REVIEWS

Notices:

Surname Distribution Prints from the GB 1998 Electoral Roll Compared with those from Other Surnames Distributions

D. K. Tucker
Carlton University, Ottawa

Abstract

This article describes and uses a new graphical method to create *prints* of the surname distribution of the GB 1998 Electoral Roll and other national distributions using a new *occupied frequency* technique which allows these distributions to be directly compared visually.

Seeking a Visual Image

One of the difficulties in studying surname distributions is to get an overall visual image of the distribution. The UK 1998 Electoral Roll surname distribution has 781,728 types, where each type, or individual surname, has a number of tokens, or people, with that particular surname.[1] The tokens total 47,054,569 and represent the population in this distribution.

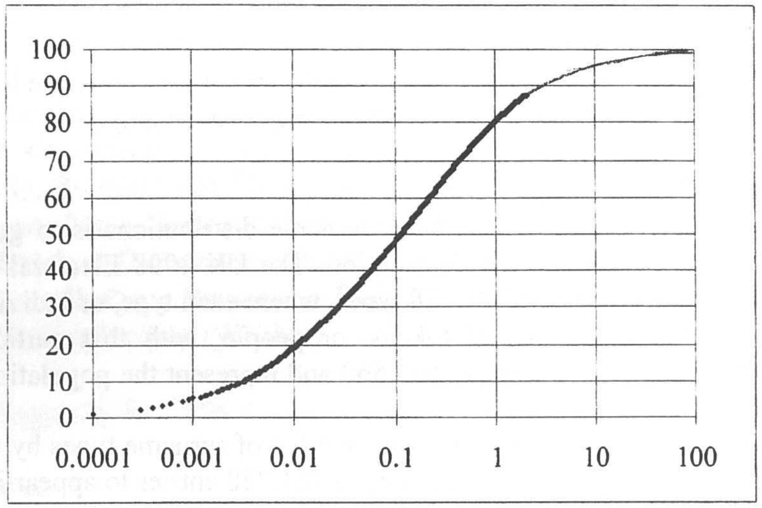
It is comparatively easy to organize a list of surname types by their counts, but the list would be too long at 781,728 entries to appear in an article such as this one. Furthermore comparing lists of such magnitude is virtually impossible. Even if we wanted to plot such information to get a picture or diagram of the distribution, there are few means to do so.

In my article I used two kinds of charts: *The Percentage of Population Against the Percentage of Names*, for example Graph 3 in the reference; reproduced here as Graph 1, and, *The Percentage of Population Against Rank*, for example Graph 7 in the reference; reproduced here as Graph 2. Both kinds have their advantages and uses, but they also have their disadvantages.

[1] D. K. Tucker, 'The forenames and surnames from the GB 1998 Electoral Roll compared with those from the UK 1881 Census',*Nomina*, 27 (2004), 5–40.

The first kind, Graph 1, is normalized and each plot line must, by definition, end at the (100,100) coordinate since 100% of the names must axiomatically accommodate 100% of the population. The plots for very different surname distributions are very similar. This is extremely useful in showing the similarities between very different distributions but not in showing those differences, if any.

Graph 1: UK 1998 Electoral Roll Surnames

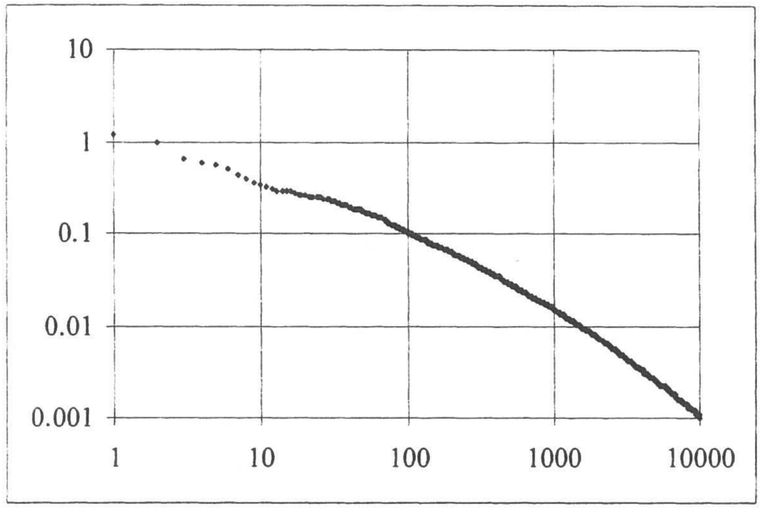


Percentage Population (Y) plotted against Percentage of Surname Types by Rank (X log scale)

Note that the heavy line represents the machine plotted line ending at (2.1, 87.6) for 16,383 plots: the maximum number of plots that Microsoft (MS) Excel can accept. The lighter line shows the curve that would result if all values could have been plotted. What the curve shows clearly is that 1% of the surname types represent 80% of the population.

