Names, Vol. 38, Nos. 1 & 2 (March-June 1990)

Computer Data Bank of Danish Names

Georg Søndergaard

Abstract

Danish onomastic research today has exciting new possibilities, thanks to an extensive data bank of personal information compiled through the government's civil registration system. With computer access to such facts as individual years and places of birth, for instance, we can undertake very detailed analyses of the distribution patterns of personal names in recent times.

*****

The CPR

In the autumn of 1985, the Danish Ministry of the Interior's Department for Civil Registration gave to the Institute of Onomastics at the University of Copenhagen an extract from the Central Civil Register (*Det Centrale Person Register—* the *CPR*). This register includes approximately six and a half million people, all of those who live or have lived in Denmark since the establishment of the CPR in 1968. A copy of some parts of the information is kept at the Data Processing Center of the University of Odense, where, together with Torben Møller Christensen, I have undertaken an examination of the surname material.

The CPR was developed by the Ministry of the Interior to standardize civil registration forms across local administrations. It is a comprehensive, systematic means of satisfying the requirements of the different public administrative sectors. Accurate, up-to-date information about individuals and groups was urgently needed for purposes of taxation, social welfare, old age pensions, and drafting for National service, to name a few. Recent addresses are important, and information about age, sex, and marital status is often valuable. For planners and legislators, such statistics about various groups are essential.

The CPR is administered independently of other public authorities, but it is the basic system for almost all of them. Its central computerized registration is automatically and directly available through an integrated network. Since its inception in 1968, the CPR has been continually improved, and today it is the most modern civil registration system in the world.

In addition to its public function, the CPR is of great value to the social sciences. On the basis of its data, which is currently systematically updated, the composition of the population can be defined at any time according to such socio-demographic parameters as age, sex, occupation, nationality, etc. Geographic breakdowns of the information can be specified to any desired degree of detail. The CPR is divided into a series of subregisters, of which the most relevant to this article is the register of persons, containing about six and a half million people. Of these, well over five million actually live in Denmark and Greenland. The rest died, emigrated, or disappeared after April 1, 1968. For each person is provided his or her name, status (alive, dead, or emigrated), address, birthplace, nationality, occupation and several other data, and cross-reference numbers to related persons, all tied together with an entry-identification number. Information about previous names, addresses, and occupations is also kept in the register, for about three years. The name information includes first name and surname, surname at birth, "search name," and an abbreviated form constructed automatically if the name contains more than thirty characters (to improve the economy of data handling).

The examination of the comprehensive data material will last for years, but some results have already been published. The most important is Pedersen and Weise's *Fornavnebogen* [*The Book of First Names*] where 12,500 first names are registered together with information about variation in use in different periods and different regions. Together with the Data Processing Center of Odense University, I have published *Oversigt over efternavne i Danmark*, a prepublication that provides surname frequencies for 10,000 surnames in thirteen regions in Denmark.

The surname material has not yet been thoroughly examined. With currently available computer facilities, a simple read-through of the data requires eighteen hours. Scholars are not accustomed to such riches.

Onomastic Use of the CPR: Some Examples

Because many people alive in 1968 and thereafter were born as early as the 1880s, and because the information about birth places is so detailed, it is possible to analyze personal naming patterns by year and by parish, from the end of the 1800s to 1985. Even the oldest material is detailed enough to be statistically representative. Figs. 1 and 2 show the results of two such studies on the name *Knud*.

The inventory of first names is considerably smaller than that of surnames, the figures being respectively about 500 and about 90,000. Of the 500, only about 100 for each sex are frequent enough for statistically significant patterns to emerge. *Knud* is such a name. Its revival is seen first, during the renaissance of Nordic names in the early nineteenth century, among the enlightened middle classes. It appears first in Copenhagen, and in the year 1900 is still typically urban, most common still in Copenhagen. From there the use spreads outward through the country, until in 1947—two generations later—the last *Knud* is baptized in Copenhagen, whereas the name has become fashionable in the Northwest of Jutland (Søndergaard, "Navne i Odense").

Naming patterns can also be analyzed for social variation. Such patterns can be found several hundred years ago as well as today. (See, for example, Søndergaard, "General Outline.")

The use of surnames does not spread and vary in the same way. Where first names change, surnames are stable. Although they are much more numerous and differentiated than first names, many surnames are firmly attached for generations to certain parts of the country. This regional stability characterizes all cultures with surname principles similar to ours.

