Computers and Research on Personal Names

Herbert Barry, III
University of Pittsburgh

Computers assist researchers by storing, arranging, selecting, and summarizing numerical and alphabetical information. Their speed, accuracy, and large handling capacity contribute to research on personal names by providing rapid and accurate summaries and analyses of large amounts of information. Computers can relate personal names to many attributes, and are thus useful for developing indices such as quantitative phonetic gender scores and in assisting in such studies as the evolution of unisex names, the properties of personal names, names in genealogies, and names in fictional and factual texts.

Computers allow researchers to manipulate and analyze information that occurs in large quantities and in diverse contexts. Personal names, since they consist of specific combinations of alphabetic characters, are particularly suited for analysis by computer.

Word Processing and Name Processing

Word processing refers to the use of a computer in the production of text. Typed text is converted into an electronic file that facilitates storage, modification, reproduction, and generation of printed versions. Word processing programs, such as WordPerfect®, Microsoft Word®, and Wordstar®, make it easy to locate selected words or phrases and to sort data sets — to group together all items with particular characteristics. Adding, deleting, rearranging and replacing information can be done easily and selectively.

Using computers for name processing applies principles of word processing to an electronic file that contains a list of names. A user can modify selected names, combine different lists of names, select a subset of all names, or sort the names into a variety of sequences, for instance

Names 43.4 (December 1995):315-324
ISSN:0027-7738

into alphabetical order or frequency, or any characteristic coded in the electronic file. Name research benefits enormously from the speed, accuracy, and data-handling capacity of computers.

Sources of Information on First Name Frequencies

Studies of the popularity of first names use information available from many states. Schwegel (1988), for example, reported the relative frequencies and rank orders of the 100 most frequent names in several states of the United States and provinces of Canada. Evans (1992) identified the rank order frequency of the 500 most frequent names of boys and girls, compiled from first name frequencies provided by a number of states. The information contained in these sources can be used to create an electronic database.

Some states make available electronic files of first names as recorded on birth certificates in a specific year.[1] I purchase the files of first name frequencies from the State Center for Health Statistics and Research, Pennsylvania Department of Health, in Harrisburg.[2] In these files, each record contains a first name and the number of individuals given the name in a specified year. Separate files contain the information on boys and girls. The files can be divided into separate files, for example, according to the ethnicity of the mother.

Creating an Electronic Database

Researchers usually need to modify the format of the files they receive. For example, if a diskette contains separate files for boys and girls and for different years, a single file with codes to identify different years and genders can be created. If the research is to be limited to the 100 most frequent names, the file can be sorted according to frequency and the 100 most frequent names can be selected.

Example of an Electronic Database

Table 1 shows the first name frequencies (listed alphabetically) of five popular names for boys and five popular names for girls given in Pennsylvania at three different times. In each line, the first name is followed by six columns of frequencies; these are the six variables, which are identified by the headings **B60, B75, B90, G60, G75,** and **G90,** indicating the number of boys and girls given that name in 1960, 1975, and 1990.

Table 1. Frequencies of some popular names given in Pennsylvania at three different time periods.

Name	B60	B75	B90	G60	G75	G90
David	5776	2261	1636	0	2	2
James	5206	2000	1503	1	1	3
Matthew	519	1766	2648	0	3	2
Michael	5791	4065	3462	6	11	5
Robert	5518	2123	1596	0	2	1
Ashley	9	12	9	3	53	2014
Christina	0	0	0	226	638	576
Elizabeth	1	0	0	1185	534	884
Sarah	0	0	2	168	317	1264
Susan	0	0	1	2922	527	81

The raw frequencies of first names are influenced by the birth rate, which in Pennsylvania was highest in 1960 and lowest in 1975. The name *David* was rank 3 for boys in both 1960 and 1975 although the frequency was more than twice as high in 1960. The frequencies of first names are also influenced by greater diversity in choice of names for girls than for boys, which results in lower frequencies for girls. The frequency of 2014 in 1990 for *Ashley*, the most frequent name of girls, is much lower than the frequency of 3462 in 1990 for *Michael*, the most frequent name of boys.

Differences in frequency between years and between genders can be controlled by converting absolute frequencies into separate rank order frequencies for boys and girls in each year. A simple technique is to sort the names according to frequency and determine the successive rank orders. Computer programs can do this easily.

