most unevenly at the bottom right.
However this is clearly not so here. Each trace descends in small
steps as far as D $\pm$ 0.5, but then the D values for the few remain-
ing items drop much more abruptly. Taking the four traces together,
the lowest item on the 'hill' (F 72 - 57, D 0.442) is separated
from the first item over the 'cliff' (F 18 - 17, D 0.262) by a gap
of 0.180, i.e. almost 20% of the full scale. The evidence of these

<u>Table 1</u>   Four 30-item samples from frequency count of French business
correspondence comprising approximately 80,000 running words:
frequency (F), subfrequencies (x_1 to x_5) and Juilland's index
of dispersion (D). The complete frequency list comprises 3497
items.

	F	x_1	x_2	x_3	x_4	x_5	D		F	x_1	x_2	x_3	x_4	x_5	D
	Frequency band 72 ≥ F ≥ 57 (around rank 150)								Frequency band 41 ≥ F ≥ 38 (around rank 250)						
cas	71	15	17	12	15	12	.932	désirer	38	7	8	7	8	8	.968
nécessaire	63	11	12	12	16	12	.931	commercial	39	8	10	8	6	7	.915
afin	59	12	13	8	12	14	.914	note	41	10	9	8	9	5	.895
effet	66	14	10	11	18	13	.895	août	39	7	10	8	9	5	.890
après	59	16	10	10	14	9	.885	mètre	39	7	9	5	10	8	.890
pièce	68	8	13	17	17	13	.878	avant	40	9	7	9	5	10	.888
entendre	60	6	12	13	14	15	.868	occasion	39	6	8	11	8	6	.883
mettre	72	12	12	12	22	14	.866	rappeler	39	7	8	8	5	11	.876
marchandise	66	15	5	15	18	13	.834	ils	40	11	5	8	6	10	.858
connaître	62	13	17	17	8	7	.828	noter	39	8	4	7	9	11	.852
sans	64	12	11	16	6	19	.827	vivement	41	5	10	12	6	8	.844
rester	65	10	8	13	22	12	.815	rapidement	39	11	9	4	6	9	.841
fournir	59	15	3	15	13	13	.810	où	41	6	12	11	7	5	.830
kilogramme	68	19	11	21	9	8	.803	mai	40	5	4	11	10	10	.819
vente	69	13	16	21	15	4	.799	aimer	38	6	10	12	4	6	.807
fin	62	7	14	12	8	21	.798	souhaiter	41	4	12	9	12	4	.781
renseignement	60	6	15	17	16	6	.794	mars	40	7	12	13	3	5	.757
prochain	67	7	23	9	14	14	.794	entretien	40	2	13	7	12	6	.747
fabrication	65	6	17	21	14	7	.779	convenir	39	4	15	10	5	5	.733
sujet	69	12	7	18	8	24	.768	lors	41	14	5	8	12	2	.732
octobre	64	18	12	5	21	8	.767	effectuer	40	6	2	7	8	17	.692
novembre	64	6	12	6	22	18	.750	général	38	13	4	9	12	0	.676
avril	57	13	22	10	6	6	.741	appareil	38	11	15	9	3	0	.643
juin	64	10	8	11	9	26	.739	règlement	39	8	0	3	10	18	.602
accuser	61	4	7	12	16	22	.738	annexe	39	8	0	1	14	16	.582
intéresser	60	8	15	23	11	3	.719	reconnaissant	41	20	11	6	4	0	.581
tonne	71	26	19	14	12	0	.698	rubrique	39	3	3	3	10	20	.572
usine	68	36	8	6	7	11	.584	proposer	40	22	10	4	2	2	.526
sincère	62	35	5	14	7	1	.514	fusil	39	0	0	3	0	36	.093
dévouer	64	41	3	10	4	6	.442	carabine	38	0	0	0	0	38	0

Table 1 (ctd.)

