Comparing corpus-based counts versus web page counts as estimates of lexical frequency

Christophe Pallier*

January 29, 2004

The "Lexique" database (http://www.lexique.org) provides textual frequencies and web page hit rates of about 130000 French words. In this note, we explore the properties and compare these two estimates of lexical frequency.

The frequency of occurence of words plays an important role in human visual and auditory word recognition processes. Psychologist usually employ lexical frequencies estimates based on corpora consisting of books, journal articles...

Using web pages as a sample is an alternative, and attractive option (Blair, 2002). Several Internet search engines report the number of pages containing a given target word (this number is called the 'web page hits'). Thus, knowing the total number of pages indexed by the engine, one can compute the proportion of web pages containing a given item. It seems to us that Internet might better reflect the current word usage than the type of texts (books and journal articles) which are included in the corpora used by Psycholinguists.

It should be obvious that page hit rate is *not* a direct estimate of the lexical frequency of the item. Firstly, repetitions on the same page are not taken into account. Second, consider the English item 'the' which probably appears in most English web pages, and therefore has a percentage of hits nearing 100%: yet, its frequency of occurrence in the language is obviously much less than 100%. For very high frequency items, therefore, one can expect the percentage of web page hits to be larger than the frequency of occurence in texts. On the other hand, very-low frequency items probably often occur in the same web pages, and therefore, their web page hits might be rather small and maybe underestimate their actual frequency.

It would be nice if Internet search engines returned the number of occurences of a given target word in all the indexed pages, but they do not. Nevertheless, intuitively, when *comparing two items*, the page hits ratio may be a good approximation of the frequency ratio. This is what we assessed using the Lexique database (cf. http://www.lexique.org). The 'Graphemes' table lists about 129000 word forms and provides, for each of them:

- "frantfreqparm" which is the frequency per million words in a subset of the Frantext corpus gathered by the ATILF (http://www.atilf.org). Our subset of the corpus contained about 31 millions words and was mostly based on novels, but also included some poetry, philosophical essays, and technical or scientific works, all

published between 1950 and 2000.[1]

- "fsfreqparm" which is the number of 'hits' per million of web pages, computed by the fast-search engine (which had scanned about 15 millions French pages in 2001).

Table 1 shows the 30 most frequent items according to each variable ('franfreqparm' abbreviated as 'fran', and 'fsfreqparm' abbreviated as 'fs). Not surprisingly, all items are function words. The aggreement between the two orderings is not too bad. A word like 'Je' is more relatively more frequent in texts (remenber ther are mostly novels) than on the web.

Details about the construction of the database are provided in the paper by New, Pallier, Ferrand and Matos (2001), available on the lexique web site. It is important to know that:

- Words were defined graphically as series of letters enclosed between punctation signs. So the various forms of a given lexical item (give, gave...) are counted separately, while homographs are not distinguished.

- Lexique includes only the legal French words from the Frantext corpus: foreign words, abbreviations, proper names... were excluded. These excluded items represented about half of the different word forms, and accounted for 18 % of the all the corpus.

- We choose to keep the frequency of occurence in the original corpus for 'frantfreqparm' and this is why its sum is 820000 rather than 1000000.

- Concerning the web page hits, it is noticeable that Graphemes does not reflect a random sampling of words in web pages: the words all came from the written corpus and then their associated hits were retrieved. In other words, there are items on the web that have not included the database.

Figure 1 and Table 2 report the distribution of the variable 'frantfreqparm'. This distribution is very much dominated by the very low frequencies items: for example, only 31000 items out of the 129000 word forms have a textual frequency larger or equal to 1 per million.

The apparent linear relationship on the plot with double logarithmic coordinates is rather striking (the slope is -0.74). It is actually related to the so-called Zipf law which states that the product of the rank of a given word and its frequency is constant (Zipf, 1935, cited in Miller, 1951). The demonstration of the relationship is given in Appendix.

The distribution of web hits, shown on Figure 2, has a rather different shape than the distribution of textual frequency, especially in the low frequency range. It is not dominated by very low frequency item, but has its maxima in the [10-100] range, meaning that most words appear on about this number of pages (per million). The explanation for the difference with the textual frequency is not obvious to us. It suggests that it is not the case that the majority of words appear only on very few web pages.

