



**HAL**  
open science

## Mapping Urban Linguistic Diversity in New York City: Motives, Methods, Tools, and Outcomes

Ross Perlin, Daniel Kaufman, Mark Turin, Maya Daurio, Sienna Craig, Jason  
Lampel

► **To cite this version:**

Ross Perlin, Daniel Kaufman, Mark Turin, Maya Daurio, Sienna Craig, et al.. Mapping Urban Linguistic Diversity in New York City: Motives, Methods, Tools, and Outcomes. *Language Documentation & Conservation*, 2021, 15, pp.458-490. halshs-03519940

**HAL Id: halshs-03519940**

**<https://shs.hal.science/halshs-03519940v1>**

Submitted on 14 Feb 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Mapping Urban Linguistic Diversity in New York City: Motives, Methods, Tools, and Outcomes

Ross Perlin

*Endangered Language Alliance*

Daniel Kaufman

*Queens College (CUNY) & Endangered Language Alliance*

Mark Turin

*University of British Columbia*

Maya Daurio

*University of British Columbia*

Sienna Craig

*Dartmouth College*

Jason Lampel

*A Better Map*

Communities around the world have distinctive ways of representing language use across space and territory. The approach to and method of mapping languages that began with nineteenth-century European dialectology and colonial boundary making is one such way. Though practiced by relatively few linguists today, language mapping has developed considerably from its roots yet remains stymied by problems of ideology, representation, and data quality. In this paper, we argue that digital language mapping in hyperdiverse cities can both contribute to overcoming these problems and bring visibility and resources to communities using Indigenous, minority, and primarily oral languages. For these communities, official surveys like the census are often inadequate, leaving a gap that communities, linguists, and mapping experts working in partnership can address. Urban language mapping as a field should make space for Indigenous, minority, and primarily oral languages through geospatial visualization – in terms that the communities themselves recognize and with a public policy agenda. As a case study, we present our ongoing efforts with LANGUAGEMAP.NYC to map the most linguistically diverse urban center in the world: New York City.

**1. Introduction<sup>1</sup>** Large-scale urbanization is a global phenomenon. According to the United Nations, more than half of the world's population now lives in urban areas, and the number is projected to rise to two-thirds by 2050 (United Nations Population Fund n.d.). Rural push factors include the dispossession of Indigenous peoples and the disruption of traditional lifeways. Urban pull factors include the cash economy, higher education, and political authority, among other attractions and amenities. The result is that cities are now home to speakers of Indigenous, minority, and primarily oral languages from nearly every corner of the globe, though this remains an insufficiently understood and underdocumented reality.

Cities are often portrayed as little more than centers or beachheads for dominant languages and cultures, but this is neither accurate nor inevitable. Linguistic and cultural change can proceed rapidly in urban environments, given the dual pressures of assimilation and disconnection from traditional lifeways, but there are also powerful examples of language maintenance and revitalization in urban settings. Though some analysts assume urban linguistic diversity is temporary, the unprecedented levels of linguistic diversity in today's cities represent a remarkable opportunity not only for those who care about languages, but also for those who care about cities.

Language documentation has traditionally privileged rural and village settings at the expense of cities. Much sociolinguistic research has privileged cities but is usually focused on larger languages, as understood through the prisms of race, class, gender, and other identities. Recently, numerous researchers have explored the multilingualism, 'metrolingualism' (Pennycook & Otsuji 2015), 'superdiversity' (Vertovec 2007; Blommaert & Rampton 2012), and linguistic landscapes (Landry & Bourhis 1997) of contemporary urban spaces. Still missing, for the most part, is a focus on the place of Indigenous, minority, and primarily oral languages in cities. Many researchers, policy makers, and other outsiders are not aware of the presence of such languages at all or need to see this presence somehow substantiated.

Communities around the world have distinctive ways of representing language use across space and territory. The method of mapping languages that began with nineteenth-century European dialectology and colonial boundary making is one such way. Though practiced by relatively few linguists today, language mapping has developed considerably from its roots yet remains stymied by problems of ideology, representation, and data quality. In this paper, we argue that digital language mapping in hyperdiverse cities can both contribute to overcoming these problems and

---

<sup>1</sup> The authors gratefully acknowledge support for this research from the Peter Wall Institute Wall Solutions Grant (University of British Columbia), the Endangered Language Alliance, and Dartmouth College's Office of the Provost SPARK Award, as well as the Claire Garber Goodman Fund within the Department of Anthropology at Dartmouth College. We also wish to acknowledge the Social Science Research Council's Rapid-Response Grants on COVID-19 and the Social Sciences, with funds provided by the SSRC, the Henry Luce Foundation, the William and Flora Hewlett Foundation, the Wenner-Gren Foundation, and the MacArthur Foundation. In addition, we thank Mapbox Community, ESRI, Sentry, and Airtable for providing their products for free or at lower cost to nonprofit organizations like our own. A special thanks to Matt Malone, Julia Schillo, and Bridget Chase for their assistance. We have benefited immensely from feedback from community members and colleagues. Full credits for the map can be found at <https://languagemap.nyc/Info/About>.

bring visibility and resources to communities using Indigenous, minority, and primarily oral languages. By working in partnership, language communities, linguists, and mapping experts can furthermore address certain inadequacies in the census and other official surveys. Our thesis here is that urban language mapping as a field should make space for Indigenous, minority, and primarily oral languages in terms that the communities themselves recognize and benefit from and with a clear public policy agenda. As a case study, we present our ongoing efforts with [LANGUAGE-MAP.NYC](#) to map the most linguistically diverse urban center in the world: New York City (Perlin et al. 2021).

Converging developments across different fields of study make this work particularly timely. We draw simultaneously on language documentation, with its increasing focus on how academic-community collaboration can increase knowledge about linguistic diversity; geolinguistics, with its interest in the spatial dimensions of language use; and new urban sociolinguistics, with its emphasis on the “plurality, variation, contingency and ambivalence” of urban language ecologies (Smakman & Heinrich 2017).<sup>2</sup> On the technology side, browser-based mapping platforms like ArcGIS Online and Mapbox are helping to make maps and data visualization more accessible than ever before (van Rees 2015). In short, the time is ripe to bring these disciplines together and re-conceive how the mapping of urban linguistic diversity can serve communities, policy makers, researchers, and the wider public.<sup>3</sup>

The New York City digital language map is the focus of this paper. In particular, we focus on our **motives** for mapping urban linguistic diversity, the **methods** by which we gathered the data, the **tools** with which we mapped that data, and the project **outcomes**. Each topic is addressed in turn in the sections below.

## 2. Motives

**2.1 Achieving visibility** Indigenous, minority, and primarily oral languages have always been present in cities, but where they have not been driven out, they have often been rendered invisible.

In the case of New York, before the arrival of European settlers, numerous varieties of the Indigenous Algonquian language now known as Lenape were spoken in dozens of settlements across what is today New York City (Weslager 1999). The establishment of New Amsterdam in the seventeenth century resulted not in a Dutch colony but in an *entrepôt* consisting of Europeans, Africans, and Native Americans, where the Jesuit Father Isaac Jogues reported in 1646 that “there may well be four or five hundred men of different sects and nations [...] [including] men of eighteen different languages” (Jameson 2010).

By the early twentieth century, New York City had absorbed massive waves of immigration from every corner of Europe, with small but growing communities

---

<sup>2</sup> Yet, in this volume, which brings together chapters on twelve such ecologies, Indigenous, minority, and primarily oral languages are almost nowhere to be found.

<sup>3</sup> For recent approaches to language mapping, see Auer & Schmidt (2010), Lameli et al. (2010), Mutter & Zacherl (2019), and Brunn & Kehrein (2020).

from the Caribbean, Latin America, and Asia. Now, in the early twenty-first century, New York City is ‘hyperdiverse,’ operating with and through intensifying and multiplying levels of cultural and linguistic differences. Communities now arrive from every corner of the globe, notably including newer arrivals from zones of deep linguistic diversity such as Mexico, Central America, the Himalaya, West Africa, South Asia, China, and Island Southeast Asia. Yet, with little of this linguistic diversity surveyed, studied, or recognized, it has remained invisible both to policy makers and to the general public.

There are undeniable dangers associated with visibility, and it requires a certain level of trust, especially for a community that has faced persecution or marginalization in the past, to raise a hand, seek recognition, and declare its presence. The earliest detailed ethnic map of New York City, drawing in part on data from the 1910 census, was compiled during a period of antiforeign hysteria by the Joint Legislative Committee for Investigating Seditious Activities, to serve as a guide to army officers should they need to put down “an organized uprising” in the city (Wallace 2017). The abuse of census data as part of the effort to target Japanese Americans in Los Angeles for internment during the Second World War (Anderson 2015) prompted the adoption of a confidentiality provision (also known as Article 13). More recently, the New York City Police Department’s infiltration of mosques in the wake of 9/11 has provided another example of how visibility can lead directly to surveillance.

We argue, however, that the consequences of invisibility, in the context of contemporary New York City within which we work, are far graver than the dangers of visibility. The choice of individuals or communities to remain invisible should be inviolable, but likewise they should have the tools and capacity to dictate the terms of their own visibility. In the context of a tolerant society, visibility can be a first step toward recognition, “a vital human need,” as argued by Taylor (1994).

The need for visibility and support has only deepened in recent years with anti-immigrant policies and the COVID-19 pandemic, which has disproportionately impacted urban immigrant communities in the United States and elsewhere (Craig et al. 2021). For these communities, the New York City language map is understood as a tool for building power, generating civic engagement, and earning recognition in a political space dominated by the claims of larger ethnic and religious formations – helping languages “increase their perceived status, both within the community and among the general public” (Gawne & Ring 2016: 195).

Like mapping, language planning and policy have typically focused on the national level, with comparatively little concern given to “the engagement of urban governance with linguistic diversity” (Christ & Thomas 2008: 3). It is becoming “inevitable that for many linguistic groups—for better or for worse—multilingual cities will become central to their languages’ survival,” while at the same time, linguistic diversity may be “a force for cohesion rather than division” for the cultural ecology of the city itself, but the literature on this is still limited (Christ & Thomas 2008: 5–9).

**2.2 Decentering the census** Until 1890, when the Census Bureau first asked about language, there were no significant attempts to collect information about the

languages spoken in New York City or any other American city. From then until 1970, various questions were asked about language use, typically about the ‘mother tongue’ of non-English speakers or the languages spoken by those residents identified as foreign-born. Since the 1970 census, a relatively stable set of questions has been asked – transferred in recent years from the decennial to the more detailed, annual, sample-based American Community Survey (ACS):

- Does this person speak a language other than English at home? (Yes/No)
- What is this language? \_\_\_\_\_ (For example: Korean, Italian, Spanish, Vietnamese)
- How well does this person speak English (very well, well, not well, not at all)? (United States Census Bureau n.d.)

