A Diagnostic Database of American Personal Names

Patrick Hanks
Oxford University Press

D. Kenneth Tucker
Esson & Tucker Management

Computational analysis of pairings between surnames and forenames in "A Diagnostic Database of American Names" reveals a number of significant associations. These can be used as evidence to help pinpoint the most likely source language of an unfamiliar and unresearched surname, and to provide a framework for historical and genealogical research. The data can also be studied geographically since a number of surnames are associated with particular states or regions. A study of present-day distribution of surnames can be correlated with historical and other evidence regarding settlement patterns.

Where do the surnames and forenames of America come from? Almost every major language and culture in the world has contributed to the great melting pot that is America. Some names accompanied first-generation immigrants; others are descended from long-established families. In some cases, the form of a name makes it easy to identify the language or culture in which it originated. *Kalkbrenner* is obviously German, and German speakers will recognize that it is from *Kalk* 'lime' plus *Brenner* 'burner'. Here the task of the onomastic lexicographer is to explain why any German-speaking person should ever have wished to burn lime, and why the practice should have given rise to a surname. The answer is that lime—calcium carbonate—is a product of great historical importance, having various long-established agricultural, domestic, and industrial applications, including fertilizing soil, treating furniture, bleaching, and making mortar. It is obtained from limestone by heating or "burning" it, so lime burners were people of some importance in earlier centuries.

In other cases, the source and explanation of the name is opaque. What, for example, is the source of the American surname *Kalla?*

Names 48.1 (March 2000):59-69
ISSN:0027-7738
© 2000 by The American Name Society

Exhaustive comparative linguistic research is required even to find out where the name might have originated, an essential preliminary to explaining what it means. The problem is further complicated by the fact that many long-established family names have changed drastically. Thus, in America *Reinwasser* became *Rainwater* and *Kirchthaler* became *Cashdollar,* among many others.

The task confronting the lexicographer of American family names is truly daunting. Attempts to obtain a systematic account of American family names from the genealogical literature have been disappointing for a number of reasons:

1. Working through the genealogical literature is impossibly slow going: digesting 60,000 genealogies at, say, four a day, would require some 15,000 days, or 60 editor-years—an unacceptably long time.

2. The yield is low: only a very small percentage of American families have been studied genealogically; for many family names, indeed most of them, primary research remains to be done.

3. A selection based on work now available would be biased towards the unusual since genealogical studies have been most successful in dealing with rare and unusual names.

4. Much of the available material is unreliable; many genealogies are compiled by enthusiastic amateurs, working on their own and without expert guidance. Errors and misconceptions abound.

5. Anglicizations of non-English names get particularly short shrift in the genealogical literature since many (perhaps most) genealogical writers are unaware of the regularities of language change.

A more robust source of data is needed, to help provide a linguistic and lexicographical overview of American personal names. Fortunately, such a source is available. Approximately 35% of all Americans are listed as telephone subscribers, and the lists are publicly available. We are in the process of analyzing data from the 1997 edition of INFOUSA ProCD Select Phone, a pack of six CDs listing almost 100 million telephone subscribers.

Fortunately, too, the majority of Americans declare their forenames in their directory entries. The task would be much harder in a country like England, where telephone subscribers often prefer to give their

initials rather than their forenames. Nevertheless, some surnames are not accompanied by forenames. Those subscribers who list themselves as *Mr. and Mrs. Turner* go into our database as "Mr. and Mrs. [Unknown] Turner." *J. Turner* goes into the database as "[Unknown] Turner." There are also some business names. *Alpha Laundry* is almost certainly the name of a business, not a person. As far as possible, we eliminated business names from our database. Some cases are not clear-cut. It is not clear, for example, whether names such as *California Baker* and *Little Dam* are the names of individuals or businesses. In such cases, we have made educated guesses. The number of doubtful cases of this kind is mercifully small, and thus does not affect the overall statistics.

