Two Projects in Onomastic Lexicography

Patrick HANKS and D. Kenneth TUCKER,
Oxford, England and Manotick Canada

Abstract
Computational analysis of pairings between surnames and forenames in 'A Diagnostic Database of American Names' ('ADDAN') reveals statistically 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. Some surnames are more 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.

1 Researching the Linguistic and Cultural Origins of Personal Names

Where do the surnames and forenames of America come from? Almost every language and culture in the world has contributed to the great melting pot that is America. Some are first-generation immigrants, others are descended from long-established families. In some cases, the form of the name makes it easy to identify the language or culture in which it originated. For example, *Kalkbrenner* is obviously German, and German speakers will recognize that it is from *Kalk* 'lime' + *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, with 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*? Exhaustive comparative linguistic research is required even to get to first base: to find out where in the world the name might have originated, an essential preliminary to explaining what it means. The problem is further complicated by the fact that long-established family names have sometimes become garbled almost beyond recognition. Thus, in America *Reinwasser* became *Rainwater***;** *Kirchthaler* became *Cashdollar*, and so on[1].

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 yielded disappointing results, for a variety of reasons:

- It 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 task, given normal constraints of budget and resources.

- The yield is low: only a very small percentage of American families have been studied genealogically at all; for many family names, indeed most of them, primary research remains to be done.
- A selection based on work now available would be biased towards the unusual: genealogical studies have most success with rare and unusual names.
- The quality is unreliable: many genealogies are compiled by enthusiastic amateurs, without expert guidance. Errors and misconceptions are legion.
- Anglicizations of non-English names get particularly short shrift in the genealogical literature. Genealogical writers do not have much sensitivity to the regularities of linguistic change.

Some more robust source of data is needed, to help provide a linguistic and lexicographical overview of the subject. Fortunately, such a source of data is available. Approximately 35% of all Americans are listed as telephone subscribers, and the lists are publicly available on CD^2.

Fortunately, too, the majority of Americans declare their forenames in their directory entries. The task would be much harder in a country like Britain, where telephone subscribers often prefer to give their initials rather than their forename. 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 enquiry. 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 taken an educated guess. The number of doubtful cases of this kind is mercifully small, and does not affect the overall statistics.

The Population of the USA	
Total population of the USA	250 million
Number of households	100 million
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	1.25 million

Figure 1: American Family Names: Some Facts and Figures

This part of the database is called AMSUR. It contains 88.7 million surnames, representing approximately 35% of the population of the USA. AMSUR is a highly representative sample: telephone subscribers are, typically, heads of household. The majority of Americans who are not listed as telephone subscribers are dependents [children and other family members]. 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 remain ex-directory]. 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 (35%); it is also a fairly representative sample of the population of the USA.

2 Types and Tokens

A digression here to illustrate terminology may be useful, in particular the distinction between *type* and *token*. A token is a single individual occurrence – in our data, the name of a 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 form, which 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 of *Schexnaydre*. There are therefore 3 types pronounced "sheksnyder", with a total of 1105 tokens.

The distribution of surnames in a 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. A few surnames have over 100,000 tokens, while many have only one token: there are, in fact, over 800,000 surname types with only one token.

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

How they got these English-seeming names is a different matter. *Johnson*, for example, has absorbed a proportion of bearers of those surnames from other European languages that are patronymics from cognates or derivatives of the Biblical personal name *Johannes*. *Taylor* has undoubtedly 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 USA have one of just 67,000 surnames, and all of them have more than 100 bearers in our sample (i.e. more than about 280 bearers in the population at large, assuming that we are right that our sample is representative). This forms the basic entry list for the *Dictionary of American Family Names*. To these are added less frequent variants, plus other family names of particular historical, linguistic, and other interest, bringing the total up to over 100,000 entries and subentries.

3 Assigning Language Values to Forenames

Let us now turn to the forenames in our database. The part of the database consisting of forenames with their frequencies, is called ADDAN ("A Diagnostic Database of American Names").

