Some Methodological Issues in Quantitative Onomastics

SASHA WEITMAN

I have been engaged for some time now in an elaborate research project devoted to the statistical analysis of the first names Israelis have been giving their children over the course of the last one hundred years.* (The first wave of returnee-settlers arrived back in 1882 in what was then Turkish Palestine.) The source on which this research is based is the Population Registry maintained by the Interior Ministry of the State of Israel. In the course of preparing this vast data-set for statistical analysis I have been confronted by various methodological problems and dilemmas, some of which are general and instructive enough to seem worth sharing with onomastic specialists at large, especially with those who have an interest in the quantitative treatment of names.

A few words about the rationale for this study before I launch into the methodological discussion proper. I undertook this research out of an interest not in first names as such, but in what first names can reveal about the people who select them to identify their children. What is more, my interest is not in the individual name-givers, but in the larger social formations to which they belong, such as social groups (classes, ethnic groups), social categories (e.g. gender), and generations. I am an historical sociologist, and my principal objective in this research is to establish on firm objective grounds those continuities and changes across generations which can be revealed by an analysis of over-time trends in patterns of name-giving. To be sure, the study of first names, even if carried out at the aggregate level, can only reveal a part, perhaps even only a small part, of the cultural characteristics of generations. But at least the information so obtained will be firm and incontrovertible, thus providing a clear factual basis from which to either challenge it or build on it. This is more than can be said for the informational bases on which

*This research was supported by the Pinchas Sapir Center for Social Development, Tel Aviv University.
My thanks to Mr. Motti Regev for his research assistance, and to Mr. Yuval Fichman for his computer programming assistance.

many learned discussions about generational changes and continuities have been conducted.

The methodological issues which I have selected to discuss in this paper are the following:

The first question concerns the type of primary source on which the quantitatively-oriented researcher of names should rely: a partial source (such as a *Who's Who* listing, a municipal telephone directory, a list of class graduates), or a comprehensive source (such as official birth registries, population censuses).

The second issue concerns whether to work with the entire Master File of names thus compiled, or with a representative sample of that file; it turns out on analysis that this apparently methodological issue resolves into a more substantive question, that of whether to limit the analysis to only those names which appear with a relatively high degree of frequency, or whether to cover all the names used by the population, including those used very rarely.

Many names — thousands and thousands of them — constitute spelling variations, or grammatical derivatives, of foreign or regional versions, or phonetic similes, of the more familiar forms of those same names. The question is where to draw the line between those variants which can be safely merged with their standard versions, and those which should be treated as distinct names in their own right.

The quantitative treatment of names raises the hoary issue of which statistic to use for purposes of inter-group and over-time comparisons — raw frequency counts, population-standardized proportional frequencies, or frequency-based rankings?

Finally and most importantly is the question of just what kinds of culturally meaningful information can be extracted from first names, and by what means. The section devoted to this question will outline the method of *symbolic decoding,* by means of which names are, as it were, de-constructed into their significant components, and thereby prepared for the kind of quantitative operations which Harold Lasswell tried to coin nicely as ''statistical semantics''.

I. PARTIAL vs COMPREHENSIVE SOURCES

The first methodological issue which confronts the statistically-oriented researcher of names concerns the choice of an appropriate primary source of onomastic data. Roughly, the choice is between, on the one hand, *partial sources,* such as lists of school graduates, telephone directories, *Who's Who* registers, lists of elected officials, and the like, and, on the other hand, *comprehensive sources,* that is, official government records covering virtually the entire population, such as birth registries, tax rolls, military draft lists, population censuses, and the like. What are the advantages, the disadvantages, and therefore the most appropriate uses of each of these two types of sources?

The outstanding advantage of partial sources is their ready availability: access to them is usually free of charge and there are no legal or bureaucratic obstructions to their use. As for their disadvantages, first, they can be costly and time-consuming to transcribe into machine-readable form. Second, and more importantly, they cannot well serve as bases for generalizing to the general population or even to some meaningful segment of it. (What can be made of the fact that, say, 15 per cent of the parties listed in the Tel Aviv-Jaffa telephone directory bear Biblical first names?)

