

Dialect surveys and student-led data collection

Laurel MacKenzie

One of the most rewarding aspects of teaching at the university level, for both teacher and student, is the opportunity to engage students in research. In addition to the educational benefits (on which see, e.g., Van Herk, 2008; Wanjema et al., 2013), it's just plain fun, carrying the thrill of finding something novel in an assignment that could turn out differently every year. And the field benefits, too: big classes can collect big data sets, so setting a cohort of undergraduates on an understudied variable can produce publishable results.

Though some of the issues in student-led data collection (methods, ethics...) are the same no matter who's involved, others are unique to the presence of a large number of enthusiastic but inexperienced assistants. This vignette summarizes what I learned from four years of teaching a large (90–130 students) undergraduate introduction to variationist sociolinguistics, with a fairly major research component, at The University of Manchester.

The experiences I report here are primarily based on one particular data collection assignment that I repeated from year to year. As part of a unit on dialect geography, students administered to friends and family a dialect survey targeting phonological, lexical, and morphosyntactic variables in British English. For many of these variables, their geographic distribution is well studied (for instance, the FOOT-STRUT and TRAP-BATH phonemic splits; see Wells, 1982), but I also made a point to include one or two unstudied variables by student request (in 2016, we had the many local British words for chewing gum, among them *chewy*, *chuddy*, and *chud*).

So, I used a survey, but pretty much any method of gathering data that professional sociolinguists make use of can translate to the undergraduate classroom. Perhaps the simplest approach is to have your students be the research subjects: Wanjema et al. (2013) describe a number of small undergraduate projects where students record themselves and submit the recordings for their professor to analyze. Other researchers have students carry out rapid and anonymous surveys (Durian et al., 2009) or collect internet language (Van Herk, 2008).

Pedagogically, student-led data collection has a lot of potential. If there are a couple dozen students in your class, tasking each one with collecting ten or so observations will give you a fairly large class data set, which you can then use to teach the students quantitative analysis techniques. My own students brought in around a thousand total responses to our dialect survey each year, and 35% of their course grade came from an assignment where they mapped our findings using Google Maps and quantified the linguistic differences between regions of the UK using descriptive and inferential statistics. Though you could certainly teach the same skills using someone else's data, I think students are more invested in the exercise when they have a personal connection to the data they're analyzing.

If quantitative analysis isn't part of your course curriculum, the data collection process can still serve as a teaching tool in itself. I've found that the section of our dialect survey that targets lexical variables can be a good source of class discussion. This is because we let respondents choose as many variants as they feel they use regularly, and students are often unsure what it means when speakers have selected more than one option. Getting students to think about what someone's motivations for selecting multiple variants might have been can engender discussion about the social and situational factors that might lead a speaker to make use of multiple linguistic options. More generally, students' experiences gathering their data can inform discussions about the pros and cons of different data collection techniques, ways of mitigating the observer's paradox, and principles of ethical data collection.

Speaking of ethics: they of course can't be ignored when students are the ones doing the work. You'll need to check the policy of your university's ethics board on student-collected research. Even when the university doesn't require ethical approval of undergraduate work, I lead my students through the motions anyway. Our dialect survey clearly explains who we are and what we're using the data for, and asks respondents to consent to participating.

Also relevant to this topic are the ethical implications of making data collection a required component of your students' grades. It's tempting to want to do this, of course, because it increases the likelihood that students will do what you ask. But, where your students are your research subjects, requiring them to participate can be construed as coercion (Gurung and Schwartz, 2009). Even if your students themselves are not serving as the research subjects, requiring their grade to be contingent on their interviewing a certain number of speakers, or getting some number of friends to fill out a survey, can still be a form of coercion: those friends or family members may feel pressured to participate so as not to damage the student's chances at a good grade. Later in this vignette I present a solution to this issue that I've used successfully in my own teaching.