Error Rates in the CPR

The advantage of storing this complete corpus of information in machine readable form is somewhat offset by the necessity to transform, or encode, certain types of data, for more efficient handling, and also by the significant error rate. Whereas the numerical information can be subjected to some kinds of error-detection routines, onomastic data is more unpredictable: the difference between typing mistakes and spelling variations, for instance, is impossible to define. Manual data entry, performed nationwide by people with widely differing levels of training and experience, creates many anomalies. The names themselves in the CPR are of minimal importance, as each entry is recognized internally by its unique identifying number. Some deficiencies are due to hardware, which requires a character set that provides no diacritics: ä and ö cannot be represented.

In practice, many of the problems could be avoided by preprocessing the data to weed out persons not born in Denmark or without Danish citizenship; persons born in Greenland; names with a certain set of automatically identifiable errors (such as misspelled suffixes); and nonce names, which comprise about one third of the total.

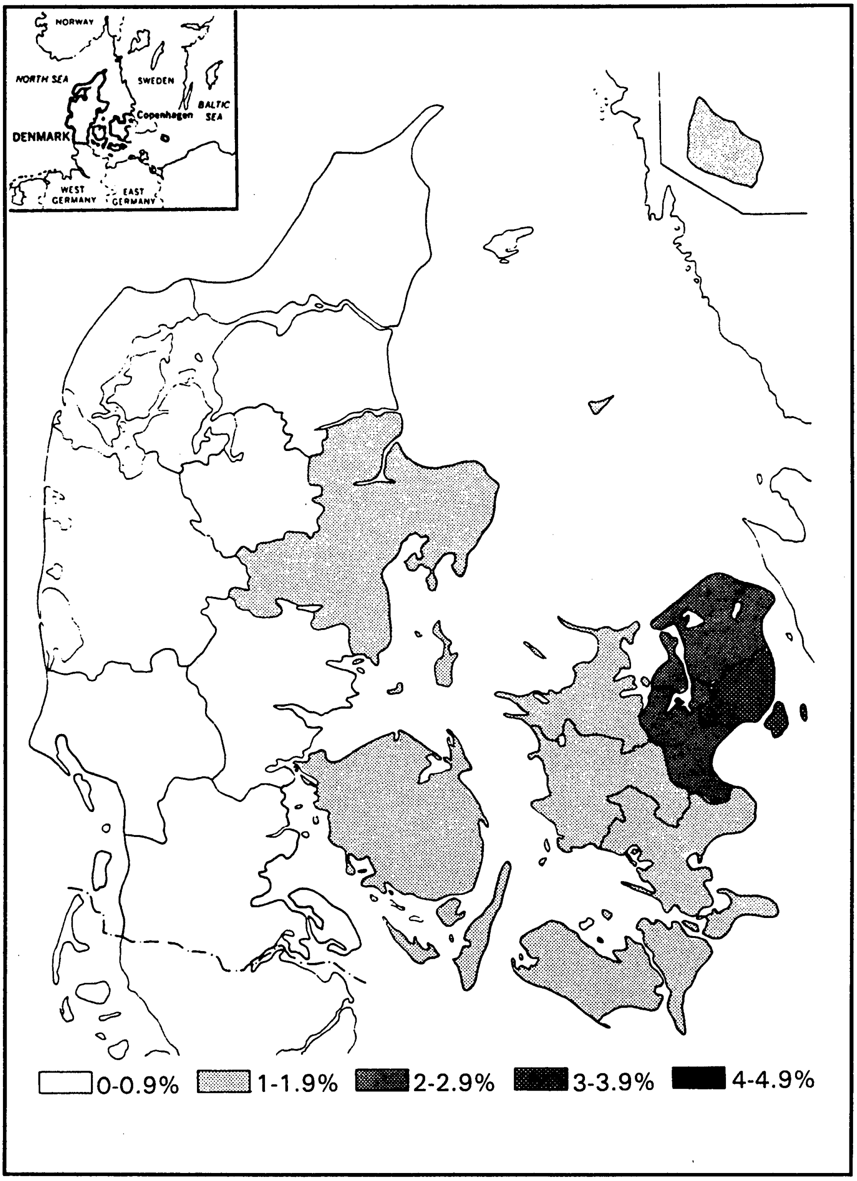

Fig. 1. Regional distribution of the name *Knud* in the year 1900. After the nineteenth-century revival of interest in Scandinavian names, *Knud* became fashionable among towndwellers, primarily in Copenhagen.

