My proposal takes advantage of the publically-available American Community Survey data to track changes in foreign languages spoken in NYC over time, so that companies and agencies can make informed communications decisions. So for example, let’s take a look at the top 10 languages spoken in Manhattan since 2005:

<https://drive.google.com/file/d/0BxlTe9m4Pa1hbWE4QWx4NFFhejQ/view?usp=sharing>

This plot looks pretty static. So if as a company, your goal is to reach the most people possible anywhere in Manhattan, while supporting as few different languages as possible, this is probably all you need to know – Spanish, Chinese, and French are the top 3 languages, and their populations have remained steady, so if you have finite resources for language support, they’re a safe bet.

But imagine say, you want to do a more targeted advertising campaign. Maybe you’re opening a new store in the Lower East Side, so you want to know specifically what languages are spoken there. Well fortunately the ACS data is broken down at roughly the neighborhood level, so we can look at that.

<https://drive.google.com/file/d/0BxlTe9m4Pa1hM2x0TXU3aGI0Mms/view?usp=sharing>

This plot shows the top 10 languages spoken in the LES, and you might notice that there are some differences from Manhattan as a whole. In the last plot, we didn’t see Bengali for example. Ok, so now we know that if we want to market in the LES, we’d be better off focusing on Spanish, Chinese, and Cantonese, rather than French.

Where this analysis gets interesting though, is in its ability to make predictions about which languages will be important in the future. So sticking with our LES example, we can maybe see some trends in this plot, but it’s not obvious which ones are statistically robust. However, if we build a linear regression model, where we predict the log-scaled speaker estimate as a function of the interaction between survey year and language, we find that there are two significant changes over time. Namely, Polish has increased since 2005, while the “Other Asian” category has decreased. Both these changes are significant at the p < .001 level. So that tells us that if we really want to invest in the LES, we probably want to consider adding Polish to our language support systems, but probably don’t need to worry about the Other Asian languages.

And we can do this type of analysis for all the neighborhoods – to give you a quick overview some of the other significant findings from this model, I found that there’s been an influx of French Creole speakers in Washington Heights (p=.022), and that use of Italian has been declining in Chelsea and the Upper West Side.

[Right now I’ve only done this for Manhattan, but obviously I’d like to extend it to the rest of NYC, and actually I imagine the other boroughs might be more interesting than Manhattan, at least in terms of linguistic diversity. Another factor that I think will be really important to add to the analysis ultimately is the self-report on English fluency included in the ACS data – because companies maybe don’t need to prioritize language communities with lots of strong English speakers.]

So how I envision this analysis being used, is in the form of a web app, where people can interact with a map of NYC, and plot the variables of interest to them. It might look something like this:

<https://drive.google.com/file/d/0BxlTe9m4Pa1hUWVVZlZYSGVmUXc/view?usp=sharing>

Ideally people using this could pick neighborhoods and demographics that they’re interested in, and then plot the languages that are most common, and which ones are trending or declining.

UES: +Portuguese, +Russian (p=.030, .041)

Chelsea: -Italian (p=.004), +Hindi (p=.037)

SoHo: +Mandarin (p=.016)

UWS: -Italian (p=.001)

Washington Heights: +French Creole (p=.022)

<https://drive.google.com/file/d/0BxlTe9m4Pa1hWGdhOEZpV0pFNUU/view?usp=sharing>

Central Harlem: -Mande (p=.029)

<https://drive.google.com/file/d/0BxlTe9m4Pa1hMm1mYkdUdGVjb1k/view?usp=sharing>