This method of obtaining language data, despite the reach and resources of the Census Bureau, has consistently failed to do justice to the full breadth of linguistic diversity in the United States. Indeed, any information gathered by the census about linguistic diversity is perhaps best understood as almost incidental, with the main intent coded in the first and third parts of the question: to gauge segments of the population with low English proficiency. The five-year 2009–2013 ACS, a particularly deep dive representing “the most comprehensive data ever released by the Census Bureau on languages,” estimated “at least 192 languages” spoken at home in the New York City metropolitan area. In a typical year, the numbers are even less granular. For example, the most recent five-year ACS data available (2015–2019) breaks out and tabulates just over a hundred “languages,” of which around one-fifth are groupings such as “Other Specified Native American,” with no further information available.

We argue that Indigenous, minority, and primarily oral languages are systematically undercounted for both historical reasons and on account of implicit biases of the survey instrument itself, with major implications for smaller language communities in urban centers. Turin (2014) offers a bracing reminder of how such exercises are “fraught with taxonomic, political, and ideological problems, often compressing complex and highly local ethnolinguistic identities into standardized checkboxes” (380). In this sense, the context of Sikkim, India, described by Turin is not actually so different from New York City.

Although the US census is supposed to enumerate every individual living in the country, and the ACS is supposed to provide a reasonable sample of the same, there are many reasons why recent, undocumented, and non-English-speaking immigrants in particular might not be aware of, able, or willing to take the census. There is in fact significant overlap between areas of consistent undercount – in 2020, the response rate in New York City was approximately 62% (NYC Census 2020 n.d.) – and areas of high ethnic and linguistic diversity, with ethnolinguistic communities in New York City known to number in the thousands either not taking the census at all, not identifying themselves as such, or being lumped in with other groups.

**a. Problems with the instrument** An obvious contributing factor is that the census instrument itself is only available in the most commonly spoken languages. The 2020 census, the best-supported so far in terms of language access to date, was only available in thirteen major languages, though short informational guides were provided in fifty-nine languages.<sup>4</sup> The question of how responses are collected is also significant. Whether online (advocated for strongly by the Bureau in 2020), by phone, by mail, or with an enumerator, the means at the Census Bureau's disposal have not inspired confidence in vulnerable and marginalized populations. During the lead-up to the 2020 census, the Trump administration's attempt to insert a 'citizenship question,' at a time of accelerating activity against undocumented immigrants, led to a further erosion of trust.

Even if someone were to overcome understandable mistrust *and* happen to be selected in the sample for the annual ACS, they would find just two questions about their language. At first glance, these appear innocent enough: "Does this person speak a language other than English at home?" and "What is this language?" Yet, as language professionals know, the word *language* itself and its various translations are loaded terms for many segments of any population. Though the terminology may vary, colonial notions of 'language' (official, standardized, and written) versus 'dialect' (no official status, unstandardized, primarily oral) are very much alive in immigrant communities from all over the world. 'Language' is thus mapped on to emic distinctions made in immigrants' societies of origin (e.g., *lengua/lenguaje* versus *dialecto* in Mexico [Kaufman forthcoming], *lingua* versus *dialetto* in Italian contexts [Andriani et al. forthcoming], 语言 *yǔyán* versus 方言 *fāngyán* [and sometimes 土话 *tǔhuà*] in China [Mair 1991]). The linguistic criterion of mutual intelligibility often plays little or no role in these distinctions. In Italian and Chinese cases, for instance, where language shift from one variety to another linguistically related variety may be taking place, the mutual intelligibility criterion may be difficult to operationalize in the first place. In the context of Latin America, the lack of mutual intelligibility is clear to any speaker, but this may not prevent them from using the Spanish term *dialecto* for what any linguist would likely term an Indigenous language.

The linguistic categories employed behind the scenes to tabulate ACS responses are also problematic. Written-in responses are not made public, for privacy reasons, and the responses are instead tabulated and grouped by the Census Bureau via a complex system of heterogeneous codes, which have satisfied neither linguists nor language communities, resulting in groupings like 'Cushite,' which is, in fact, a language family rather than a language, and the vaguely geographic unit 'Kru, Ibo, Yoruba.' Recent reforms have expanded the number of codes and attempted to link them with Ethnologue ISO 639-3 codes (Gambino 2018), but as respondents are not provided with the ISO 639-3 codes, they are unlikely to make full use of them.

For instance, the ISO 639-3 recognizes Chittagonian, Sylheti, and Rangpuri as independent varieties on par with Bengali, but in Bangladesh, these are generally treated as dialects of the national language even though mutual intelligibility can be very low. Without knowing that these varieties have been given an independent

<sup>4</sup> <https://2020census.gov/en/languages.html>

status in the census, Bangladeshi respondents will typically ‘round up’ to Bengali, which they may speak or be shifting to in any case as a second language. Newer privacy concerns further dictate that even better tabulation might not result in better public data because the number of speakers volunteering the same language name on the census form will have to cross a certain national threshold to be publicly visible. Finer granularity here may only yield further invisibility.

**b. Hybridity, translanguaging, and multilingualism** There are also several reasons why hybrid language practices and ‘translanguaging’ (Garcia & Wei 2014), recognized as particularly common in urban contact settings, face the most daunting obstacles in becoming legitimate objects of enumeration and recognition. Not only are they seen by many as ‘bastardizations,’ they are often seen to be more flexibly defined than standard languages. It is only in the relatively rare cases of stable mixed languages (e.g., Michif, Media Lingua, Sri Lankan Malay, Chavacano; see Bakker & Mous 1994) that such languages become emblematic of a community’s ‘authentic’ identity. In the more canonical case, hybrid language practices are perceived to be a deficient blending of two legitimate codes rather than a legitimate code in and of itself.

This recalls the steep challenges faced by those advocating for the possibility of selecting multiple racial categories in the census, related by Kertzer & Arel (2001: 33–34). Not only did the campaign for accepting multiracial identity in the 2020 census not find support among ethnic and racial organizations, many saw it as a threat to their membership and long-term existence. Nonetheless, it became possible to select multiple racial categories starting in the 2000 census, and nine million Americans now do so. Interestingly, the ACS approach to language lags behind its approach to race. Tucked away in a Census Bureau working paper, we find a rare description of the language tabulation process:

Respondents who speak a language other than English at home specify the language in writing. The responses are recorded and then coded at Census Bureau headquarters, first by computer-matching responses to a compiled list of previous responses (called the ‘autocoder’), then, if there is no match, by staff who examine the write-ins and assign codes (‘clerical coding’). *If multiple languages are listed, only the first language is coded.* (Gambino 2018: 2; emphasis ours)

Thus, while the census inadvertently collects information on multilingualism, this information ends up on the cutting room floor as it does not fall within the Bureau’s interests. This procedure affects Indigenous and minority languages disproportionately, as these are spoken alongside national languages in the majority of cases. Given the enduring colonial hierarchy in which national languages take priority over all else, it is likely that respondents who fill out multiple languages will put the national language first, rendering all following languages untabulated. Despite important improvements in modernizing the language categories for purposes of



tabulation, the bias against hybrid languages conspires with the narrow focus of the census to render invisible those languages that are not primary national languages.

**c. Measuring speakerhood** Another domain whose complexity is underappreciated is the relation between an individual and their language. Moore et al. (2010) observe that “the deployment of numbers of speakers inevitably conjures up an image of the speaker as a stable – hence, countable – entity; it obscures the actual elasticity of speakerhood in real sociolinguistic life, even if that elasticity may come back with a vengeance whenever new counts have to be made” (11). In her study of language shift among a Scottish Gaelic speaking community, Dorian (1981) gives a central place to the role of the ‘semi-speaker.’ The term has, since then, been commonly used to refer to those with some working knowledge of their heritage language but who speak in a manner distinct from that of conservative fluent speakers. The term *semispeaker* remains controversial on account of its imagined deficit-based interpretation, which presupposes an idealized if often unattainable ‘complete speaker.’

Lacking an explicit assessment of an individual’s fluency, the category ‘heritage speaker’ appears more neutral but in practice alludes to speakers whose speech shows contact effects of simplification in their heritage language and is implicitly contrasted with a full native speaker, also sometimes referred to as a ‘conservative speaker.’ For several reasons, those who might be deemed heritage speakers rarely, if ever, claim their language in official enumerations, especially when the questionnaire targets the language of daily life, the home, or ‘main language.’ The utilitarian spirit of the census and ACS leads respondents to offer only information they perceive to be of use to the state.

A language justice perspective requires us to consider other types of speakers in addition to the above. Grinevald (2003) offers a typology consisting of fluent speakers, semispeakers, terminal speakers, and rememberers. Grinevald’s ‘terminal speakers’ are “speakers of the dominant language who may know some phrases, or simply some words of the endangered language” (Grinevald 2003: 65). The final category, ‘rememberers,’ are described by Grinevald as “speakers who once in their life-time had a better knowledge of the language” (Grinevald 2003: 66). While ‘rememberer’ appears to be an increasingly common type of Indigenous language speaker worldwide, it is particularly prevalent in contexts where immigration leads to a rapid abandonment of a language, even if the attrition can be reversed in some cases.

A further category, not taken into account by Grinevald, who approaches the problem from the perspective of language documentation, is that of the new learner of any ‘dormant,’ ‘sleeping,’ or otherwise revitalizing language (Leonard 2008), which can be further subdivided into categories such as young learners, adult learners, and heritage learners. One of the most important aspects the census might additionally

capture is information on the ongoing revitalization of Indigenous languages, now supported with at least some federal funding by the US government.<sup>5</sup>

In a similar way, the narrow focus of the census on a certain type of ‘speakerhood’ renders invisible the large population that signs, rather than speaks, its languages. Indeed, signed languages have only belatedly been recognized as languages by hearing linguists. Although a variety of estimates place the US-based community of American Sign Language (ASL) users in the hundreds of thousands, if not more, this information cannot be gleaned from the census, where neither Deaf ASL users nor hearing ASL users (e.g., Children of Deaf Adults, sometimes called CODAs) register at all. There is even less hope of learning anything about the prevalence of other signed languages such as Hawaii Sign Language (Perlin 2016) or Mexican Sign Language (Quinto-Pozos 2008). Given that there is ASL support for responding to the census, it is possible that many signers do respond and answer “ASL” to the language question, but according to Mitchell et al. (2006), “in the initial data processing phase, the census codes any mention of an American signed language as English” (309). This raises the question as to whether the bureau believes, as many hearing people do, that ASL is ‘merely’ a signed form of English, unaware that it is a distinct language in its own right bearing no relation to English. In its official explanation, the Bureau states that “current question design” was narrowly conceived to support the 1975 amendment to the Voting Rights Act, which sought to end discrimination at the polls against speakers of the largest few minority languages (notably Spanish). For this narrow purpose, the assumption is that ASL users can read English.

