Methods in Dialectology 2022
Special Session Proposal

New Methods for a 21st Century Linguistic Atlas Project

Introduction

The Linguistic Atlas Project (LAP) began in 1929 and consists of a series of regional surveys that began with interviews for the Linguistic Atlas of New England (LANE) between 1931 and 1933 and the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS) in the 1930s and '40s. In 2018, the LAP moved from its long-time host (University of Georgia) to its now permanent home at the University of Kentucky. LAP has been reinvigorated by this move, which has created a renewed sense of purpose for the project and has underscored the need for LAP to showcase the breadth, value, and continued usefulness of its data.

The ultimate goal of early American linguistic geography was the delineation of dialect boundaries, an exercise interwoven with the creation of maps of individual linguistic features, published in works such as Kurath's *Word Geography of the Eastern United States* (1949) and Kurath and McDavid's *Pronunciation of English in the Atlantic States* (1961). In addition to studies of lexical features, scholarly investigations of Atlas data during this era included discussion of the distribution – social as well as geographic – of some grammatical features (e.g. Atwood 1953; V. McDavid 1956) and extensive work on pronunciation (Kurath & R. McDavid 1961) and the use of specific words and phrases (e.g. *shivaree* Davis & R. McDavid 1949; *hoosier* McDavid 1967; *Civil War* McDavid & McDavid 1969; *ain't* McDavid 1941).

While the geographic and social distributions of responses remain important, more recent investigations of LAP data have looked at what the numerical distribution of responses reveals about the nature of variation and of language itself. The ability to assess Atlas responses in database form as "big data" has allowed numerous innovative statistical approaches, some suggesting theoretical stances on the nature of variation and change itself. Kretzschmar (e.g. 2009, 2015), for example, suggests that language is a complex system that reveals a pattern of variation found in a great deal of Atlas data – that of the asymptotic hyperbolic curve (A-curve) – across data sets organized by region or state, by sex, or by ethnicity, and for all types of data: lexical, phonological and grammatical (Kretzschmar, as above; Burkette 2011, 2012, 2013).

It is also safe to say that to some extent Atlas data has been used by 'word people' who have written numerous articles on variation within a specific set of responses (Burkette 2001, 2009, 2011, 2012, 2013, 2017; Antieau 2012; Antieau & Darwin 2013), but the dataset has been underutilized by contemporary variationist sociolinguists. Although LAP data are not completely absent from sociolinguistics (see Labov 1963; Wolfram 1977) and the study of language variation (e.g. Montgomery 1995, 1988a, 1988b), the LAP regional surveys are a largely untapped resource for sociolinguistics, and one of the hoped-for results of this panel is to encourage greater use by demonstrating the application of new methods and new perspectives to the Linguistic Atlas collection.

1. Doing Sociophonetics with Linguistic Atlas Project Data

Josef Fruehwald, University of Kentucky

Recent trends in sociophonetic research have seen us expanding the size of our data sets through the use of semiautomatic techniques like forced alignment and automatic vowel formant analysis. In some cases, canny use of archival recordings has allowed for dueling time depths to analyses: "100" or "130" years of sound change (Labov et al, 2013; Hay et al, 2015). The use of Linguistic Atlas data has the potential to unlock similar time depths for sociophonetic research across North America, but LAP data, like any other historical data, does pose a certain technological hurdle for

researchers wishing to unlock this potential. The currently most common tools to use for these semiautomated techniques (the Monteal Forced Aligner (McAullife et al 2018) and the FAVE suite (Rosenfelder et al 2015) currently require full time-aligned orthographic transcripts as input. For this presentation, we carry out a feasibility study of utilizing fully automated speech-to-text systems on LANCS data. We will explore how currently available systems such as DARLA (Reddy & Stanford, 2014), CLOx (Wassink et al 2018) and wav2vec-U (Baevski et al 2021) perform on archival data. The word error rate will be calculated to evaluate the accuracy of these automated transcriptions, and the transcripts will be corrected for forced alignment with MFA and vowel formant analysis with FAVE-extract. Any identifiable errors or shortcomings of these automated systems, especially as they relate to the age or original media of recording will be tracked and documented, again to provide best practice recommendations.