As for the comprehensive sources, their outstanding advantages include the following: (a) they are readily transcribable, effortlessly and errorlessly, into the researcher's own electronic storage facility (e.g. magnetic tapes); (b) since registration into these sources is mandated by law, their coverage of the population is virtually complete; and (c) valuable additional bits of information are usually recorded alongside each name, thus making it possible for the researcher to classify and compare names according to such criteria as sex, year of birth, geographical origin, social class membership, etc. The disadvantages of comprehensive sources are, first, that gaining access to them can be quite trying, even for the most bona fide academic researcher with the most unassailable credentials. In my case, for example, no less than 8 (eight) months elapsed between the time I received the informal O.K. by the Deputy-Minister responsible for the Population Registry to use this source of data, and the time I could at last pick up the computer tapes onto which the registry records had been transcribed . . . The second disadvantage of working with comprehensive sources is that they confront the researcher with such a massive quantity of data that he soon finds himself beset by all sorts of problems for which he is as unprepared temperamentally as he is ill-equipped methodologically. Many of the issues taken up in the pages which follow stem directly from the challenges that arise in the course of working with a massive comprehensive source such as the Population Registry.

Which of these two kinds of sources should the researcher favor? The answer, of course, depends on the main objectives of the research. The use of partial sources seems to be called for (a) when the research is undertaken as a "pilot study", e.g. for the purpose of exploring the feasibility or the profitableness of undertaking a much more ambitious piece of research, based on a comprehensive source, and (b) when the objective of the research is mainly classificatory, that is, for the purpose of establishing a nomenclature of names.

When, however, the objectives of the research are mainly *statistical* — that is, when the object is to use onomastic statistics as quantitative indicators of differences between groups or of changes over time — then partial sources are virtually useless, mainly because of the unknown composition of the population on which they are based. For statistical purposes, it is imperative to base the research on comprehensive sources, whatever the difficulties involved in such a choice.

II. WORKING WITH THE MASTER-FILE OR WITH A SAMPLE, OR WORKING WITH ALL THE NAMES OR ONLY WITH HIGH-FREQUENCY NAMES

Suppose the researcher has decided (as I did) to acquire a comprehensive source of onomastic data, and that he has had it electronically transcribed and stored at his own computer facility. Faced now with this overwhelming mass of data, numbering in the hundreds of thousands if not in the millions of cases, our researcher is immediately and somewhat rudely pressed into making a most strategic choice: to work with the entire Master File, or to settle for working with a randomly drawn sample of that file. Which should he choose?

In the last analysis, of course, the answer depends, as always, on the objectives the researcher has set for himself. In my own case, however, knowing that my objective was to use first names as cultural indicators did not provide me with a ready answer to see myself clear out of the dilemma. Some of the benefits to be gained for working with a sample of manageable size, as opposed to working with the Master File, were quite obvious at the outset. But what were the hidden costs of such a decision? What research options would I thereby be foregoing? Worse still, might such a decision systematically bias my results in one way or another? To explore these questions, let us discuss in general terms the pros and cons of working with a sample vs working with the Master File.

What are the arguments, first of all, against working with the Master File?

First and most obviously, the larger the data file, the more difficult it is, technically, to process it efficiently, even by means of high-speed computers equipped with large memory capacities, and with the assistance of skillful programmers.

Second, the larger the file, the larger the number of things that can go wrong along the way, and usually do. The upshot is a seemingly never-ending plague of snags and wastages of valuable time.

Third, the larger the file, the greater the computer-related costs, including costs of machine time, costs of expert programming services, and costs of material supplies — in the case of my own research I was required to purchase no less than 25 computer tapes. . .

Fourth, and because of what precedes, the larger the file, the more it requires of the researcher that he plan and design every computer run most elaborately, so as to obtain the very maximum that can be squeezed from that single run. This elaborate planning, besides being quite time-consuming, significantly reduces the researcher's opportunity to interact with his data, that is, to adjust his operations and analyses in response to the results he obtains as he goes along.

By contrast, a relatively small file, such as could be had by drawing a sample from the Master File, can be handled more efficiently, more playfully, at much lesser expense, by making use of already existing library packages of data-processing programs, such as SPSS (Statistical Package for the Social Sciences). Modern computer-aided sampling methods, using random number generators, are extremely reliable in constructing representative samples according to the exact specifications set by the researcher.

