Distribution of Forenames, Surnames, and Forename-Surname Pairs in Canada

D. K. Tucker

Esson & Tucker Management

Using information from the 1996 INFOUSA ProCD CanadaPhone, I present graphically frequency distributions of forenames, surnames, and forename-surname pairs in Canada and compare each distribution with the equivalents in the United States. Minor cultural differences between Canada and the U.S. in the use of forenames are noted. Lists of the most popular forenames, surnames, and forename-surname pairs in Canada are presented. Although the source provides more information than previously available, several deficiencies are noted, in particular the absence of diacritics does not allow all French names to be correctly rendered, and the method of listing results in a gross under-representation of female forenames.

Introduction

In this article I describe the distribution of personal names in Canada and generally follow the pattern set in my earlier article (Tucker 2001), which described the distributions of forenames, surnames, and forename-surname pairs in the United States. Here, as there, I use *name,* unless otherwise noted, to mean *personal name*, either a forename or surname.

Assuming that everyone in Canada has a personal name which is comprised, as a minimum, of a forename and a surname, we can say that if the population is X, then there are X surnames, X forenames used as first names, and X forename-surname pairs. My son's name, Jason Andrew Tucker, would have *Jason* as the forename, *Tucker* as the surname, and *Jason Tucker* as the forename-surname pair. *Andrew* does not feature further in this discussion, as it is almost impossible to obtain information on "middle" names on a grand scale, whereas the other information is readily available from CD-based telephone directories,

Names 50.2 (June 2002):105-132
ISSN:0027-7738

albeit with their well known limitations, as discussed in Hanks and Tucker (2000).

Most people share their forename and/or surname with others; some have a unique forename or surname, or both. Each name that is different from all other names is called a name *type*. Each example of that name is a name *token*. Among Canadian telephone subscribers, there are 7,274 tokens of the type *Jason*, 3,003 tokens of the type *Tucker*, and 2 tokens of the type *Jason Tucker*. *Jason*, *Tucker,* and *Jason Tucker*, represent three name classes: forename, surname, and forename-surname pair, respectively. We know that the total number of tokens of all the types within a type-class equals the population, but what is not obvious is the relationship between types and tokens. Here I will provide an answer to that general question and a few more, but in turn I will raise additional questions for others to answer.

The source data I used was the 1996 Edition 4 of INFOUSA ProCD CanadaPhone, a CD listing 12 million Canadian telephone subscribers. Using the standard export function supplied with the product and the greater than 50,000 records export facility authorized by an unlock code from ProCD, the subscriber names for all residential (as opposed to business) listings, were extracted. (It should be noted that in giving frequency counts we are counting telephone lines and not people. However, the sample is about 35% of the population and is a highly representative sample, as discussed in Hanks and Tucker [2000].)

The extract was subjected to extensive analysis to remove such remaining non-residential listings as municipalities, universities, services, hospitals, religious houses, utilities, military, and others. The compound names were repaired where necessary[1] and the individual forenames extracted and extraneous qualifiers, such as *Real Estate, The Big One,* and *Physiotherapist,* were removed. Unfortunately, the data contained no diacritical marks, which is the usual case with computer files of names of this sort, which were generated with older technology and where the data has not been recaptured.

In the tables below, *Unknown* means that a forename was shown to exist but the actual forename was unknown. An entry such as *Mr & Mrs Frank Churchill* shows one forename but the other is unknown; the two forename-surname pairs from this entry are thus *Frank Churchill* and *unknown Churchill.* In the case given we know that *unknown Churchill* is a female. In the case of *Mr & Mrs F Churchill* we get two *unknown Churchill* forename-surname pairs: one female and the other male. In the case of an entry such as *Frank & Jane Churchill* we see two known forenames, with the two forename-surname pairs: *Frank Churchill & Jane Churchill.* In the case of an entry such as *Mr Frank Churchill* we know that there is one entry; we know no more and anything else would be conjecture. Finally we get entries of the form *Frank Churchill & Jane Fairfax* where again we have two forename-surname pairs.

The extraction and analysis revealed the statistics that are shown in tables 1 and 2.

Table 1. Number of Name Types and Tokens—Canada and U.S.—in Millions.

	U.S.	Canada	Ratio
Surname Tokens	88.7	11.0	8.1
Unknown Forename Tokens	15.7	5.6	2.8
Known Forename Tokens	73.0	5.5	13.3
Forename-Surname Pairs Tokens	73.0	5.5	13.3
Surname Types	1.75	0.52	3.4
Forename Types	1.25	0.15	8.3
Forename-Surname Pairs Types	27.3	2.87	9.5

For completeness the mean and standard deviation of tokens per type for each class is given in table 2 for both Canada and the U.S., but as we shall see, because of the skewness of the distribution, these measures offer little value.

Table 2. Means and Standard Deviations of Tokens per Type.

Type	Mean		Standard Deviation	
	U.S.	Canada	U.S.	Canada
Surname	51	21	1544	281
Forename	58	36	4703	946
Pairs	3	1.9	70	5.4

From tables 1 and 2, we can get an indication of the relative variety of names in Canada and in the U.S. Table 1 shows that the U.S. has more tokens, and types, of forenames, surnames, and forename-surname pairs than does Canada. The mean values found in table 2 show that—relatively speaking—there are fewer tokens per type for all categories in Canada. Put another way, there is greater relative variety of forenames, surnames and forename-surname pairs in Canada than in the U.S. and the Canadian distributions are less skewed as well, as we will see below.

Assuming that telephone penetration in the U.S. and Canada is roughly comparable, we can get a measure of the relative populations from the surnames ratio shown in table 1. We might expect that all the ratios would be roughly in this order; however there are two major factors at work which make this expectation non-viable. The first factor is that in the U.S. 82% of people with a telephone listing also list their forenames along with their surnames, while only 50% do so in Canada.[2] (The use of forenames versus initials presents an interesting cultural difference between Canada and the U.S., which needs to be explained. Another cultural difference concerns the number of entries that are for two or more people. In the U.S. there were 10% such entries, whereas in Canada there were less than 3%.)

The second factor is the greater variety of surname types in Canada than in the U.S. This fact needs to be explained. It may have to do with Canada's smaller population, with different immigration policies and practices in the two countries, or with other factors. While intriguing and while bearing upon the issue of names, further consideration is beyond the scope of this article.

In Canada there are as many unknown forenames as there are known forenames; it is likely that if we knew these unknown forenames, many—but not all—would probably be subsumed within the 0.15 million forename types. Thus the forenames types and surname-forename types ratios in table 1 are likely to be overstated. In the U.S. we were able to deduce that the majority of unknown forenames are forenames of females but we cannot say that for the Canadian situation, although we know that women are underrepresented in the source data.

The number of known forename-surname pairs, 2.87 million, is surprisingly low as the number of different surnames and the number of different forenames would allow (0.52*10^6)(0.15*10^6), or 78 billion unique forename-surname pairs; more than enough to allow every Canadian to have not only one but many unique names. Since this is not the case I suspect that there is order in the naming of people; a notion that I will return to below.

