SURNAMES, FORENAMES, AND CORRELATIONS

Some Facts and Figures

D. Kenneth Tucker

Surname Frequencies and Selection of Entries

DAFN is the first surname dictionary for which the selection of entries is based on a study of frequencies of actual surnames as they exist. In the past, content of surname dictionaries has been selected by the author with no systematic account of present-day frequencies or indication of the percentage of the population represented by the content. In his preface to the first edition of *A Dictionary of English Surnames*, for example, P. H. Reaney wrote, that, for space reasons, he had to delete 4,000 surnames from his first draft: "The great majority of those eliminated are local surnames such as Manchester, Wakefield, Essex, etc." We thus see that Reaney selected, or rather deselected, not by coverage but by class. Similarly, in the preface to the second edition of *The Penguin Dictionary of Surnames*, also a dictionary of English surnames, Basil Cottle wrote about the increase in surnames from 12,000 in the first edition to 16,000 in the second edition: "This new word-hoard was largely built up by my mother . . . who . . . went on listing, from the local and national newspaper, all names not in my first volume." Selection of at least these 4,000 was therefore clearly arbitrary. We know that both Reaney (and later Wilson) and Cottle worked with what they had, with the tools then available. The authors of these important dictionaries were unable to discuss the coverage of their works in relation to the population, because the information to do this was unavailable to them.

DAFN makes use of the technology now available to do this. After each surname headword in DAFN there is a statement of the comparative frequency of each surname. Furthermore, entries were selected in large measure on the basis of computational analysis of comparative frequencies.

The challenge in generating a dictionary of surnames has always been source material. Ideally, information from the U.S. Census Bureau information might have been used, but this was not available, so other sources had to be used. A preliminary study was carried out in 1990–91, using data extracted from a list of 7 million names provided by Donnelley Marketing. Name frequencies in this list were studied by research scientists at AT&T Bell Laboratories in the 1980s, mainly for purposes of machine speech recognition and synthesis. Ken Church of Bell Labs went on to process the data for lexicographic purposes, providing a list of 60,000 different surname forms, representing the surnames borne by approximately 224 million Americans, or 80% of the population.

For the main headword list of DAFN, the Bell-Donnelley list was replaced in 1998 by data excerpted from a machine-readable phone directory. This was done partly in order to obtain independent confirmation of the names on the Bell-Donnelley list, and partly in order to obtain a more up-to-date and complete list.

The source data used was the 1997 edition of InfoUSA's ProCD Select Phone product, which lists almost 100 million telephone subscribers. Using the standard export function supplied with the product and the "greater than 50,000 records" export facility authorized by an unlock key from ProCD, the subscriber name(s) for all residential listings were extracted. The extracted data was filtered to remove nonresidential listings such as municipalities, universities, business services, religious organizations, utilities, etc. Listings for "summer residences" and other multiple listings were also removed, as were listings for children's lines, except where separate forenames for these were given.

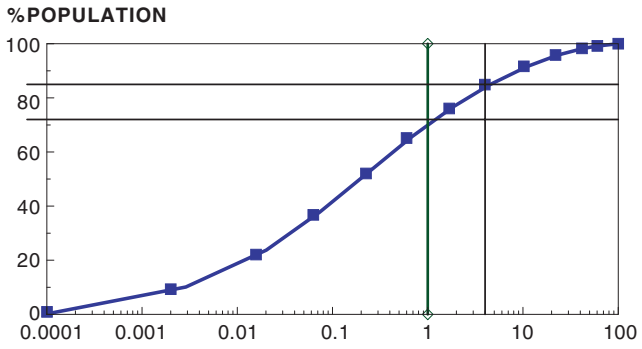
In about 10% of the listings more than one person's forename was given within an entry. These were expanded into two (or more) entries. For example, a listing for "Jones, Bill & Mary" gives two entries: "Bill Jones" and "Mary Jones." Similarly, "Jones, Mr. & Mrs. Richard" gives two entries: "Richard Jones" and "Unknown Jones." Listings such as "Jones, Fred & May Smith" were rendered as "Fred Jones" and "May Smith."

After filtering, 88.7 million forename-surname records remained. These comprise 1.75 million different surnames (types), stored in the AmSur (American Surnames) database. About 73 million (82%) of these records include an associated forename. The balance of 15.7 million records has "Unknown" for the forename.

At the time when the data used for AmSur was compiled, the U.S. Census Bureau gave the residential population of the United States as 266 million. So AmSur is a sample representing almost exactly 33% of the total population in 1997. (By the beginning of 2003, the U.S. population had climbed to over 290 million.) The sample is geographically proportionately distributed, and there is no reason to suspect that nonlisted individuals bear any particularly characteristic set of surnames.