Barry and Harper (1995) used the SPSSX® (Statistical Package for the Social Sciences) set of statistical procedures (SPSS, 1986) to associate rank order frequencies of first names with other variables. The database containing first name frequencies included several phonetic measures for each name. Table 2 shows numerical scores for number of syllables (**SYL**), number of phonemes (**PHO**), the position of the

accented syllable (**ACC**), and a letter code indicating the last phoneme (**LP**) in the name. **PS1** is the first phonetic gender score, **PS2** is the second phonetic gender score, and the overall phonetic gender score (**PGS**) is the sum of PS1 and PS2.

Table 2. Coded phonetic information for the names given in Table 1.

Name	SYL	PHO	ACC	PS1	LP	PS2	PGS
David	2	5	1	0	d	-2	-2
James	1	4	1	-1	s	-1	-2
Matthew	2	5	1	0	u	+1	+1
Michael	2	5	1	0	l	0	0
Robert	2	6	1	-2	t	-2	-4
Ashley	2	4	1	0	i	+1	+1
Christina	3	8	2	+2	X	+2	+4
Elizabeth	4	8	2	+2	T	-1	+1
Sarah	2	4	1	0	X	+2	+2
Susan	2	5	1	0	n	0	0

The first phonetic gender score was produced by the following set of SPSSX commands:

```
IF (ACC GE 2) PS1=+2

IF (SYL GE 3 AND ACC EQ 1) PS1=+1

IF (SYL EQ 2 AND ACC EQ 1 AND PHO LE 5) PS1=0

IF (SYL EQ 1 AND PHO LE 5) PS1=-1

IF (SYL LE 2 AND ACC EQ 1 AND PHO GE 6) PS1=-2
```

The second phonetic gender score was produced by SPSSX from the following commands:

```
IF (LP EQ 'X') PS2=+2

IF (LP EQ 'i' OR LP EQ 'u' OR LP EQ 'a' OR LP EQ 'A' OR LP EQ 'E' OR LP
EQ 'e' OR LP EQ 'I' OR LP EQ 'o' OR LP EQ 'U' OR LP EQ 'V' OR LP EQ 'W'
OR LP EQ 'Z' OR LP EQ 'Y') PS2=+1
```

```
IF (LP EQ 'm' OR LP EQ 'n' OR LP EQ 'G' OR LP EQ 'r' OR LP EQ 'l') PS2=0

IF (LP EQ 'f' OR LP EQ 'v' OR LP EQ 'H' OR LP EQ 's' OR LP EQ 'z' OR LP
EQ 'S' OR LP EQ 'c' OR LP EQ 'j' OR LP EQ 'J') PS2=-1

IF (LP EQ 'p' OR LP EQ 'b' OR LP EQ 't' OR LP EQ 'd'OR LP EQ 'k' OR LP
EQ 'g') PS2=-2
```

The phonetic gender score (the sum of PS1 and PS2) was obtained by the command:

```
COMPUTE PGS=PS1+PS2
```

In each of these examples, **GE** means 'greater' or 'equal' and **LE** means 'less' or 'equal.' Note that the scores are more often negative for popular names of boys and more often positive for popular names of girls.

The first phonetic summary, PS1, is +2 for names that are pronounced with the accent on the second or later syllable, such as *Christina* or *Elizabeth*. PS1 is +1 for names with three or more syllables and the accent on the first syllable, such as *Jennifer*. PS1 is 0 for names with two syllables, the accent on the first syllable, and five or fewer phonemes, such as *David* or *Ashley*. PS1 is -1 for names with one syllable and five or fewer phonemes, such as *James*. PS1 is -2 for names with two or fewer syllables, the accent on the first syllable, and six or more phonemes, such as *Robert*.

The second phonetic summary, PS2, refers to the pronunciation of the last phoneme. The phonetic symbols are lower-case or upper-case alphabetical letters, shown for the ten names in table 2. Standard phonetic symbols that are not alphabetical letters were converted into letters, such as **X** for the unstressed, mid-central vowel (schwa). PS2 is +2 for the schwa of *Christina* and *Sarah*. PS2 is +1 for any other vowel, such as *Ashley* and *Matthew*. PS2 is 0 for a sonorant consonant, either nasal, as in *Susan* and *William*, or resonant, as in *Michael* and *Amber*. PS2 is -1 for a consonant that is either fricative, as in *James*, *Elizabeth*, and *Joseph,* or affricate, as in *Mitch* and *George*. PS2 is -2 for a plosive consonant, as in the final phoneme of *David*, *Robert*, *Philip*, *Jeb*, *Mark*, or *Greg*.