We are currently working steadily through all the forenames with more than ten tokens in the database, adding new fields under each name and addressing the following questions:

1. Is the name diagnostic or nondiagnostic? A 'diagnostic' name is one that is so strongly associated with a particular language or culture that it provides a very strong clue for the individual's ethnicity – a clue that, taken with other similarly diagnostic names, adds up to a certainty. For example, if someone is called *Declan* or *Niamh*, they are almost certainly of Irish extraction. These two names, then, are 'diagnostically' Irish. On the other hand, for names such as *Patrick* or *Kevin*, there is a definite association with Ireland and Irish culture, but the signal is too weak to be regarded as diagnostic. Many people with no Irish blood whatsoever in their veins are called *Patrick* and *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, whereas on the other hand finding *Balazs*, *Gabor*, *Laszlo*, *Sandor*, and *Zoltan* with the surname *Bako* points unmistakably to a Hungarian origin. These five forenames are very diagnostic.

2. Is it a female or a male name (or both)? Female forenames are by definition less diagnostic than male ones, because of the possibility of intercultural marriages.

3. What language or languages is it associated with?

   Language/culture classification is determined by naming practices within the culture, not by linguistic affinities. The Welsh language is, of course, related to Gaelic – but the personal-naming practices are almost entirely distinct in the two languages. There is very little overlap. By contrast, several Irish names are also found in Scottish Gaelic, though in some cases there are give-away distinctions in spelling: subtle distinctions that can be very helpful to the onomastic historian. At the other end of the spectrum, Czech and Slovak are languages that share a high proportion of their forenames; only a few names are distinctively Slovak as opposed to distinctively Czech.

As can be seen from Figure 2, page 217, ,some of these classifications are divided into more delicate subclasses, with the aid of DAFN (*Dictionary of American Family Names*) consultants. So, for example, the DAFN consultant Alexander Beider has subdivided Jewish names into the following classes:

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

jyd: Yiddish, explicitly Jewish (Eastern Ashkenazic).

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

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

jru: common among recent Jewish immigrants from (former) USSR. Jewish immigrants from USSR are much more numerous in the USA than ethnic Russians. Some names are shared

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

Figure 2: The Language and Culture Groups of American Forenames

in Russia by Jews and Russians, but with important differences in frequencies: *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 US they are mainly Jewish.

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

jus: Jewish American (U.S.): e.g. *Morris*, *Louis*, *Seymour*, of comparatively weak diagnostic value.

jew: other Jewish; mainly, European (German, Latin, Greek) names used by Ashkenazic Jews in Germany and USA. Sometimes they replace genuine Jewish names: e.g. the Germanic name *Bernhart* instead of Yiddish *Ber*, the Greek name *Isidor* (with variants) instead of *Isaac*, etc.

4 Using the Correlations

Even though the database is not yet complete, the correlations between forenames and surnames are already being used to make primary adjustments to the *Dictionary of American Family Names*. For example, 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, which are shown in Figure 3. European examples do occur (marked with an asterisk in Figure 3), but these turn out to be mostly Norwegian or Danish, not 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), Xuong (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, *Ole, Que, Sokhom, Song, Sun, Vang, Diep, Du, Huan, My, Tac, Thu, Thuy, Tuoi, *Helge, Nam, Phieu, *Raul, Vien, Chuong, Hien, *Jorgen, Quan, Thuc, Ton, Do, Ha, [??]Little, Nguyet, Ok, Phuong, 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, Thuong, Toha, Tram, Tran, Trang, Truc, Tuyet, *Vagn, Vu, Xa, Xuyen, Yen.*

Figure 3: Diagnostic Forenames with the Surname ‘Dam’

What is the origin of the surname *Anne*? The obvious guess is that it might be an English metronymic. A slightly less obvious guess is that it might be an English habitational name, from a place in Hampshire. However, when we look at the contemporary English telephone directory, we find that it is extremely rare as an English surname, while when we look at ADDAN, we find that 34% of the American forenames with this name have been identified as Indian (see Figure

4). It has therefore been referred to Professor Miranda, DAFN's learned consultant on Indian names, for an opinion and, if possible, an etymology.

Venkata 3, Suresh 2, Anand, Abdoulaye, Asher, Mamadou, Bose;
Rana, Aruna, Madhavi, Pramod, Ramesh, Rao, Ravindra.

Figure 4: Diagnostic Forenames with the Surname 'Anne'

Arabian, which might conceivably have been an English or French ethnic name for an Arab, turns out to be Armenian. 30% of the forenames associated with it are already identified as Armenian, and this number will certainly rise dramatically as the identification of rare forenames proceeds.