Not only is AmSur a very large sample; it is also probably as representative a sample of the population of the U.S. as it is possible to obtain. By contrast, it is more than twelve times larger than the 1990 U.S. Census Bureau sample used to create its first name and last name distribution tables (www.census.gov/genealogy/names).

Figure 1: *Surname Distribution in the United States*



One of the aims of DAFN is to enable the largest number of people to find an entry for their surname in the dictionary, within limits of what is manageable and publishable. We thus want to know the maximum population represented by given number of surnames. Each surname that is different from any other surname is called a *type*. Every example of that surname is called a *token*. We know for each surname type how much of the population it covers-the number of tokens.

Our sample population is 88.7 million, which we arrange by surname in descending frequency order. *Smith* is the most frequent surname with a count of 831,783 tokens and represents almost 1% of the population: in fact 0.937749%. However, it is only one name type in 1.75 million types, or 0.000057% of the surname types. This point (0.937749; 0.000057) is the origin of our graph shown in Figure 1. We can now add the next most frequent name, *Johnson,* with a count of 610,104 tokens to the list where the cumulative effect of adding *Johnson* to *Smith* is to generate a point at 1.625577; 0.000114. We can continue to do this until we have added all the name types in descending frequency order and arrived at the final point 100; 100. This point says that 100% of the surname types represent 100% of the name tokens, or population, for all the surnames in AmSur. Note that: the "Percentage of Surnames" scale is logarithmic in order to show the shape of the distribution; the graph is cumulative in order to gives a direct read of maximum coverage for a given set of surnames. Note also that the scales have been normalized, so that similar graphs may be compared.

The curve becomes very flat as it approaches the point 100,100 (top right). This is because there are many surnames with few bearers. There are 706,771 surnames with a count of one in the sample. They include *Ficalowych, Hataisutitum, Kapoakun, Larbaig, Mdududi, Onwubu, Papalz, Quinores, Tuvamontolrat, Xerokostas, Yaddanapudi,* and *Znorkowski.* Some of the singleton entries may be transcription errors, but most of them really are surnames, borne by real people. Of course, since AmSur is a sample of the population, not the whole population, we cannot conclude that all these surnames are unique. For one thing, a single phone listing could represent a family of two or more people. Furthermore, no doubt there are other rare and unique surnames for which no telephone subscriber listings exist. Rare and unique surnames deserve further study, beyond the scope of the present work.

The criteria for selection of entries in DAFN were threefold: frequency, historical importance, and etymological importance. All surnames with a frequency in AmSur greater than 99 were automatically given an entry and researched as far as possible. Names with a frequency less than 100 do not receive an entry in DAFN unless they are of particular historical or etymological importance. In practice, it turned out that very few names with an AmSur frequency lower than 100 are historically or etymologically important. Examples of names entered because of their historical importance are **Faniel** (AmSur frequency 91), an altered form of *Faneuil,* the name of a historically significant Boston family, and **Stuyvesant** (AmSur frequency 74), which was the surname of the director general of New Netherland in 1647–64. Examples of names entered for etymological reasons are **Apostolos** (AmSur frequency 89) and **Brabazon** (AmSur frequency 84), both of which serve as anchor names for cross-references from other entries. Several Chinese names of low frequency carry the main explanation even though they are rare, because of the editorial decision to put the main explanation for Chinese names at the Pinyin romanization, rather than at any of the many folk romanizations. A few Polish and other names with AmSur frequencies lower than 100 survive from earlier phases of the editing: these had been researched, with useful information, and there did not seem to be any good reason to delete them. There is nothing particularly significant about the figure 100; it is nothing more than a convenient cutoff point for editorial purposes. All surnames with a count of 100 or more have been included, with the result that there are 70,315 entries in DAFN. This represents little more than 4% of surname types, but it represents the surnames of over 85% of the population of the U.S.

Normalization of Multiple-Format Surnames

Some surnames-in particular those formed with an initial preposition or definite article, for example names that begin with *La, Le, Li, De, Di, Mac, Mc, Van, Van der,* and so on-exist in a number of different formats as regards spacing and capitalization. For example **De Vito** is also found in the formats *de Vito, Devito, DeVito,* and *deVito*. It would be misleading as well as wasteful to list each of these forms as a different name. The approach adopted was to combine the counts for all forms, rendered in the most appropriate format as determined by the editors from linguistic norms. (There are occasional instances where the same concatenated string represents names of two different roots. In such cases the names are, of course, treated as two separate, independent surnames.)