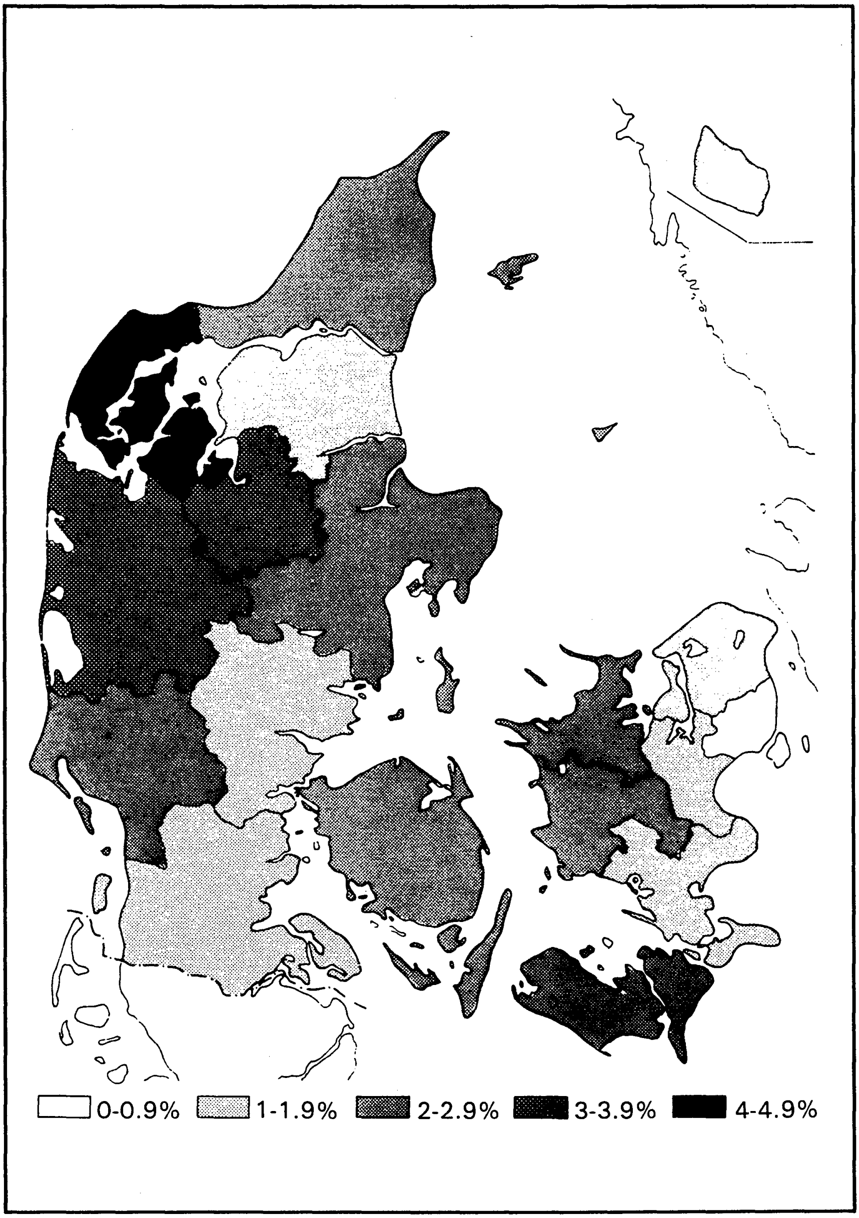

Fig. 2. Distribution of *Knud* in 1947, the first year it was not used at all in Copenhagen. By then it had become frequent in Northwest Jutland, with a use equalling that in Copenhagen in 1900.

A Sample Project: Geographical Distribution of the Surname *Hansen*

Once the total occurrences of a name in the corpus are counted, its relative distribution is easy to determine. Figures are also available for total populations in the various districts, allowing us to calculate the proportional frequency of a name within any locality of interest, nationwide, within one or more of the twenty-four counties (as in Fig. 3), or even by individual parishes.

Fig. 3 maps regional distribution of the name *Hansen*. Frequency is expressed as a deviation from the name's normal distribution. Two components comprise the deviation index: first, the frequency of the name among the total nationwide population (in the case of *Hansen* 5.9286%); and second, the proportion of all bearers of the name living in the region of interest. The percentage thus calculated is its normal distribution, or 1.00. A distribution index of .49 for some area means that the name occurs there at nearly one half its frequency overall. A rate of 2.00, on the other hand, indicates twice the normal frequency.

The map is based on figures which show that 17.4% of all Hansens live on the island of Funen whereas the corresponding number for North Jutland is 5.1%. The explanation is that the personal name *Hans*, from which, of course, the patronym *Hansen* is derived has been for many hundreds of years the most common male name on the island. Although the name is not particularly common in the overall Danish population, its bearers are located with twice the expected concentration in the island of Funen (see Table 1).