Figure 3 shows the relationship between logtext and logweb, the log in base 10 of the two original variables (Thus an item with 'logtext' equal to 0 has a textual frequency of 1 per million; an item with 'logweb' equal to 2 appears in 100 pages over 1 million...).

The 'band structure' on the left and the bottom of are due do the discretisation due to the finite size of the corpora (e.g. textual frequencies were obtained

[1] In comparison the previous French lexical database Brulex was based on a corpus of 23.5 million words, with similar type of texts published from 1919 and 1964.

sort according to franfreq			sort according to fsfreq		
graph	fran	fs	graph	fran	fs
de	37524.35	867040.78	de	37524.35	867040.78
la	23889.00	734542.05	la	23889.00	734542.05
et	18621.71	709526.96	et	18621.71	709526.96
le	17901.87	698139.78	d’	12502.19	699171.08
à	16994.68	671758.44	le	17901.87	698139.78
les	16011.00	662602.52	à	16994.68	671758.44
d’	12502.19	699171.08	des	12299.45	662666.40
des	12299.45	662666.40	les	16011.00	662602.52
il	12021.52	330532.85	du	7141.45	622528.00
un	11468.61	543474.20	en	10644.13	621800.77
en	10644.13	621800.77	pour	5332.48	555476.39
que	9208.19	361979.81	un	11468.61	543474.20
une	8972.77	494661.72	sur	4209.61	507840.34
est	8745.94	458286.00	une	8972.77	494661.72
dans	7480.90	469398.81	a	4294.90	473902.87
du	7141.45	622528.00	au	4962.61	473442.11
qui	7121.97	368291.01	par	4773.39	470143.83
pas	6372.55	304400.04	dans	7480.90	469398.81
qu’	6360.81	242926.37	est	8745.94	458286.00
je	6260.29	124006.43	s’	4547.23	373409.51
pour	5332.48	555476.39	qui	7121.97	368291.01
ne	5218.58	251732.69	que	9208.19	361979.81
se	5031.00	267711.77	avec	3019.71	358380.68
au	4962.61	473442.11	ce	4696.55	356992.64
elle	4931.74	150930.07	ou	2954.74	355446.02
par	4773.39	470143.83	plus	4519.81	331528.28
ce	4696.55	356992.64	il	12021.52	330532.85
s’	4547.23	373409.51	n’	4255.13	311097.04
plus	4519.81	331528.28	pas	6372.55	304400.04
on	4364.48	216109.55	aux	1989.35	301894.26

Table 1: The 30 most frequent words according to franfreq and fsfreq

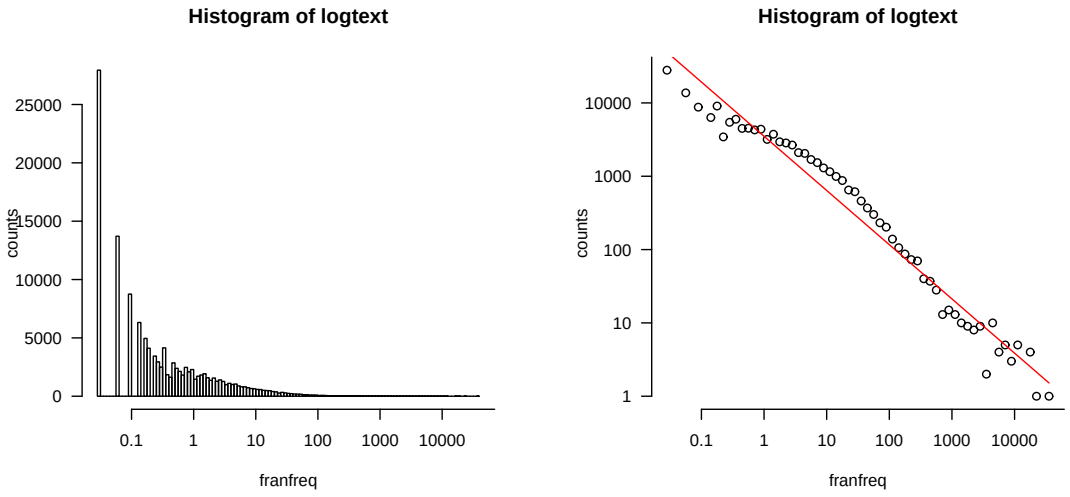

Figure 1: Distribution of textual frequency (left: linear scale on y-axis; right: logarithmic scale on y-axis)