Graphic Representation of the Data

Cumulative Curves—Tokens Plotted Against Types

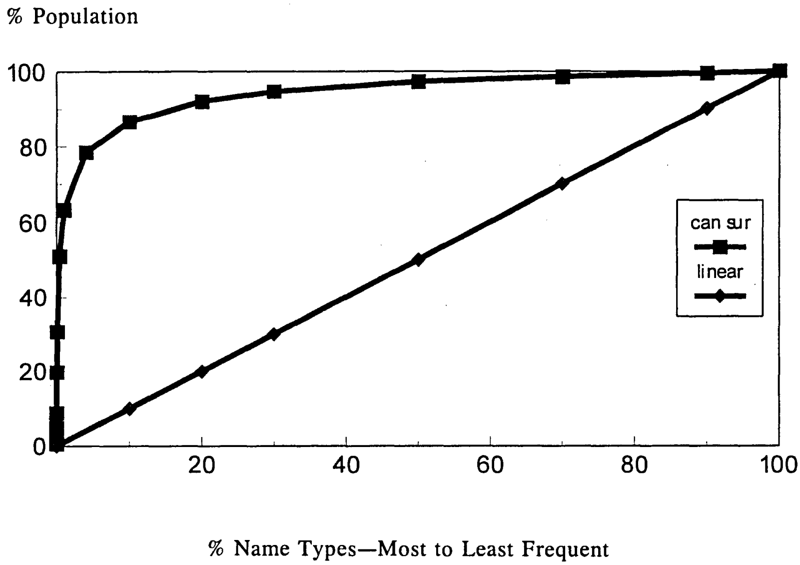
A cursory look at practically any telephone directory would convince most observers that there are many people with popular names and many more with rare names, but we need to be more specific than this. A dictionary publisher might ask, in order to optimize the market: "What is the smallest number of name types required to cover 75% of the population's names in the dictionary?" Since the names and their frequencies are available, this can be calculated. We can order the names, say surnames, in the order of descending frequency—where frequency simply simply means the number of occurrences, or count, of a particular name. For example, the surname *Smith* is the most common surname in Canada (and also in the U.S.), with a count of 61,854.

We know that the sum of the counts for all name types must equal the total name tokens, which is the population, so we know for each name type what portion of the population it covers. Our sample population of Canadian surname tokens is 11 million, so *Smith* represents 0.56% of the population. (In the U.S. *Smith* is almost 1%.) However, it is only one name type among 520k name types, or 0.00019% of all name types. This information—0.56; 0.00019 provides the point of origin for a graph of surname distribution. We can now add the second most popular surname, *Brown*, with a count of 35,316, to the list. The cumulative effect of adding *Brown* to *Smith* is to generate a point at 0.88; 0.00038. The third most popular surname, *Tremblay,* with a count of 34,787, when added to *Smith* and *Brown,* generates a point at 1.2; 0.00058. We can continue in this manner until we have added the count of all the name types in descending frequency order until we arrive at the final point, 100; 100, where all the name types (100%) represent all the name tokens, or the population (100%).

If we plotted the results on a normal graph with linear scales for population and names, we would get a graph shown in figure 1. The graph starts near the 0; 0 point, rapidly rises to about 85; 10 and then slowly rises to 100; 100. It is difficult to see a pattern in this distribution since nearly all the activity seems to take place for low values of name types. The distribution is thus said to be skewed. A non-skewed, linear distribution with, say, a population of 10 million and 1 million name types each with 10 tokens, is also shown in figure 1 and will be useful later for comparison.

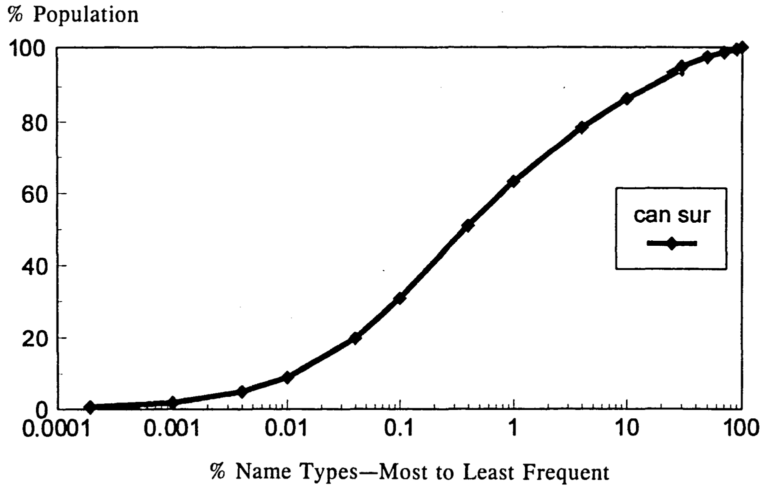
A linear distribution is one where all names have the same frequency so that x% of the name types represents x% of the population. This arrangement is one of most unorderedness—or high entropy—to be contrasted with the other theoretical extreme where the population all has the same name, a condition of low entropy. Entropy in systems—and naming people is a system—is a measure of their ability to accommodate change. These are concepts that require further study, especially in respect to links between names and other social and physical phenomena such as genetics. Suffice to say at this point that the linear curve shows the high entropy condition to which naming curves may be compared.

Figure 1. Distribution of Canadian Surnames—Linear Plot.



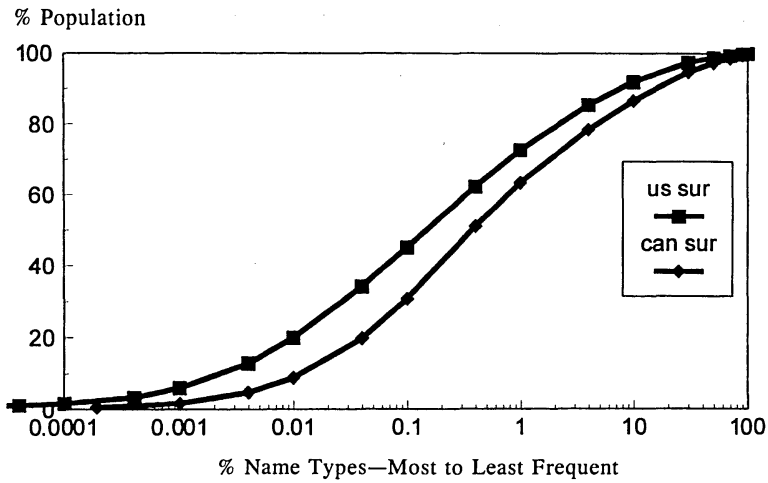
The graph shown in figure 1 has linear scales for both its axes; it is therefore *lin-lin*. A linear scale is one where the increment is constant, such as 10. Starting at, say, 0, the scale goes to 10, 20 and so on up to, say, 100. A logarithmic scale is one where the increment is the power of a base number. Consider a base number of 10. We might start at 10^-4, which is 0.0001, and increase by 10 times each increment: to10^-3, which is 0.001, and so on up to 10^2, which is 100. If we use such a scale for the x-axis (% of name types) we get the plot shown in figure 2, where the graph, a *log-lin*, or *semi-log* plot, allows us to see much more detail.

Figure 2. Distribution of Canadian Surnames—Semi-Log Plot.



We can see, for example, that the most popular 1% of name types accommodate more than 60% of the Canadian surnames and that 90% of the name types, from 10% to 100%, the rare name types, accommodate a mere 12% of the surnames. The answer to the dictionary publisher's question posed earlier is that the top 2.84% of surname types, 14,774 surnames, cover 75% of the population. The distribution of surnames in Canada is thus seen to be highly skewed, although a little less so than that of the U.S., as can be seen in figure 3.