The second kind, Graph 2, again as demonstrated in the same article, is very useful, but readers will recall that only the first thousand ranked names were shown. A thousand plots are not enough to get the feel for the whole distribution to compare it with another. What is required ideally is a picture that accommodates all the types and tokens in a meaningful way generated by a commonly available graphing device.

Graph 2: UK 1998 Electoral Roll Surnames



Population (Y log scale) plotted against Surname Types by Rank (X log scale)

Zipf's Law and Power Laws

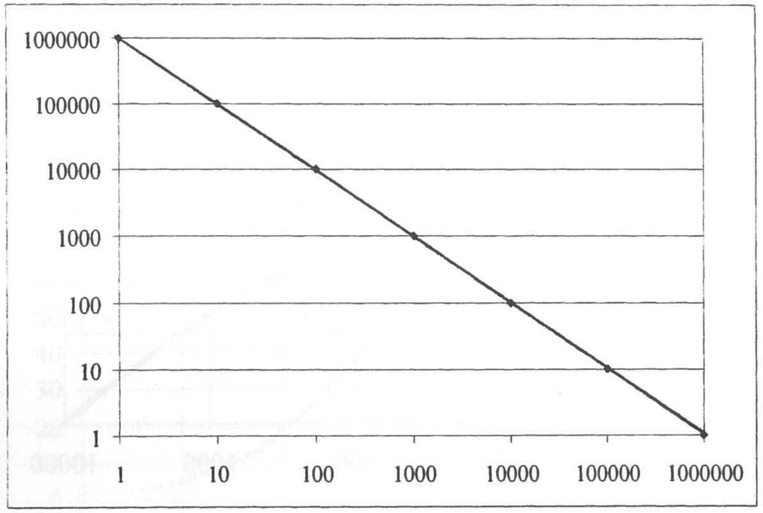
Surname and forename distributions follow the empirical law known as Zipf's Law (defined at http://en.wikipedia.org/wiki/Zipf's_law). The law is often explained as a power law relationship of the form X times Y equals a constant K: algebraically, X*Y=K. When the logarithm of Y is plotted against the logarithm of X the result is a straight line descending from the left to the right.

When applied to surname and forename distributions X represents the individual names: the types, and Y represents the number of tokens. Note that both X and Y can only be integers: it is impossible to have 1.6 of a name, or 56.3 people.

Graphical Constraints

Graph 3 shows a hypothetical curve for X*Y=1,000,000. Only seven plots are given as we know the formula and thus know that the curve is in this case a straight line.

Graph 3: Zipf Curve for X*Y=1,000,000



Both Y and X axes have log scales

If we wanted to plot a data set that we supposed was Zipfian, (i.e. that it followed Zipf's Law) with a common desktop tool such as MS Excel we would find that the X range is constrained to a little over 16,000. We could not plot the surname distribution discussed earlier using MS Excel as the distribution requires over 781,000 entries.

Even if the plotting problem could be solved what would be the utility of the result? We probably could not visually compare distributions with X values in the hundreds of thousands. What is required is to restructure the data without losing any information. This can be done as demonstrated in the following section.

Integer Values, Identical Counts and Multiple Occupancy

The relationship X*Y=1,000,000 has a unique value of Y for every X value, and there can be no same Y values for different X values. The Y value need not be an integer, even if X values are constrained to be integers, e.g. when X=3, Y is not an integer. However, as previously mentioned there can be no fractions of people, or occupied frequencies, so they are both constrained to integers; this is how multiple counts appear. For example if X=1032, Y=968.99=968 as an integer; when X=1033, Y=968.05=968 as an integer. Hence there is a double occupancy of 1936 at occupied frequency 1032. The same argument applies for all other multiple occupancies.