Censuses have commonly used language as a proxy for ethnicity but have historically employed three disparate types of identification criteria: ‘native’ language, language of home, and language of daily use (Kertzer & Arel 2001: 26). With ‘language of daily use’ as the criterion by which to enumerate ethnolinguistic groups, significant populations are rendered invisible when they are dependent on a more dominant ethnolinguistic group. A case in point were Czech speakers under the Austro-Hungarian empire at the close of the nineteenth century, many of whom worked for German-speaking families and who used German for the better part of the day. A campaign within the Czech community advocated for a ‘backwards’-looking view on language based on ‘mother tongue,’ understood as including what the first language ‘should have been’ given ethnic origins. Kertzer & Arel (2001) term this case and others like it *census primordialism* and discuss how it has been exploited by various national projects throughout the last two hundred years. Clearly, the choice of ‘language of daily use’ over ‘native language,’ especially when construed as above, has the potential to affirm or erase the presence not only of a language but of an entire ethnolinguistic group in the census.

The first language-related question on the ACS, “Does this person speak a language other than English at home?” leans more toward language of daily use than

---

<sup>5</sup> In addition to the Esther Martinez Native American Languages Preservation Act, funded since 2008, the \$1.9 trillion stimulus package signed into law in March 2021 authorized \$20 million for Native languages (out of \$31 billion for tribal governments and other programs for Native American communities).

mother tongue. While the home is undoubtedly a primary domain for smaller languages that may not have a public presence, the question bypasses other levels of more latent language competence and use. Its wording additionally excludes Grinevald's 'rememberers,' also referred to as 'silent speakers,' who have nobody to speak their language with. Our interviews with dozens of Indigenous Mexican families in New York City suggest that the Indigenous language is often primarily a language of the telephone, while Spanish is the dominant language of the home even if the head of the household is far more comfortable and fluent in the Indigenous language rather than Spanish (Kaufman forthcoming).

### 3. Methods

**3.1 Partnerships** The Endangered Language Alliance (ELA) is an urban language organization with a mission to support linguistic diversity in New York City and beyond (Kaufman & Perlin 2018). Through its day-to-day operations collaborating with speakers and communities on language documentation, revitalization programs, policy work, classes, and so on, the ELA has been collecting information about the languages of New York City since its founding in 2010. More purposeful data collection by coauthors Perlin and Kaufman began in 2016 with an invitation to contribute a language map focused on the heavily immigrant borough of Queens to a popular, subversive atlas of New York City (Solnit & Jelly-Schapiro 2016). This in turn led to a standalone print map of the entire New York metropolitan area (Perlin & Kaufman 2020), designed by cartographer Molly Roy and widely covered by the media upon release. While a print map has some advantages, there was strong demand for a multilayered, interactive digital version of the map, which ultimately became [LANGUAGEMAP.NYC](http://LANGUAGEMAP.NYC), and attempts to represent every distinct 'communalect' in the city.

The transition from print to digital was made possible by a partnership between the ELA and the University of British Columbia and involving coauthors Turin and Daurio, with support from the Peter Wall Institute for Advanced Studies and also including Dartmouth-based anthropologist Sienna Craig. This led in turn to the hiring of Jason Lampel of *A Better Map*, a designer and developer of interactive maps for the Web. Complex digital projects require resources, collaboration, and very specific decisions, all of which may limit their reproducibility, but efforts in other cities indicate that there is considerable interest in the potential of urban language mapping. We believe that motivated teams anywhere can improve on what little currently exists.

Linguist-community collaboration, in the ELA's experience, has also entailed an ever-increasing focus on tangible benefits for speakers who are among the most marginalized in New York City (and in the United States) in terms of health, education, housing, income, and access to information and interpretation. Moreover, the ELA's existing partnerships with city agencies such as New York City's Department of Health and Mental Hygiene and the Mayor's Office of Immigrant Affairs, and the encouragement received from sympathetic individuals at these agencies, have encouraged us to work toward a map that is comprehensible and useful in terms of public service delivery and urban language policy.

The most distinctive feature of our approach is an emphasis on data gathering by linguists and communities working in concert, bypassing inadequate official data sources like the census. The need to gather our own data grew directly out of the ELA's community relationships because none of the communities with whom we planned to partner were recorded as even *existing* in the census or any other data set. Many of the Indigenous Latin American, Himalayan, Pamiri, Middle Eastern, and other language communities with whom the ELA has now worked comprise thousands or even tens of thousands of individuals across the city, but they were, and remain, invisible in all existing official data sources. This problem is not unique to New York City or the United States – indeed, we have yet to encounter a census or survey of *any* city's languages *anywhere* that fully represents the Indigenous, minority, and primarily oral language varieties that are recognized by communities themselves and constitute the majority of entries in databases such as Ethnologue or Glottolog.

At most, data may be available at the national level, where large-scale surveys (e.g., the official, federal Linguistic Survey of India or the more ground-up, subversive, and unofficial People's Linguistic Survey of India run by a nonprofit organization) have been undertaken, but this may only reinforce a static view of one language per homeland, failing to account for the mass migration of speakers away from traditional homelands to cities, both domestically and internationally. Fully representing and supporting deep linguistic diversity is simply not a goal for most government agencies, and this is reflected in their data-gathering activities. As described above, the ACS and other such surveys fall short in terms of whom they ask, how they ask, and what they do with the answers. On the other hand, our commitment to representing the Indigenous, minority, and primarily oral languages that have neither public visibility nor official support made it essential to address each of these challenges.

Other linguists working in cities, for lack of a better alternative, have typically relied on official census data, or similar sources, and started their analysis there.<sup>6</sup> Gaiser & Matras (2016) present an innovative crowd-sourced, language landscape map of Manchester, England, facilitated by a mobile app. The map is a rare exception – a fine-grained portrait of public multilingual signage that uncovers patterns of settlement as well as language use. While the project does not restrict its scope to national languages, there exists an inherent bias in focusing on written languages, holding little promise for those communities whose languages are spoken but not written (Daurio et al. 2020; Daurio & Turin 2020).

**3.2 Inverting the census** In terms of *whom* we asked, our bottom-up method of gathering language data for the map was in some ways the inverse of what the census does. With limited resources, there was never any way that our small team could

---

<sup>6</sup> See, for example, Veselinova & Booza (2009) for Detroit, Willis (2013) for Houston, Van der Merwe (1993) for Cape Town, Vandenbroucke (2020) for Brussels, and Musgrave & Hajek (2010) and Shari-fian & Mugrave (2013) for Melbourne. Some, such as Multilingual Manchester (Matras 2018), have also drawn creatively on a variety of data sets, but none to our knowledge have undertaken a years-long 'language census' on this scale.

hope to survey a sufficiently large sample of New York City’s approximately nine million people or the more than twenty million in the metropolitan area. Nor did we need to, for it is unlikely this would have addressed any of the issues identified above, and in any case, the ACS appears to be broadly accurate when it comes to the largest national languages. By design, the larger languages are underrepresented in our data, and many of the city’s varieties of American English, too pervasive to locate precisely, have been left out or only very selectively represented.<sup>7</sup> We worked instead through an approach much more akin to ‘snowball sampling,’ a method often used for ‘hard to count’ populations, spreading the word through the ELA’s already substantial network, which covers precisely those groups least likely to respond to the census.

Before we began to formalize data for the map, the ELA had already recorded speakers of nearly a hundred language varieties spoken in New York City, only a handful known to the census, so we were immediately aware of a large number of communities whom it was important to map and where we already had contacts. Kaufman & Perlin (2018) describe in brief how the ELA’s network came together through a similar kind of ‘snowball’ effect whereby those already involved introduced others in their communities and related communities. At the same time, regular publicity, an accessible online presence, a diverse array of projects, and an office right in the middle of the city make it quite easy for speakers and community leaders to reach out to the ELA. In terms of *what* we asked, unlike the Census Bureau, our goal was never to enumerate authoritatively the number of speakers in every census tract (or any other unit of territory). In any case, this would be an impossible task with our limited resources. Rather than undertake a formal, generic survey, we have held thousands of interviews and discussions with community leaders, speakers, and other experts. The essential point throughout this work has been to combine community expertise with linguistic knowledge. In other words, ours is not a ‘normal,’ clear-cut data set such as would be familiar to most data scientists, nor do we strive to be comprehensive.

The first priority in all conversations was to establish that a language variety is or was used in the New York metropolitan area at all and to specify where the language is or was used the most. While some partners and respondents cited the names of neighborhoods or towns without being able to provide further specifics, some were able to specify residential clusters in certain blocks, intersections, or smaller areas. Many mentioned community centers, religious institutions, hometown (or regional) associations, restaurants, and other gathering places. Some groups would universally cite a single community center (e.g., the Sherpa Temple in Elmhurst for the Sherpa community). For others, choosing a significant site for the language community was less clear-cut, as we describe below.

**3.3 Significant sites** Through conversations with community members, we decided to focus on the modest but achievable goal of mapping significant sites that

---

<sup>7</sup> Other maps have recognized the need to strip away at least English and Spanish in order to ‘see’ other languages (e.g., Hubley 2019). We have simply gone further and applied this logic to all larger languages. See Museum of the City of New York (n.d.).

could be precisely located to serve as representative of a language's presence in the city. Given that most communities cannot be reduced to a single site or area, we decided to represent most languages at more than one site and to tag secondary neighborhoods (without points on the map). With the exception of Lenape, we also decided to limit the number of significant sites for any language to just seven so that varieties of English, Spanish, Chinese, and others would not dominate the map, given our goal of emphasizing the presence of Indigenous, minority, and primarily oral languages. This choice to focus on significant sites merits further explanation. At least in the context of New York City, the approach has other advantages besides being achievable, including the fact that the names and locations of such sites are usually already public information. Most communities do have such sites, and we learned that most community members supported their inclusion as primary data for the map. Moreover, a census of these sites, rather than atomized individual respondents, can itself yield important findings about how communities choose to organize and represent themselves.