The database and information on the CD pack provide the following overview of American personal names:

Number of listed residential phone subscribers	88.7 million
Number of "unknown" forenames	15.7 million
Number of surname-forename pairs	73 million
Number of different surname types	1.75 million
Number of different forename types	1.25 million

A part of the database is called AMSUR (for American Surnames) and it contains 88.7 million surnames, representing approximately 35% of the population of the United States. AMSUR is a highly representative sample: telephone subscribers are, typically, heads of households. The majority of Americans who are not listed as telephone subscribers are children or other dependents. Other types of people who are not listed include "telephonophobes" (people who do not have and/or do not want a telephone), "dropouts" (people who do not have a home, let alone a telephone), and "secretives" (people who have a telephone, but prefer to be unlisted). There is no reason to believe that these non-listed individuals bear any particularly characteristic set of surnames. In other words, AMSUR is not only a very large sample, it is also probably as representative a sample of the population of the United States as it is possible to obtain. By contrast, it is more than 12 times larger than the 1990 US Census Bureau sample used to create its first name and last name distribution tables (www.census.gov/geneology/names/).

In dealing with American personal names, we make use of the distinction between "type" and "token." A token is an individual occurrence in our data, the name of a single person. There is only one individual in the database called *Curley Schexnayder;* that is, a single token. There are three individuals called *Murphy Schexnayder;* that is, three tokens of the type *Murphy Shexnayder.*

A type is an individual spelling of a name form, and it may occur once or many times. There are in our data 796 tokens of the surname type *Schexnayder,* a further 182 tokens of the spelling *Schexnider,* and 127 tokens of the spelling *Schexnaydre.* Together these constitute 3 types with a total of 1105 tokens.

The distribution of surnames across the population is very uneven. Altogether there are 1.25 million different surname types in AMSUR. In a database of 88.7 million tokens, if the distribution were completely even, each surname would have 71 tokens. This is very far from being the case, however. A few surnames have over 100,000 tokens, while more than 800,000 types are unique, with only one token each.

The ten most frequent surnames account for 4.45% of the population; over 11 million individuals in the United States are called *Smith, Johnson, Jones, Miller, Williams, Brown, Davis, Anderson, Wilson,* or *Taylor.*

Many English-seeming names derive from non-English sources. *Johnson,* for example, includes a number of bearers of surnames from other European languages that are patronymics from cognates or derivatives of the Biblical personal name *Johannes. Taylor* has surely absorbed numerous cases of *Schneider, Kravitz, Krawczyk, Sutter, Hüller, Szabó,* and other occupational names that mean 'tailor'.

Just over 5% of the surname types (67,000 different surnames) account for 90% of the tokens. For each of these surname types, there are over 100 tokens. That is, 90% of the population of the United States have one of only 67,000 surnames, and each of these names has more than 100 bearers in our sample. These names form the basic entry list for the *Dictionary of American Family Names* (DAFN). To these are added less frequent variants, plus other family names of particular historical, linguistic, and other interest, bringing the total to over 100,000 entries and sub-entries.

The part of our database consisting of forenames with their frequencies is called ADDAN: "A Diagnostic Database of American Names." We are currently working through all the forenames with more than nine tokens in the database, adding new fields as needed and addressing the following questions:

1. Is the name diagnostic or nondiagnostic? A diagnostic forename is one that is so strongly associated with a particular language or culture that it provides a very strong clue to the individual's ethnicity. For example, people named *Declan* or *Niamh* are almost certainly of Irish extraction. These two names, then, are diagnostic for Irish. On the other hand, for names such as *Patrick* or *Kevin,* there is a definite association with Ireland and Irish culture, but the association is too weak to be regarded as diagnostic. Many people with no Irish blood whatsoever are called *Patrick* or *Kevin,* so these are classified as nondiagnostic names, although the possible Irish connection is still recorded. Names such as *Thomas, Robert, Sara,* and *Margaret* are utterly nondiagnostic, while *Balazs, Gabor, Laszlo, Sandor,* and *Zoltan* are highly diagnostic. Finding one of these forenames paired with the surname *Bako*, for example, points unmistakably to a Hungarian origin.

2. Is the name a female name, a male name, or can it be either? Female forenames are by consensus assumed to be less diagnostic than male forenames, largely because of the possibility of intercultural marriages.

3. With which language(s) or culture(s) is the name associated? Language and/or culture classification is determined by naming practices within a culture rather than by linguistic affinities. The Welsh language is, of course, related to Gaelic, but the personal-naming practices are almost entirely distinct in the two cultures; there is very little overlap. By contrast, several Irish names are also found in Scottish Gaelic, though in some cases there are distinctions in spelling that can be very helpful to the onomastic historian. By contrast, Czech and Slovak are languages that share a high proportion of their forenames, with only a few distinctively Slovak and only a few distinctively Czech.

Table 1 shows the classification system used the Diagnostic Database of Forenames.

