The Oregon File of the Geographic Names Information System: A Macro Corpus Typology

Lewis L. McArthur
Portland, Oregon

The Oregon file of the national Geographic Names Information System (GNIS), containing some 51,500 geographic names, is being classified by type. The first part of this article discusses approximately 9,000 of these names, about which many basic facts are known and have been recorded. The second part covers a few general suppositions about various kinds of names, and discusses how certain of these suppositions compare statistically with the researched corpus. It concludes with some comments on name classification problems and suggestions as to how a large-scale study could be of use to various disciplines.

Comprehensive studies of geographic names in the United States probably start with Henry Gannett's 1905 compilation for the Department of the Interior, *The Origin of Certain Place Names in the United States*. It covers the entire country and the bibliography includes accounts of early major exploring expeditions, many well-known state histories of the time and a few individual state monographs, such as Will Steel's *The Mountains of Oregon* (1890). By the 1920s toponymic publications covering the names of individual states appeared; these included Edmund Meany's *Origin of Washington Geographic Names* (1923) and the first edition of *Oregon Geographic Names* (1928), by the late Lewis A. McArthur. In 1945 George R. Stewart's *Names on the Land* first revealed the variety and complexity of geographic names in the United States. Between 1919 and 1948 literary and social critic H.L. Mencken provided a vast number of toponyms in the several editions of *The American Language* and its two supplements (1936).

In the late 1940s, George R. Stewart and Robert Ramsay, founders of the American Name Society (ANS), held many discussions about the

Names 48.3/4 (September/December 2000):227-232
ISSN:0027-7738

design parameters of a national gazetteer of the United States. Twenty years later, these discussions resulted in the establishment, by the American Name Society, of a committee of scholars to plan and oversee a project to be called the *Place Name Survey of the United States* (PLANSUS). However, the ambitious plans for the PLANSUS gazetteer were never brought to fruition, and in the 1970s, the task was taken over by the National Mapping Division of the U.S. Geological Survey (USGS). The responsibility for developing and maintaining this gazetteer was assigned to Don Orth, who had been appointed Chief of the Office of Geographic Names of the USGS in 1960. The first outline of a national Geographic Names Information System (GNIS) was prepared by Sam Stulberg, and was continued by Roger Payne after 1979.

The geographic names database, which is one element of the National Digital Mapping Database, is designed to include all past and present geographic names, and their variants, used in the United States. Initially, this database comprised all names found on the largest scale USGS topographic maps. After Arizona, Oregon and North and South Dakota were surveyed, the scope was expanded to include U.S. Forest Service maps and the Office of Coast Survey nautical charts. Following this initial phase of data compilation, cooperative programs with most states have added a large number of names found on maps produced by state and local agencies, and on commercial maps, both current and historical. In addition, names found in textual sources, such as books and documents, were added. The Oregon cooperative project added both the omitted Forest Service and Coast Survey maps and charts as well as the complete lists of names of post offices and railroad stations, both past and present.

The current file for the United States includes approximately 2,000,000 names and continues to receive additions, modifications, and changes on a daily basis. There will always be new names applied to previously unnamed features, and as more old maps and documents are studied, names for long forgotten places, along with variant names for existing features, will be revealed. There are also large numbers of administrative names including those of political subdivisions, state, county and city parks, airports, cemeteries, schools, and Forest Service administrative and recreational facilities.

The computer made the national GNIS possible and it has also permitted the segregation and classification of names within the

database. A typical GNIS record includes a unique identification number, the official feature name, the feature type (or class), the county name, latitude and longitude, and a bibliographic reference code. It also includes the name of the USGS topographic map or maps on which the feature is located, and it may also include other fields of immediate importance and interest to toponymic scholars. Names approved by the U.S. Board on Geographic Names will usually have the history of name usage or application, date of first use of the name, and a bibliographic reference for this associated information.

The national GNIS does not, however, address in a systematic nature a matter of great interest to onomasts: a typology of geographic names. In 1928, Lewis A. McArthur, author of *Oregon Geographic Names*, classified the entries in the Oregon names file into five types: 1) descriptive, 2) complimentary, 3) arbitrary, 4) honorary, and 5) unknown. Eight years later, H.L. Mencken, in the fourth edition of *The American Language,* described eight general classes of names: 1) those embodying personal names, 2) those transferred from other places, 3) Indian names, 4) European names, 5) Biblical and mythological names, 6) descriptive names, 7) names derived from flora, fauna, and geology, and 8) fanciful or arbitrary names. Unfortunately, neither of these systems was fully satisfactory: McArthur's was too general and Mencken's mixed application with language. In most studies, linguistic origin is regarded as distinct from the reason for application. In 1954, Stewart proposed a more complete classification. His ten classes were: 1) descriptive, 2) associative, 3) incident, 4) possessive, 5) commemorative, 6) commendatory, 7) folk etymological, 8) manufactured, 9) mistake, and 10) shifts. However, this system is only satisfactory if one knows something about either the origin or the history of the name being classified. All three systems fall short when dealing with the computerized lists of the 2,000,000 or more names in the United States, or even the more than 50,000 names in the Oregon file.