There was some 'smudging' in Graph 2 at around X=10,000. What we see is the integer effect described in the above paragraph. If Graph 2 were plotted for all values we would see, not a straight line but a line as in Graph 2 which becomes what looks like an off-set stack of horizontal lines of ever increasing length. The smudging effect noted above is the beginning of this phenomenon: if we could plot the X values for which the Y value is one, there would be 374,271 of them. The relationship X*Y=K (in this case K=1,000,000) thus only describes the overall form of the discrete plots when the objects represented by X and Y can only have integer values, and the usefulness of such a description is rapidly curtailed with multiple plots of the same value for different X values.

The Occupied Frequency Approach

Consider a hypothetical surname distribution that follows Zipf's Law and has the form X times Y equals 1,000,000—the same as for Graph 3.

The initial values of X and Y are thus:

X	Y
1	1,000,000
2	500,000
3	333,333 (only integers are allowed)
4	250,000
5	200,000

For convenience we can give the X values hypothetical names such as *Ssmith, Jjones, Wwilliams, Ttaylor* and *Ddavies.*

X	Name	Y
1	Ssmith	1,000,000
2	Jjones	500,000
3	Wwilliams	333,333
4	Ttaylor	250,000
5	Ddavies	200,000

The count for *Ssmith* is 1,000,000; we may also say that the frequency of *Ssmith* is 1,000,000, and that of *Jjones* is 500,000. In this hypothetical surname distribution the first occupied frequency is thus 1,000,000 and the second occupied frequency is 500,000; frequencies 499,999 to 999,999 inclusive are thus unoccupied frequencies; there are no names at these frequencies and they will be ignored.

We could now plot the population, Y, against the occupied frequencies, X. For *Ssmith* to *Ddavies* these plots are:

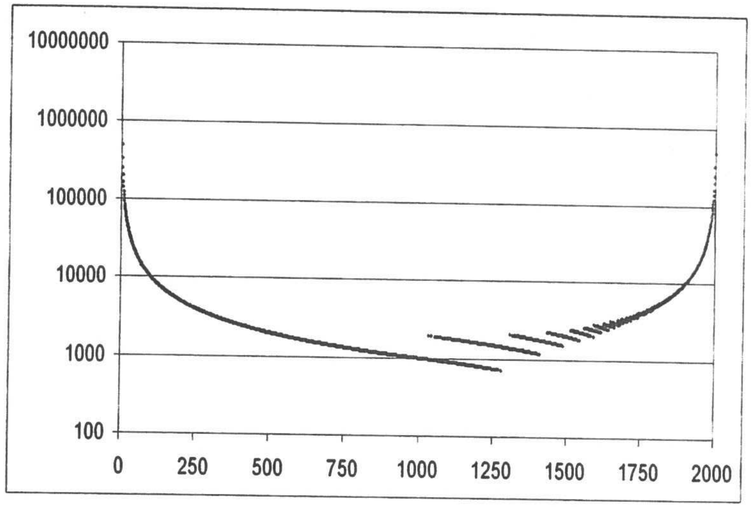
Name	Population	Occupied Frequency
Ssmith	1,000,000	1
Jjones	500,000	2
Wwilliams	333,333	3
Ttaylor	250,000	4
Ddavies	200,000	5

We continue plotting the individual counts until we reach a situation where two surnames have the same count. This occurs at the 1044th occupied frequency where the counts are 956. The population at this occupied frequency is thus 2 x 956 =1912 and this is the plot value at X=1044. We continue plotting single and double occupancies, until we find three names with the same number of tokens, 683, at occupied frequency 1,317. The plot would be 2,049 at occupied frequency 1,317. This process continues with multiple occupancies until the last occupied frequency where 500,000 names each have a count of 1.

The Occupied Frequency Graphs

The total number of occupied frequencies is under 2,000 and can easily be plotted using MS Excel. For convenience we will use a log scale for the Y axis as the values range from 1,000,000 to just under 1,000, but there is no need for a log scale with the X axis. The results of this exercise are shown in Graph 4 which is most unusual.

Graph 4: Idealized Zipfian Surname Distribution



Population (Y log scale) Plotted Against Occupied Frequency Number (X)

The graph has what appear to be two maxima, and a minimum. The first maxima is self explanatory but the second is not. It is caused primarily by the 'integers only' effect described earlier. The minimum is also caused by the integer effect.