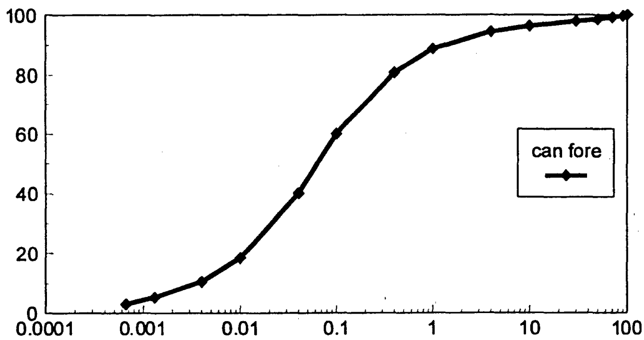
Figure 3. Distribution of Surnames in Canada and the U.S.



It should be noted that although Canada has only one tenth the population of the U.S., it is possible to plot both distributions on the same curve since using percentages has made the distributions comparable. (If we plotted a similar curve for the UK, it would lie somewhere between those of Canada and the U.S. This general similarity suggests that the shape of the curve is intrinsic to at least some linguistic or cultural surname distributions.)

We can treat forenames similarly, as shown in figure 4.

Figure 4. Distribution of Canadian Forenames.

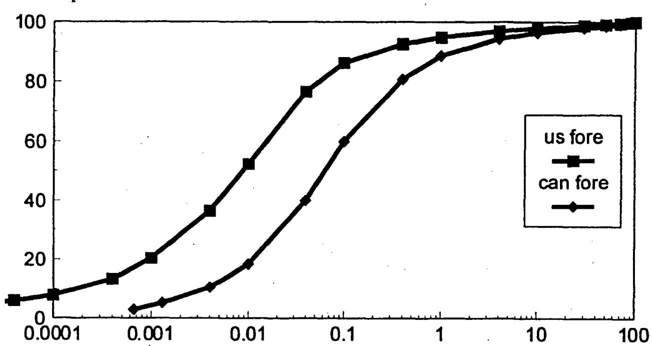


% Name Types—Most to Least Frequent

This curve rises faster than the surname curve, is more skewed, and shows that the most popular 1% of forename types accommodates about 89% of the population. This means that 99% of forename types are found in only 11% of the population. For comparison, the U.S. curve, which is more skewed than the Canadian curve is shown in figure 5.

Figure 5. Forename Distribution in Canada and the U.S.

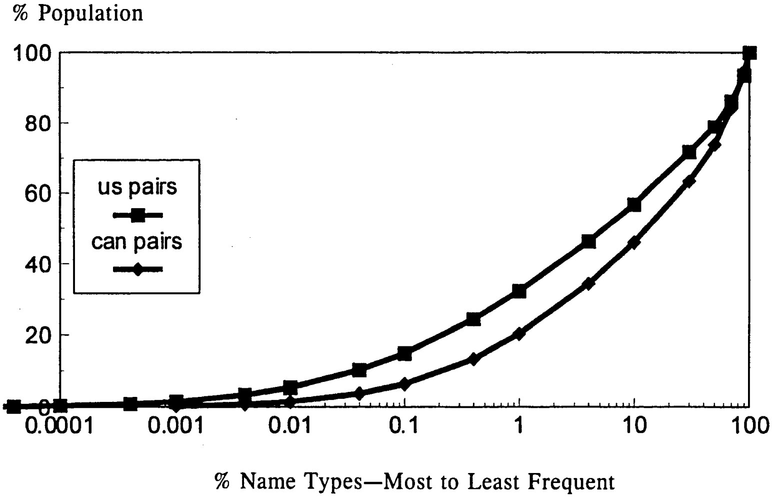
% Population



% Name Types—Most to Least Frequent

The next plot in the series is of types of forename-surname pairs for both Canada and the U.S.; these are shown in figure 6.

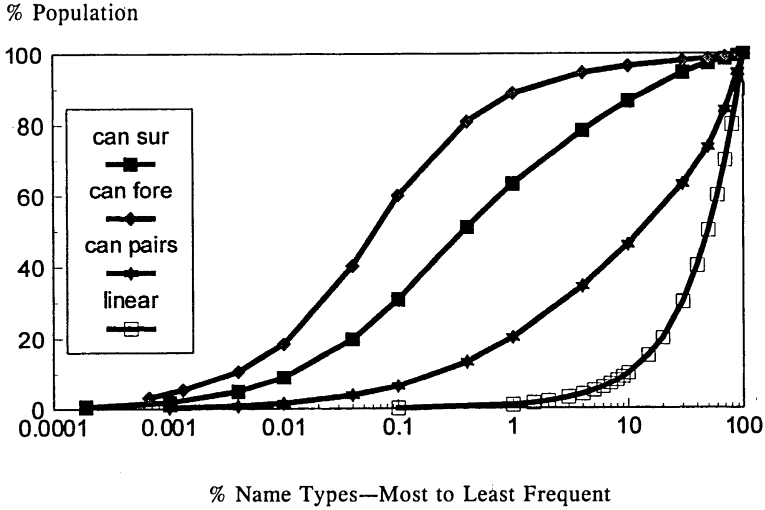
Figure 6. Distribution of Forename-Surname Pairs in Canada and the U.S.



The forename-surnames pairs curves are less skewed than those for the surname distribution, but still skewed, and again the Canadian distribution is a little less skewed than the U.S. distribution. One percent of the forename-surname pair types accommodates about 20% of the population in Canada, while in the U.S. it accommodates nearly a third.

The final plot, figure 7, shows Canadian forename, surname, and forename-surname pair distribution, and, for comparison, the linear, totally unskewed distribution. We know that the population has about three times as many surnames to choose from as forenames: *Smith*, the most common surname, has a count of 61,854 while *John*, the most common forename, has a count of 162,690, notwithstanding the fact that only 50% of forenames are disclosed in the telephone listings. Clearly, forenaming, as evidenced in telephone listings, is more highly ordered than surnaming. However, this evidence lags the current trends because few younger people get their names in the telephone lists. Schwegel, for instance, shows that many are working hard to expand the number of forenames. From these distributions we can conclude that the forename curve appears as the most ordered curve followed by the surname curve with the forename-surname pairs curve nearest to the completely unskewed distribution.

Figure 7. Canadian Forename, Surname, Forename-Surname Pairs, and Unskewed Distribution.



Non-Cumulative Representations—Types Against Tokens Per Type

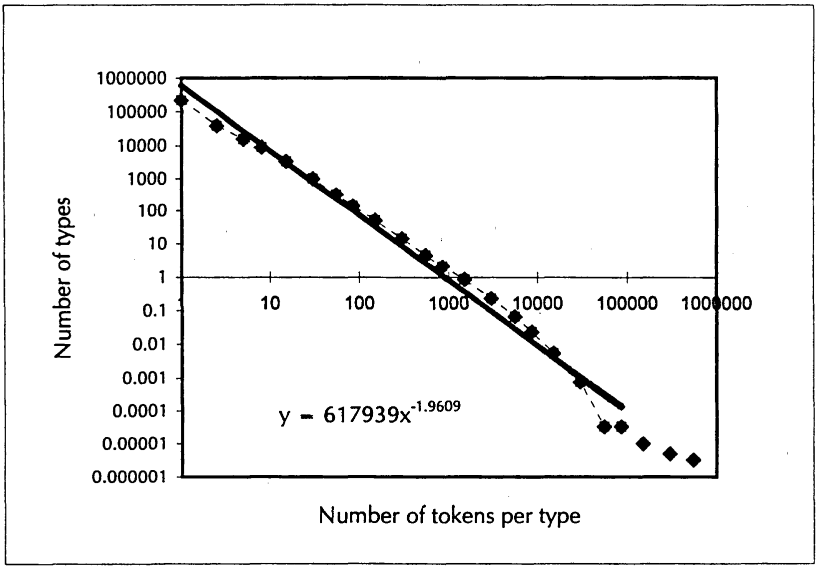
The cumulative curves are useful in describing the distributions in everyday terms, but other researchers (e.g., Ogden 1998) have attempted to identify how many name types there are for a given count, or frequency; in other words, the number of types which occur with a given number of tokens, the object being to determine the relationship (if any) between types and tokens.