Given this impressive array of arguments, what arguments could there be against working with a sample? There is only one such argument, but it is a weighty one. It is based on the proposition that the probability that any given name will appear in the sample is directly proportional to the frequency of appearance (i.e. the "popularity") of that name in the population. It follows that, in the last analysis, the decision to working only with popular names, and to ignoring all the other, less popular and unpopular names. This would seem a tolerable risk to take, on condition that unpopular names constitute but a minority of all names. What if, however, there is a very large number of such names? Can we still afford to ignore them for the sake of methodological convenience?

The question before us, then, is an empirical one: how many unpopular names are there relative to the number of popular names? Fortunately, in the case of my own research, I have managed to locate a rough quantitative answer to this question. Thus, two years ago, the programming staff of the Population Registry produced a detailed frequency distribution of the first names of all the persons listed in the Registry[1], indicating for each interval of this distribution (a) the number of names in it, (b) the proportion these names constitute over all first names, and (c) the proportion of the population (in the Registry) bearing such names. Table 1 below is based on a recomputation of

[1] Note that these figures are based on a data-set much larger than the one on which my own study is based. First, it includes all the cases in the Registry — some 4.6 million of them — while mine includes "only" those 2 million Jews who were born here. Second, the number of distinct names indicated by these statistics is inflated, because the Registry researchers made no attempt to consolidate under their standard version the many variants of each name. (More on this subject in section 3 below).

these figures, where the frequency distribution has been collapsed into 4 intervals.

The figures in this table are nothing short of spectacular. What they show is that high-frequency names, defined here as those borne by more than 1,000 persons, account for less than 1 per cent of all names, whereas rarely used names, defined here as those borne by 10 or fewer persons, account for a little over 90 per cent of all first names! Even more startling is the finding that this tiny one per cent fraction of frequent names has been borne by nearly three-quarters of the entire population (by 73 per cent to be exact), while the overwhelming majority of rare names have been borne by only 3 per cent of the population. The reader is also asked to note the absolute figures involved: there are 658 frequently borne names, and 91,774 rarely borne names. . .

Coming back now to the issue at hand, we see that the question of whether to rely on a sample or to work with the entire Master File boils down, in the last analysis, to whether or not the researcher is prepared to ignore infrequently borne names, even though these constitute an overwhelming majority of all first names.

At the time when I had to make this decision — a time when I had not yet begun to imagine the difficulties, frustrations, and delays which lay ahead — I took the decision to work with the entire Master File. My reasoning at the time was that perhaps, just perhaps, a key to understanding the differences between successive generations, or different groups, or social categories, lay precisely in the mass of first names which they innovate, and most of which never do "take off", rather than in those names which are the most popular. This was but a hypothesis, to be sure, but it was a hypothesis which I found intriguing enough not to want to commit myself to a strategy, that of sampling, which would rob me of the opportunity to test it.

In retrospect, the armed with the proverbial wisdom of hindsight, I now realize what my mistake was. My mistake was in thinking that I had to opt between these two alternatives. Instead, I should have simply planned the research to proceed in two successive stages, the first based on the sample, the second one based on the Master File. in this way I would have reached much earlier the stage of obtaining results and analyzing them, rather than spending well over a year in protracted efforts in debugging, reblocking, and otherwise preparing the Master File for analysis. The real danger in proceeding as I did is the danger of burning oneself out before reaching the pay-off stage. This danger is very real, and researchers should be wary of it.

III. BONA FIDE NAMES VS VARIANTS

A problem which may seem trivial — and actually would be were it not for the fact that it requires weeks of painstaking work to bring it under control — stems from the presence in the Master File of numerous names which closely resemble one another, or are closely related to one another. As far as the computer is concerned, it goes without saying, all these names are treated as distinct. Given that there are, literally, thousands and thousands of such variants, one of the first tasks of the researcher is to streamline his list of names by instructing the computer to merge the variants with their standard versions. To do this, the machine must be provided with a lexicon of variants. It is in constructing such a lexicon that the researcher is forced to confront the problem of exactly where and how to draw the line between those variants which may be safely ignored and merged with their respective standard versions, and those variants which should be considered as names in their own right.