Barry and Harper (1995) used this information in reporting that the average phonetic gender score indicated a preference for phonetically

female names for the 25 most frequent names in 1990, in contrast with the popular names of 1960. For example, *David*, *James*, and *Robert*, with negative phonetic gender scores, were preferred in 1960. *Christina* and *Sarah*, with positive scores higher than +1, were preferred in 1990 (table 1).

Formats for Electronic Databases

The large data-handling capacity of computers enables many more names and many more measures of each name to be tested than was previously possible. The electronic database of popular and unisex names used by Barry and Harper (1993, 1995) contains several hundred names and codes for more than 30 variables. In our case, each record consists of the name (reproduced on two lines) followed by a code of 1 or 2 which designates the record number. Each variable is located in a specific column or group of columns on one of the two lines. The length of the lines is limited to 80 columns to permit visual display and printing of each line.

Barry and Harper (1993, 1995) used SPSSX (SPSS, 1986), which can manipulate data (sort, alter, etc.) and can apply a wide variety of statistical tests as well. Other commonly-used statistical packages that perform the same general functions are SAS® (Statistical Analysis System) and BMDP® (Biomedical Package).

Alternative formats for the kind of database discussed here are spreadsheet programs, such as those provided by Excel®, Quatro®, Lotus 1-2-3®, and dBase®. These are widely used and in some ways are easier to learn than SPSSX. They can be organized into rows of different names and columns that contain the alphabetical spelling of the name and the other variables.

Using Genealogies

In addition to frequencies, other sources may provide useful information on personal names. Genealogies contain information on names of large numbers of people over extensive periods of time. The vital statistics on individuals and the relationships among them provide much useful information.

Table 3 shows a sample of names in a database that summarizes a portion of the genealogy of the Roosevelt family, obtained from standard sources such as Whittelsey (1902) and Mosley (1993). Each line contains

information on one individual. A series of five or six two-digit numbers indicates the birth order of the child in each generation following the original couple, Nicholas Roosevelt, born in 1658, and his wife, Heyltje Jans Kunst, born in 1664.

Theodore Roosevelt was a member of the sixth generation. His ancestor in the first generation was the fourth child of Nicholas and Heyltje. Franklin Delano Roosevelt was also a member of the sixth generation. His ancestor in the first generation was the sixth child of Nicholas and Heyltje. The two presidents, therefore, were fifth cousins.

Following the two-digit numerical codes, the sex of the individual is designated by an alphabetical code **B** (boy) or **G** (girl). Other codes are for spouses: **W** (first wife), **X** (second wife), **H** (first husband), **I** (second husband). A second letter following **B** or **G** designates whether the boy or girl is the child of the first spouse (**W** or **H**) or second spouse (**X** or **I**).

Alphabetical letters reproduce the first name, middle name, and surname of the individual. Numerical codes indicate the year (four digits), month, and day of birth, the year (last two digits), month, and day of marriage, and the year (last two digits), month, and day of death. Blank spaces indicate the information was not available.

Table 3 shows a small sample of a very large electronic database, which contains information on descendants and their spouses in each of six generations. Programming commands can generate new measures on each individual, such as age of marriage and death. Characteristics of the names can be coded according to various criteria, such as length, phonetic attributes, frequency in the population, and frequency among ancestors.

One of the variables of interest concerns middle names. A change in custom is indicated by information on the presidents of the United States. Among 17 presidents born before 1820, only three (John Quincy Adams, William Henry Harrison, James Knox Polk) had a middle name. Among 24 presidents born subsequently, including William Jefferson Clinton, only three (Benjamin Harrison, William McKinley, Theodore Roosevelt) did not have a middle name or, in the case of Harry S. Truman, a middle initial.

Table 3. Genealogical information on two generations of Roosevelts.