Aram (2), Omid (2), Zohrab, Ara, Artin, Davood, Harout, Nerses,
Armand, Haig, Siri, Ali, Angel, Ani, Bedros, Daryoush, Gaspar, Grigor,
Hovsep, Kevork, Nishan, Ohannes, Panios, Sarkis, Vartkes, Zaven.

Figure 5: Some Diagnostic Forenames with the Surname 'Arabian'

All Armenian surnames end in *-ian*, but it does not follow that all surnames ending in *-ian* are Armenian. So, for example, *Subramian* is Indian, not Armenian.

Not all names are interesting as this, of course. A large number of American surnames are associated only with forenames such as *John*, *Richard*, *Mark*, *Dwight*, *Earl*, *Margaret*, *Mary*, *Billy-Jo*, and other typically English or American nondiagnostic forenames. 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 back in the 17th or 18th century.

5 Degrees of Naturalization

ADDAN also functions as a barometer of the degree of naturalization of a family name. Americans tend to favor forenames that were borne by their forefathers, long after they have ceased to use the language of those forefathers. Residual cultural loyalties are slow to die. Forenames therefore provide useful evidence for the degree of assimilation as opposed to recentness of ethnic origins.

So, for example, the Spanish surname *Archuleta* is associated not only with Spanish forenames such as *Jose*, but also quite strongly with English forenames such as *Charles* and *John*. This suggests that it is better established in English-speaking America than some other, equally common Spanish surnames, for which the forenames are all (or almost all) Spanish in form.

The results are usually but not always etymologically predictable. For example, it comes as no surprise that *Zbigniew* occurs exclusively with surnames of Polish origin. More surprising is the fact that there is a statistically 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 *Stanislaw*.

Similarly, *Louis* tends to pick out Italian surnames, even though it is French in form. Italian and Spanish Americans also use the forenames *Anthony*, *Frank*, and *Joseph* with a frequency greater than chance.

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

6 Geographical distribution

Because they contain addresses, the telephone listings also provide valuable information about where the names are found. 87% of all bearers of the surname *Cancienne,* for example, are found in Louisiana. The location confirms the French spelling form, for of

course 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 forenames (*Alcee, Cecile, Celeste, Emile, Estelle, Louis, Madeleine, Michelle, Sybille*) to know that this is a French name, even though the etymology is unknown. And yet, and yet... how distinctly French are its bearers? Or, to put it another way, how naturalized is it as an American name? In fact, only a comparatively small proportion of the forenames are diagnostically French. Many of the forenames are thoroughly American, not French.

Brent, Cindy, Cleveland, Craig, Donald, George, Harold, Henry, Inez, Kevin, Kimberly, Larry, Linda, Lloyd, Michael, Norman, Owen, Rhoda.

Figure 6: Nondiagnostic Forenames with the Surname ‘Cancienne’

This pattern indicates that at least some bearers of the Louisiana French surname*Canciennes* have become assimilated into the general run of English-speaking American culture.

Nevertheless, there are also plenty of diagnostically French forenames, as one would expect in Louisiana:

Alcee, Cecile, Celeste, Emile, Estelle, Louis, Madeleine, Michelle, Sybille.

Figure 7: Diagnostic Forenames with the Surname ‘Cancienne’

Thus, the combination of data about a surname from location and forenames can be very suggestive for researchers. To illustrate this further, I close with an example from *Schexnayder*, in its various spellings.

Like the Canciennes, the Schexnayders are strongly associated with Louisiana: 78% of them live there. From its form, this looks as if it might be an Americanization of a German name, although Louisiana is not noted as an area of German settlement. 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, Simo.

Figure 8: Forenames with the Surname 'Schexnayder' and Variants

Taken by itself, no one of these forenames is conclusive. But taken together, they provide empirical 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. But collectively, the forenames point to a much greater degree of Frenchness for *Schexnayder* than for *Canciennes*, notwithstanding the fact that morphologically the latter looks more French.

7 Conclusion

These correlations and distribution patterns are facts demanding explanation, rather than explanations in their own right. It is up 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, these tools will provide valuable resources for onomastic, historical, genealogical, and demographic research.

Notes

[1]Additional examples are given in Hanks and Hardcastle (1996): *The Multiplicitous Origins of American Family Names*, in Proceedings of ICOS 1996, University of Aberdeen (ed. Nicolaisen)

[2]The authors of the present paper 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.