At the same time, there are also limitations and downsides to mapping languages via sites. Most maps have fallen back on representing languages as geolocated points, polygons (representing a larger area), or some combination of the two (Drude 2018), despite the fact that neither dots nor polygons can do justice to lived linguistic realities. We have not overcome this difficulty, but we are aware of it, and we believe that the complexity, interactivity, and rich contextualization of our map offers a modest step in the right direction. Ours is also a project designed with communities and the general public in mind, as well as researchers and policy makers. If the representation of deep linguistic diversity is a paramount aim, it is necessary to work with limited data and strive to bring Indigenous and minority languages to the front, notwithstanding the wealth of information available for mapping the use of larger ones. The immediate visual impression given by the language map, as seen in Figure 1, is of a mass of dots, each typically representing a significant site, with dots color-coded by world region.

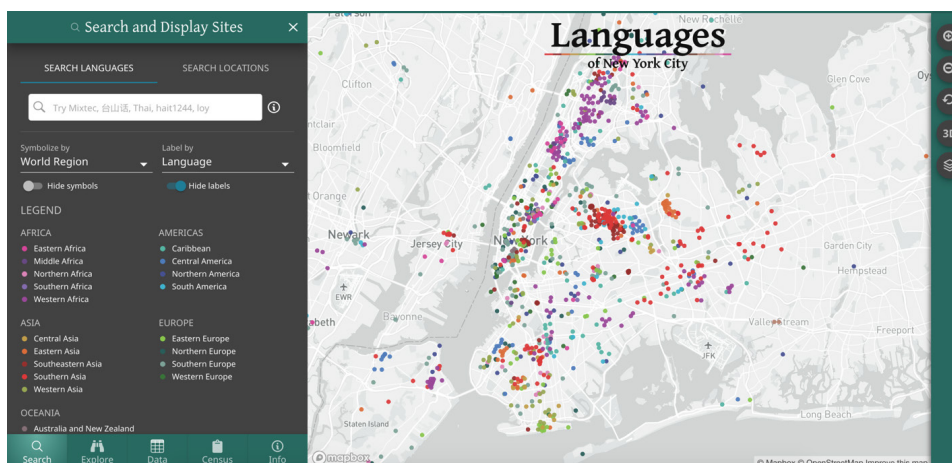


Figure 1

In cases where speakers of a language did not know of or name any significant sites, information on residential patterns was used. In most cases, this was less precise, with information at the level of a neighborhood or a bounded area within a neighborhood. In some cases, significant sites do not fall within the areas where people live now, and this involved some judgment calls, though often both significant sites and residential clusters were used and tagged as such. Placement was undertaken as precisely as possible with the information available, and there were relatively few languages with neither significant sites nor residential clustering (i.e., where speakers are completely atomized). In a handful of these cases, a speaker's own 'fuzzed' home location (see below) had to be used.

In terms of multilingualism and significant sites shared by speakers of multiple languages, we did not find any ready solution to the problem of how to 'stack' multiple languages onto a single significant site, whether it be a matter of 'vertical diversity' in an apartment building (an important issue in a vertical city like New York City) or the multilingualism of a single community center. Any solution (e.g., a heat map to show density) had its tradeoffs in terms of also honoring distinct communities.

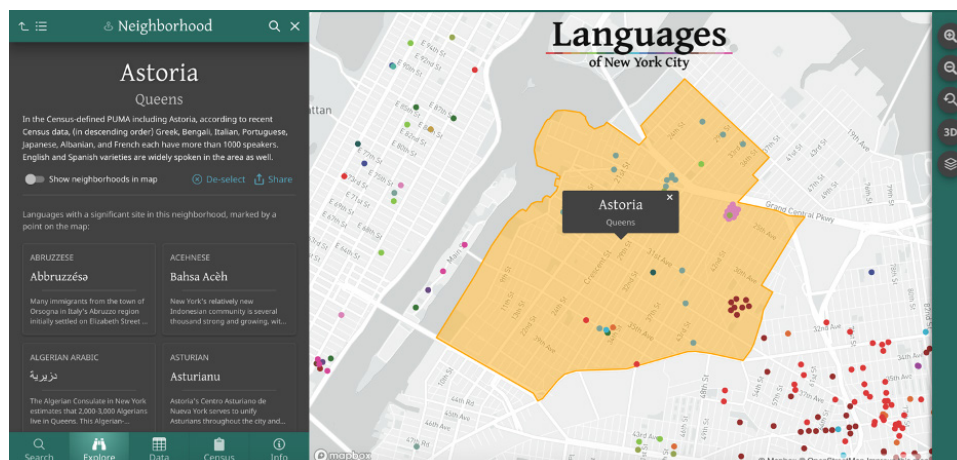


Figure 2

For example, the Al-Hikmah Mosque in Astoria is a hub for numerous Indonesian languages, with Indonesian itself a lingua franca and Arabic the liturgical language, but we have arranged the relevant dots (in brown in Figure 2), discretely, around the mosque to allow for reasonable visibility at common zoom levels. Unfortunately, there is no clear way to indicate that the Al-Hikmah cluster differs from another nearby cluster (in pink in Figure 2), which represents an actual stretch of several adjacent blocks of Steinway Street where speakers of North African languages are concentrated. Manually adjusting ('fuzzing') the latitude and longitude coordinates in cases like the Al-Hikmah cluster was also deemed important, even essential, for privacy in case a site's address was not already public information and

in particular for residential addresses. The value of the map lies not in specifying the exact address where speakers gather or live, but in giving visibility to the overall structure of linguistic diversity in the context of geospatial realities: individual neighborhoods and the city as a whole.

In addition to the focus on significant sites, we asked community members to offer an estimate for the size of the community and to provide any other information about its history, its present-day makeup, and its language practices. These narratives contributed to short, qualitative descriptions, which are included in the digital map for every language group, or at least macrolinguistic group. While not presented as authoritative, these descriptions are perhaps the heart of what is in some ways an in-depth ‘story map,’ a multimedia whole that combines mapping with other elements to produce visually rich narratives. Other elements also include audio and video recordings made in New York City by the ELA wherever possible.<sup>8</sup> Unlike the census, the ELA database is built on what speakers themselves say about their own languages and thus foregrounds the names most commonly accepted by the speakers themselves, or endonyms (sometimes known as autoglossonyms), in the appropriate orthography, while also giving common English names for the use of researchers and the public.

**3.4 Every kind of speaker, every kind of language** With regard to the relation between an individual and their language, we took a domain-neutral approach, collecting information on any and all kinds of language use by various kinds of speakers.

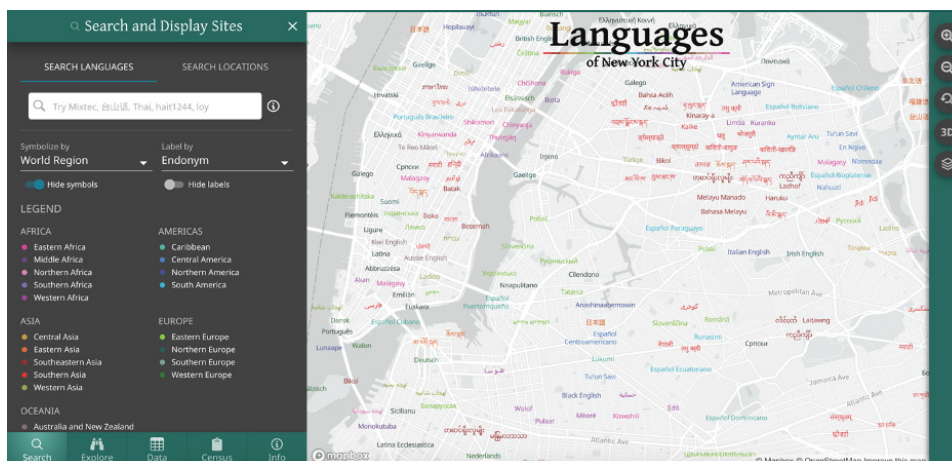


Figure 3

<sup>8</sup> Among the corpora of recordings that can also be experienced via sites on the map is Voices of the Himalaya, a video storytelling project launched in 2016 that documents the extraordinary diversity and vitality of Himalayan and Tibetan New York City (Gurung et al. 2018). We soon hope also to make available in a similar way a corpus of COVID-19 diaries and interviews created by Himalayan New Yorkers (Gurung et al. 2020).



To give two examples of how we tried to gather data on language revitalization and new learners, our map takes the position that both Lenape and Taíno – as distinct linguistic codes practiced by learners and promulgated by activists in the Lenape and Puerto Rican diaspora and linked by a thin but unbroken thread to the pre-colonial languages of Lenapehoking and Boríken, respectively – are justifiably considered ‘languages of New York City.’ Moreover, the map can provide visibility not just for these languages but for these language movements by including information about the various types of new users and new resources for language learning and language community-building.

Lenape is the Indigenous language of Lenapehoking, which includes parts of what are today New York, New Jersey, and Pennsylvania. Often acknowledged as one of “the last fluent Lenape speakers,” Weenjipahkihelexkwe – more widely known as Nora Thompson Dean – died in Oklahoma in 1984, although there are also said to be one or two remaining speakers of the language in Moraviantown, Ontario, whither part of the community had been displaced. There have not been speakers in New York, in any traditional sense, for well over 200 years, and yet a Lenape woman from Ontario, Karen Mosko, has been coming to the city once a month to teach her language – which she learned from other revitalizers – to an eager cohort of students with Indigenous local ancestry together with non-Indigenous students. What then is the status of Lenape in New York City? While it is not spoken as a first language by anyone in the city, people of Lenape ancestry are actively reawakening and identifying with the language.

Taíno, the Indigenous Arawak language of what is today Puerto Rico, has not been spoken as a mother tongue for at least four centuries. Yet here in New York, several Puerto Ricans of Taíno descent are working on reconstructing the language from historical sources and through comparison with related Arawakan languages, including Garífuna, which is spoken by a large population in the city. Unlike Lenape, Taíno is only preserved in place names, a rudimentary vocabulary list, and several words that survived in Puerto Rican Spanish (Granberry & Vesceius 2004). Practically no full sentences were recorded by settlers or other voyagers, and thus the reconstruction of the grammar relies completely on related languages. Feliciano-Santos (2017) discusses how different Puerto Rican indigenist groups take very different approaches to the language, with some proceeding in the manner of reconstruction as in traditional historical linguistics, others creating folk etymologies based on connections to Indigenous languages of Mesoamerica, and yet others taking the intriguing position that everything spoken by indigenist Puerto Ricans, whether labeled by outsiders as ‘English’ or ‘Spanish,’ is at its core Taíno. Where does this leave the status of Taíno in New York City? The map answers this question by recognizing the unbroken linguistic thread that connects today’s Taíno activists to their precolonial ancestors as well as their ongoing efforts to reinforce that connection through reclamation and revitalization.