In the 1980s, a committee of the revitalized PLANSUS devised a system that would permit the classification of large numbers of names in the GNIS, irrespective of known origin or history. Compilations based on the GNIS would automatically include the identification number (IDNO), the feature name, the feature type, and the county in which the feature was located. The system also added four additional required elements: type of name (TYPE), how the name is used

(USED), the date of first use (DATE), and bibliographic reference(s) (BIBLIO). The first element, TYPE, is further categorized as biographic (BIOG), physical (PHYS), biological (BIOL), activity (ACTI), coinage (COIN), or miscellaneous (MISC). USED is further divided by primary (PRIM), shift (SHFT), or transfer (TRFR). DATE indicates a record of the first known use of the name and BIBLIO provides a reference or code identifying the source of the information used to classify the name (Smith 1992). Names whose origins cannot be classified with any reasonable certainty are listed as unknown (UNKN).

In Oregon, a second database has been established which includes the four GNIS fields (identification number, feature name, feature class, and county), copied directly from the existing Oregon file, with the addition of the four added PLANSUS fields of DATE, BIBLIO, USED, and TYPE. In addition, another second level field, RELATION, has been added in the TYPE element to further define the type of name.

The Oregon file has some 51,500 names, or records, not including variants, such as Chemeketa, a Calapooya name for the site of Salem. Approximately 3,000 names contain modifiers such as North, South, Middle, Big, or Little, added to the name of the primary feature. For fifteen years, Oregon has been adding names from many sources not found on current maps, and it is estimated that the state file now contains 85% to 90% of the probable total. If one disregards the 3,000 adjunct names mentioned above, there are 48,500 records to be classified. This figure includes some duplicate records that are being eliminated, so that the final study group will likely comprise some 48,000 names. In addition, there are 5,300 variant names left for later study.

The text of *Oregon Geographic Names* is stored in digital format, with constant additions, revisions, and corrections. It now has 6,100 entries, with an additional 2,600 names embedded in the text of these entries. All 8,700 names have been classified according to TYPE, USED, and DATE. Within the TYPE category, 4,800 records are classified as BIOG. PHYS accounts for another 1,200 names, BIOL for 450, ACTI for 800, COIN for 300 and UNKN for 200. Within the USED category, TRFR names total 300 and are not further classified. Lastly, there are 600 records classified as MISC, the majority consisting of Native American names applied by Europeans. The intention is to reclassify this latter group, as part of a study of indigenous names, prior

to the publication of the seventh edition of *Oregon Geographic Names*, scheduled for 2001. In summary, this edition of *Oregon Geographic Names* is 55% BIOG, 15% PHYS, 10% ACTI, and 5% BIOL. The remaining 15% include those names classified as COIN, MISC, UNKN, or TRFR.

How well does this sample represent the whole? Names from A through K, comprising 24,500 names, or half the total file, have been scanned. There has been a very general classification, based on supposition only, of 10,800 unresearched names. These will have no bibliographic entries and will be sorted separately from the researched names. The obvious BIOG total is 6,450 (60%); BIOL, 2,450 (23%); PHYS, 1,250 (12%); and ACTI, 500 (5%).

BIOG, PHYS, and ACTI correlate well between known and unknown origins. The large difference in the number of BIOL names comes from the great number of fauna names (e.g., bear, beaver, deer, and fish) given by unknown persons for unknown reasons in times long ago. The biggest potential problem is with names including *green, white, brown,* and *black,* which could be either BIOG for persons, or PHYS for color. There are many such names in the A through K total but not included in the above summations. The few COIN or MISC that were identifiable are less than 1% of the total. This was to be expected, as such names are determined only after specific study.

While every name in the nation was applied to a geographic feature by one or more known or unknown individuals, the BIOG, ACTI, COIN, and TRFR names can be largely attributed to specific individuals. Almost all of the BIOG names are those of early settlers, postmasters, mill operators, or others connected closely with the actual locales. The BIOL, PHYS, MISC, and UNKN are generally not so directly connected.

As far I know, this is the first attempt to classify a large and inclusive body of geographic names. As the examined corpus has become larger, inconsistencies have increased. For example, how is *Bear Creek* to be classified when the name was applied because of a specific incident of bear ACTI? Should it still be BIOL? Should subsequent uses within a particular state of TRFR names continue as TRFR, or should they be classified SHFT? Should all commercial and industrial names be ACTI even if they comprise a BIOG name such as Weyerhaeuser?

An overview of 48,000 names in one state could have many uses and provide much interesting information. For example, it is possible to correlate the number of names with the population totals of each of Oregon's 36 counties. BIOL names should distinguish the flora and fauna of each of Oregon's diverse climatic regions. Since all the names of present and historic post offices and railroad stations are dated, their densities can be plotted by decade, or any other desired period of time. Many of the other researched names are also dated, at least to the decade, and this information should allow comparisons of name density with population in different periods.

As this work of classifying the typology of names on a very large scale has been undertaken without any precedent upon which to rely, it is hoped that other onomasts will take this opportunity to suggest other uses or ways to organize and utilize this information.

References

McArthur, Lewis A. 1928. *Oregon Geographic Names.* Portland, OR: Oregon Historical Society P.

Meany, Edmond S. 1923. *Origin of Washington Geographic Names.* Seattle: U of Washington P.

Mencken, H.L. 1936. *The American Language.* 4th ed. New York: Alfred A. Knopf.

Smith, Grant. 1992. "Describing Types of Placename Information." *Names* 40:299-304.

Steele, William G. 1890. *The Mountains of Oregon.* Portland: D. Steele.

Stewart, George R. 1954. "A Classification of Place Names." *Names* 2:1-13.

_____. 1975. *Names on the Globe.* New York: Oxford U P.