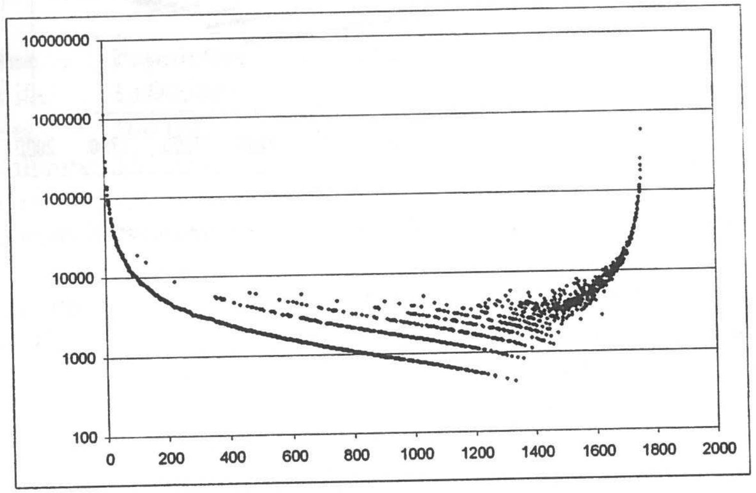
The plots are so close together that it seems that there are continuous lines; but there is no continuity between plots. However, there are sets of plots that appear as a distinct set (quasi-lines) and these quasi-lines (QLs) overlap. The major QL from the top left to about x=1,300 is when there is one name in the occupied frequency; this QL comes to a minimum at plot (1279,721). The second QL represents the case where there are two names that occupy the same frequency. It starts at plot (1044,1912) and ends at plot (1410,1180). It is overlapped by the third QL. This series of overlapping QLs continues until the QLs become single plots, and so on to the last occupied frequency where there are 500,000 names in occupancy, each with a count of one.

The hypothetical population covered by the graph is 13,970,034 people.

A Less Idealized Graph

This idealized graph is very distinctive. However, it is an idealized graph and to make it more realistic each plot contributing to an occupied frequency was subject to a random variation factor between 0.80 and 1.20 (i.e. ±20%) drawn freely from www.random.org. The population per occupied frequency was then re-plotted as per Graph 5.

Graph 5: Typical Zipfian Surname Distribution



Population (Y log scale) Plotted Against Occupied Frequency Number (X)

As expected the clean QLs of Graph 4 have been replaced by broken and fuzzier QLs, but the overall shape has been preserved. There are fewer occupied frequencies in Graph 5 than Graph 4 as the randomization has forced a net increase in the multiple occupancy. Graph 5 is an easily recognizable kind which represents a pure Zipfian curve subject to a ±20% randomization. Graphs of real surname distributions can be compared with Graph 5 to see if they are Zipfian.

The 1998 GB Electoral Roll

The 1998 GB Electoral Roll data and the UK 1881 Census data are taken from my previous article.[2] Consider the first five names in the ranked list of surnames from the UK 1998 Electoral Roll; they are, with their counts:

Types	Tokens
Smith	569,914
Jones	448,306
Williams	304,938
Taylor	264,905
Davies	232,247

The count for *Smith* is 569,914; we may also say that the frequency of *Smith* is 569,914, and that of *Jones* is 448,306. In the UK 1998 Electoral Roll the first occupied frequency is thus 569,914, and the second occupied frequency is 448,306; frequencies 448,307 to 569,913 inclusive, are thus unoccupied frequencies.

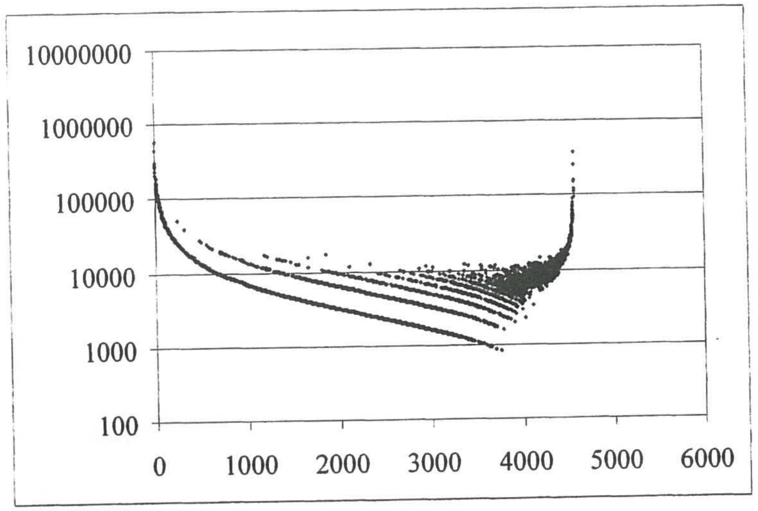
We could now plot the population, Y, against the occupied frequencies, X. For *Smith* to *Davies* these plots are:

Name	Population	Occupied Frequency
Smith	569,914	1
Jones	448,306	2
Williams	304,938	3
Taylor	264905	4
Davies	232247	5

[2] Tucker, 'The forenames and surnames from the GB 1998 Electoral Roll compared with those from the UK 1881 Census'.