In a preliminary and similar manner, we attempt to include hybrid language practices, ethnolects, and other language varieties that have historically been marginalized or subsumed into other categories but have at least some recognition within communities and the linguistic literature. Adequate representation of multilingualism

has been much more challenging. Although we can imagine potential visualizations of the extraordinarily plurilingual character of many New York City communities, we have so far primarily represented this in the text descriptions that accompany each dot on the map. Extensive tagging also links each language community directly to others of the same neighborhood, country, world region, language family, and even ‘macrocommunity’ (as shown in Figure 4 below, to cover active cultural, religious, historical, and other connections not captured by other categories).

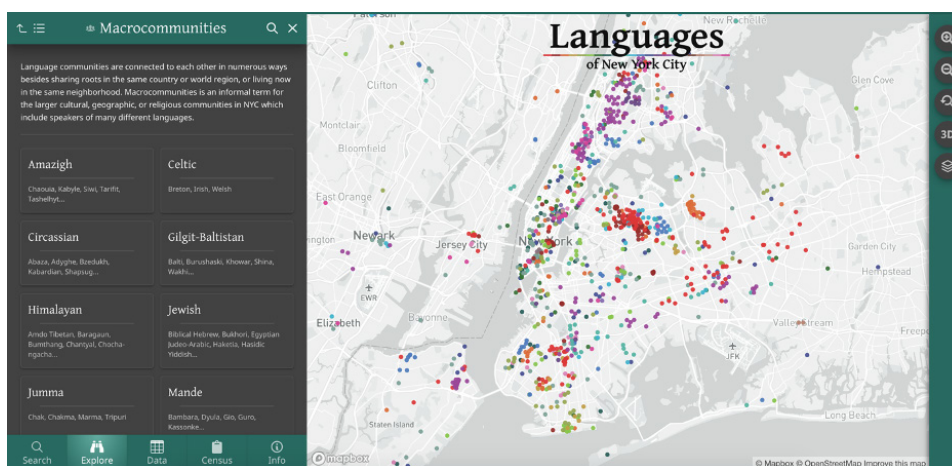


Figure 4

As for the analysis and representation of the responses, we did not tabulate or cluster language varieties into larger groupings for statistical or privacy reasons, as the Census Bureau does. Our tendency, rather, has been toward greater granularity, repeatedly inquiring as to the most specific (mother tongue or heritage) variety used by members of a community.

Although we use the shorthand word *languages* in the map’s name and a few other places, we took a neutral approach to collecting information not only about hybrid language varieties, but also to what are often considered as dialects and ethnolects. We have not used terms like *languoid*, *doculect*, or *glossonym* (Cysouw & Good 2013), in part because in a public-facing map of this kind, the recognition of Indigenous and minority languages *as languages* is of crucial importance and also because such academic terminology is alienating for all but the most specialist audiences. Varying levels of knowledge, both about the New York City communities and in some cases about the languages in question more generally, have also played a role in the deliberate ‘unevenness’ of a schema that sometimes breaks out very specific varieties (e.g., Casamassimese from Italy or Nar-Phu from Nepal) while also using macrolanguage terms that patently await further disaggregation, like Tu’un Savi (*Mixtec*) from Mexico or Fulani from West Africa.

Nonetheless, recognizing the importance of connecting our very particular data set to existing information sources, we sought to match the varieties attested through our process to ISO codes and Glottocodes, even though this was not possible for all cases. Likewise, using those codes, we drew on the Ethnologue for tagging languages whenever possible for the world region and a few major countries where it is spoken, as well as the total number of speakers globally; and we drew on Glottolog to tag by (top-level) language family. In all of these cases, the motive was to make our data set maximally informative and useful while recognizing the limitations of all these sources. Likewise, in tagging languages by neighborhood, we had to grapple with the fact that the neighborhoods of New York City, despite their social and cultural importance to New Yorkers, are not actually official administrative units with formalized boundaries. Because no authoritative schema exists, we had to find a consistent, transparent schema that best connected with community intuitions about neighborhoods. What we ultimately selected (and slightly modified) was a schema that grew out of the city's 2020 census outreach efforts, of which the ELA was a part.

Conversations with community partners revealed the many fascinating ways that the city's diversity is linked to its increasingly diverse suburban and sometimes even rural hinterlands. This, in turn, raised a further issue: whether to ignore the city's 'commute shed,' areas where people who come to the city to work actually live. Following the thinking of urban planners, we widened our scope to include the thirty-one-county metropolitan area (including twenty-six counties outside New York City), as defined and explained by the Department of City Planning (NYC Department of City Planning n.d.) and as highlighted in Figure 5, while maintaining a clear focus on the city itself.

**4. Tools** Many of the technical decisions we made, although familiar to those well-versed in geographic information systems (GISs), were new and unfamiliar to the linguists and community members on our team.

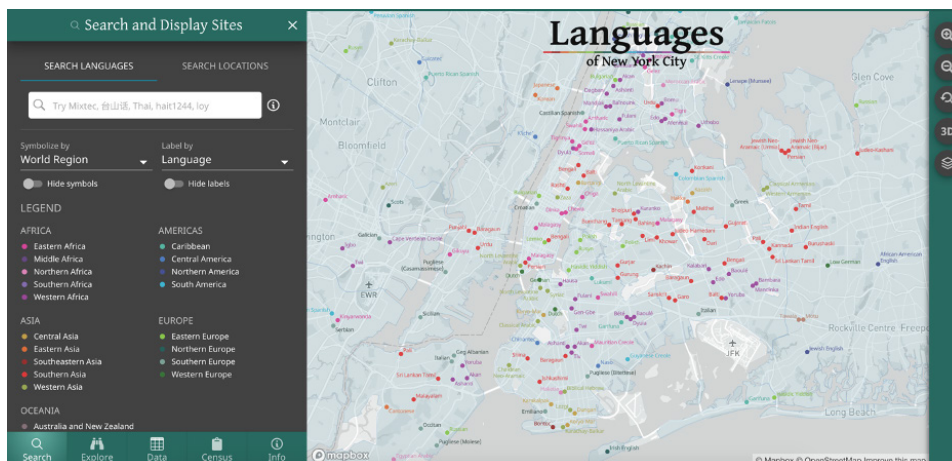


Figure 5

The sheer volume of technical decisions was at times overwhelming, swamping our earnest attempts to document our process (which itself was very time-consuming). Wherever possible – and we should clarify that it was not always possible – our goal was to follow a noncommercial, open-source ethos. We aimed for whatever methods, tools, or schemas we used and developed to be transparent, generalizable, and community-oriented. Already challenging when it came to collecting and working with the data, this commitment to open source and open access would prove even more difficult when it came to selecting and working with tools for visualization of and interaction with the data on the map itself.

There are numerous ‘off-the-shelf’ options of the kind described by Gawne & Ring (2016) that allow linguists to begin mapping their data themselves with relatively little training. However, after a lengthy exploratory period during which we tested existing tools and looked at different models, we decided that the functionality, flexibility, durability, and design we were aiming for required a degree of customization and technical expertise. We decided it was worth investing up-front in a skilled web developer who could draw on publicly available technologies, write original code as needed, and document their workflow in a public-facing repository, creating an accessible codebase and setting a standard for future urban language-mapping projects. To this end, we posted the position on Code for America, a hub for public service developers, and this led to the hiring of coauthor Jason Lampel on a short-term project basis. The first goal was simply to ensure that a standalone map would be up and running, offering users the ability to query and interact with our data, enriched with audio and video recordings for certain languages. Longer term, we wished to create an open-source toolkit documenting the workflows and tools used so that others could potentially undertake language mapping in their own communities. We established that the map would be housed at the ELA after development and maintained by nontechnical personnel. Having few illusions about the life cycle of digital projects, which depend on so many moving parts (cf. Turin 2021), we thought it essential that the data themselves and as much of the other content as possible be managed, validated, shared, and updated (in real time) in a user-friendly way through the cloud. At first, this meant a single spreadsheet in Google Sheets, where we had gathered our data. Later, we moved to the commercial platform *Airtable*, a similar service that is more oriented to the making and maintenance of relational databases. Our public-facing version of the database is downloadable, and while the *Airtable* data is technically accessible to anyone with an API (application programming interface) key, we did not attempt to create an API to serve it up to other sites and applications (ELA n.d.).

While data gathering for the most part preceded map design, there was a need for continual polishing and wrangling of the data even as we iterated on the design aspects, in part because of our inexperience and in part because the data continually needed to be outputted and shared in different and derivative forms due to technical requirements. Nor did we always know, or understand exactly, what we wanted or what was possible. Discussions about representation were continual and iterative until a ‘schema’ was finalized, specifying (for instance) that every language must be ‘tagged’ with at least one country and world region but that other fields not always

available (e.g., total number of speakers globally) would be ‘optional.’ Compromises were usually made in the direction of clarity and consistency of representation. For instance, we selected the top-level language family listed in Glottolog rather than attempting in any way to replicate Glottolog trees, let alone representing distribution of branches or terminology. ‘Clean’ data, designed to be filtered and queried, have little room for this kind of nuance.

Our workflow was designed to make it easy to add further points by crowdsourcing information through a feedback form, although we do not allow users to make such changes themselves. We initially had (for the most part) street addresses for all of the significant sites, and we used a geocoding service to assign latitude and longitude coordinates to these addresses in bulk. Esri is a commercial GIS mapping software company that offers such services; there are also free alternatives, such as the one offered by the City of New York. Some addresses required troubleshooting. For example, in cases where the site named by a community member was imprecise (e.g., a park), points had to be edited manually. GIS expertise, along with the desktop GIS software QGIS and ArcGIS Pro, was crucial for this, as it was subsequently also for adding polygon layers that reflected different relevant geographic units: neighborhoods, counties, census tracts, and Public Use Microdata Areas (PUMAs), which delineate geographic areas with no fewer than 100,000 people. A single digital data layer, such as the census tracts, contained names and coordinates for the boundaries of over 2,000 polygons (for New York City alone). In some cases, manual editing of polygons in the ‘Neighborhood’ layer was required as well (e.g., redrawing neighborhoods to exclude cemeteries or parks that might otherwise appear to be residential areas and matching the shoreline to the official geographic units mentioned above).