Before resolving this problem, it must be taken into account that there are several types of name variants, depending on how they were produced. The main types I found in the Population Registry of Israel are the following:

(1) Legitimate spelling variants, allowed for by the rules of orthography of modern Hebrew. Most of these variants are cases of names spelled in their ''expanded'' or in their ''contracted'' versions (ktiv malé vs. ktiv chaser). Other subtypes of variants in this category stem from ambiguities in the transliteration into Hebrew of foreign names.

(2) Spelling variants stemming from errors in spelling — errors on the part of parents, but also and especially errors on the part of the recording clerks in the maternity hospitals who fill the birth certificate forms which are then transmitted to the Interior Ministry and the Population Registry.

(3) Spelling aberrations stemming from mechanical errors in transcription by clerks or by keypunch operators. This problem is particularly aggravating in the case of Hebrew, since only the consonants appear in the written version of the language, while the vowels are left for the reader to fill in mentally. (The name DAVID, for example, would appear DVD.) As a result, and especially in the case of short names consisting of 5 consonants or less, a spelling aberration often produces what I have called a Weirdonym. (A Weirdonym is to the researcher of names what an UFO is to the U.S. Air Force.) Given the ethnic diversity of the population of Israel, any given weirdonym could be a keypuncher's aberration of a familiar name, but it could also be a bona fide first name unknown to the researcher, e.g. a Kurdish nickname. Eventually, and at the cost of many hours of detective work, hundreds of aberrations were rescued from the list of weirdonyms and included in the lexicon of variants.

(4) Foreign variants, i.e. variants resulting from the various ways in which a name is spelled or pronounced in different languages. Examples: Robert vs. Roberto, Solomon vs. Shlomo.

(5) Morphological variants, i.e. variants resulting from new names being derived from older names by means of the addition of a suffix, or of a prefix, or by some other morphological alteration. Examples are diminutives and nicknames such as Sarah vs. Sarala, Tsachi vs. Yitschak, Udie vs. Ehud and the like.

(6) Middle names or initials, i.e. variants which result from the presence in the Maste File of "middle names". Formally, there are no middle names in Israel, only "private names" and "family names". Accordingly, the birth registration forms do not provide a space for recording middle names, even if the parents want to give such a name. This formal restriction has not prevented, of course, a minority of Israelis — a minority in 2 million cases amounts to several tens of thousands of cases! — from squeezing in anyway a middle name into the space reserved for first names. These middle names do appear in the Master File after the first name, either in parentheses, or linked by a dash to the first name, or separated from it by a single blank space, or reduced to their first initial letter(s). As far as the computer is concerned, of course, each of these instances is regarded as a distinct first name "in its own write", to use an expression of the late John Lennon.

(7) Truncated names, i.e. variants which have been systematically truncated down to 7-letter names, and on occasion even to 5-letter names, compliments of the programming staff of the Population Registry, for reasons known only to them. In a number of instances the original names can be reconstructed by a simple visual inspection (e.g. Trumpel is clearly a truncation of Trumpeldor). In others, they produce yet another source of weirdonyms.

Let us return now to the problem with which we started, that of whether to consolidate the variants of a name under its "standard" version, or to treat each of them separately. The advantage of consolidation is that it simplifies and streamlines the repertoire of names. Its disadvantage, of course, is that, crudely conceived and executed, it systematically biases the onomastic repertoire in favor of the standard versions of names, at the expense of those finely nuanced variants which are deliberately produced by the name-giving public.

In my own research, I resolved this problem by means of a strategy of *partial consolidation*. Given that my interest is in names regarded as sign-vehicles, as symbol-concentrates, and that my objective is to extract from them all the public meanings that are condensed in them, it follows that the task of consolidation must be limited strictly to those variants, and only to those variants, which constitute *semantically trivial deviations from their standard*. The work of consolidation, therefore, applies "only" to those variants — there are nearly 10,000 of them! — which can be reasonably attributed to types 1 (legitimate spelling variants), 2 (spelling errors), 3 (errors in transcription), 6 (presence of middle names or initials), and 7 (computer truncations). As for those variants which carry a semantically-relevant difference from the standard version, however minor, such as those covered by types 4 and 5 above, they have been treated as distinct names in their own right.