Forenames and Diagnostic Forenames

How many forenames (given names) are there in the United States, and what is the nature of the association between forename and surname? Recall that 73 million records in AmSur have an associated forename. These comprise 1.25 million discrete forenames (types). The number of forename types is therefore quite similar to the number of surname types, but the distribution, although of the same overall form, is very different. There are fewer forenames than surnames, but the most frequent forenames are very much more frequent than the most frequent surnames, and there are many more singletons. Thus, the most frequent surname is *Smith,* with a count of 831,783, whereas the most popular forename, *John*, has a count of 2,229,952. Furthermore, whereas there are over 70,000 surnames with a frequency greater than 100 in AmSur, there are under 14,000 forenames with a frequency greater than 100. This suggests that fore-naming has been more constrained than sur-naming. If this was true in the past, there is now an observable thrust to invent new forenames in the United States. The distribution curve for U.S. forenames is shown in Figure 2.

Figure 2: *Forename Distribution in the United States*

FIGURE TK

In comparing this curve with the Figure 1-the surname curve-we see that its origin is higher and it rises more steeply and has a long, flat run from about 0.1% of forename types. Furthermore, 95% of the population share 1% of the forenames, or seen from the other end of the telescope, 5% of the population share 99% of the forenames. A forename dictionary the same size as DAFN, i.e. one with over 70,000 entries, would have to include all forenames down to a count of 11, but that would be still less than 6% of the total number of forenames.

By 1996, research for DAFN had already been carried out by the editorial team on over 50,000 surnames, but for more than 20% of those entries it was impossible even to state the language of origin with any confidence, which of course made it impossible to research the etymology. It seemed probable that there was a chance of getting a bearing on the etymology of some of these more obscure surnames if the language of some of the forenames that are associated with them were known. Clearly, many forenames are strongly indicative of cultural, ethnic, or linguistic group (CELG). Even when a surname has been Americanized beyond recognition, the choice of forenames for children is often traditional and reflects the languages of the mother country. A list of all the forenames in the database was therefore sent to the editors, who (with specialist advice where necessary) made the following judgments about each forename:

Is it male or female, both, or unknown?
Is it associated with one or more particular CELG?
Is it diagnostic or non-diagnostic, for one or more languages?

These judgments were recorded in a database and used for cluster analysis of forenames. Some forenames are strongly diagnostic in that they are rarely if ever used outside a particular CELG. In other cases, a forename may be weakly diagnostic because it is favored by a particular CELG, even though it may actually be from another language. So, for example, the English forename *Stanley* is favored by Polish Americans, presumably because it is reminiscent of the common Polish forename *Stanisław.* In our terminology, *Stanley* is associated with the Polish CELG and is therefore weakly diagnostic. In the course of cluster analysis, weakly diagnostic names were scored at half the values of strongly diagnostic ones.

Examples of diagnostic and nondiagnostic forenames may be given. *Declan* and *Niamh* are diagnostic for Irish. *Patrick* and *Bernadette* have a statistically significant association with Irish American people, but they cannot be classed as diagnostic, because they are also borne by very large numbers of non-Irish people. *Giuseppe* is diagnostic for Italian. By contrast, it is undeniable that *Antonio* and *Maria* are associated with Italian, but they are associated also with Spanish, Portuguese, and other languages, so these two names are not strongly diagnostic for Italian.

For determining the likely CELG of a surname, all fore-

names associated with that surname are reviewed and each forename is scored as to whether it is diagnostic, male or female, by CELG, and by count. At the end of the process the total score is normalized to 100, so that we may see results like:

English 72%, Polish 17%, Spanish 4%, Jewish 3%,

with a long tail of very rare other CELG possibilities. English is the default (the object of the exercise being to detect non-English surnames), so the English results are discarded, as are the very low-scoring CELGs. The results in this example point to a likely Polish origin, not an English one, for the surname under consideration.

One problem is that many of the results have quite prominent scores for Spanish where the surname is known not to be Spanish. We believe that this is because of the growing pervasiveness of Spanish naming culture; in some parts of the U.S., Spanish, not English, is the default CELG. Our process was adjusted to accommodate this, by treating Spanish names as diagnostic but weighting the results against Spanish and in favor of rarer CELGs such as Latvian or Dutch.

Finally, the results of the forenames cluster analysis were further processed for separation of CELGs sharing several names in common, of which the most striking two are Italian and Portuguese. Some filtering of results was done, for example to delete low-scoring results for Spanish in the context of high-scoring results for Italian, and vice versa.