Alongside a number of text-rich Google Docs, most of the workflow was documented in GitHub, an increasingly standard repository hosting service used by programmers for software development and version control that can also help to replace email and other forms of communication as a project collaboration tool. In effect, the entire development process for our project has been public by default. Anyone who would like to ‘listen in,’ learn, replicate, or simply understand any particular decision or feature can consult our Git repository for the project (*NYC Language Mapping* n.d.) and will find 844 Git ‘commits’ (changes to the repository) and over 200 team-created ‘issues’ (task management) to date, on everything from tiny features to major design questions. Likewise, anyone can contribute to our 100,000 lines of code by logging an ‘issue’ in our repository, and any user can clone our code without restriction for their own language-mapping project. Such a task would be nontrivial but at the same time, quite feasible. Having everything in GitHub is an important part of our strategy for lightening the technical load should others be interested in reusing the tool. Through GitHub, we have provided access to our code, made our workflow and development process entirely open, and hosted all of this in an open digital platform. In the same spirit, the code has been released under a common, permissive MIT license, which allows modification and distribution as well as private and, in theory, commercial use. In our thinking, there is little to fear and much to be gained from maximal openness (*MIT License* n.d.).

No less consequential for the presentation of the map itself was the choice of Mapbox, a location data platform that has supported and contributed to a number of open-source mapping libraries and applications. For this project, with the support of the Mapbox Community team, which aids organizations using its tools for positive impact, we used various Mapbox services (e.g., Mapbox Geocoding API and Studio) to take advantage of several APIs for creating maps and querying data. We also used Mapbox GL JS, a JavaScript library that facilitates the interactivity of our map.

In effect, there is an important division of labor between Mapbox and Airtable. The coordinates for all sites on the map and any information that must be symbolized on the map itself must be uploaded into Mapbox as a ‘tileset’ (necessitating a multistep workflow to update).<sup>9</sup> Any information in the panels that accompany the map (the user interface [UI] for searching, filtering, etc.) is pulled ‘on the fly’ from Airtable, allowing for optimal performance and easier editing.

The points on the map can be supplemented by labels showing either the endonym or the most widely used name of the language in English. The color-coding of dots by world region follows the United Nations geoscheme,<sup>10</sup> implicitly emphasizing the geographic diversity of languages as a major takeaway for even the most casual user. We also provide options, through a prominent drop-down menu, to symbolize the dots, as shown through a toggleable legend, in terms of local community size (based on a five-point scale we devised) or local status (using the five categories Residential, Community, Liturgical, Historical, and Reviving).<sup>11</sup>

We can only outline here some of the ancillary, publicly available tools that were deemed important for the map to function according to our expectations, which coauthor Lampel used to good effect during development (for more details, see *Languages of New York City Map* n.d.). React is a JavaScript library for building UIs – in other words, a set of models to draw on ‘to get stuff to do stuff’ on the map, with the computer language TypeScript (a superset of JavaScript). In addition, we benefitted from Material-UI, an open-source React framework that offers a simple and customizable component library, as the basis of our UI.

While a full-on content management system would be desirable, we are currently managing between Airtable (for parts of the UI) and the relatively user-friendly blog-

---

<sup>9</sup> Reflecting a similar division, the UI panels do not always interact seamlessly with the map itself. In particular, it proved challenging to account for and ‘remember’ all the ways an individual user might want to query and filter the data across all the different parts of the site, and sometimes the result is not intuitive.

<sup>10</sup> A schema devised by the United Nations Statistics, for statistical purposes, which divides countries and territories into six regions and twenty-two subregions, plus Antarctica. See <https://unstats.un.org/unsd/methodology/m49/>.

<sup>11</sup> While the present is our principal focus, and diversity has mostly increased over time, we mapped seventy-two historical language in communities, drawing on both community testimony and relevant historical sources. We also looked at earlier census data, considering more graphically interesting ways of representing change over time, but significant decade-to-decade differences in census methodologies add to what is already a nontrivial technical challenge.

ging platform WordPress for a few pieces of long-form text, notably the About and Help sections. In terms of media, we use YouTube to host all the videos and playlists, which are discoverable from the records of individual language communities in the map. We use the Internet Archive to host all the audio. The ELA was previously already using both platforms because they are free services allowing speedy and potentially unlimited uploading, and they offer efficient provision of media to viewers. Both YouTube and the Internet Archive have handy embeddable players as well as APIs that allow us to ‘call’ some of the metadata associated with the information on those sites with the media files (such as the title and description). This basic metadata then loads directly in the modal dialog windows in the map where the audio and video files open, such that upon closing, users do not lose their place in the map.

Given the importance of using community orthographies, at least for endonyms,<sup>12</sup> font support for a full range of Unicode characters was another important focus. We settled on an approach using the fonts in Google’s Noto font family (Google n.d.). Through exhaustive trial and error, we found it best to load all the less common fonts into the user’s browser, having defined them in the code. Ensuring that endonyms appear in the correct fonts as labels on the map itself demanded an additional step of uploading the actual fonts into Mapbox while indicating via an Airtable database which fonts are required in which cases. In a handful of cases where we could find no appropriate font with encodable characters, we made do by showing users an SVG (scalable vector graphic) image of the relevant endonym, with the SVG format ensuring that the quality of the image remains constant at any image size.

Full support of the orthographies is restricted, for practical reasons, to the two most recent versions of major browsers, on both mobile and desktop, but significant effort was put into making the map work across the spectrum of devices and screen sizes.<sup>13</sup> With all the different (especially older) systems on users’ phones and computers, it may be impossible for all scripts to display properly, but we believe ours is one of the few sites or maps able to support such a wide range of orthographic systems both in terms of display and input (searching and filtering). Overall, it is striking how few websites are optimized for the orthographic and typographic needs of multiple Indigenous languages (cf. Schillo & Turin 2020).

**4.1 Incorporating census data** Having completed a working prototype based on ELA data, we decided to incorporate census data in addition to the ELA data, at least around language (and in the future potentially around important social variables such as health, income, or housing). As noted earlier, the disparity between the ELA’s mapping work and the census count, in nearly every respect, is too large to ignore, especially when the two data sets can be superimposed on one another in the digital

---

<sup>12</sup> The entire website that hosts the map is in English. For Google Translate-style versions in other languages, a browser extension can provide a basic functional translation. Many of the individual tools, like Mapbox, support other (larger) languages, but a full, professional translation would be a significant undertaking, given all the components (ours and others’) that are involved and the tens of thousands of words of descriptive text.

<sup>13</sup> From initial analytics, half of all users access the map on mobile devices.

map (ELA points on census polygons). For example, a comparison of ACS data on Spanish-speaking census tracts with ELA data on significant sites for the Indigenous languages of Latin America suggest that many who are counted as Spanish speakers also have an Indigenous language as their mother tongue. A single category in the census data like Mande, referring to a large group of related languages spoken across West Africa with varying degrees of mutual intelligibility, is reflected in the ELA data set as twenty distinct languages. This includes both widely spoken languages such as Bambara and Dyula and those with much more limited distribution such as Marka and Vai, which nonetheless have speakers or even substantial communities in New York City. Census categories that speakers would hardly recognize, such as ‘Niger-Congo’ (a language family with over 50,000 ‘speakers’ in New York City) can be analyzed visually, by comparison with ELA data, as potentially involving dozens of languages, but almost certainly featuring Akan, Igbo, Wolof, and Yoruba as major components, as in Figure 6. Nor is this kind of reverse-engineering of the census tabulation process, which obscures actual responses, restricted to the African languages for which the census’s problems are now notorious. Comparison may also help clarify census information that is otherwise unclear or hard to use. Zooming in on an area of Eastern Queens with a concentration of Other Indo-Iranian’ speakers according to the census, we can surmise that many are likely to be speakers of Tajik or Bukhori.

**5. Outcomes** As of mid-2021, ELA data-gathering efforts, as described above, have confirmed just over 700 languages in the New York City metro area, mapped to over 1,200 significant sites. Some of the language varieties shown are heritage languages only known to individuals or small groups, but in the majority of cases, they are spoken by communities of at least several hundred, if not thousands of, people. There

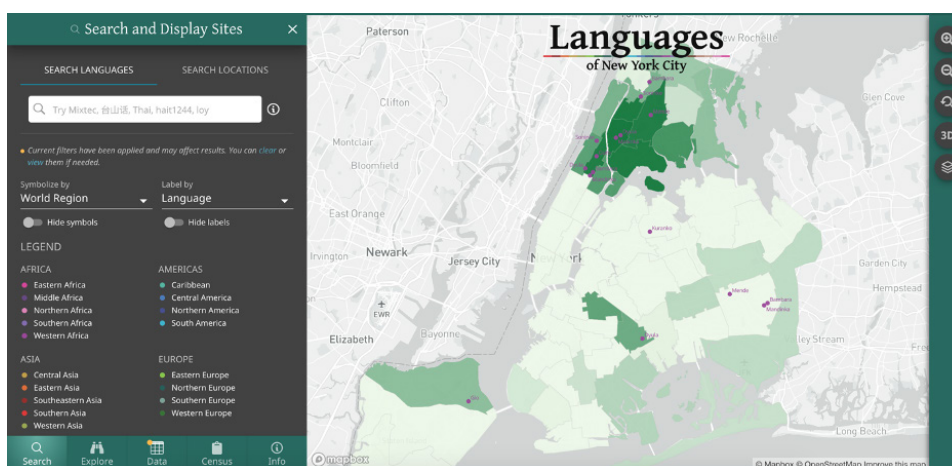


Figure 6



is, of course, considerable variation in terms of size, settlement patterns, and degree of organization.

Moreover, cross-comparison between our data and census data suggests that the latter is only consistently reliable and recognizable for approximately sixty languages, almost all of which are major languages with official status in their countries of origin.

Instead of describing the resulting site in further detail, we invite readers to spend some time exploring <https://languagemap.nyc> and consulting the detailed documentation at <https://languagemap.nyc/info>. Our hope is that this information, in addition to the source code, the Git repository, and this article, will work together to provide at least a roadmap and a content-rich example for those wishing to create similar maps in different localities. Though LANGUAGEMAP.NYC is not by any means straightforwardly replicable in an ‘out of the box’ way, our team remains committed to sharing our experiences and supporting researchers around the world interested to work with and hopefully improve on our approach.

Designed to feel relatively straightforward and intuitive with five panels and just a few on-map tools, our digital language map is nevertheless packed with features, enabling users to interact with the data in multiple ways. There is no space in this article for a discussion of the purely aesthetic and stylistic elements, many of which derive from the earlier print map. Instead, and only briefly, we focus here on what mapping New York’s linguistic diversity in this way has revealed and on some of the practical outcomes it may lead to.