IV. THE CHOICE OF AN APPROPRIATE QUANTITATIVE INDICATOR — SIMPLE FREQUENCIES? PROPORTIONS? RANKS?

The commitment to treat names quantitatively confronts the researcher with the problem of which of the following measures to rely on if he wishes to compare, say, successive periods in terms of their onomastic characteristics. Should he use simple frequencies (F), percentages (P), or ranks (R) based on relative popularity? Let me define these terms in the context of my own research, using the name David as an example, and the symbol B to stand for the number of births (in this case, male births) in any given year:

F = the number of male infants born that year who were given the name David

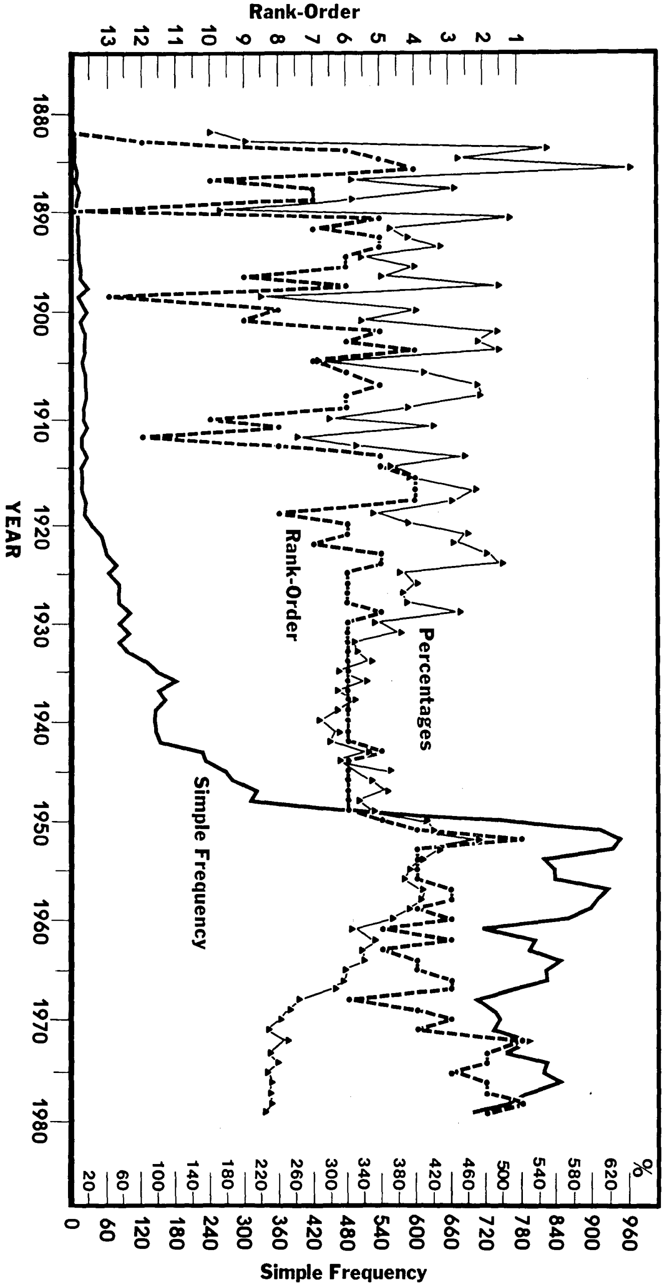
P = 100 (F/B). That is, P is equal to F (above) expressed as a percentage of B, the total number of male births recorded that year

R = the rank of the name David in relation to that of all the names in use that year

The question before us, to repeat, is which of these measures to use in a comparative-longitudinal study of names.

Concerning F (simple frequencies) it may be said a priori that it constitutes a relatively poor statistic because it is substantially confounded with population size. Thus, an increase over the years in the yearly number of infants in Israel who are given the name David *may* reflect an increase in the popularity of this name, but first and foremost this increase reflects simply a corresponding increase in the size of the population in this country over those same years — or, to be precise, it reflects an increase in the number of infants born each year. (Note, incidentally, that while simple frequencies are of no great use in the kind of research envisaged here, they may have direct *practical* use, for example, to help set production quotas for manufacturers of name-bearing personalized items such as I.D. bracelets, key-chains, necklaces, coffee mugs, birthday greeting cards, and the like.)