Another issue inherent in student-led data collection is that of quality control: how do we ensure that student-collected data is clean, reliable, and, ultimately, usable? This issue is tied up with the logistics of how

students record their data. When I first started teaching, data collection for the dialect survey was fairly low-tech: each student administered paper surveys and coded the responses into an Excel spreadsheet which they sent me; I compiled everyone's findings into one big spreadsheet, which I sent back to them for analysis. (I was grading students on their data collection at the time, and this method allowed me to keep track of which students participated.) But having students code their data by hand inevitably introduces typos and other errors.

A solution to this that I've used is as follows: rather than having students code their raw data into a spreadsheet, have them enter it into an online form, submitting the form once for each observation. The form needs one question for each variable students were asked to code for, and wherever possible, the options should be hard-coded into the form. So, for instance, rather than reporting a respondent's sex by typing 'm' or 'f' into a cell—which sounds simple, until you realize that 'm' and 'M', not to mention 'm ', are treated differently by some spreadsheet programs, plus someone will inevitably typo and enter 'n'—the form has a dropdown menu for Sex where data collectors just choose one of the available options. Form-based data submission not only results in much cleaner data, but removes the hassle of having to download and save dozens of individual spreadsheets: the online form compiles responses in real time which you can download once the data collection period ends. You can still keep track of which students have participated, too, by having students report their name with each observation. Moreover, because some online survey software (e.g. Google Forms) lets you share live results, students who are expected to provide a certain number of responses can easily keep track of how many submissions are recorded under their name.

For the dialect survey project I've mentioned here, it made the most sense to take the online form even further: I had students post a link to it on social media and ask their friends and family to share it. In doing this, I eliminated any means by which I could keep track of how many data points each student brought in—the point of the crowdsourcing was to extend the survey beyond my students' immediate social circle, so I couldn't ask "Which student sent you this survey?"—and thus couldn't make data collection required. There was thus a chance that the survey would flop: if students weren't required to send it out, would any of them do it at all? I'm pleased to report that we ended up with close to 900 responses, or ten per student on average: just the number I would have required had I gone the old-fashioned route.

A final aspect of quality control concerns errors introduced by respondents. With a "hands-off" survey like mine, where data collectors have no contact with their subjects, there's always a chance that subjects won't respond authentically, or, even worse, will maliciously give bogus responses. We can't do much about this (apart from deleting obvious joke responses, of course), but I've taken heart in the fact that, year after year, the results we get show clear and consistent regional patterning which broadly matches what other researchers have found.

I mentioned earlier that my students make dialect maps with the results of our survey. This has turned out to be one of the most rewarding, and enduring, aspects of the project. We've made our maps public (and we do ask respondents explicitly in the survey whether we can include their responses in our public maps) as part of a website for general audiences, "Our Dialects." We put out a press release when we launched the site, which brought in over 10,000 page views and got several students mentioned in national newspapers. And we run a community outreach program whereby students in my class go into local secondary schools and use the maps to teach lessons on sociolinguistics and dialect variation. What started as a class project has thus burst out of the classroom, creating not just an academic learning experience for students but also a means of engagement with the wider community.

References

- Durian, David, Julia Porter Papke, and Salena Sampson. 2009. Exploring social, regional, and ethnic variation in the undergraduate classroom. *American Speech* 84:227–238.
- Gurung, Regan A.R., and Beth M. Schwartz. 2009. *Optimizing Teaching and Learning: Practicing Pedagogical Research*. Malden, MA: Wiley-Blackwell.
- Van Herk, Gerard. 2008. The very big class project: Collaborative language research in large undergraduate classes. *American Speech* 83:222–230.
- Wanjema, Shontael, Katie Carmichael, Abby Walker, and Kathryn Campbell-Kibler. 2013. The OhioSpeaks Project: Engaging undergraduates in sociolinguistic research. *American Speech* 88:223–235.
- Wells, J.C. 1982. *Accents of English*, volume 2: The British Isles. Cambridge: Cambridge University Press.