We have defined frequency as count—the number of tokens of a particular name type. Our surname data ranges from a frequency of 1, which is the count for each unique name, to 61,854, which is the count for *Smith*. As we have seen with unique types, it is not uncommon for a number of types, especially those at low frequencies, to have the same count. There are 223,929 surname types with a frequency of 1 (2% of the population), 74,895 with a frequency of 2, and 43,223 with a frequency of 3. Ranked by increasing frequency, the type count is generally descending, but there are exceptions. The first occurs at a frequency of 39 which is shared by 576 types, but slightly more, 578 types, share a frequency of 40.

Additionally, some frequencies contain no name types. The general decline in the number of types sharing a frequency continues with increase in frequency until the number of types sharing a frequency reaches zero. The first empty frequency we find is at 497; there are no surname types with 497 tokens. The gaps get bigger with increase in frequency; there are 26,538 empty frequencies between 35,316 (Brown) and 61,854 (Smith). There are thus two related events as we increase frequency: reduction in the number of types at a given frequency, and increase in the number of empty frequencies, hence more and bigger gaps.

We are now in a position to plot name types against tokens per type. This is best plotted on a *log-log* scale which results in the curve shown in figure 8.

Figure 8. Canadian Surname Distribution.



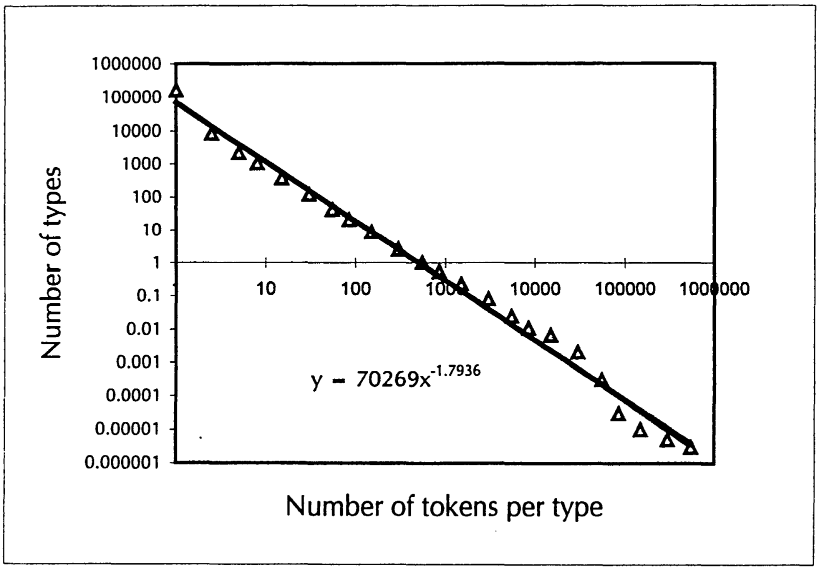
The data approximates to a power law curve with a best fit of Y=617939*x^(-1.961).

However, with surnames, forenames, and forename-surname pairs, the unique frequencies tend to be overstated because that is where the typos and other detritus settle. Nothing short of eyeballing the data for non-names and having knowledge of all legitimate forms is necessary to separate the names from the non-names. Unfortunately, this knowledge is not available currently and the problem of determining whether or not a particular sequence of characters is a name, while anything but a trivial task, is often insoluble.[3, 4]

Population can be derived from the non-cumulative curves. The population at a particular frequency is P=y*x, where y is the frequency and x is the population count at that frequency. Taking the surnames as an example, P=617939*x^(-1.961)*x, which approximates to P= 617939/x, where x has the range from 1 (*Smith*) to 223,929 (the unique surnames). This is a standard inverse curve, which is a reasonable fit, although it gives a value of just over 3 instead of 1 for x=223,929 and overstates the value for x=1 as 617,939 instead of 223,929. To get the population over a range of x it is necessary to integrate y.dx over the desired range.

The curve for forenames, plotted in the same way as for surnames, is also a power law curve, and is shown in figure 9.

Figure 9. Canadian Forename Distribution.



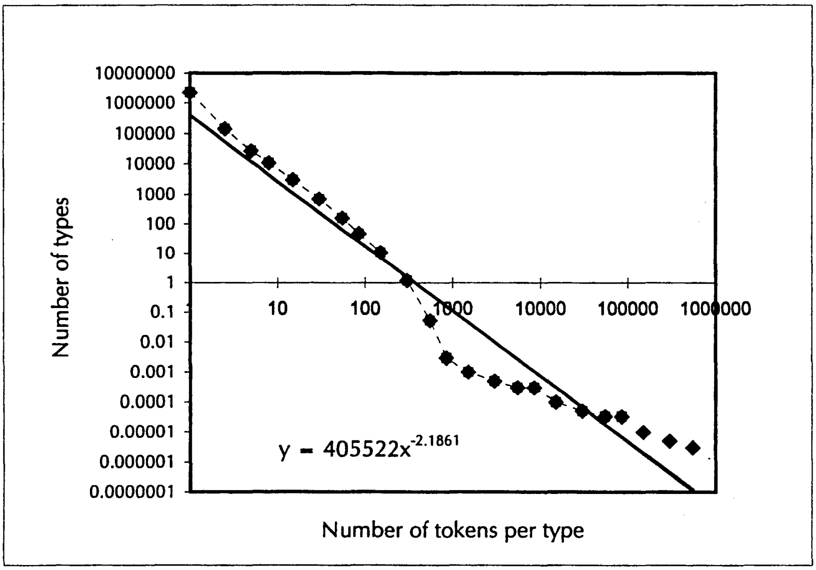
The best fit for the complete series is a simple power law relationship: Y=70,269*x^(-1.794), which is very similar to the U.S. forenames relationship: Y=339,550*x^(-1.734).

This underestimates the number of unique forenames as 70,269, whereas the actual sample number is 102,596. However, the sample number itself is overstated since this is where the flotsam and jetsam gravitates, mainly typographical errors.

The curve for forename-surname pairs plotted in the same way and shown in figure 10, is also, as for forenames, a power law curve: Y= 405,522*x^(-2.186). It is very similar to the U.S. relationship, which is Y = 25,783.821*x^(-2.380).

The Canadian curve underestimates the number of unique forename-surname pairs at 405,522, whereas the actual number is 2.87 million, but the estimate is in the same general range.

Figure 10. Canadian Forename-Surname Pairs.