Next let us consider Proportions (P) and Ranks (R). Both of these have the double advantage of constituting familiar ways of expressing popularity on the one hand, and of being unaffected by population size on the other. Which of them should the researcher use? Initially, my inclination was to prefer Proportions over Ranks, because ranks can be derived from proportions, but


Rank-Order
1
2
3
4
5
6
7
8
9
10
11
12
13
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
YEAR
Percentages
Rank-Order
Simple Frequency
%
20
60
100
140
180
220
260
300
340
380
420
460
500
540
580
620
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
900
960
Simple Frequency

not vice-versa, and also because proportions are expressed in cardinal numbers and are thus amenable to arithmetic operations such as adding and subtracting, while this is not the case with ranks since these are expressed in ordinal numbers.

Nonetheless, before committing myself one way or the other, I decided to make a visual inspection of the actual *behavior* of each of these indexes over time. For this purpose, I graphed the name David, and the over-time trajectories I obtained for F, P, and R are reproduced in Figure A.

The first thing to note is that, roughly through the first quarter of the 20th century, both the P and R measures fluctuate rather wildly from year to year, while the corresponding yearly changes in F are quite moderate and monotonoüs. It is only after the 1st quarter of the 20th century that the P and R curves become smooth and behave clearly and coherently.

The second thing to note concerns the period covering roughly the last 25 years. Here we see one index, R, showing that David has been becoming increasingly popular with the population, climbing steadily from 6th place to 1st place, while the other index, P, shows the same name becoming less and less popular over the very same period of time!

The resolution of both these apparent paradoxes lies in a factor which I had originally neglected to take into account, namely, the changing size of the yearly *repertoire of names*. Thus, on reflection it turns out that:

(a) when the repertoire of names is relatively small, as it was up to the end of the first quarter of this century, even small fluctuations in F can (and often do) result in large fluctuations in P and R; and

(b) as the repertoire of names becomes increasingly large, the proportional share P of any given name is liable to decline, because of the ever-growing number of other names with which it has to compete, as it were, for the favors of the public. This can and does occur even in those cases, like that of David, when a name is clearly becoming more and more popular with the public, as indicated by its steady climb up the rankings chart.

What are some of the operational lessons to be drawn from the above?

The first lesson is that for each year it is necessary to record not only B, the number of births, but also the size of the name repertoire[2].

The second lesson is that both P and R are relatively useless measures of onomastic popularity for those periods when the repertoire of names is small.

[2] A distinction should perhaps be drawn between the [3]repertoire of names in-use, composed of all the names presently in use at least nce by a given population in any given year, and the *repertoire of names ever-used*, composed of all the names above plus all those which had been used in the past but have since fallen into desuetude.

In my own research, this has led me to graph and to investigate the behavior of P and R only after the first quarter of the century.

Third, and most importantly, regarding whether to rely on P or on R as my measure of onomastic popularity, I decided to reverse my original tentative decision, and to graph onomastic popularity in terms of R rather than in terms of P. Both measures, we have seen, are functionally dependent on nr, the size of the name repertoire. But when a name, like David for example, is gaining increasing favor with the public relative to other names — which is largely what we have in mind when we speak of its ''popularity'' — the R-curve reflects this well by showing it climbing up the scale of ranks, while the P-curve shows it slipping down the P-scale for the same period.

Fourth, suppose the researcher is interested in tracing the trajectory not of any particular name, but of that of a *class of names*, where the class is defined by some characteristic(s) commonly held by all the names in that class. In that case, the measure to be used is, clearly, P and not R, for P lends itself to arithmetic operations in ways in which R does not. For example, suppose I wish to trace changes over time in names which denote Nature in one way or another, like tree-names, flower-names, lake-names, river-names, star-names, and the like. Combining all such names into an analytical scale, I would then be in a position to compute for each generation an Identification-with-Nature score and thus to monitor cross-generational fluctuations and inter-group differences in this type of identification. An operation of this type *requires* the use of cardinal numbers, hence of the P index, since ordinal numbers (R) do not lend themselves to the above scale-scoring operations.

Let us sum up. For most research purposes, F (simple frequencies) is inappropriate as an index, because changes in its value are more likely to reflect changes in the size of the underlying population than onomastic changing proper. R (the rank-order of popularity) is the most appropriate index for tracing changes in the popularity of individual names. As for P (the popular-standardization frequency), it is most appropriate for the computation of analytical scores, such as Nature-scores, Power-scores, etc., whereby one compares successive generations or parallel social groups and categories along these broad cultural dimensions.