	F	x_1	x_2	x_3	x_4	x_5	D		x_1	x_2	x_3	x_4	x_5	D
			Frequency band $18 \geqslant F \geqslant 17$ (around rank 500)							Frequency 10 ($\leqslant$ample) (around rank 750)				
commander	18	4	3	4	4	3	.932	absolument	2	2	2	1	3	.842
réaliser	18	3	5	4	4	2	.859	contenir	2	2	3	1	2	.842
côté	17	4	3	5	2	3	.850	faible	2	2	1	3	2	.842
droit	17	5	2	4	2	4	.824	simplement	2	2	1	2	3	.842
mieux	18	3	5	5	4	1	.792	position	3	3	1	2	1	.777
poids	17	2	3	6	4	2	.780	allemand	2	2	3	3	0	.726
stock	18	4	4	6	1	3	.775	naturellement	2	0	4	2	2	.684
peut-être	18	3	6	5	3	1	.758	placer	3	3	0	3	1	.684
actuel	18	4	3	2	7	2	.743	supérieur	3	1	3	3	0	.684
ci-après	18	2	3	4	7	2	.743	licence	2	3	1	0	4	.647
port	18	3	1	3	7	4	.728	journée	1	5	1	2	1	.613
réclamation	18	2	2	2	6	6	.728	supplément	1	1	1	2	5	.613
direction	17	7	2	3	2	3	.727	administration	4	3	0	3	0	.582
elles	18	7	1	4	2	4	.714	américain	1	4	4	0	1	.582
caractéristique	17	6	1	4	5	1	.697	bâtiment	0	2	2	5	1	.582
figurer	17	2	1	4	3	7	.697	cadre	0	1	1	4	4	.582
intention	17	2	4	1	3	7	.697	pair	0	0	4	3	3	.582
poser	17	2	6	5	4	0	.683	programme	1	4	4	1	0	.582
avenir	18	7	2	1	6	2	.665	communication	1	5	1	3	0	.553
ceci	18	7	4	0	2	5	.665	effort	5	1	3	1	0	.553
début	18	1	2	7	2	6	.665	exécuter	4	0	0	4	2	.553
considération	18	0	6	3	7	2	.642	perdre	4	0	0	4	2	.553
différence	17	8	1	5	1	2	.599	profiter	4	2	0	0	4	.553
malheureusement	18	3	9	3	3	0	.592	arrivée	0	1	2	6	1	.476
an	17	1	4	9	2	1	.558	concurrence	5	0	4	1	0	.476
modification	17	4	2	0	9	2	.548	distribuer	1	6	0	1	2	.476
cliché	17	1	1	0	7	8	.503	haut	6	1	0	2	1	.476
chambre	18	2	11	0	4	1	.455	rouleau	2	6	0	2	0	.453
litre	18	0	14	1	3	0	.262	écrou	0	8	1	1	0	.242
four	18	17	1	0	0	0	.068	arlequin	0	0	10	0	0	0

four samples suggests that the high and medium frequency items in the corpus (say F $\geqslant$ 10) are not simply more or less even in their distribution, but rather that they fall into two discrete categories, evenly and unevenly distributed items, with the vast majority belonging to the former category.

The next question is whether this quantitative dichotomy corresponds to any qualitative difference between the evenly and unevenly distributed items. In Figure 2 all those items having D < 0.6 are displayed in such a way that the vertical spacing in the four columns corresponds to the distances between their D values. It therefore enables us to see at a glance the identities of the items whose D values constitute the lower part of the 'hills' and the whole of the 'cliffs' in Figure 1. (The items with D > 0.6 are of course omitted simply to save space.)

There does in fact seem to be a qualitative difference. The six 'cliff' items may be said to be ones whose relatively high frequency in the corpus, indeed whose very presence in the corpus, may be unequivocally attributed to the specialized technical activity of a few particular branches of industry. In short, they are items like the troublesome crème, brique, tube, calendrier, which we met earlier. Fusil and carabine betray the fact that one of my sources was a firearms manufacturer. Four and arlequin relate to trade in refractory bricks and foodstuffs respectively. Litre is only found in letters concerned with trade on liquid products and écrou 'nut' with trade in hardware. The less specialized nature of the latter two 'branches of industry' explains the slightly higher D values of the two items concerned. In contrast to these six items, which I will refer to as technical for short, the other items displayed in the upper part of Figure 2 (as well as all the items with D >.6 listed in Table 1) seem to be genuinely commercial, i.e. not tied to any particular branch of industry but equally at home in letters from firms engaged in all types of commercial activities – in other words, my target register.