**5.1 Findings** In geographic terms, approximately 38% of the languages in the ELA database are from Asia, 24% from Africa, 19% from Europe, 16% from the Americas, and the rest from Oceania and the Pacific. Some patterns emerge, such as the dense clustering of West African languages in Harlem and the Bronx, the presence of Indigenous languages in areas usually just considered ‘Spanish-speaking,’ the deep and multifaceted Asian-language diversity of Queens, to name a few top-level insights. These patterns hint at the complexity of the city’s linguistic diversity in ways that ACS data miss or distort. To the extent that the map is updated or similar efforts are undertaken in the future, we may be able to trace change over time, including language loss, maintenance, and revitalization.

Communities undergoing language shift that are likely to have large numbers of heritage speakers, ‘semi-speakers’ in the sense of Dorian (1981), or ‘rememberers’ in the sense of Grinevald (2003), are shown as such, albeit with precedence given to the heritage language. For example, we identify the Crimean Tatar community, which has largely shifted to using Turkish and now English, or the Bukharian Jews who shifted to Russian and now English. The ELA database also contains information about over a dozen ‘liturgical’ languages now used primarily in religious contexts (e.g., Latin, Coptic, Ge’ez), several ongoing cases of language revival (especially Indigenous) such as Lenape and Taíno, numerous ethnolects and dialects, as well as a few dozen languages that were historically used by communities but were never officially recorded as such. By no means does the map attempt to be comprehensive, and it surely represents an undercount and a crude reification of linguistic realities

that are much more complex on the ground. However, we hope that accompanying textual descriptions, videos, and audio recordings bring us somewhat closer to an accurate and representative understanding of the linguistic complexity of New York City.

The map makes visible not only hundreds of speech communities missed by the census, but also a whole range of settlement patterns and interaction zones that are integral to the city's linguistic ecology. One immediate conclusion, also borne out by census data on race and ethnicity, is that there are few 'true' enclaves, understood as monoethnic (or nearly monoethnic) areas of residential settlement sealed off from other groups. It is possible, judging from the evidence in New York, that the idea of the enclave is in fact just as much of an idealization as the notion of a monoethnic homeland. Even in areas strongly associated with a particular group (from Manhattan's Chinatown to Richmond Hill's Little Guyana), the titular group may not even represent a clear majority of the population, or itself be substantially internally diverse in ways that language use reflects. There are many domiciles that we might describe as 'UN buildings,' where speakers of dozens of languages live, with or without much direct contact.

At the same time, there are *vertical villages*, our chosen informal term for individual buildings where members of the same ethnolinguistic group, or even close kin, have managed to settle at least temporarily. While we have not formalized or delved deeply into this concept, we have identified several cases where dozens of members of a single small ethnolinguistic community have managed to rent, or in at least one case even own, adjacent or nearby apartments in the same building. Such buildings become well-known community hubs in and of themselves, where new immigrants may more easily get their start, where responsibilities like childcare can be shared, and where a language may find domains of use beyond the individual family home. Given the volatile realities of real estate in a city like New York, these arrangements may be fragile indeed, but the very existence of vertical villages reflects how urban linguistic diversity can operate even at the most granular levels.

Where residential concentrations exist, there is typically not just one but several, albeit with important linguistic differences. Where the census simply identifies Arabic-speaking tracts in Brooklyn, Queens, and the Bronx, the ELA map makes clear (as community members know) that those in Bay Ridge, Brooklyn, are somewhat more likely to speak forms of Levantine Arabic; those in Queens speak more Moroccan and Egyptian Arabic; and those in the Bronx mostly speak forms of Yemeni Arabic. 'Coterritorial' settlement patterns highlight the ways in which one group, for any number of reasons, tends to settle (sometimes in a kind of succession pattern) with or near another to which it has linguistic, historical, cultural, religious, or other connections. For example, throughout the city, Albanian neighborhoods have formed in Italian areas in large part because many Albanians are proficient, for historical reasons, in Italian – which the settlement pattern in New York only strengthens. In other cases, whole microcosms of world regions can form, as in the post-Soviet world of south Brooklyn, where Russophones from across the Soviet Union (especially Central Asia) may find themselves using Russian more than either Uzbek (for instance) or English. In some cases, we find no pattern at all, with individuals simply settling

where they can or wish for reasons of work, convenience, cost, and so forth. In others, communities that had initial nodes in the first generation experience dispersal, especially with suburbanization, and this may be associated with a shift to English and absorption in the wider society. At least some of these histories of mobility and migration are either, to a degree, latent in the map or explicitly captured in the text descriptions, but there is much that could potentially be visualized about the intense and constant mobility of language groups within a metropolitan area.

Patterns of language shift and change already underway in a home region often continue or accelerate with migration (itself very often a multistop process that involves continual linguistic adjustments). Much depends on how movement and settlement bring speakers into contact with other groups, but the map makes clear that Indigenous Mexicans live within a Spanish-speaking matrix, just as Fujianese speakers live within a Mandarin matrix and Loke speakers are surrounded by Nepali and Tibetan, not to mention Urdu and Bengali. Far from a traditional model representing ‘Americanization’ as a straightforward, intergenerational shift from a mother language to English, we find a complex patchwork of multiple assimilations, based on differential settlement patterns in the city, often leading at least initially to high degrees of multilingualism and mixing.

Applying GIS analysis to census data (with an awareness of its limitations) on the eighty-eight languages of Metro Detroit, Veselinova & Booza (2009: 151) note that a divide between languages that form “clear density areas” (e.g., Syriac) and those that do not (e.g., Polish), “correlated with recency of immigration” or possibly with reason for immigration (Veselinova & Booza 2009: 152). They also note “clustering of completely unrelated language groups”, including speakers of Arabic, Urdu, and Bengali for whom Muslim belonging is a common denominator (Veselinova & Booza 2009: 153). Developing related ideas, our data on New York City indicate a much more complex and multifactorial set of patterns awaiting analysis.

**5.2 A place on the map, a place in the city** Despite all of our caveats and disclaimers, a professionally designed map, whether analog or digital, is an artifact that carries a certain authority, much like that of a book, a law, or, in some cases, the printed word itself. Over several years of ‘road-testing’ first the print and then the digital map, we have found that people consistently look for their language(s) where they think they should be. Overwhelmingly, the initial response from speakers of small languages is satisfaction at being represented, especially at seeing a name, particularly an endonym, which in many cases they have never seen printed (at least outside the community), put on the same plane as languages like English, Spanish, and Chinese.

In some cases, visibility and recognition can come almost as something of a shock. While displaying an enlarged version of the map at a festival in Prospect Park, Brooklyn, we were approached by a young Senegalese-French man who had recently moved to the area and was visibly astonished to find his heritage language, Bâinounk, shown in the very first place on the map he looked, among the Senegalese languages spoken in the Bronx. He eagerly called over his wife, telling us that she was a speaker of Monokutuba from the Republic of Congo-Brazzaville – a language

not then on the map but which both were happy to see added, even if she was the only speaker in the city that they knew of.

There was no small irony in it being a speaker of Bāinounk who searched for and found his language community on the map that day, as Bāinounk has been held up as a particularly thorny case of sociolinguistic complexity; defining the language itself is a challenge due to the extreme multilingualism and language contact found in Casamance, Senegal, as can be noted from the very title of Lüpke 2010: *Language and identity in flux: In search of Bāinounk* (see also Lüpke & Storch 2013). But this betrays an important truth: While linguists and other specialists have been anxiously pondering the identification and demarcation of languages (as well as their invention and ‘disinvention,’ cf. Errington 2007; Makoni & Pennycook 2007), in the meantime the labels in question, whatever their provenance, gain significant traction ‘on the ground.’ Both the linguist and the speaker are in search of Bāinounk in their own ways. Linguists may try to document and describe a language by sorting through layers of multilingualism, while native speakers may be more concerned with locating themselves in the multilingual diaspora city.

In another public exposition of the map, this time on a street corner in the South Bronx, a child, roughly ten-years-old, approached and began scanning the language names intently. He was trying to remember the name of his parents’ (perhaps heritage) language, he told us. “It begins with a G,” he said, starting to look in the section of the map representing where we stood (a major center for Garifuna people, see England 2006) and coming upon the name with the force of discovery: “Garifuna!” In this case, the map had unexpectedly served as both a reminder and as validation of a buried heritage language.

These engagements with the map are not outliers. Lacking any census data about their communities, a group of Indigenous Latin American language activists in New York City who have recently formed a group called *El Consejo de Pueblos Originarios Viviendo en Nueva York* (‘The Council of Indigenous Peoples Living in New York’) have asserted that the map will be one of the most powerful tools at their disposal for lobbying for recognition and resources from the city government. An Armenian New Yorker was delighted to find that not just ‘Armenian’ is displayed, but also (endangered) Western Armenian, (the national language) Eastern Armenian, and (the liturgical language) Classical Armenian. An activist for the West African script N’ko proudly noted the correct use and encoding of the script in the endonym for the Mandinka language. No community or individual has asked for their language to be removed from the map, though eyebrows have been raised about ethnolects included such as ‘Jewish English’ or ‘Mexican Spanish,’ reflecting sensitivity that these may somehow be nonstandard or insufficiently distinctive variants with some social stigma related to perceptions of particular groups.

For journalists, whose coverage by its very nature brings visibility but who have also evolved safeguards to protect individual identities, the map also serves as a reference that can lead them to ask sharper questions and discover that their sources may be Indigenous. Major news stories, from immigration to COVID-19 to the organizing of food delivery workers, have vital Indigenous dimensions that have been consistently overlooked because of invisibility (Craig et al. 2021). Articles like Hol-

puch 2020 cite the map and center Indigenous voices in Corona, Queens, one of the neighborhoods hardest hit by the pandemic in the country, while similar and otherwise exemplary reports discuss the struggles of Indigenous Latin Americans in New York City without acknowledging their identities beyond Guatemalan and Mexican.

For policy makers at the city level, the map is already serving as a sorely needed guide to glaring blind spots. The ELA has now had several years' experience working with the city's 2020 census outreach team, the NYC Department of Health, and the Mayor's Office of Immigrant Affairs. In policy-making environments, city resources simply cannot be allocated to communities without some justification drawing on a published source, ideally statistical. The map provides a starting point, a validation from a linguistic point of view of what at least some community leaders and organizers already know – something tangible (if digital) that they can point to. With resources, we envision future mapping projects or extensions of this map specifically designed to support policy makers and to understand spatial language data in relation to other data, both around health and other issues. Daurio et al. (2020) describe our initial attempt to map COVID-19 case data at the height of the pandemic in New York City onto the language data set, observing how the city's most multilingual communities, for a variety of reasons, were among the most affected by the first wave of the pandemic.