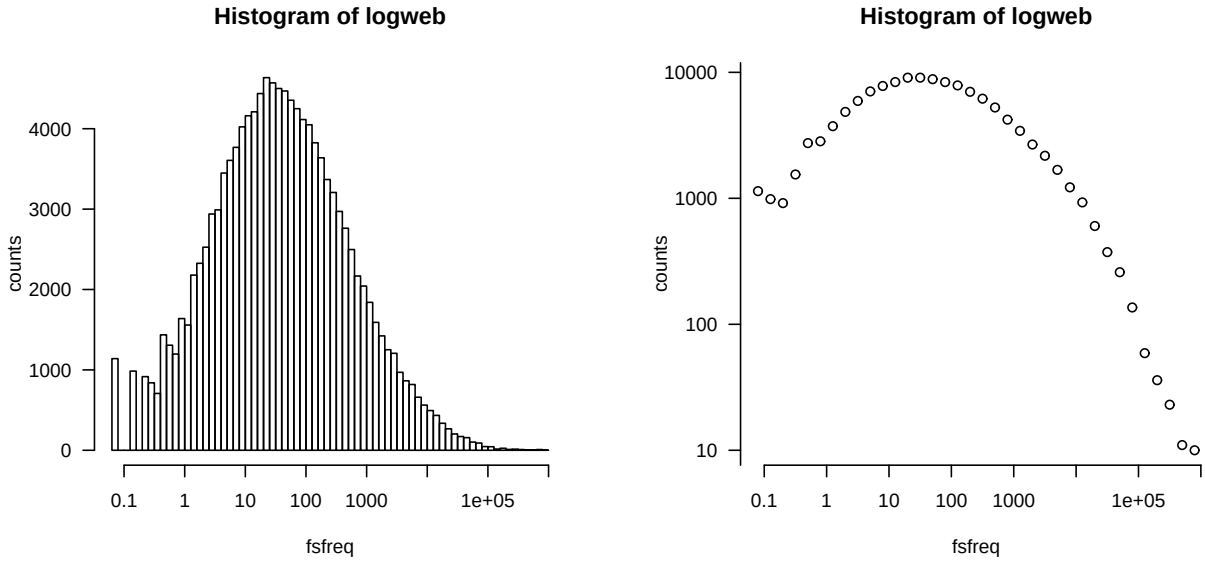

Figure 2: Distribution of web hit rates (left: linear scale on y-axis; right: logarithmic scale on y-axis)

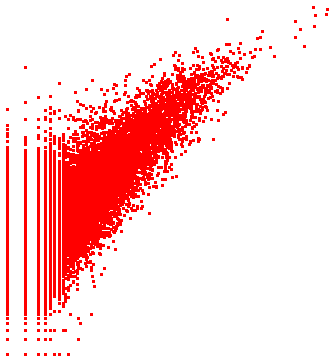

Figure 3: Relationship between logtext and logweb

Frequency range	Count	Proportion
<0.01	0.00	0.00
$0.01 < . < 0.02$	27941.	0.22
$0.02 < . < 0.05$	22462.	0.17
$0.05 < . < 0.10$	18825.	0.15
$0.10 < . < 0.25$	15918.	0.12
$0.25 < . < 0.50$	13182.	0.10
$0.50 < . < 1.$	19522.	0.15
$1. < . < 5.$	4523.	0.04
$5. < . < 10.$	3029.	0.02
$10. < . < 20.$	2085.	0.02
$20. < . < 50.$	735.	0.01
$50. < . < 100.$	334.	0.00
$100. < . < 200.$	218.	0.00
$200. < . < 500.$	56.	0.00
$500. < . < 1000.$	32.	0.00
$1000. < . < 2000.$	29.	0.00
$2000. < . < 5000.$	12.	0.00
$5000. < . < 10000.$	9.	0.00
$10000. < . < 20000.$	2.	0.00
$20000. < . < 500000.$	0.	0.00
$500000 < . < 1000000.$	0.	0.00

Table 2: Distribution of franfreqparm

by dividing a count per 31).

The regression line between logtext and logweb has equation:

$$\text{logweb} = 1.09\text{logtext} + 2.18 \qquad (1)$$

The 2.18 constant reflects the fact that the web hits counts are larger (by a factor 150!) than the frequencies of occurrence.