As mentioned earlier, D values have been calculated for only a proportion of the 3500 items in my frequency list. This is because the items are lemmata and their total frequencies have to be obtained by summing the occurrences of their inflectional variants 'manually'. The labour involved in finding the five subfrequencies for each lemma manually from a concordance is very considerable. The indications are (Lyne 1981:225-8) that of the 134 highest frequency items (F $\geqslant$ 72), i.e. those above my first 30-item sample, only one item has a D value below 0.442, viz. brique with D = 0. However, were we to process more 30-item sets in the same way as the four already described, it would not be surprising if some items fell in the present tidy gap between 0.442 and 0.262. It would be too much to hope that evenly and unevenly distributed items were in two completely watertight compartments, particularly where an item is polysemous (e.g. chambre as in ___ de commerce, ___ d'hôtel, ___ à air etc.). Nevertheless, provided only a very few items turned out to have intermediate values (compared with the total number of items examined), this would suffice to make the 'hill and cliff' a legitimate model for the distribution of D values in this corpus.

Fig.1 Graph showing dispersion coefficients (Juilland's 'D') for 4 x 30-item samples drawn from word-count of French business correspondence.
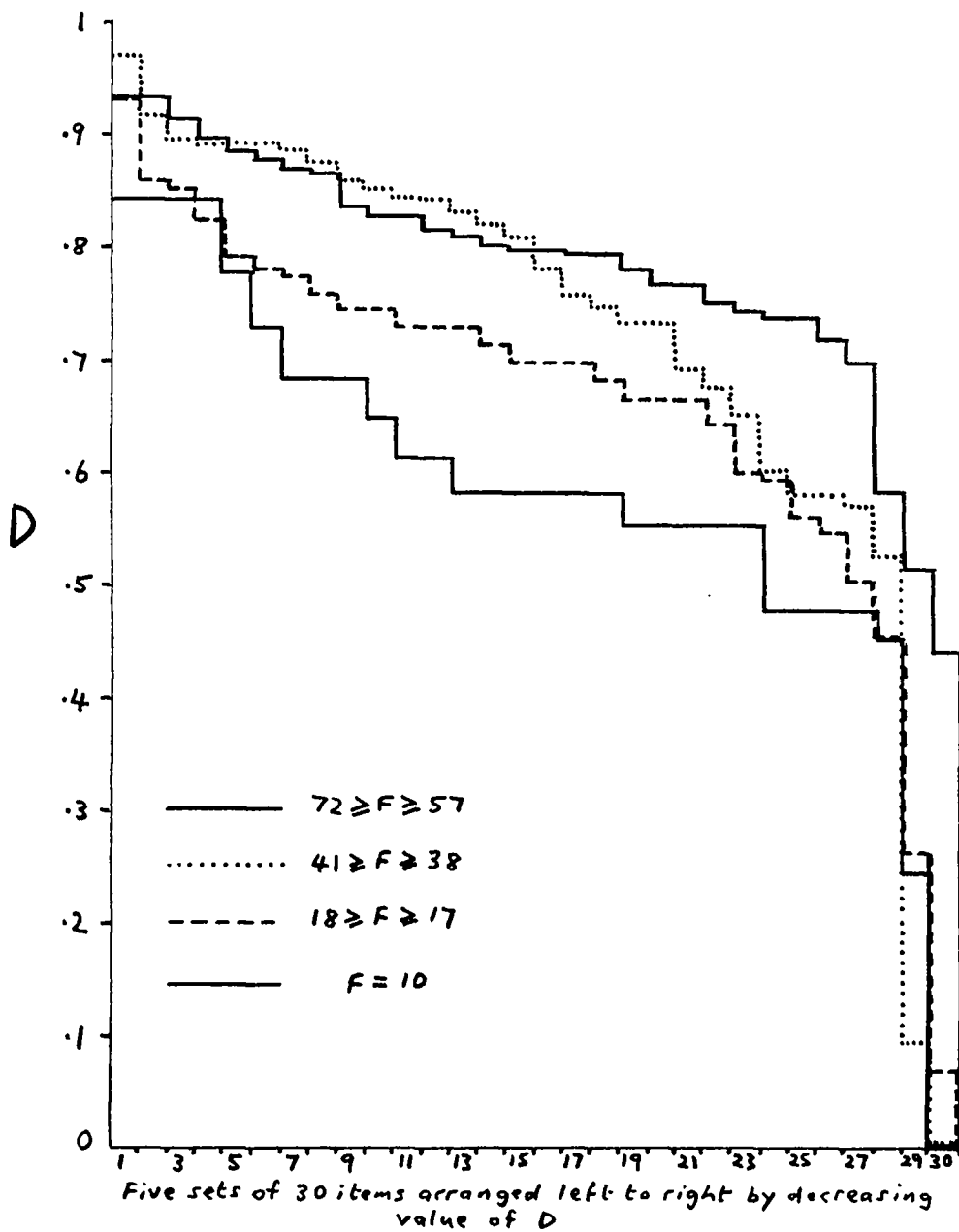
Fig.2   Lower 'hill' and all 'cliff' items displayed.