The complete curve is shown in Graph 6.

Graph 6: UK 1998 Electoral Roll Surnames—The Occupied Frequencies



Population (Y log scale) Plotted Against Occupied Frequency Number (X)

This graph is identical in overall format to Graph 5 and shows that the surname distribution obeys Zipf's Law. This graph contains all the data for the 47,054,569 individuals in the surname distribution of the UK 1998 Electoral Roll—nothing is omitted. It is a complete graphical representation of the distribution, and is plotted using MS Excel on a desktop machine. It could be described as looking like a Viking ship, or like a fingerprint.[3]

The graph shows that it is almost as common to have a unique surname, for example *Eklu* (a component of the second maxima), as it is to have the most common surname, *Smith* (the first maxima).

[3] An early precursor of this kind of graph appeared in D. K. Tucker, 'Distribution of forenames, surnames, and forename-surname pairs in the United States', *Names*, 49 (2001), 69–96, although readers may not recognise it.

However, inspection of the listing of the surnames with a count of one shows many surnames which do not match English orthography, such as *Ekmpthorne*. It is probable that this is a *transposed letters* typographical error for *Kempthorne*. We can say with certainty that there are typographical errors present and that some break English orthographical rules which a good English speaker could spot given the time and motive. However, there are many that satisfy the orthographical rules. Is *Eklu* a surname, or is it a typo for *Kelu*? Perhaps both are legitimate surnames, or perhaps both are typos. Without some reference it is impossible to say. *Eklu* is a surname with many Google hits; *Kelu* appears to be a forename. What can be said is that the second maxima is inflated by typographical errors.

The minimum of the first line, and that of the graph, is at the 3,748th occupied frequency with a count of 823; the surname is *Brosnan*. The minimum for the second line is at the 3,782nd occupied frequency with a count of 1,578, This trend of increasing occupied frequency with increasing count for the nth line is generally true, and it culminates at the last occupied frequency at 4,570 with a count of 374,271. This is the number of unique surnames, and typos, and compares in number with the second plot of the graph, the second most popular surname which has a count of 448,366.

It is argued that this graph style, although strange to look at initially, allows the reader-viewer a better feel for the distribution; every count of every surname is built into the graph: nothing is omitted and it can easily be plotted using standard desktop-laptop tools.

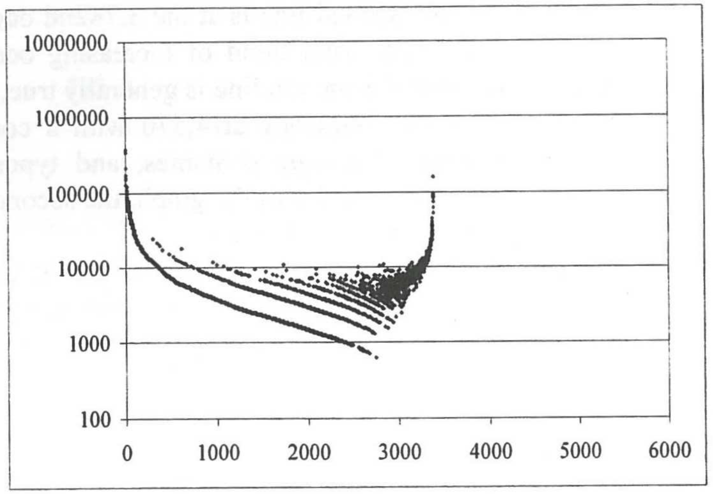
Comparison by Country

Table 1 lists the type and token data for the distributions which are offered for comparison. All these graphs have identical formats and scales: thus the size of the collections is reflected in the size of the plot and may be compared directly.