V. FROM FIRST NAMES TO CULTURAL DATA BY MEANS OF SYMBOLIC DECODING

In this section I intend to touch on — since space limitations prevent me from elaborating on — the subject of specifically what kinds of, and how much, and by what methods meaningful information can be extracted validly and reliably from first names. Clearly, I have come to the conclusion that substantial amounts of such information can be culled from names, or else I would not have engaged in this mammoth research project or be writing this article. But, to be of scientific value, three important sets of conditions must be fulfilled.

First, the objects of cultural inference must be very clear to the researcher. Thus, inferences from names must be to the givers of these names, *not* to their bearers. What is more, inferences must always be to sociological formations (such as social classes, ethnic groups, historical generations, and the like), *not* to individual name-givers.

Second, and closely related to the first condition, the researcher must be committed to working with large numbers of names. It is when he is confronted with the multifarious colorful legions of names produced and worn by real populations, and only then, that he can begin to appreciate the extraordinary wealth and variety of cultural information which is buried in these seemingly trivial materials.

Third, the method for extracting culturally meaningful information from names must be rigorous, methodical, comprehensive, and explicitly spelled out down to the most minute detail, even if at the risk sometimes of sounding as though belaboring the obvious. In this kind of research, when the materials are so extraordinarily suggestive and give such wings to the imagination, that the researcher must be prepared, not to shackle his imagination, but to jettison those of his ideas which he cannot translate into operational terms so explicit that nay other competent researcher of names can apply them to the same materials and come up with virtually the same results.

I now turn to a brief description of the method I developed for the exploitation of names for purposes of cultural analysis.

The Method of Symbolic Decoding. The translation of communications materials into social science data is usually done by means of content analysis. Actually, however, content analysis was developed for the purpose of extracting social science data from *discursive flows of verbal communication*, e.g., speeches, editorial columns, newsreels, propaganda programs, even entire novels. In all these instances, content analysis serves to extract from these rich flows of text only those items which the scientist requires for his research. There is, however, a whole other category of communications materials, whichmay be called *poetic*, whose distinctive property is that, like poems, they manage to pack and to deliver rather considerable amounts of information in relatively few words. First names are prime examples of this category of "poetic" forms of communication. (Others are advertising slogans, icons, emblems, flagts, etc.) For these materils, the method for translating them into social science data has, in one sense, the opposite task of that of conventional content analysis: rather than *reduce* the flow of text down to a necessary minimum, the task of the method in this case is to, as it were, explode every symbol with a view to recording all the distinct culturally standardized meanings that are condensed in it.

I call this method *symbolic decoding*. Applied to Hebrew first names, it consists of systematically extracting and recording information for each of the following aspects of the name:

(a) *its denotation(s)*: most Hebrew first names denote one referent, and in a substantial number of cases more than one of them.[3]

(b) *the connotations of the above denotation(s)*: here I have in mind the culturally standardized meanings we associate with the denoted referent. *Alon*, for example, denotes an oak tree, and an oak tree in turn connotes steadfastness, towering height, etc.

(c) *its morphological characteristics*, including its grammatical form (e.g., possessive, diminutive) and the form of its construction (e.g., verb + suffix ''el'', meaning God). Morphological characteristics are recorded because they too are liable to express discernible cultural values and properties. For example, girls' names ending with the it-suffix like Sarit, Ronit, Sigalit, Vardit — may connote, in addition to the feminine gender of their bearers, a diminutive cuteness not unlike that communicated by the corresponding French suffix – *ette*, as in Georgette, Linette, Annette, etc.

(d) *its phonetic characteristics*, such as whether the name is monosyllabic or bisyllabic (communicating simplicity, terseness, and informality) or polysyllabic (communicating, elaborateness, discursiveness, and the like); whether it contains certain tell-tale combinations of consonnants and of vowels.

(e) *its relation to foreign names*, such as whether it is homonymous to certain foreign names, like (Tom), (Shirley), (Roy), and the like, thereby expressing possibly a certain Anglo-Saxon-oriented cosmopolitanism.

(f) *its source*, which may be Biblical, Mishnaic, Medieval, and Contemporary in the case of Hebrew names, or may be from a foreign onomasticon, or from one of the dialects spoken by Jews in the Diaspora (Yiddish, Ladino, etc.).