Table 1. Classification of Names

	Abb	Sub	Grp		Abb	Sub	Grp
African	afr	afr	afr	Serbian	srb	ssl	sla
Albanian	alb	alb	alb	Bulgarian	bul	sla	sla
American	ame	ame	ame	German	ger	ger	ger
Black Am.	bla	ame	ame	North Ger.	nge	ger	ger
Arabic	ara	ara	mus	Greek	gre	gre	gre
Armenian	arm	arm	arm	Hawaiian	haw	haw	haw
Chinese	chi	asi	asi	Hispanic	his	his	his
Korean	kor	asi	asi	Spanish	spa	his	his
Vietnamese	vie	asi	asi	Catalan	cat	spa	his
Baltic	bal	bal	bal	Galician	gal	spa	his
Latvian	lat	bal	bal	Mexican	mex	his	his
Lithuanian	lit	bal	bal	Portuguese	por	his	his
Basque	bas	bas	bas	Hungarian	hun	hun	hun
Breton	bre	bre	bre	Indian	ind	ind	ind
Cambodian	cam	cam	cam	Italian	ita	ita	ita
Dutch	dut	dut	dut	Japanese	jap	jap	jap
Frisian	fri	dut	dut	Jewish	jew	jew	jew
English	eng	eng	eng	Jew Amer	jus	jew	jew
Ethiopian	eth	eth	eth	Jew Biblical	jbi	jew	jew
Finnish	fin	fin	fin	Jew Israeli	jis	jew	jew
Estonian	est	est	est	Jew Hebrew	jhe	jew	jew
French	fre	fre	fre	Jew Russia	jru	jew	jew
Scottish	sco	sco	sco	Jew Hungar	jhu	jew	jew
Scots Gaelic	sga	gee	gae	Jew Sefardic	jse	jew	jew
Irish	iri	gee	gae	Jew Ukranian	jkr	jew	jew
Welsh	wel	wel	wel	Muslim	mus	mus	mus
Slavic	sla	sla	sla	Persian	per	per	per
West Slavic	wsl	wsl	sla	Romanian	rom	rom	rom
Czech, Slovak	csl	csl	wsl	Scandinavian	sca	sca	sca
Czech	cze	csl	wsl	Danish	dan	sca	sca
Slovak	slk	csl	wsl	Norwegian	nor	sca	sca
Polish	pol	pol	wsl	Swedish	swe	sca	sca
East Slavic	esl	esl	sla	Icelandic	ice	sca	sca
Russian	rus	esl	sla	Turkish	tur	tur	mus
Ukrainian	ukr	esl	sla	Distinctive	dis	dis	dis
South Slavic	ssl	ssl	sla	Unknown	unk	unk	unk
Croatian	cro	ssl	sla	Unclassified	xxx	xxx	xxx

Key: Abb=abbrevation, Sub=subgroup, Grp=Group

The abbreviation column shows how some of these major classifications can be divided into more subtle subclasses with the help of consultants to DAFN. For example, DAFN consultant Alexander Beider has subdivided Jewish names into the following groups:

jbi: Biblical from Old Testament. (Could be Jewish or not Jewish).

jyd: Yiddish, explicitly Jewish (Eastern Ashkenazic).

jis: Israeli. (Newly invented names or older biblical names whose spelling clearly shows that they were transliterated from Hebrew). In some cases they can also belong to American Jewish families who are imitating the naming choices of Israeli Jews.

jhe: Hebrew (explicitly not Christian) form of a Biblical name; or post-Biblical Hebrew name; explicitly Jewish; some are new Israeli.

jru: Jewish Russian. Common among recent Jewish immigrants from the former USSR. There are more Jewish immigrants from USSR in the United States than there are ethnic Russians. Some names are shared in Russia by Jews and Russians, but with important differences in frequency: *Efim, Semyon/Semen, Yuly, Ilya, Arkady, Grigory,* and *Lev* are mainly Jewish; *Boris* and *Leonid* are more often Jewish than Russian. In Russia *Sergey, Yuriy, Vladimir, Mikhail, Oleg, Igor,* and *Vadim* are mainly Russian, but in the United States they are primarily Jewish.

jse: Sefardic Jewish. Jewish names typical of the Mediterranean and the Middle East.

jus: Jewish American (United States). Names such as *Morris, Louis,* and *Seymour,* of comparatively weak diagnostic value.

jew: Other Jewish, especially European (German, Latin, Greek) names used by Ashkenazic Jews in Germany and the United States. Sometimes they replace genuine Jewish names, e.g., the Germanic name *Bernhart* instead of Yiddish *Ber,* the Greek name *Isidor* (or variants) instead of *Isaac.*

Even though the database is not yet complete, the correlations between forenames and surnames are already being used to make primary adjustments to the entries in the *Dictionary of American Family Names.* For instance, the surname *Dam,* which on etymological grounds

had been classified as a Dutch topographic name, a shortening of *Van Dam*, has now been re-classified as mainly Vietnamese, on account of the forenames which co-occur with it (shown below). European examples do occur (marked here with an asterisk), but these turn out to be mostly Norwegian or Danish rather than Dutch.