Such statistics may confirm our life-long impressions, or they may reveal facts hitherto unsuspected. But the availability of statistics serves to turn speculation into science. We cannot emphasize strongly enough

Table 1. Distribution of *Hansens.*

	Percent of Total Danish Population	Percent of All Hansens	Distribution Index
Southern Funen	2.49	5.08	2.04
Copenhagen area	26.28	24.28	.92

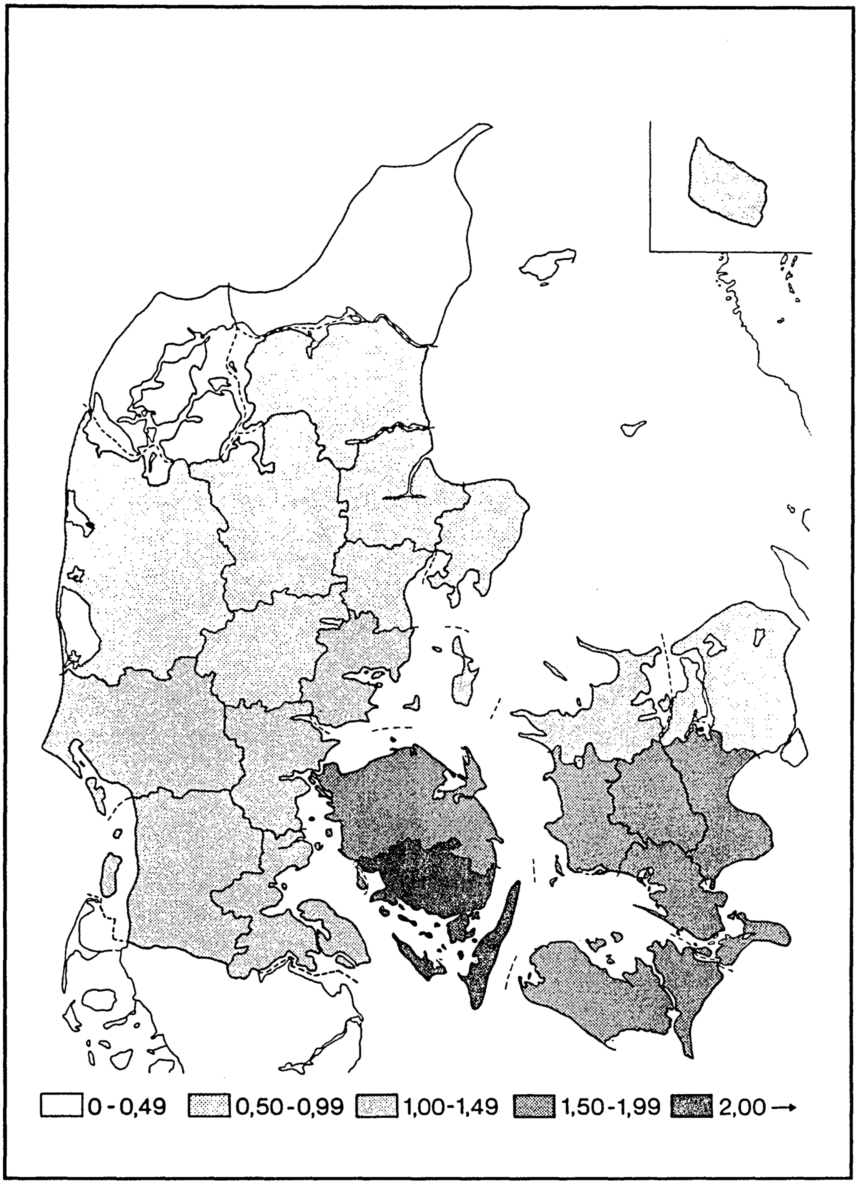

Fig. 3. The distribution of the surname *Hansen* shown in a fine-meshed network corresponding to the division of Denmark into counties used in Denmark until 1968. The figures show the distribution index for each county as explained on p.26.

that computerized storage and processing are absolutely necessary to handle the quantity and scope of the data. We have only begun to investigate the extent to which computer applications and statistical methods can be effective against certain intractable Danish onomastic problems.

Of course, not every problem lends itself to quantification. The use of personal names has many complex psychological and sociological aspects that we cannot define precisely. Analysis of the objective data may throw new light on an underlying "reality," though we must always remember that the analytical process itself introduces subjective factors (cf. Øyslebø). Indeed, the original decisions about which data to analyze are matters of interpretation.