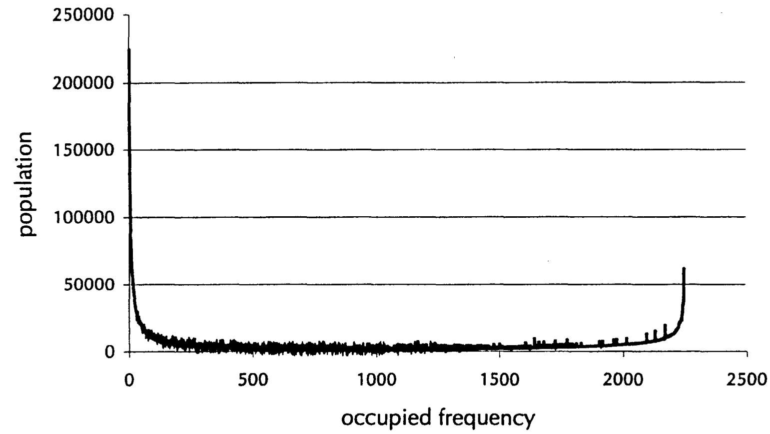


Population Curves for Occupied Frequencies.

In the cumulative curves shown above, population was plotted against name types. In the non-cumulative curves we plotted number of name types against frequency, or tokens per type. In this section we will consider another way to calculate population, by looking at the occupied frequencies only.

For Canadian surnames the frequencies range from 1, the unique names, to 61,854, the highest frequency surname, *Smith*. There are thus 61,854 available frequencies, but only 2,244 are occupied: about 3.6%, which is higher than the U.S. figure of less than 1%. The sum of the product of frequency times number at that frequency gives the population. These are generally U-shaped curves with the majority of the population at the uprights. The resulting plot is shown in figure 11.

Figure 11. Canadian Surnames by Occupied Frequency.



Going back to our surname data we know that there are about 223, 929 surname types that are unique; each contains only a single token. Since there are only 0.52 million surname types to begin with, about 43% of types are unique, which is similar to the U.S., where about 40% of all surname types are unique. Readers may care to refer to figure 3 to see the value of the Percentage of Population for 60% (100%-40%) of the surname types. It is difficult to read but may be easily calculated. Since the names are unique, the number of tokens (population) equals the number of types, which is 223,929. The population point as a percentage is thus (11M-223,929)*100/11M or 98%.

It is thus surprising but no less true that in Canada it is four times rarer to have the most common surname, *Smith*, than to have a unique surname; in the U.S. the ratio is slightly in favor of the unique surnames. Thus the almost matched uprights of the U in the U.S. case give way to the higher leading upright in the Canadian case.

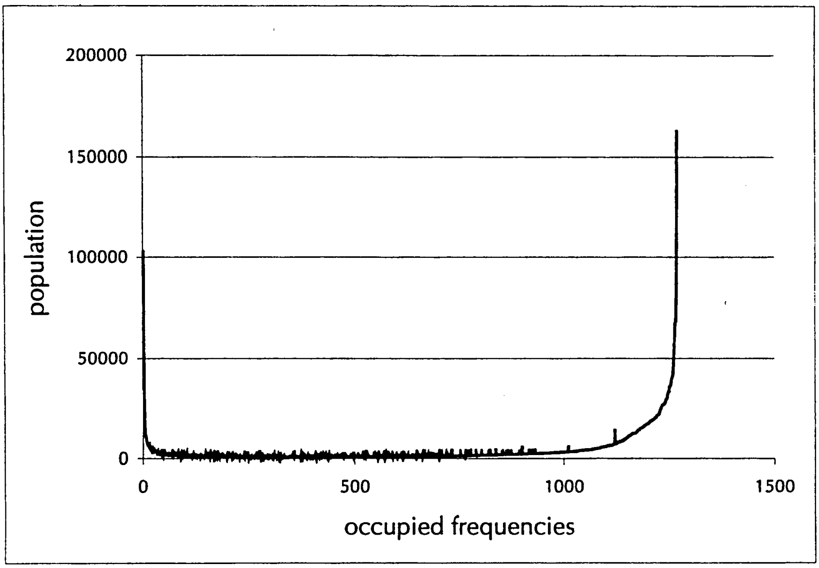
The origin and finish are not easy to see but they are the population value of 223,929 for the first occupied frequency, and 61,854 for the last at 2,244 (figure 11). The curve descends from the origin with an overall reduction in the number of types. However, the number of types vibrates about this downward trend, which gives the first part of the curve its fuzziness. This is to be contrasted with the smoothness of the finish of the curve where there is only one type.

The curve reaches a minimum at a frequency of 457, the first that has just one type. This is the frequency that represents minimum population. (The surname, incidentally, is *Bonnell.*) The line of *one types* can be followed from there along the bottom line of the curve until reaching the end. Descending from the end we can see the last of the *two types* just above at 2169 on the x axis. Moving left along the x axis we can see the density of the *two types* grow and see the first *three types* emerge at 1,773, until further increases in types are lost in the detail.

A similar plot can be made for forenames. There are 102,596 forename types that are unique. Since there are only 150,952 forename types to begin with this means that about 68% of types are unique, quite similar to the U.S. distribution where about 70% of all forename types are unique. In Canada it is rarer to have a unique forename than to be called *John* or *Robert*, the most popular and second most popular forenames in Canada, but it is rarer to be called *David*, the third most popular, than to have a unique forename. In the U.S. the break is after the seventh most popular forename.

The forename plot is shown in figure 12, which follows the same overall shape as figure 11 except in this case the trailing upright of the U curve is dominant. The minimum is at frequency 180, *Elmo.*

Figure 12. Canadian Forenames by Occupied Frequency.



Popular Name Types

Surnames

Table 3 lists in descending frequency order the 100 most popular surnames of the 520k surname types in Canada. The count given is out of the sample population of 11 million tokens. I have divided the names into two major language groups and have encoded the entries, using as reference *A Dictionary of Surnames* (Hanks and Hodges 1988), as follows:

A etymologically not French
F etymologically French
B etymologically both French and not French

This procedure gives us an indication of the relative strength of the French surnames compared with the relative strength of the Francophone population in Canada (6.6 million or 23%, according to the 1996 census). Four of the 10 most common surnames are etymologically French. *Tremblay*, the most common French surname, is third on the list, just behind *Brown*. One name in the top 10, *Martin*, is both; the others are not French. If we count the names which are both French and not French as half French and half not French, then 36.5% of the most common 100 surnames are etymologically French. If we measure population of the 100 top surnames we find that 35.8% have etymologically French surnames.

Of course not every person with a French name is a Francophone, or vise versa. Pierre-Marc Johnson, the former Premier of Quebec, is a Francophone with an etymologically English surname. If we assume, as a rough guide, that we can take the number of people with etymologically French surnames as an indicator of the number of Francophones, then it would seem that there are relatively fewer surname types used in the Francophone community than outside it, giving high token counts for those French surnames. On the other hand, it could be that a significant proportion of the bearers of French surnames are not Francophones. Further and more detailed coding of Canadian surnames would be illuminating.

In table 3, the rank is followed by the surname, the frequency (count) and the language designation.

Table 3. The 100 Most Popular Surnames in Canada.