The concrete product of this symbolic decoding operation is a comprehen-

[3] Unlike most of the first names used in contemporary Europe and in the Americas, Hebrew first names are unquestionably denotative. Their meanings are known to those who are competent in the use of the Hebrew language, while the less competent can look them up in a dictionary. Even when a particular name is chosen for ''they way it sounds'' more than for what it means, its denotative meaning is generally known.

sive dictionary of Israeli first names, where for each name are recorded, in print as well as in an electronic, machine-readable archive, its spelling variants, its various denotations, its connotations, its morphological characteristics, its phonetic properties, whether it is homonymous to foreign names, and its sources.

RECAPITULATION AND RECOMMENDATIONS.

First, onomastic research with a genuine quantitative orientation should draw its data from a comprehensive source. Partial sources are inappropriate for quantitative objectives because the composition of the population from which they are drawn is unknown.

Second, concerning the decision of whether to base one's research on the entire Master File or on a random sample of that file, it was shown, on the one hand, that working with the entire Master File is so rife with technical difficulties, costs, and delays that the researcher runs the real risk of running out of steam, so-to-speak, before even reaching the stage at which he has any results to analyze. On the other hand, working with a sample of that file, though operationally more convenient and effective, is tantamount to limiting one's research to those names only which are relatively popular with the population, even though such names represent but a small minority of all the names in use, and even though they are not necessarily the most tell-tale names. One reasonable solution to this dilemma is to plan the research to proceed in two overlapping stages, the first to be devoted mainly to working on a sample of the Master File, the second to be devoted mainly to preparing and analyzing the Master File itself.

Third, concerning the myriad names that are variants of other, more standard versions of these names, my recommendation is to merge with their standards only those variants which represent symbolically trivial deviations, that is, deviations which the researcher may safely ignore without risking the loss of any symbolic information.

Fourth, concerning the choice of an onomastic index for longitudinal and for cross-sectional comparisons, it has been suggested that raw frequency counts (C) are virtually useless as an index for research purposes, that rank-orders of popularity (R) are best suited for researching individual names, and that proportional frequencies (P) are best suited for the purpose of constructing analytical scales and assigning numeric scale scores to the populations under examination.

Fifth, I have suggested and outlined a method of symbolic decoding of names for the systematic extraction of the culturally standardized meanings that are condensed in them: this method consists of recording for each name its denotation(s), its connotation(s), its morphology, those of its phonetic properties which are symbolically-loaded, and its source(s).

In closing, I should like to add a few practical, indeed mundane, recommendations to the prospective quantitatively-oriented student of names:

(a) It is advisable to have a research assistant, but in selecting one, be sure that his or her anal compulsiveness is at least as high as your own, and preferably higher than your own. Otherwise you are liable to end up feeling as though you have yet another albatross hanging from your neck.

(b) In addition to suitably large computer facilities, be sure you also have at your disposal large amounts of computer time, a sufficient number of magnetic tapes and, most importantly, expert programming assistance for the processing of enormous data sets.

(c) It is advisable to keep in frequent touch with colleagues from neighboring disciplines. In my own case, for example, it was helpful to discover, thanks to Professor Elhanan Helpman from the Economics Department, that the problem discussed in section IV of this paper is homologous to the much belabored problem of how to measure such thigs as the competitiveness of a given market, the turnover of firms, the performance of firms over time, and the like.

(d) Finally, I have found it helpful in more ways than one to have another research hobby-horse to ride when I was experiencing serious delays and difficulties in the onomastic research.

Tel Aviv University

Table 1. FREQUENCIES OF FIRST NAMES, expressed as (a) absolute numbers, (b) proportions of all names, (c) proportions of the whole population.

Frequency of Appearance	Number of Names	Proportion of All Names	Proportion of All Population
Frequent (1,000 or more)	658	0.65%	73%
Less Frequent (101 to 1,000)	2,308	2.28%	18%
Infrequent (11-100)	7,508	7.42%	6%
Rare (1-10)	91,744	90.65%	3%
TOTALS	101,218	100.00% (N = 101,218)	100% (N = 4,618,948)

Source: Israel Population Registry internal memorandum, gracefully provided by Mr. Nissim Elissaf.