On afterthought, this is not too surprising: fsfreqperm provides the hits per million web *pages*, and this is related to the average number of words on web pages. As an estimate of word frequency, one would rather want the number of hits per million *words* on the web! If we knew the mean number of words in web pages, we could normalize fsfreqperm by this factor.

If Grapheme included all the words found on the web, and if all words occured at most once on each web page, than sum(fsfreqparm) would provide the total number of words per million web pages. This is equal to 140 millions (and log10(140)=2.1). Though this underestimates the real number of word on the web page, this figure can be compared with the factor 150 (log10(150)=2.18) found above. A slightly more realistic approach is to assume that the items in Graphemes cover only 82% of the web as for the Frantext Corpus (remember that sum(fsfreqparm)=820,000). Then, one could define, for each item,

$$\text{wfsfreqparm} = \frac{\text{fsfreqparm}}{\text{sum(freqparm)/sum(fsfreqparm)}}$$

which would provide an estimate of the frequency per million *words* on the Net (ignoring the problem of repetition of the same word in the same pages). This amount to use the same scale for franfreq and fsfreq, and therefore, by design, there would be the relationship:

$$\log(\text{wfsfreqparm}) = 1.09\log(\text{franfreqparm}) \quad (2)$$

Let us now consider the slope in the equation relating logweb and logtext. Its value (1.09) is not to far from 1, suggesting that the ratio of web hit estimates is about the same as the ratio of the textual frequency, a good news suggesting that this former ratio can be used to assess the relative frequencies of items.

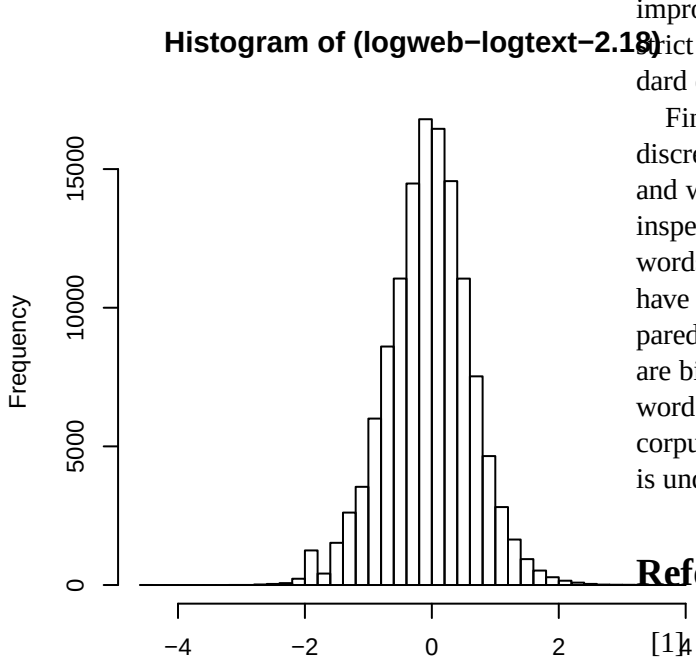

Figure 4: Distribution of logweb-logtext

Yet, let's examine ratios in more details: if we select two items, is the ratio of hits (fsfreqparm) similar to the ratio of frequencies estimates on the Frantext corpus (franfreqparm)? We said above that the 0.9 factor in the regression indicates that the ratios are roughly similar. But this is only true on average. Figure 4 shows the centered distribution of logweblogtext, which is equivalent to the log of the ratio between fsfreq and frantfreq. The width of this distribution provides an idea of the variability of the ratio fsfreq/franfreq. The standard deviation of this distribution is about 0.66.

Note, however, that this distribution includes the very low frequency items. As the left panel of Figure 3 shows, when the frequency increases, the agreement between the web estimates and frantext improves. So do the ratios. For example, if we restrict this analysis to items with fsfreq$>$10, the standard deviation narrows to 0.41.

Finally, Tables 3 and 4 display items for which the discrepancy between the two informations, textual and web frequencies, is maximal. Not surprinsingly, inspection of Table 3 reveals mostly Internet-raltaed words. Table 4, on the other hand, show items which have a relatively large textual frequency when compared to their web hits. This table suggests that there are biases in the original Frantext corpus, with some word over-represented. Recalling that the Frantext corpus mostly consists of a fez hundred books, this is understandable.

References

[1] Blair, I. V., Urland, G. R., E. Jennifer. Using Internet search engines to estimate word frequency. Behavior Research Methods, Instruments & Computers, 34 (2), 286–290.