> *Hung (7), Ngoc (6), Vinh (6), Tuan (5), Hoa (5), Nu (5), Minh (5), Binh (4), Duc (4), Thanh (4), Chi (3), Cuong (3), *Erik (3), Anh (3), Hiep (3), Hong (3), To (3), Tien (3), Chanh (3), Bich (2), Chung (2), Kinh (2), Naim (2), Qui (2), Quyen (2), *Soren (2), Tam (2), Linh (2), Mai (2), Thang (2), Trung (2), Tu (2), Hue (2), Oanh (2), Bao (2), Chan (2), Nhat (2), Xuang (2), Lien (2), Long (2), Phuoc (2), Son (2), Thi (2), Toan (2), Tuong (2), Vy (2), Buu, Dien, Hanh, *Hans, Huong, *Javier, Jie, *Levi, Manh, Ngan, *Ore, Que, Sokhom, Song, Sun, yang, Diep, Du, Huan, My, Tac, Thu, Thuy, Tuoi, *Helge, Nam, Phieu, *Raul, Vien, Chuong, Hien, *Jorgen, Quan, Thuc, Ton, Do, Ha, Little, Nguyet, Ok, Phuang, Thai, Than, Thien, Ba, Chay, Chieu, Dai, Danh, Diem, Dieu, Dong, Dung, Giang, Giap, Hai, Ham, Han, Hao, *Harald, Hau, Hieu, Hoang, Hon, *Jacobus, Khanh, Khuong, Kien, Kieu, Loi Mi, Moeun, Muoi, Nga, Nghiep, Nguu, Nhung, Nhut, Nien, Nuha, Oi, *Per, Phat, Pho, Phuoc, Phuc, Phung, *Pierre, Quy, Quynh, Sang, Tay, Tho, Thaong, Toha, Tram, Tran, Trang, Truc, Tuyet, *Vagn, Vu, Xa, Xuyen, Yen.*

As a second example of the relation between name and culture group, we ask: What is the origin of the surname *Anne*? An obvious answer is that it is an English metronymic. Less obviously, it might be an English habitation name, from a place name in Hampshire. However, when we look at a contemporary English telephone directory, we find that it is extremely rare as an English surname, but when we look at ADDAN, we find that 34% of the American forenames listed with this surname have been identified as Indian. These include *Venkata (3), Suresh (2), Anand, Abdoulaye, Alioune, Asher, Mamadou, Bose, Rana, Aruna, Madhavi, Pramod, Ramesh, Rao*, and *Ravindra*. This name has been referred to Professor Rocky Miranda, DAFN's consultant on Indian names, for an opinion and, if possible, an etymology.

Finally, *Arabian*, which conceivably could be an English or French ethnic name for an Arab, turns out to be Armenian; 30% of the forenames associated with *Arabian* have already been identified as Armenian, and this number will certainly rise as the identification of rare and previously unidentified forenames proceeds. Diagnostic fore-

names with the surname *Arabian* include *Aram (2), Omid (2), Zohrab, Ara, Artin, Davood, Harout, Nerses, Armand, Haig, Siri, Ali, Angel, Ani, Bedros, Daryoush, Gaspar, Hovsep, Kevork, Nishan, Ohannes, Panios, Sarkis, Varthes*, and *Zaven.*

Not all names are as interesting as these, of course. A large number of American surnames are associated only with forenames like *Mary, Richard, Mark, John, Earl, Margaret, Dwight, Billy-Jo,* and other typically English or American nondiagnostics. These are the surnames that are thoroughly assimilated into America's English-speaking culture. A few of them are recent arrivals from England, Australia, or elsewhere in the English-speaking world, but the vast majority of them arrived in America in the 17th or 18th centuries.