Table 1 Types and Tokens by Country

Country	Year	Types	Tokens	Graph
GB	1997	781,728	47,054,569	6
UK	1881	401,198	26,124,584	7
France	2005	31,367	36,200,159	8
USA	1997	1,742,074	89,036,422	9
Canada	1997	520,222	11,036,145	10
Australia	2000c	285,509	5,978,875	11
New Zealand	2000c	25,631	755,483	12

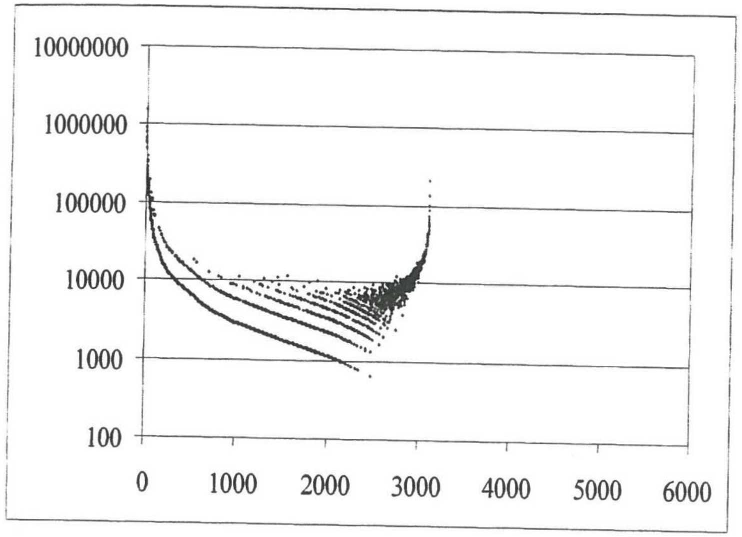
Graph 7: UK 1881 Census Surnames—The Occupied Frequencies



Population (Y log scale) Plotted against Occupied Frequency Number (X)

Graph 7 has fewer occupied frequencies and lower counts than Graph 6 as one would suspect for a smaller population.

Graph 8: French Literary Corpus 2005—The Occupied Frequencies



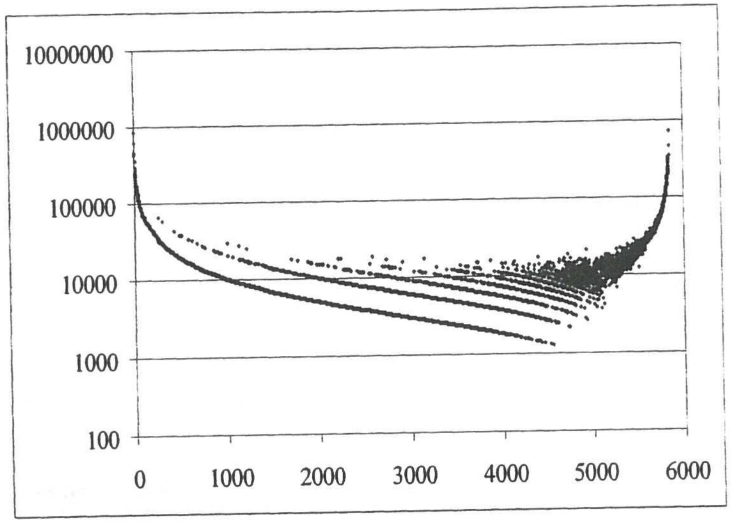
Population (Y log scale) Plotted against Occupied Frequency Number (X)

Graph 8 was generated using French data and it too has the identical form to the previous surname distributions. This is very interesting since it is not a collection of French surnames but it is from the 36 million word French literary corpus kindly supplied by Dr Boris New, Maître de Conférence, Laboratoire de Psychologie Expérimentale, Boulogne. It is not surprising to find that a literary corpus follows Zipf's Law in common with the surname distributions; the graph forms merely underline that surname distributions are Zipfian. It is almost identical in size and form to Graph 7—The UK 1881 Census Surnames.

Graph 9 is the US surname data from an earlier article which was based on the 1997 telephone directory.[4] Again the graph is of identical format but the area covered represents the larger population encompassed.

[4] Tucker, 'Distribution of forenames, surnames, and forename-surname pairs in the United States'.