Work on an Inventory of Danish Surnames

One of the first results from the new CPR material will be a much-needed register for the entire Danish surname corpus, containing information about the frequency and geographical distribution in Denmark of the individual names. Only a very few similar registries exist elsewhere; e.g. the Swiss *Familiennamenbuch der Schweiz,* the Belgian *Répertoire belge des noms de famille,* and the largest and most exhaustive *Nederlands repertorium van familiennamen.* This last, based on a 1947 census, provides in several volumes surname frequencies for each Dutch province and borough. Thirteen volumes have been published to date. This unique material has attracted a rich supporting literature over its twenty years of existence – a condition which we would hope to duplicate in Denmark. We see growing interest here, among both researchers and general public, in subjects such as recent population migration as well as sociological and genealogical studies.

As planned, the main part of the Danish book consists of a list of the ten thousand most common surnames, with information about their occurrence in thirteen geographic regions, a division that by and large follows the present Danish county districts. For those listed names which show a characteristic regional distribution, a separate section gives more details, in relation to population density. A third section lists all of the approximately 64,000 surnames found in the country. The book, a quarto of about five hundred pages, includes an introduction covering the history of surnames in Denmark and the results of modern surname research. The future publication of further studies is anticipated.

Processing the Data for Publication

The basic register of six and a half million people is available in raw form, entirely unanalyzed. In addition, various kinds of preprocessing have been performed to make the data more useful for expected onomastic purposes. There are sortings by surname (including maiden names), with first names removed; by parish code; by frequency, with nonce names removed—about 64,000 names are listed in this form; by births per parish to provide totals; and by geographic regions of six different sizes, thereby enabling analyses to several degrees of fineness.

The Special Problem of Middle Names

Almost all boys' pre-names occur as surnames too. When these appear as second, or middle, names, how should they be classified? In the data, a correction factor determines the answer, based on the frequency of each kind of use. It is clear that computerization greatly facilitates the calculation of these factors.

Conclusion

Research into personal names occupies a challenging borderland between the classic philology and history and more recent fields such as sociology and psychology. We suppose that the most frequent applications of the new CPR material will concern relatively uncomplicated studies of change: frequencies will be counted, distributions mapped, and chronologies compared. Then, built on these basic analyses, programs of interpretation will develop. Demographers will be able to investigate mobility patterns and regional attachment, especially evident with relatively rare names, often characteristic of only one region of the country. Their work can be supported quickly and thoroughly, at any appropriate level of detail. The support will constitute statistical documentation rather than impressionistic hypothesis. We can hope for unexpected clarity in some hitherto dark corners of Danish surname study.

Odense University, Odense, Denmark

Acknowledgment

I am very much indebted to Associate Editor Marion Harris, Bell Communications Research, for her cooperation in adapting my paper for the readership of *Names* and giving my paper a very congenial presentation.

Works Cited

Jodogne, Omer. *Répertoire belge des noms de famille.* 2 vols. Louvain, Belgium: Editions Nauwelaerts, 1956-64.

Meertens, P.J. *Nederlands repertorium van familiennamen.* 5 vols. to date. Assen, the Netherlands: Van Gorcum, 1963-67.

Pedersen, Birte Hjorth, and Lis Weise. *Fornavnebogen: 12500 navne på danske statsborgere i 150 år.* Copenhagen: C.A. Reitzel, 1989.

Søndergaard, Georg. "General Outline of a Computational Investigation of Danish Naming Practice." *Onoma* 23 (1979): 1-32.

___. "Navne i Odense gennem 200 år." *Profiler: Nordisk institut 1966-86.* Odense: Odense universitet, 1986. 183-96.

___."Egnskarakteristiske slægtsnavne i Danmark." *Studia Anthroponymica Scandinavica* 4(1986): 103-23.

___. *Oversigt over efternavne i Danmark: Tabel over forekomst og regional fordeling af de 10.000 almindeligste efternavne i Danmark. Odense: Odense universitet, Nordisk institut, 1987.*

Øyslebø, Olaf. *Stil- og språkbruksanalyses.* Oslo: Universitets forlaset, 1978.

*****

CHRISTOPHER COLUMBUS 1492-1992

October 12, 1992 will mark the 500th anniversary of Columbus' landfall in the New World. *Names* will join the rest of America in recognizing this significant occasion with a special issue (September 1992) devoted to onomastic topics relating to Columbus and his legacy. Probable deadline for submission of papers: November 1991. For further information contact

Thomas J. Gasque, Editor
Names
Department of English
University of South Dakota
Vermillion, SD 57069