ADDAN also measures the degree of naturalization of family names. Since cultural loyalties are slow to die, Americans tend to favor forenames that were borne by their ancestors, long after they have ceased to use the language of their ancestors. Forenames therefore provide useful evidence of the degree of cultural assimilation. For example, the Spanish surname *Archuleta* is associated not only with Spanish forenames such as *José,* but also with English forenames such as *Charles* and *John*, suggesting that the name is longer established in English-speaking America than some other, equally common Spanish surnames, for which the forenames are nearly all Spanish.

The results are often—but not always—etymologically predictable. It comes as no surprise that *Zbigniew* occurs exclusively with surnames of Polish origin. More surprising is the fact that there is a significant association between the forename *Stanley* and Polish surnames. *Stanley* is not a Polish forename, but it is chosen by Polish Americans as a forename for their children, presumably because they associate it with the Polish forename *Stanisław.* Similarly, *Louis* tends to co-occur with Italian surnames, even though in this spelling it is French. Italian and Spanish Americans also use the forenames *Anthony, Frank,* and *Joseph* with a frequency greater than one would expect.

The database can help to indicate where an etymological or genealogical search should start. For example, there are 140 occurrences of the surname *Kalla* in our database. Where do they come from? Various etymologies are possible and even plausible. *Kalla* is found as a surname in Poland, Finland, Estonia, Lithuania, and elsewhere.

However, the associated forenames in America (*Ashwan, Kamala, Keshav, Mahmoud, Moiez, Ravi, Ribhi, Shantharam, Subhi, Vijay*) point unmistakably to an origin on the Indian subcontinent.

Because they contain addresses, the telephone listings also provide valuable information about where the names are found. Of all bearers of the surname *Cancienne,* for example, 87% are found in Louisiana. This location confirms the French spelling, since Louisiana is an area of strong French settlement, both directly from France, and (in the form of the Cajuns) indirectly from Eastern Canada. We do not even need to look at the associated forenames to know that this is a French name, even though the etymology is unknown. And yet, how distinctively French are its bearers? Or, to put it another way, how naturalized is it as an American name when only a comparatively small number of the forenames are diagnostically French? Many of them, such as *Brent, Cindy, Cleveland, Craig, Donald, George, Harold, Henry, Inez, Kevin, Kimberly, Larry, Linda, Lloyd, Michael, Norman, Owen, Rhoda* are thoroughly American.

These patterns suggest that at least some bearers of the Louisiana French surname *Cancienne* have become thoroughly assimilated into the general English-speaking American culture. Nevertheless, there is a considerable number of diagnostically French forenames which co-occur with *Cancienne,* as would be expected in Louisiana, e.g., *Alcee, Cecile, Celeste, Emile, Estelle, Louis, Madeleine, Michelle*, and *Sybille.*

The combination of data about a surname gathered from location and associated forenames can be very suggestive. To illustrate this important point further, we will conclude with an example involving *Schexnayder,* in its various spellings.

Like the Canciennes, the Schexnayders are strongly associated with Louisiana: 78% of them live there. From its form, it looks as if *Schexnayder* might be an Americanization of a German name. The main associated forenames (other than nondiagnostics) include: *Murphy (3), Alcee (2), Andrus (2), Kurt (2), Desire (2), Emile (2), Marcel (2), Alphonse, Amedee, Benoit, Calice, Camille, Cecile, Curley, Damien, Elva, Felicien, Fernest, Francois, Gaston, Jaime, Leonce, Manfred, Nolton, Oleus, Odilon, Pierre, Remy, Ricardo, Saul, Seva*, and *Simo.*

Taken individually, no one of these forenames provides conclusive evidence, but taken together they provide a great deal of support for the

hypothesis that this is a surname of German origin which has become thoroughly Frenchified over many decades or even centuries in the French-speaking culture of Louisiana. German traces survive in *Kurt* and *Manfred,* reminding researchers that there were in fact a few pockets of German settlement in Louisiana two hundred years ago. Overall, of course, the forenames point to a much greater degree of Frenchness for *Schexnayder* than for *Canciennes,* notwithstanding the fact that graphically the latter appears more French.

These correlations and distribution patterns are facts which require explanation, rather than being explanations in their own right. It is left to historians and genealogists to provide thoroughly researched explanations for the tantalizing snippets of data that ADDAN provides about American family names and forenames. ADDAN and AMSUR are tools still being developed. It is hoped that, in years to come, they will provide valuable resources for onomastic, historical, genealogical, and demographic research.