The results range from 99% for a few names with CELGs such as Ethiopian, Japanese, Muslim, and East Indian to as low as 4% for some Scandinavian surnames. The high scores show low absorption by the CELG of the mainstream fore-naming patterns, i.e. great cultural and linguistic distinctiveness, whereas the low scores occur either with surnames of English linguistic origin or, more commonly, show almost total absorption into English-American patterns of forename choices, with only an occasional flicker of the heritage. The latter case is typical of surnames that have been in English-speaking North America for some considerable time.

Some normalizing of the results by CELG need to be done. For example it is intuitively easy to understand that a 94% Japanese prediction is very positive, but it is less obvious that a 15% Norwegian result is equally positive. However, despite the fact that much work remains to be done in order to refine this process, it has already yielded very useful results. CELG analysis of forenames enabled the editors to reduce the percentage of "unidentified" names, where the language of origin is unknown, from over 20% to fewer than 3%, where a CELG result could be confirmed from genealogical or linguistic sources in the language or culture indicated.

In DAFN, confidence measures based on forename cluster analysis are declared in the Given Names sections by certain entries. Absence of a Given Names section for a particular surname indicates that not enough diagnostic forenames were found to associate the name with any particular CELG.

A word of caution is necessary concerning the interpretation of the percentages printed in the Given Names section for some entries. The percentages are a measure of confidence in the evidence, not of ethnic origin. For example, "Norwegian 15%" does *not* mean that 15% of all bearers of this surname are of Norwegian extraction. It means that, solely on the basis of forename evidence, we can have confidence that the surname is Norwegian; 15% for Norwegian, as stated above, gives great confidence that the surname is Norwegian. The confidence threshold percentage is different for different CELGs; readers may assume that where the level is published in the dictionary, that level, for that CELG, represents a high level of confidence. When put together with linguistic, historical, and genealogical evidence, this confidence may strengthen or weaken. The choice of forenames is only one of many pieces in the puzzle of identifying the CELG of a particular surname.

Female forenames are, unfortunately, severely underrepresented in the Given Names sections. The main reason for this is that they are severely underrepresented in the source material. Noticeably fewer women than men are listed as telephone subscribers; women are often represented as in "Mr. & Mrs. Richard Jones" or "J. Jones," which tells us nothing about the female forename. Of course, women listing themselves individually as telephone subscribers also sometimes suppress their forenames for security reasons. It should also be noted that female names do not cluster quite so neatly into CELGs as male names; no doubt this is partly because a diagnostic female forename of a woman who has married a man outside her CELG and has taken her husband's surname (neither of which are unusual events) creates a red herring. Such events contribute to the low-scoring "tail" of CELG identification, not printed here.

Data Creation and Data Security

Compiling a book of over 70,000 names over a ten-year period, with many contributors and numerous editorial stages, research, and checking procedures, is a large undertaking needing controlled editing. The data was compiled as a structured text file, using HTML-compatible tags to show information types, fonts, accented characters and special symbols, etc.

Each of these had its own "on" and "off" code: for example the surname at the head of the entry, the headword, began with "<N>" and ended "</N>." A validation routine was developed so that at the end of any editing session unbalanced tagging could be rectified immediately, and this was expanded to include correct sequencing of senses, balancing of brackets, nesting of data, and identification of cross-references. If an entry was deleted in the course of editing, or if the spelling of the name was altered, the editor-in-chief was asked to confirm that the alteration was deliberate. The text-file method used by the compilers had the advantage that any part of the text could be changed quickly and easily without constraint, for example using macros for rapid execution of repetitive or conditional operations. However, the other side of the coin is that free text

tends to be insecure and prone to damage caused by human error. Frequent backup into a secure database meant that the DAFN team was able to get the best of both worlds: fast and flexible compilation at the same time as secure backup and structural validation.

Text would be zipped and sent from Oxford, England (later from Boston, Massachusetts), to Ottawa, Ontario, where it was converted into an MS Visual FoxPro database and processed. The text would then be automatically regenerated and sent back for continued editing. Especially toward the end of the process, a major part of the task was ensuring the integrity of the text during final editing. There was an agreed master list of headwords (the 70,315 surnames) and structural elements; deviations were only permitted following explicit confirmation from the editor-in-chief. This proved to be an invaluable aid, especially as the text for the whole book could be validated and turned around in less than 30 minutes.

Bibliography

Hanks, Patrick, and D. Kenneth Tucker (2000): 'A Diagnostic Database of American Personal

Names'. *Names, the Journal of the American Name Society,* volume 48, no.1.