1	Smith	61854	A	51	Richard	11917	B
2	Brown	35316	A	52	Girard	11879	F
3	Tremblay	34787	F	53	Murray	11856	A
4	Martin	29727	B	54	Davis	11578	A
5	Roy	27472	F	55	Simard	11402	F
6	Wilson	27231	A	56	Graham	11323	A
7	Gagnon	26535	F	57	Clarke	11242	A
8	Johnson	25193	A	58	Beaulieu	11163	F
9	Campbell	23024	A	59	Fraser	11061	A
10	Cote	22989	F	60	Jackson	11013	A
11	Taylor	22970	A	61	Kelly	10990	A
12	Macdonald	22869	A	62	Caron	10692	F
13	Anderson	22188	A	63	Mcdonald	10656	A
14	Jones	21334	A	64	Baker	10602	A
15	Lee	21327	A	65	Hall	10581	A
16	Leblanc	21233	F	66	Fournier	10563	F
17	Miller	20549	A	67	Harris	10449	A
18	Thompson	20282	A	68	Wood	10414	A
19	Williams	20127	A	69	Hill	10409	A
20	Gauthier	20055	F	70	Lewis	10392	A
21	White	18223	A	71	Bell	10194	A
22	Bouchard	17194	F	72	Robertson	10157	A
23	Young	17059	A	73	Lefebvre	10154	F
24	Scott	16718	A	74	Lapointe	10011	F
25	Morin	16669	F	75	Roberts	9877	A
26	Stewart	16129	A	76	Watson	9874	A
27	Wong	16086	A	77	Green	9874	A
28	Belanger	15587	F	78	Ouellet	9860	F
29	Pelletier	15450	F	79	Dube	9701	F
30	Lavoie	15329	F	80	Kennedy	9659	A
31	Levesque	15038	F	81	Allen	9615	A
32	Moore	14976	A	82	Cloutier	9565	F
33	Robinson	14853	A	83	Hebert	9516	F
34	Gagne	13961	F	84	Desjardins	9447	F
35	Johnston	13840	A	85	Hamilton	9226	A
36	Clark	13814	A	86	Cameron	9148	A
37	Reid	13707	A	87	Armstrong	8994	A
38	Fortin	13501	F	88	Evans	8993	A

39	Ross	13483	A	89	Adams	8990	A
40	Walker	13223	A	90	Morrison	8959	A
41	Bergeron	12800	F	91	Martel	8847	F
42	Boucher	12656	F	92	Michaud	8838	F
43	Chan	12596	A	93	Grant	8711	A
44	Poirier	12435	F	94	Bedard	8677	F
45	King	12387	A	95	Phillips	8634	A
46	Murphy	12252	A	96	Cook	8604	A
47	Landry	12229	F	97	Ferguson	8560	A
48	Wright	12178	A	98	Turner	8374	A
49	Mitchell	12119	A	99	Cormier	8272	F
50	Thomas	11982	A	100	Parent	8270	F

Forenames

Table 4 lists in descending frequency order the 100 most popular forenames of the 151k types in Canada. The count given is out of the sample population of 5.5 million tokens.

The forenames listed, it must be pointed out, are self-declared. To some people a number of the forenames may appear to be contractions, diminutives, nicknames, or even non-names, but this is the way the subscribers have listed themselves. I have made no attempt to "correct" the name form with the exception of expanding standard abbreviations such as *Ewd* to *Edward, Robt* to *Robert* and *Wm* to *William.*

Where the forename is made of multiple segments, all segments are included in the name, even when there is no hyphen. It can be argued that in some cases the person is merely listing their forenames such as in *John Robert* or *Jean Pierre* (as opposed to *Jean-Pierre*); perhaps so, perhaps not.

There is no gender information in the data used here. I am aware of the dangers of discussing "male" and "female" name lists, especially as there is considerable evidence (e.g., from Schwegel 1997) that girls are being given forenames that were previously considered to be exclusively male, such as *John, Robert, William, James,* and *David.*[3] However, on the assumption that the vast majority of usage of these names is still for males, I include these names as male. I also consider *unisex* forenames. By *unisex* I mean names that are currently recognized as being given more or less regularly to both males and females. I admit the fuzziness of this definition. Unisex names include *Leslie*, *Bunny,* and *Dominique.* They also include forenames like *Jean* and *Carol* that are normally used for males in the Francophone community but for females outside it.

Table 4. The 100 Most Popular Forenames in Canada.

1	John	162690	51	Alain	20189
2	Robert	129436	52	Albert	20187
3	David	87004	53	Maurice	20168
4	James	68252	54	Rene	19884
5	William	68026	55	Doug	19433
6	Paul	67275	56	Joe	19292
7	Richard	61878	57	Dave	19290
8	George	56768	58	Gerard	19079
9	Peter	52895	59	Walter	19020
10	Michael	47510	60	Terry	18758
11	Michel	43082	61	Harold	18596
12	Andre	41294	62	Marc	18518
13	Brian	40592	63	Tom	18482
14	Donald	39777	64	Allan	18244
15	Jean	39132	65	Mario	18077
16	Pierre	37537	66	Martin	17889
17	Claude	36458	67	Arthur	17560
18	Frank	36300	68	Stephen	17397
19	Roger	35601	69	Chris	17232
20	Daniel	35528	70	Roy	17100
21	Jacques	33180	71	Kenneth	17032
22	Gordon	33007	72	Dennis	17019
23	Ken	32739	73	Bernard	16884
24	Raymond	31674	74	Patrick	16775
25	Denis	30462	75	Louis	16656
26	Gilles	29632	76	Keith	16311
27	Don	29290	77	Dan	16257
28	Joseph	29032	78	Norman	16255
29	Mike	28081	79	Leo	15790
30	Gary	28069	80	Serge	15705
31	Jim	27341	81	Henry	15676
32	Ron	27291	82	Yvon	15431
33	Wayne	26910	83	Barry	15117
34	Marcel	26888	84	Yves	15091
35	Gerald	26854	85	Harry	14996
36	Charles	26619	86	Bob	14991
37	Guy	26526	87	Ray	14852
38	Ronald	25823	88	Andrew	14827
39	Mark	25001	89	Rick	14408
40	Fred	24965	90	Tony	14079
41	Bruce	24194	91	Scott	13910
42	Edward	23408	92	Roland	13893
43	Jack	23209	93	Francois	13810
44	Bill	21924	94	Greg	13794
45	Douglas	21779	95	Luc	13355
46	Steve	21594	96	Alan	13015
47	Eric	21551	97	Lloyd	12808
48	Kevin	21424	98	Alex	12562
49	Thomas	21274	99	Dale	12468
50	Larry	21114	100	Ralph	12352

Table 5. The 100 Most Popular Female Forenames in Canada.

1	Mary	8605	51	Wendy	2225
2	Linda	6134	52	Betty	2198
3	Diane	5202	53	Jennifer	2163
4	Karen	4535	54	Janet	2128
5	Nancy	4399	55	Claire	2117
6	Marie	4392	56	Cecile	2055
7	Margaret	4342	57	Cindy	2055
8	Louise	4254	58	Anna	2044
9	Susan	4230	59	Chantal	2039
10	Denise	3912	60	Carole	2030
11	Helen	3845	61	Carmen	1993
12	Nicole	3571	62	Laura	1986
13	Anne	3551	63	Kerry	1945
14	Donna	3531	64	Cathy	1933
15	Lisa	3384	65	Joyce	1920
16	Lise	3209	66	Catherine	1867
17	Sandra	3165	67	Yvonne	1864
18	Irene	3135	68	Tracy	1863
19	Sylvie	3123	69	Lorraine	1856
20	Brenda	3063	70	Jacqueline	1839
21	Therese	3045	71	Francine	1836
22	Julie	3016	72	Jeanne	1772
23	Rita	2988	73	Cheryl	1746
24	Elizabeth	2933	74	Florence	1730
25	Shirley	2906	75	Isabelle	1725
26	Sharon	2903	76	Lucie	1707
27	Joan	2865	77	Lori	1704
28	Patricia	2797	78	Josee	1693
29	Michelle	2729	79	Tammy	1683
30	Debbie	2708	80	Manon	1658
31	Barbara	2655	81	Angela	1653
32	Lynn	2646	82	Marion	1648
33	Suzanne	2557	83	Jane	1644
34	Doris	2543	84	Ginette	1636
35	Christine	2537	85	Elaine	1612
36	Dorothy	2530	86	Gail	1583
37	Judy	2508	87	Eva	1581
38	Heather	2475	88	Darlene	1581
39	Annie	2470	89	Marilyn	1547
40	Alice	2465	90	Bev	1540
41	Rose	2452	91	Edith	1539
42	Ann	2428	92	Gisele	1536
43	Pauline	2370	93	Evelyn	1515
44	Nathalie	2364	94	Sheila	1511
45	Helene	2354	95	Michele	1508
46	Ruth	2354	96	Frances	1489
47	Maria	2351	97	Danielle	1485
48	Kathy	2345	98	Barb	1485
49	Joanne	2307	99	Bonnie	1414
50	Monique	2296	100	Janice	1401