Graph 9: US Surnames 1997—The Occupied Frequencies



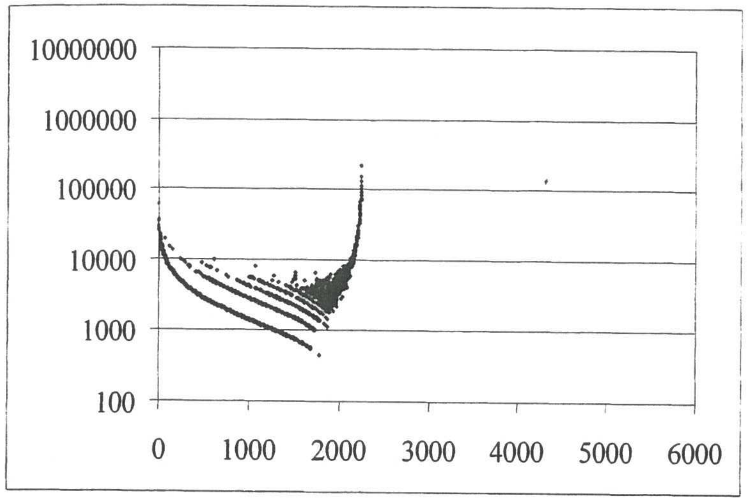
Population (Y log scale) Plotted against Occupied Frequency Number (X)

Graph 10 is Canadian surname data from an earlier article which was based on the 1997 telephone directory.[5] It maintains the standard overall shape, but with a minor variation and that is that the first maxima is smaller than the second. This probably has something to do with the two major colonizing nations: British and French, which share equally the top 10 surnames: *Smith, Brown, Tremblay, Martin, Roy, Wilson, Gagnon, Johnson, Campbell*, and *Côté*.

Graph 11 is taken from the circa 2000 Australian Electoral Roll kindly provided to me through the good offices of Professor Richard Webber, University College London (UCL). The graph is about half the size of the Canadian Graph 10 but preserves the overall shape.

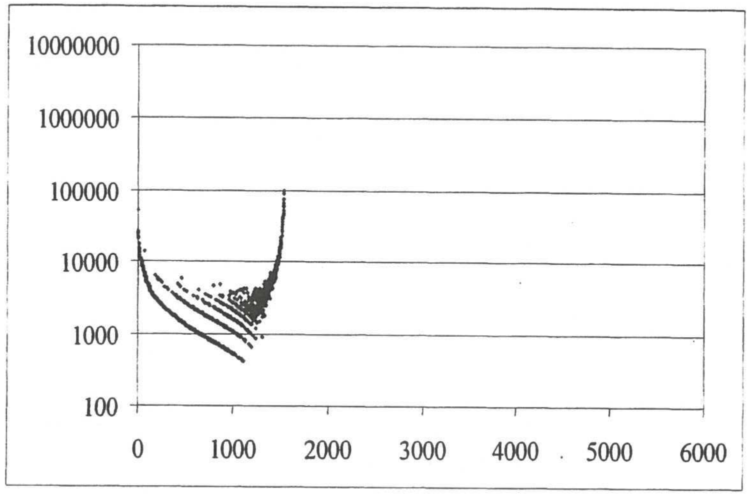
[5] D. K. Tucker, 'Distribution of forenames, surnames, and forename-surname pairs in Canada', *Names*, 50 (2002), 105–32.

Graph 10: Canadian Surnames 1997—The Occupied Frequencies



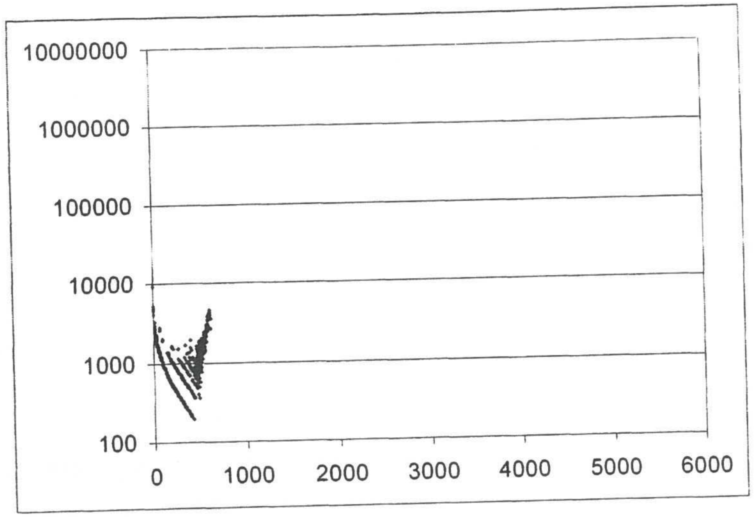
Population (Y log scale) Plotted against Occupied Frequency Number (X)

Graph 11: Australian Surnames—The Occupied Frequencies



Population (Y log scale) Plotted against Occupied Frequency Number (X)

Graph 12: New Zealand Surnames—The Occupied Frequencies



Population (Y log scale) Plotted against Occupied Frequency Number (X)

Graph 12 is taken from the circa 2000 New Zealand Electoral Roll again kindly provided to me through the good offices of Professor Richard Webber, UCL. The population of New Zealand is comparatively small and this is reflected in the graph. The second maxima is much less defined than on all the other graphs and this may have something to do with the data source.

