Assessing Geo-Location and Gender Information in Han Chinese Personal Names

Bruce Brown and Deryle Lonsdale
Brigham Young University

Brigham Young University is a particularly optimal academic institution for comparative cross-cultural onomastic research. There are over 65 nations represented by five or more students at BYU, and in fall semester of 2004, 10252 of the 29729 students had lived for two years in a foreign country.

This paper is the preliminary report of the beginning of a five year program of comparing the onomastic characteristics of family names and given names across these available national/cultural groups. Beginning in the fall of 2005, a rotating group of students will work on this cross-cultural name ontologies program as their university capstone research experience. The repeatable paradigm involves (1) creating the initial name ontologies by obtaining at least 250 names (fifty persons personally known to five or more respondents) from a given nation, (2) extending those name ontologies by having the respondent identify collateral information about the persons (age, gender, geo-location, ethnicity, religious affiliation, etc.), (3) statistical evaluation of the connections between names and the collateral information, and (4) evaluation of the accuracy with which native respondents can identify collateral information on the basis of name alone.

The pilot study in this series focused on the geo-location and gender information contained in Han Chinese names in comparison to comparable information in American names. ChART (Chinese Anthroponomastic Reporting Tool), a database application was created for gathering *onomastic contingency* data of three kinds: quantitative, categorical, and textual commentary. A team of six skilled native Chinese informants used this *ChART* application to identify on the basis of name alone the likely gender and geo-location for each of 269 Han Chinese names and indicated their level of confidence.

Identifications of gender judged on the basis of given name ranged between 65% and 83%, which was found to be substantially lower than a comparable study with American names. On the other hand, geo-location identifications from family names were surprisingly high, ranging from 19.6% to 54.5%, where 10% would be the expected level for random guessing. The probability of 19.6% accuracy occurring by chance is .000000001.

In a baseline comparison study using American names, accuracy in identifying gender from given names was found to be over 98%, but the identification of location was at chance level. In other words, American names have almost no information about geo-location, which is not surprising, given that we are a highly mobile immigrant society.

A second study consisted of a direct statistical/graphical analysis of the geo-location information found in Han family names. Of the 2180 names in the *Directory of Officials and Organizations in China*, 1258 contain geo-location information. A statistical cross-tabulation of names by province was analyzed with correspondence analysis mathematics and transformed into three-dimensional scatterplots through Metrika. The thirty provinces represented in these names grouped clearly by region, and family names were identified that are most characteristic of each region. This confirms

the finding of the first study, that there is substantial geo-location information contained in Han family names.

Figure 1. Geo-Location of the Han Names: The Thirty-two Provinces of China Grouped into Nine Regions

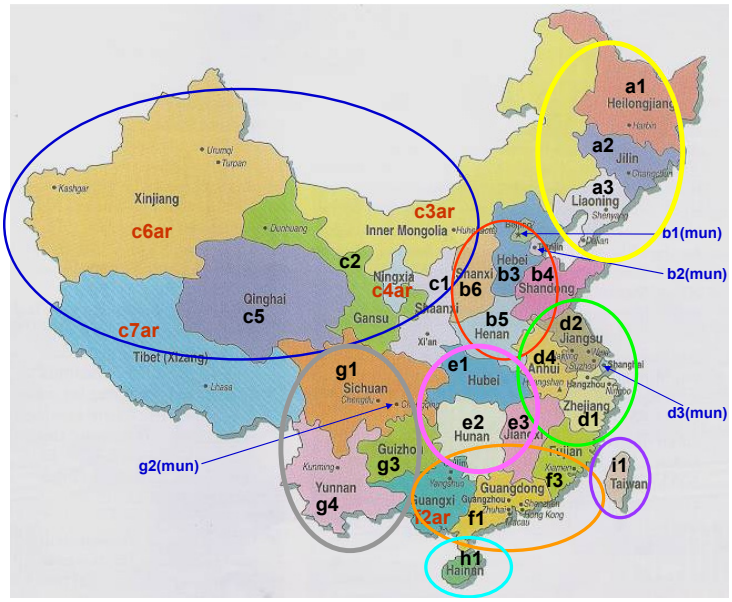

Figure 2. Metrika Vector Plot of Thirty Chinese Provinces in the Anthroponomastic Space of the Thirty-Two Most Common of the 2180 Names, Colored According to Region