New ways in onomastics: Digitised analysis and visualisation of the areal distribution of family names. The example of the onomastic landscape within the historical borders of Luxembourg.

W. Amaru Flores Flores
Johannes Gutenberg-Universität, Mainz
flores@uni-mainz.de

There is an enormous number of extensive sources (historical and recent) on family names that can serve as basis for digital research, some of which even are already digitised and/or geo-referenced according to actual name bearers, thus making analyses of the areal distribution of family names possible. I will argue that although there have been numerous invaluable studies (partially of considerable scale) in that area, the true potential of this research has not yet been exploited, because all studies are, principally, based on one of only two methodical approaches with specific shortcomings. I will demonstrate, however, a method allowing to combine these two approaches delivering new and fruitful insights into the emergence and development of onymic spaces and borders.

1) In the first part of this paper, different sources on family names (e.g. birth certificates, parish registers, war casualty lists, personal columns of newspapers, telephone registries and civil registries) and the possibilities they offer for research will be presented. A special focus will lie on the more systematic, digitised and geo-referenced sources which can be the base for studies on the areal distribution of family names.

2) In the second part, I will introduce (major) previous studies analysing these sources for the areal distribution of family names. In principle, they employ two different methods: distributional maps visualising the distribution of specific (groups of) names (using GIS systems), and measurement of isonymy, i.e. the calculation of the amount of identical names for two or more different name inventories.

Distributional maps have been used in some smaller works and single articles, mainly to compare the name inventories of two or more different countries (cf. e.g. Farø/Kürschner 2007, Flores Flores 2014, Marynissen/Nübling 2010, Nübling 1997), or to describe regional peculiarities of family names (cf. e.g. Kollmann 2014, 2012, Muller 2014, Schmuck 2012), but they have also been used in large-scale projects aiming at the comprehensive description of the distribution of name inventories of whole countries from an intra-linguistic and an extra-linguistic perspective. Probably the largest and most influential project in that area is the “Deutscher Familiennamenatlas” (DFA) [German family name atlas] that will contain 2000+ maps in 6 volumes (cf. Kunze/Nübling 2009-2016).
Areal distribution of the family names derived from the profession of the butcher (source: DFA). Each colour represents not only a single name (token) but a set of related (graphematic) variants. These name types are selected using regular expressions. Theoretically, there is no limitation of the number of different types shown in a single map (nor the number of single name tokens merged into a single type).

All these studies and tools have in common, that they implement the enormous full data sets to calculate the (relative) frequencies of the mapped names at each locality. The selection of the presented names, however, isn’t based on a statistical analysis of the name inventory but follows a hermeneutical approach: they are mainly selected based on the operating experience and onomastic knowledge of the investigators.

The isonymy based studies, on the other hand, fully employ the statistical opportunities of the complete data sets. They are mainly based on the so-called Lasker’s coefficient. Lasker (1985) originally introduced his formula to calculate the genetic overlap of two given populations by analysing their family names but it was soon adopted for onomastic studies as a measure of similarity for two (or more) name inventories, based on the amount of names (and number of their bearers) both of them share. In short: the more names two inventories share, the bigger their Lasker’s coefficient.

\[
R = \frac{\sum \overline{d} \overline{d}}{2}
\]

By entering the coefficients for a given set of localities into a matrix, it is possible to analyse them for areal patterns using e.g. hierarchical cluster analyses or Multi-Dimensional Scaling, and to plot maps showing the onymic areas, borders and transitional spaces.
All visualisations show analyses of the recent name data in the historical Luxembourgish territory.

Left: Dendrogram of a hierarchical cluster analysis using Ward’s algorithm ($\Delta(A, B) = \frac{1}{n_A + n_B} \left[ \sum_{i \in A} - \sum_{j \in B} \right] ^2$), 10 clusters coloured, equivalent 2-dimensional Scaling exhibits a correlation of $r=0.94$;
Top right: Map representation of these clusters;
Bottom right: Multi-dimensional Scaling; 3 dimensions mapped in the RGB colour space, similar colour represent similar name inventories, correlation $r=0.93$.

On the downside however, this method can only show the (non-)existence of differences within the name space but lacks any explanatory potential, i.e. one can neither tell why these occur nor which names produce them. Additionally, as many of these studies compare large and predefined areas, the determined borders can only arise where they were expected by the investigators.

Especially in the last decades however, an increasing amount of isonymy studies were published to analyse the differences of name inventories mostly in a larger scale, i.e. comparing the inventories of several countries or at least bigger areas within a country (cf. e.g. Barrai et al 1996, 1987, Cheshire/Longley/Mateos 2010, Cheshire/Mateos/Longley 2009, Crow/Mange 1965). They exploit the fact that isonymy analyses are conducted rather quickly and derive their areal patterns from the full data-sets, but don’t try to explain which names are responsible for the onymic landscape.

3) The third part is finally devoted to describe the method and main results of my dissertational project on the areal distribution of family names in the historical Luxembourgish area (Flores Flores 2016). For this study, I developed an approach to combine the two presented methods exploiting the advantages of both while minimising their disadvantages:
I used an isonymy analysis to determine the overall distribution of the family names in the investigated area based on the full data set (cf. Figure 3). By splitting the data into a large number
(127) of sub inventories evenly spreading over the whole investigation area the outcome is free of any bias by predefined self-contained areas.

For this study, the concept of an “Index of difference in frequency” was developed to deduce the names responsible for their formation directly from the onymic spaces and borders detected in the isonymy analysis. It builds on the idea that, according to the concept of isonymy, a given name’s influence on a linguistic border should be the bigger the bigger the difference of its frequency on both side of that border is. In this manner, it is possible to link the statistically identified onymic landscape with concrete groups of names creating them, and thus giving answers to the questions how, when and why it developed.

\[
\text{Index of difference in frequency} = \frac{\sum \Delta f_i}{\sum f_i}
\]

**Figure 4:** Formula for the “Index of difference in frequency” (Flores Flores 2016)

In a last step, those names can then be documented exhaustively in their distributional patterns using distributional maps following the example of the DFA (cf. Figure 5). The results can serve as a basis for further investigations on genealogy, historical linguistics, socio-history, and the etymology of names, and also as a role model for further onomastic studies.

**Figure 5:** Example of a map used to document the names involved in the formation of the onymic landscape in the historical Luxembourgish territory: The most prominent onymic border is built up mainly by the linguistic border between the Romance (green, red) and Germanic (blue, yellow) languages. The map shows the distribution derived from the French and German denominations of the profession of the tailor. Each colour represents multiple graphematic variants of certain denominations, defined by regular expressions.

Blue: Schneider: ^Schneide(rs?|sch)$ (Schneider 2180, Schneiders 128, Schneidesch 27)
Red: Tailleur/Couturier: ^Le|[Tt]ailleur[Cou]turi$ (Couturier 59, Lecouturier 13, Tailleur 7)
Yellow: Schroeder: ^Schr(o|oe)(d|t)ers?$ (Schroeder 926, Schröder 849, Schröders 2, Schroeders 2, Schröter 13, Schrötter 13)
Green: Parmentier: ^De?P(a|e)rm(a|e)ntier$ (Parmentier 360, Parmantier 18, Permentier 1, Depermentier 1)
References


Cf. also the project’s web page: www.namenforschung.net/dfa2