We see few unisex forenames and no exclusively female forenames in table 4. The low count for forenames of females is a function of the source. Women listed with men are often in the form of *Mr & Mrs John Smith* or sometimes not listed in the household entry. Furthermore, some women tend not to use their forenames in phone listings for security reasons—especially solo women—so the lower counts are to be expected.

Table 5 shows the 100 more frequent forenames used by females. The table was obtained by considering the ordered list of all forenames and selecting the female forenames. I have attempted to exclude unisex names as no one seems to know them all and if they were to be included in the female group they would give unrepresentational numbers. I fully realize the dangers in using such a procedure, but this seems to me to be the best choice available at this time.

This table shows not only the relative popularity of the various female forenames, but also the low counts compared to the male forenames, and thus the gross under-representation of females in the source data. *Mary* is the most popular name and has a count of 8,605, which is 5% of that of the most popular male forename, *John*; *Marie*, at about half the count of *Mary*, is 6th most popular, and *Maria,* at about half the count of *Mary*, is 47th. In the U.S., *Mary* was also the most popular female forename, with a count of 451,437, which is 20% of the most popular male forename in the U.S., also *John*.

If we compare the frequency of the most popular forename, *John*, in the U.S. and in Canada, and ignoring the propensity for more people in the U.S. to list their forenames, the ratio is 13.7:1, not too far from the rule of thumb ratio of 10:1. However, when comparing the most popular female forename, *Mary,* we find a ratio of 52.5:1, which suggests again that female forenames are greatly under-represented in the Canadian data.

Forename-Surname Pairs

Table 6 shows the 100 most common forename-surname pairs. There are no forenames used by women in this list for reasons previously discussed. And surprisingly there is only one French name among the 18 most frequently-occurring pairs, *Michel Tremblay*, at number 11. *John Smith,* the combination of the most popular forename with the most popular surname is not number 1, as we might expect, but number 3. I have repeated the coding used in table 4 for convenience; the coding applies to the pair, and to the surname, but not to the forename. For example, there is no suggestion that *Robert* is etymologically not French but that the combination *Robert Smith* is not French.

Table 6. The 100 Most Popular Forename-Surname Pairs in Canada.

1	Robert	Smith	985	A
2	David	Smith	831	A
3	John	Smith	737	A
4	John	Macdonald	715	A
5	Robert	Brown	677	A
6	James	Smith	671	A
7	Donald	Smith	564	A
8	John	Wilson	561	A
9	Robert	Wilson	535	A
10	John	Campbell	532	A
11	Michel	Tremblay	531	F
12	David	Brown	528	A
13	Robert	Taylor	473	A
14	William	Smith	473	A
15	George	Smith	458	A
16	John	Taylor	445	A
17	James	Brown	444	A
18	John	Brown	441	A
19	Jean	Roy	437	F
20	Brian	Smith	432	A
21	Jacques	Tremblay	427	F
22	Gordon	Smith	417	A
23	John	Martin	416	A
24	Andre	Roy	411	F
25	Robert	Martin	411	B
26	Michel	Gagnon	411	F
27	Richard	Smith	404	A
28	Pierre	Tremblay	403	F
29	Michel	Cote	396	F
30	Robert	Jones	395	A
31	Andre	Tremblay	394	F
32	John	Anderson	393	A
33	Claude	Tremblay	389	F
34	Jean	Tremblay	388	F
35	Paul	Smith	382	A
36	Andre	Gagnon	382	F
37	Gilles	Tremblay	381	F
38	Michel	Roy	375	F
39	Michael	Smith	373	A
40	John	Thompson	370	A
41	John	Miller	369	A
42	David	Wilson	368	A
43	Denis	Tremblay	368	F
44	John	Stewart	368	A
45	John	Scott	367	A
46	Robert	Young	365	A
47	Robert	Anderson	362	A
48	Andre	Cote	358	F
49	Robert	Campbell	353	A

50	Robert	Thompson	351	A
51	John	Williams	348	A
52	David	Jones	345	A
53	Robert	Johnson	344	A
54	Peter	Smith	342	A
55	James	Wilson	341	A
56	Douglas	Smith	340	A
57	William	Brown	339	A
8	Pierre	Gagnon	335	F
59	John	Moore	333	A
60	Robert	Miller	332	A
61	Ronald	Smith	329	A
62	Robert	Scott	328	A
63	Jean	Gagnon	328	F
64	Robert	White	327	A
65	John	White	327	A
66	John	Walker	324	A
67	John	Murphy	322	A
68	Robert	Macdonald	317	A
69	John	Young	316	A
70	Denis	Roy	315	F
71	David	Williams	314	A
72	Pierre	Cote	313	F
73	Alain	Tremblay	311	F
74	Wayne	Smith	310	A
75	Guy	Tremblay	309	F
76	Jacques	Gagnon	308	F
77	Gilles	Gagnon	307	F
78	Jacques	Cote	307	F
79	Marcel	Tremblay	306	F
80	Robert	Tremblay	305	F
81	James	Macdonald	305	A
82	Claude	Gagnon	305	F
83	Gary	Smith	303	A
84	Robert	Reid	296	A
85	John	Reid	293	A
86	Marc	Tremblay	292	F
87	Michel	Gauthier	292	F
88	Richard	Tremblay	290	F
89	Daniel	Tremblay	290	F
90	Robert	Stewart	289	A
91	Robert	Roy	289	B
92	George	Brown	287	A
93	James	Stewart	286	A
94	Robert	Gagnon	286	F
95	Roger	Roy	285	B
96	John	Clark	285	A
97	David	Johnson	284	A
98	Claude	Roy	282	F
99	John	Ross	282	A
100	Denis	Gagnon	281	F

Table 7. The 100 Most Popular Female Forename-Surname Pairs in Canada.