2. Leaner, cleaner, and full of attitude: The new Atlas interview

Allison Burkette, University of Kentucky
Lamont Antieau, University of Kentucky

Originally, the LAP fieldworkers conducted 6- to 8-hour-long interviews to elicit lexical, phonological, and grammatical targets. They were trained to ask questions that were usually in the form of descriptive phrases that tasked the speaker to name the item being described (these were often framed by the phrase "what would you call [description]") or fill-in-the-blank questions that asked the speaker to supply the target as the missing word ("if a glass fell on a hard floor and shattered, you would say the glass _____"). For the earliest surveys, responses were written down in the International Phonetic Alphabet (IPA), but later interviews were tape-recorded and the responses transcribed form the tapes.

The ultimate goal of the earliest surveys was the mapping of dialect boundaries, particularly the mapping of individual linguistic features. Over time, however, the goals and structure of the LAP interview have changed. The overarching goal of LAP interviews has shifted from an interest in isoglosses and dialect boundaries to a desire to record variation and to look at the correlations between that variation and specific social and regional groups.

Today's LAP interviews take the form of conversational interviews that still seek names for specific foods, animals, weather phenomena, etc., as well as morphological and syntactic forms, but also address the contemporary sociolinguistic interest in perceptions and attitudes. This paper discusses and demonstrates details of the new 'hybrid' Linguistic Atlas interview, one whose format offers the best of all worlds: a set of dialectological targets that will facilitate comparisons across the LAP surveys and through time, the free-flowing conversation prized by traditional sociolinguistics for grammatical and phonological analysis, and questions tuned to get at interviewees' attitudes and beliefs about language use in the area.

3) **The Fractal Structure of Language: How Many Dialects?**

William A. Kretzschmar, Jr., University of Georgia

In previous study of dialect survey data (e.g. Kretzschmar 2009, 2015), the frequency profiles of variant lexical responses to the same cue are all patterned in nonlinear A-curves. Phonetic transcriptions from the Atlas also have the same A-curve frequency profiles, as have the distributions of measurements of vowels in F1/F2 space. Moreover, these frequency profiles are scale-free, or fractal, in that the same A-curve patterns occur at every level of scale. A-curve patterns describe the distribution of all linguistic features we have observed—lexical, phonetic, and grammatical—for a survey overall, for different groups of speakers, for individual speakers, and

even for separate environments in which vowels occur. These findings challenge the boundaries that linguists have traditionally drawn for dialects, whether geographic, social, or phonological, and demand that we use a new model for understanding language variation. Instead of using statistics to try to match our generally shared impressions of dialects, we should realize that there is an unlimited number of dialects, and our choice to focus on one or another of them should follow, not from popular perception, but clear definitions of the population of speakers we wish to observe.

4) **Use all the LAP data!: Moving toward inductive discovery of patterns and connections in the data of the Linguistic Atlas Project.**

Mark Richard Lauersdorf, University of Kentucky

A call for data-driven, inductive discovery of patterns and connections in large datasets would not strike anyone as novel in the world of "big data" (cf. Kitchin 2014) and it really isn't even new in the study of language variation (see, for example, 35 years ago in Horvath 1985, Horvath and Sankoff 1987). It can be argued, however, that data-driven approaches have not achieved the status of "standard tool" in the toolkit of many linguists, even while such approaches present enticing new opportunities for investigating language variation in its social context (e.g. Lauersdorf 2018). In making an argument for data-driven approaches to the study of language variation, it is also important to emphasize the need to "use all the data", both the obvious and the "hidden" data, in order to maximize the effectiveness of the inductive process (Lauersdorf 2018).

The data of the Linguistic Atlas Project (LAP) is truly "big data", unstructured, semi-structured, and structured, and with all of the "volume, velocity, and variety" that one would expect (Laney 2001). The rich data in the LAP includes: language data at all structural levels (phonetic/phonological, morphological, syntactic, semantic, lexical); socio-cultural data concerning not only the interviewees, but also the interviewers; image data representing objects and physical contexts; and time-space (geospatial and temporal location) data; stored in a full range of audio-visual media and objects from raw audio recordings, to textual fieldnotes with image illustrations, to formally published atlas volumes and cartographic representations. And this rich data environment exists iteratively across the multiple regional atlas projects that constitute the LAP. With this level of data complexity and the potential for related pieces of information to be scattered across the various data types and their media manifestations, the LAP presents an ideal testbed for deploying a data-driven, inductive approach that uses *all* the data – the entirety of the linguistic and socio-cultural materials assembled and produced in the atlas process – to facilitate discovery of complex patterns of language variation.

This presentation will provide a structured overview of the LAP as "big data" and will sketch a "big-picture" conceptualization of a data-driven approach to mining and analyzing that data for (historical) dialectology and (historical) sociolinguistics.