D	Frequency Bands			
	72 – 57	41 – 38	18 – 17	10
0.600	usine	annexe reconnaissant rubrique	différence malheureusement	administration, pair, américain, bâtiment, cadre, programme
			an modification	communication, effort exécuter, profiter, perdre
	sincère	proposer		
0.500			cliché	
				arrivée, concurrence, haut, distribuer
	dévouer		chambre	rouleau
0.400				
0.300				
			litre	écrou
0.200				
0.100		fusil		
			four	
0		carabine		arlequin

It would be premature to claim that the 'hill and cliff' model has been proved correct conclusively for the FBC corpus, still less for other corpora. If it is applicable to other counts, it seems likely that they will have to be ones based, like mine, on a homogeneous corpus with undifferentiated sections (as explained earlier).

The potential practical benefits are not inconsiderable. The word-counter already routinely tempers raw frequencies by combining them with dispersion coefficients to yield a list by decreasing 'usage' (Juilland 1964) or the like. But even so, some rogue items like my fusil, four, etc., provided their D is greater than zero, may still achieve ranks which common sense tells us are too high.[2] What the 'hill and cliff' model promises is a possible way of objectifying the investigator's intuitions, so that he may then take whatever action seems most appropriate to prevent the eventual user of the count being misled by them. He is thus enabled to remedy any lack of representativity of his corpus vis-à-vis the target register, a flaw which, as we have seen, is well-nigh impossible to avoid and which may be revealed only after the count has been completed.

Notes

[1] This measure of dispersion was first presented in Juilland (1964) but the presentation in the Introduction to that work contains errors, as pointed out by Huddleston (1967). A clear exposition, in French, is available in Muller (1965). For an appraisal of Juilland's D vis-à-vis Range, Carroll's D_2 and Rosengren's S cf. Lyne (1981, Ch.9).

[2] This problem is exacerbated in a list based on 'registral value', i.e. on the degree to which each item's probability of occurrence in the register under investigation is higher than in the language in general (Lyne 1983).

References

Huddleston, R. (1967) "Review of Juilland & Chang FREQUENCY DIC-TIONARY" Journal of Linguistics 3, 1: 165-166

Lyne, A.A. (1975) "A word-frequency count of French business corre-spondence based on a corpus of approximately 80,000 running words" IRAL 13, 2: 95-110

Lyne, A.A. (1981) A Lexicometric Approach to the Description of a Language Variety: French Business Correspondence. Ph.D. thesis, University of Sheffield

Lyne, A.A. (1983) "Word-frequency counts: their particular relevance to the description of languages for special purposes and a tech-nique for enhancing their usefulness" Nottingham Linguistic Circular 12, 2: 130-140

Muller, Ch. (1965) "Fréquence, dispersion et usage: à propos des dictionnaires de fréquence" Cahiers de lexicologie 7, 2: 33-42

Muller, Ch. (1977) Principes et méthodes de statistique lexicale. Paris: Hachette-Université