1	Mary	Macdonald	81	A
2	Mary	Smith	74	A
3	Linda	Smith	52	A
4	Susan	Smith	52	A
5	Lise	Roy	52	F
6	Karen	Smith	47	A
7	Sylvie	Tremblay	45	F
8	Marie	Tremblay	45	F
9	Denise	Tremblay	45	F
10	Nicole	Leblanc	44	F
11	Therese	Tremblay	43	F
12	Manon	Tremblay	41	F
13	Nancy	Roy	41	F
14	Margaret	Smith	40	A
15	Donna	Smith	40	A
16	Diane	Tremblay	39	F
17	Denise	Leblanc	39	F
18	Debbie	Smith	38	A
19	Lise	Tremblay	37	F
20	Louise	Tremblay	37	F
21	Helen	Smith	37	A
22	Sylvie	Roy	37	F
23	Louise	Gagnon	37	F
24	Dorothy	Smith	36	A
25	Nathalie	Roy	36	F
26	Diane	Roy	36	F
27	Sylvie	Levesque	36	F
28	Diane	Leblanc	36	F
29	Mary	Campbell	36	A
30	Nathalie	Tremblay	35	F
31	Louise	Leblanc	35	F
32	Lise	Gagnon	35	F
33	Therese	Gagnon	35	F
34	Denise	Roy	34	F
35	Sylvie	Gagnon	34	F
36	Ginette	Tremblay	33	F
37	Helene	Tremblay	33	F
38	Louise	Roy	33	F
39	Linda	Martin	33	A
40	Margaret	Macdonald	33	A
41	Nathalie	Cote	33	F
42	Josee	Tremblay	32	F
43	Lise	Cote	32	F
44	Isabelle	Tremblay	31	F
45	Cecile	Tremblay	31	F
46	Therese	Roy	31	F
47	Denise	Cote	31	F
48	Lise	Bouchard	31	F
49	Shirley	Smith	30	A

50	Marie	Leblanc	30	F
51	Rita	Leblanc	30	F
52	Lise	Gauthier	30	F
53	Helene	Gagnon	30	F
54	Mary	White	29	A
55	Joan	Smith	29	A
56	Patricia	Smith	29	A
57	Judy	Smith	29	A
58	Diane	Levesque	29	F
59	Monique	Levesque	29	F
60	Lise	Gagne	29	F
61	Therese	Cote	29	F
62	Nicole	Tremblay	28	F
63	Jeannine	Tremblay	28	F
64	Sandra	Smith	28	A
65	Joyce	Smith	28	A
66	Heather	Smith	28	A
67	Sharon	Smith	28	A
68	Lynn	Smith	28	A
69	Kathy	Smith	28	A
70	Ruth	Smith	28	A
71	Josee	Roy	28	F
72	Sylvie	Pelletier	28	F
73	Mary	Murphy	28	A
74	Yvonne	Leblanc	28	F
75	Sylvie	Lavoie	28	F
76	Mary	Johnson	28	A
77	Diane	Gagnon	28	F
78	Mary	Brown	28	A
79	Helene	Bouchard	28	F
80	Diane	Bouchard	28	F
81	Julie	Tremblay	27	F
82	Barbara	Smith	27	A
83	Nicole	Roy	27	F
84	Mary	Martin	27	A
85	Nicole	Gagnon	27	F
86	Marie	Gagnon	27	F
87	Louise	Cote	27	F
88	Suzanne	Cote	27	F
89	Susan	Brown	27	A
90	Louise	Bouchard	27	F
91	Mary	Young	26	A
92	Nancy	Tremblay	26	F
93	Lisa	Smith	26	A
94	Nicole	Pelletier	26	F
95	Diane	Morin	26	F
96	Joanne	Leblanc	26	F
97	Nathalie	Gagnon	26	F
98	Isabelle	Gagnon	26	F
99	Sylvie	Cote	26	F
100	Diane	Cote	26	F

Table 7 has been extracted in sequence to include only forenames for women in the forename-surname pairs. The frequencies shown in table 7 are quite different from those found in table 6. Here, there are 12 Francophone names among the 20 most frequent name pairs, with *Lise Roy* the most common at number 5, showing the greater relative strength of the Francophone representation of female names compared to male names within the most frequent forename-surname pairs. The distribution suggests that there are comparatively fewer French female forenames in the popular group than there are in the French male forenames group and in the Non-French females forenames group, or, alternatively, female Francophones may be more likely to list their forenames than are non-Francophones. Having said that, I must point out that the counts for the female pairs are only a tenth that of the male pairs because of the underepresentation of females in the source data. We need for this reason, and others, a better source of name data for Canada.

Conclusion

Two graphic methods of representing the forename, surname, and forename-surname pairs data culled from the Canadian telephone directory have been demonstrated. The cumulative curve method allows immediate apprehension of the severe skew of the distributions of forenames, surnames, and forename-surname pairs, particularly that of forenames. One can read from the forename curve (figure 4) that the most frequent 0.25% of forenames represent 75% of the population.

The non-cumulative or frequency method allows the derivation of algebraic expressions, basically power law expressions, for the various name classes. What now needs to be done is to explain why these curves are the shape they are and what the parameters in the algebra mean in the world of names.

From the algebraic expressions we can calculate the sample population. Population can also be drawn directly from the occupied frequencies. These population curves have maxima at the high and low ends of both the forename and the surname distributions and lead to the paradox that it is rarer to be called by the most popular surname (in this instance, *Smith*) than it is to have a unique surname.

The lists of popular surname and forename types brought out the under-representation of women in the source data and the strong showing of French forename-surname pairs. While telephone directories have the appeal of immediacy, further study of the personal names of Canada requires access to data that is currently outside the public domain. Personal name research is in the public interest, from genealogy to genetics and beyond. Extracts from the public records could be made available to serious researchers with no degradation in the privacy of the people involved. This is an issue for the *Canadian Society for the Study of Names* and other interested parties to champion.

Notes

1. The source data treats a string after a space as a new name and generally assumes that within a given sequence of names the first will be the surname and the remainder will be the forename(s). With a name string like *Kets De Vrie Manfred,* it assumes the surname is *Kets* and the forenames are *De Vrie Manfred.* This has to be repaired to surname *Kets De Vrie* and forename *Manfred.* Similarly, *Many Fingers John* is presented as surname *Many* and forenames *Fingers John.* This is repaired to surname *Many Fingers* and forename *John.* The practice of many married couples to use both their surnames also presents a problem. If Bill Smith and Mary Jones decide to use *Smith Jones* as their surname, the string will be *Smith Jones Bill and Mary,* which will be presented as surname *Smith* and forenames *Jones Bill and Mary.* This must be repaired to one entry, *Smith Jones, Bill,* and a second entry, *Smith Jones, Mary.*

2. By way of comparison, in the UK so few people list their forenames in the telephone directory that an analysis of forenames by this method which had been planned had to be abandoned. Canada thus sits in the middle between the U.S., where most people prefer to list their forenames, and the UK where most prefer to list only their initials. Some have argued that the telephone directory officials constrain the UK entries to initials only, but that would mean that no forenames would be listed, which is not the case.

3. Readers who would like to test this assertion for themselves are invited to consider the following: *Eri, Eri'c, Eric, Erica, Ericc, Erics, Erijc, Eriq.* Are these real forenames, typographical errors, intentional alterations, or other?

4. These names are all "in-use" boy's forenames and are listed in *The Baby Name Countdown* (Schwegel 1997). This reference draws name data from 28 U.S. states, 6 Canadian provinces, and one Canadian territory; the names are compiled primarily from 1994 and 1995.

References

Hanks, Patrick, and Kenneth Tucker. 2000. "A Diagnostic Database of American Personal Names." *Names* 48: 59-69.

Hanks, Patrick, and Flavia Hodges. 1988. *A Dictionary of Surnames.* New York: Oxford Univ. Press.

Ogden, Trevor. "How Rare Are Surnames?" 1998. *The Journal of One-Name Studies.* April. Pp. 119-124.

Schwegel, Janet. 1997. *The Baby Name Countdown*, 4th Ed. New York: Marlowe.

Tucker, Kenneth. 2001. "Distribution of Forenames, Surnames and Forename-Surname Pairs in the U.S." *Names* 49: 69-96.