IDENTITY		FIRST NAME	MIDDLE NAME	SURNAME	YEAR, MONTH, DAY BORN	WED	DIED
0409070105	B	Theodore		Roosevelt	18310922	531222	780209
0409070105	W	Martha		Bulloch	18340708	531222	840212
040907010501	G	Anna		Roosevelt	18550107	951125	310825
040907010501	H	William	Sheffield	Cowles	18460801	951125	230501
040907010502	B	Theodore		Roosevelt	18581027	801027	190106
040907010502	W	Alice	Hathaway	Lee	18610729	801027	840214
040907010502	X	Edith	Kermit	Carow	18610806	861202	480930
040907010503	B	Elliott		Roosevelt	18600228	831201	940814
040907010503	W	Anna	Rebecca	Hall	1863	831201	921207
040907010504	G	Corinne		Roosevelt	18610927	820429	330217
040907010504	H	Douglas		Robinson	18550103	820429	180912
0607050301	B	James		Roosevelt	18280716	530000	001208
0607050301	W	Rebecca	Brien	Howland	18310115	530000	760821
0607050301	X	Sarah		Delano	18540921	801007	410907
060705030101	BW	James		Roosevelt	18540327	781118	270507
060705030101	W	Helen	Schermerhorn	Astor	18540327	781118	93
060705030102	BX	Franklin	Delano	Roosevelt	18820130	050317	450412
060705030102	W	Anna	Eleanor	Roosevelt	18841011	050317	621107

Coded characteristics of names can be related to other measures of the individuals as well. For example, people with popular or unusual names can be related to such demographic variables as marriage, age at marriage, number of children, and longevity.

Genealogies also provide information on relationships among family members. Barry (1984) reported that among 39 presidents of the United States, all except six had fathers whose first names were given either to the president or to a brother of the president. The father's first names were given to the older brother of President Bush and to President Clinton. Barry (1984) also reported greater longevity for presidents who were named after their fathers than for presidents who had brothers named after their fathers. Data files on genealogies can be used to test whether these findings apply to other families as well.

Other Onomastic Research

In addition to the examples shown here, computers are useful for other types of research on names as well. Electronic databases can also reproduce factual texts such as newspapers and magazines, and fictional texts such as novels and plays. Each occurrence of selected names can be identified and related to its context by using key words or other names in the adjacent text. Computers thus provide a powerful new technique for literary onomastic research and for analyses of names in factual documents.

Conclusion

Electronic databases and computer programs greatly enhance onomastic research. An important advantage of computers over paper and pencil is their ability to store and analyze enormous amounts of information, much more than otherwise could be analyzed in many human lifetimes. New measures can be generated and related to each other rapidly and accurately and new hypotheses can be tested. Computers are not limited to making traditional research on names quicker and easier. Rather, they allow a remarkable increase in the scope of the research. Larger samples of names can be studied, multiple samples of names can be compared with each other, and more extensive information on names can be analyzed and reported easily, elegantly, and error-free.

Notes

I gratefully acknowledge the help of Jerry Orris at the Pennsylvania State Center for Health Statistics and Eileen S. Kopchik at the University of Pittsburgh Computer Center.

1. The availability and form of the data vary from state to state. Some states will not provide the information at all; others will provide it only on magnetic tape or in print. The price varies as well, but is usually reasonable; the Pennsylvania data cost $75 for each year requested plus $25 for each diskette (which can hold the files for several years).

2. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

References

Barry, Herbert, III. 1984. "Predictors of Longevity of United States Presidents." *Omega* 14:315-321.

Barry, Herbert, III and Aylene S. Harper. 1993. "Feminization of Unisex Names from 1960 to 1990." *Names* 41:228-238.

_____. 1995. "Increased Choice of Female Phonetic Attributes in First Names." *Sex Roles* 32:809-819.

Evans, Cleveland Kent. 1992. *Unusual & Most Popular Baby Names*. Lincolnwood, Il: Publications International, Ltd.

Mosley, Charles, ed. 1993. *American Presidential Families*. New York: Macmillan.

Schwegel, Janet. 1988. *The Baby Name Countdown: Popularity and Meaning of Today's Baby Names*. Edmonton, Alberta, Canada: Personal Publishing.

SPSS Inc. 1986. *SPSSX User's Guide*, 2nd edition. Chicago: SPSS.

Whittelsey, C. B. 1902. *The Roosevelt Genealogy*. Hartford, Conn: J. B. Burr & Co.