[2] New, B., Pallier, C., Ferrand, L., Matos, R. (2001) Une base de données lexicale du francais comtemporain sur Internet: Lexique. L'Année Psychologique, 101, p.447–462.

[3] Zipf, G. K. (1935)The psychobiology of language. Boston: Houghton Mifflin.

graph	fran	fs
mail	1.29	106833.62
informatique	1.06	62732.96
site	5.52	198122.45
by	1.68	53753.07
forum	1.90	49081.95
vid´eo	1.26	31939.88
home	2.16	49468.39
fi chiers	1.58	29623.51
sites	3.71	67040.89
top	1.29	22029.56
tv	1.52	24157.15
r´eserv´es	6.32	80520.22
navigateur	1.32	16401.67
bienvenue	3.55	42860.90
formulaire	1.48	17573.48
envoyez	1.42	16557.74
actualit´es	2.06	23837.60
choisissez	1.03	11865.39
group	1.13	12066.21
h´esitez	1.16	11907.21
programmation	1.48	14639.79
serveur	2.55	25110.77
utilisateur	1.94	18876.84
accueil	17.32	167381.20
concern´es	1.13	10855.74
noël	1.19	11176.06
sommaire	7.90	72918.71
annonces	3.52	31513.79
technologique	1.26	11224.11

Table 3: Items with the largest fs/fran ratio for fran>1

graph	fran	fs
boches	11.74	39.72
g´eronte	6.81	22.70
caoua	1.10	3.57
tabès	1.71	5.18
boghei	1.26	3.78
jacter	1.90	5.32
vioque	2.48	6.72
folliculine	1.32	3.43
tabulatrice	1.13	2.87
superphosphate	4.81	12.19
guitounes	1.03	2.52
feldwebel	1.13	2.73
entrevision	1.16	2.66
chondriome	1.06	2.38
micheton	1.74	3.78
loufi at	2.55	5.32
zanzi	4.10	8.48
openfi eld	3.55	7.07
vape	1.10	2.10
nochère	1.10	1.82
exogamiques	1.23	1.96
frontin	15.74	24.10
c´ezigue	1.68	2.38
frichti	1.97	2.45
brocanteuse	1.52	1.89
catilinaire	2.23	2.66
syntol	1.74	1.75
drifter	18.13	17.72
bignole	2.45	2.10
maximation	2.97	1.82
d´ecarrade	1.16	0.70
soupier	1.23	0.63
passettes	1.26	0.63
lardus	1.13	0.56
rouquemoute	1.87	0.49
sevrais	16.77	0.07

Table 4: Items with large fran/fs and fran>1

[4] Miller, G. (1951) Language and Communication. New York: McGraw-Hill

Appendix

Zipf considered the function Z which associates rank to frequency:

$$Z : \text{rank}(w) \rightarrow \text{freq}(w)$$

The cumulative distribution of lexical frequency, F, associates to the frequency of a given word w, a value which is (roughly) equal to $1 - \text{rank}(w)/N$, where N is the total number of words. Therefore:

$$F = 1 - Z^{-1}/N$$

The histogram of frequencies has the same shape of the density of probability which we call f. f is the derivative of F. Therefore:

$$f = -(Z^{-1})'/N$$

Zipf' law states that the product fr is constant:

$$Z(r) = k/r$$

Therefore:

$$f(x) = k_2/x^2$$

Which shows that the density of probability (and therefore the histogram) is linear on a log-log diagram.

Figure 5 show this relationship for the items with a frequency larger than 100.

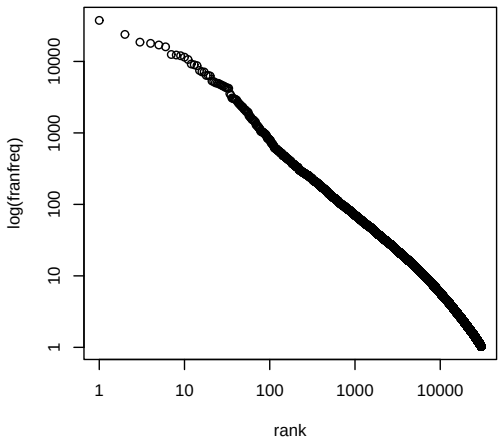

Figure 5: Zipf's law predicts a linear relation ship between rank and log(frequency). Here the relationship is plotted for items with franfreqparm>100.