The French Data and Typographical Errors

The French data allows an insight into the probable typographical error rate in the surname collections if we assume that the same method of data entry—key entry—was used in all cases. As previously argued it is very difficult to identify these errors in a surname or forename distribution. *Jfffrson* (a real case) is clearly an error for *Jefferson*, but is *Micklewhite* an error for *Mickelwhite*, vice-versa, or neither? These latter problems are insoluble by inspection as opposed to the *Jfffrson* type. Perhaps if we could get Carter he could tell us.[6] What we need is a register of all surnames and forenames.

With the corpus, as opposed to forename or surname collections, the words are known and they all appear in a dictionary. Thus all 'obvious to this author' typos were removed from the French Literary Corpus numbers, as shown in Table 2.

Table 2 French Literary Corpus

Version	Types	Words in Millions
Original	895,253	34.8
Revised	313,671	36.0
Reduction	65%	4.33%

Table 2 clearly shows the impact of typographical errors in such distributions. The revised number of types shows a reduction of 65% but represents only 4.33% of the words; as with surnames and forenames, typos have a huge impact on the number of types whereas the impact on the number of words is minimal.

The problem of typos can be eliminated. For example *The Globe and Mail* Canadian newspaper of Thursday, August 3, 2006, reports that a map was incorrectly catalogued in the *Library and Archives Canada*, as *Isola Di Ferra Nuova* rather than *Isola Di Terra Nuova.* This problem can be largely avoided by only allowing words, or phrases, not letters, to be keyed. This is achieved by using a controlled register and using the keys to point to the word required; *Terra Nuova* would be in the register; *Ferra Nuova* would not be. It is safer and faster than typing. Exactly the same method can be used for surnames and forenames.

Conclusion

The *Occupied Frequencies* graphs provide a meaningful way of representing large and small surname distributions, using standard desktop

[6] Michael Caine is the screen name for Maurice Micklewhite. *Get Carter,* 1970, is a popular film of Caine's, who played Carter.

tools, and allowing direct comparison of different distributions. Surname distributions are Zipfian as established from first principles, and by direct comparison with the French Literary Corpus.

The method provides a recognizable image, the size of which is proportional to the population represented. It is thus easy to differentiate the Australian from the US, although their overall forms are identical. If the image of a Viking boat is used: the larger the boat the higher it floats.

Shoreditch and Car Dyke: Two Allusions to Romano-British Built Features in Later Names containing OE *dīc*, with Reflections on Variable Place-Name Structure

Richard Coates
University of the West of England

Shoreditch, Middlesex

Shoreditch (Middlesex) should be regarded as having an obscure first element. The English Place-Name Survey for Middlesex offers no solution;[1] Mills,[2] following Field,[3] suggests 'ditch by a steep bank or slope', from a hypothetical Old English (OE) **scora*, but adds: "Not surprisingly [presumably because of urbanization, RC], the precise topographical features originally referred to are no longer evident". The spellings in the record for Shoreditch alternate between having medial <e>, medial <es> or medial zero, and Mills's suggestion does not account for the range seen in the alternation. Ekwall[4] and Watts[5] identify the feature as the shore, but Shoreditch is not adjacent to the Thames waterfront, and the eastern part of the City of London intervenes; no ditch leading to the shore, as Ekwall and Watts would have it, is in evidence but that which formed part of the Roman city wall complex itself. Houndsditch, parallel with a section of the wall, was not made till the reign of John (1199–1216), and the name of Shoreditch is on record from *c.*1148. What follows suggests that it is very much older.

[1] J. E. B. Gover, A. Mawer and F. M. Stenton, with the collaboration of S. J. Madge, *The Place-Names of Middlesex apart from the City of London*, EPNS, 18 (Cambridge, 1942), pp. 145–46.

[2] A. D. Mills, *A Dictionary of London Place Names* (Oxford, 2001), p. 208.

[3] J. Field, *Place-Names of Greater London* (London, 1980), p. 85.

[4] E. Ekwall, *The Concise Oxford Dictionary of English Place-Names*, 4th edn (Oxford, 1960), p. 419.

[5] *The Cambridge Dictionary of English Place-Names*, edited by V. Watts (Cambridge, 2004), p. 547.