Having only recently launched LANGUAGEMAP.NYC in late April 2021, we anticipate analyzing in more depth the various reactions and uses that emerge. An initial wave of over 10,000 users within just a few days speaks to the considerable interest a public-facing language map like this can generate. Tellingly, the most common type of feedback from users, at least so far, has been the request to add their language to the map.

## References

- Anderson, Margo J. 2015. *The American census: A social history*. 2nd edn. New Haven: Yale University Press.
- Andriani, Luigi, Ross Perlin, & Daniel Kaufman. Forthcoming. Dialects in diaspora: Preservation and loss in Italian New York. In Hajek, John & Francesco Goglia (eds.), *Italian(s) abroad: Italian language and migration in cities of the world*. Berlin: Mouton de Gruyter.
- Auer, Peter & Jürgen Erich Schmidt (eds.) 2010. *Mapping language and space: An international handbook of linguistic variation*, vol. 1. Berlin: Walter de Gruyter GmbH & Co. KG.
- Bakker, Peter & Maarten Mous. 1994. *Mixed languages: 15 case studies in language intertwining*. Amsterdam: Institute for Functional Research into Language and Language Use.
- Blommaert, Jan & Ben Rampton. 2012. Language and superdiversity. In *MMG working papers*, 1–36. Göttingen: Max-Planck Institute for the Study of Religious and Ethnic Diversity. (MMG Working Paper 12-09.) (<http://www.mmg.mpg.de/59855/wp-12-09>) (Accessed 2021-05-13.)

- Brunn, Stanley D. & Ronald Kehrein (eds.). 2020. *Handbook of the changing world language map*. Cham: Springer Nature Switzerland AG.
- Christ, Diarmait Mac Giolla & Huw Thomas. 2008. Linguistic diversity and the city: Some reflections, and a research agenda. *International Planning Studies* 13(1). 1–11.
- Craig, Sienna, Maya Daurio, Daniel Kaufman, Ross Perlin, & Mark Turin. 2021. The unequal effects of COVID-19 on multilingual immigrant communities. In *RSC COVID-19 Series*. Ottawa: Royal Society of Canada. (Publication #97.) (<https://rsc-src.ca/en/voices/unequal-effects-covid-19-multilingual-immigrant-communities>) (Accessed 2021-05-13.)
- Daurio, Maya & Mark Turin. 2020. “Langscapes” and language borders: Linguistic boundary-making in northern South Asia. *Eurasia Border Review* 10(1). 21–42.
- Daurio, Maya, Sienna R. Craig, Daniel Kaufman, Ross Perlin, & Mark Turin. 2020. Subversive maps: How digital language mapping can support biocultural diversity—and help track a pandemic. *Langscape Magazine* 9. 8–13.
- Dorian, Nancy C. 1981. *Language death: The life cycle of a Scottish Gaelic dialect*. Philadelphia: University of Pennsylvania Press.
- Drude, Sebastian. 2018. Why we need better language maps, and what they could look like. In Drude, Sebastian, Nicholas Ostler, & Marielle Moser (eds.), *Endangered languages and the land: Mapping landscapes of multilingualism*, 33–40. London: FEL & EL Publishing. (Proceedings of FEL XXII/2018, Reykjavík, Iceland.)
- ELA, Endangered Language Alliance. n.d. (Airtable of database for *Languages of New York City* map.) (<https://airtable.com/shrqQo5FJHvhKtffs/tblHBOmrPVk-0WJGZ1>) (Accessed 2021-05-19.)
- England, Sarah. 2006. *Afro Central Americans in New York City: Garifuna tales of transnational movements in racialized space*. Gainesville: University Press of Florida.
- Errington, Joseph. 2007. *Linguistics in a colonial world: A story of language, meaning, and power*. Malden: Blackwell Publishing.
- Feliciano-Santos, Sherina. 2017. How do you speak Taíno?: Indigenous activism and linguistic practices in Puerto Rico. *Journal of Linguistic Anthropology* 27(1). 4–21.
- Gaiser, Leonie & Yaron Matras. 2016. *The spatial construction of civic identities: A study of Manchester’s linguistic landscapes*. Manchester, UK: Multilingual Manchester, University of Manchester. (<http://mlm.humanities.manchester.ac.uk/wp-content/uploads/2016/12/ManchesterLinguisticLandscapes.pdf>) (Accessed 2021-05-10.)
- Gambino, Christine. 2018. *American community survey redesign of language-spoken-at-home data, 2016* (Social, Economic, and Housing Statistics Division Working Papers, 2018-31). Suitland: U.S. Census Bureau. (<https://www.census.gov/content/dam/Census/library/working-papers/2018/demo/SEHSD-WP2018-31.pdf>) (Accessed 2021-05-13.)
- Garcia, Ofelia & Li Wei. 2014. *Translanguaging: Language, bilingualism and education*. London: Palgrave Macmillan.

- Gawne, Lauren & Hiram Ring. 2016. Mapmaking for language documentation and description. *Language Documentation & Conservation* 10. 188–242. (<http://hdl.handle.net/10125/24692>)
- Good, Jeff & Michael Cysouw. 2013. Languoid, doculect, and glossonym: Formalizing the notion “language.” *Language Documentation & Conservation* 7. 331–359. (<http://hdl.handle.net/10125/4606>)
- Google. n.d. *Google Noto fonts*. (<https://www.google.com/get/noto/>) (Accessed 2021-05-19.)
- Granberry, Julian & Gary Vescelius. 2004. *Languages of the pre-Columbian Antilles*. Tuscaloosa: The University of Alabama Press.
- Grinevald, Colette. 2003. Speakers and documentation of endangered languages. In Austin, Peter K. (ed.), *Language documentation and description*, vol. 1, 52–72. London: SOAS.
- Gurung, Nawang, Ross Perlin, Daniel Kaufman, Mark Turin, & Sienna R. Craig. 2018. Orality and mobility: Documenting Himalayan voices in New York City. *Verge: Studies in Global Asias* 4(2). 64–80.
- Gurung, Nawang Tsering, Ross Perlin, Mark Turin, Sienna R. Craig, Maya Daurio, & Daniel Kaufman. 2020. Himalayan New Yorkers tell stories of COVID-19. *Nepali Times*, June 6. (<https://www.nepalitimes.com/here-now/himalayan-new-yorkers-tell-stories-of-covid-19/>) (Accessed 2021-05-13.)
- Holpuch, Amanda. 2020. Corona in Corona: Deadly toll in a New York neighborhood tells a story of race, poverty and inequality. *The Guardian*, June 15. (<https://www.theguardian.com/us-news/2020/jun/15/coronavirus-corona-queens-ny-virus-shook-neighborhood>) (Accessed 2021-05-24.)
- Hubley, Jill. 2016. Languages of NYC: Most frequently spoken language at home, excluding English and Spanish, by census tract. (Digital map.) (<https://www.jill-hubley.com/project/nylanguages/>) (Accessed 2021-05-24.)
- Hull-House maps and papers, a presentation of nationalities and wages in a congested district of Chicago, together with comments and essays on problems growing out of the social conditions*. 1895. New York: T. Y. Crowell & co. (Map.) (<https://lcn.loc.gov/04003818>) (Accessed 2021-05-18.)
- Jameson, John Franklin (ed.). 2010. *Narratives of New Netherland*. New York: Cosimo Classics.
- Kaufman, Daniel. Forthcoming. The Mixtec language in New York: Vitality, discrimination and identity. In Hajek, John, Catrin Norrby, Heinz L. Kretzenbacher, & Doris Schuepbach (eds.), *Multilingualism and pluricentricity: A tale of many cities*. Berlin: De Gruyter Mouton.
- Kaufman, Daniel & Ross Perlin. 2018. Language documentation in diaspora communities. In Regh, Kenneth L. & Lyle Campbell (eds.), *Oxford handbook of endangered languages*, 398–418. Oxford: Oxford University Press.
- Kertzer, David I. & Dominique Arel (eds.). 2001. *Census and identity: The politics of race, ethnicity, and language in national censuses*. Cambridge: Cambridge University Press.

- Lameli, Alfred, Roland Kehrein, & Stefan Rabanus (eds.). 2010. *Language and space: An international handbook of linguistic variation*, vol. 2: Language mapping part 1. Berlin: De Gruyter Mouton.
- Lampel, Jason. n.d. *A Better Map*. (<https://www.abettermap.com/>) (Accessed 2021-05-18.)
- Landry, Rodrigue & Richard Y. Bourhis. 1997. Linguistic landscape and ethnolinguistic vitality: An empirical study. *Journal of Language and Social Psychology* 16(1). 23–49.
- Languages of New York City Map*. n.d. (GitHub README.md) (<https://github.com/Language-Mapping/language-map#readme>) (Accessed 2021-05-19.)
- Leonard, Wesley Y. 2008. When is an “extinct language” not extinct?: Miami, a formerly sleeping language. In King, Kendall A., Natalie Schilling-Estes, Jia Jackie Lou, Lyn Fogle, & Barbara Soukup (eds.), *Sustaining linguistic diversity: Endangered and minority languages and language varieties*, 23–33. Washington, DC: Georgetown University Press.
- Lüpke, Friederike. 2010. Language and identity in flux: In search of Bainouk. *Journal of Language Contact* 3(1). 155–174.
- Lüpke, Freiderike & Anne Storch. 2013. *Repertoires and choices in African languages*. Berlin: De Gruyter Mouton.
- Mair, Victor. 1991. What is a Chinese “dialect/topolect”? Reflections on some key Sino-English linguistic terms. *Sino-Platonic Papers* 29. 1–31.
- Makoni, Sinfree & Alastair Pennycook (eds.). 2007. *Disinventing and reconstituting languages*. Clevedon: Multilingual Matters.
- Matras, Yaron. 2018. The Multilingual Manchester research model: An integrated approach to urban language diversity. *Acta Linguistica Petropolitana* 14(3). 248–274.
- MIT License. n.d. (<https://choosealicense.com/licenses/mit/>) (Accessed 2021-05-19.)
- Mitchell, Ross E., Travas A. Young, Bellamie Bachleda, & Michael A. Karchmer. 2006. How many people use ASL in the United States?: Why estimates need updating. *Sign Language Studies* 6(3). 306–335.
- Moore, Robert E., Sari Pietikäinen, & Jan Blommaert. 2010. Counting the losses: Numbers as the language of language endangerment. *Sociolinguistic Studies* 4(1). 1–26.
- Museum of the City of New York. n.d. Who we are: Visualizing NYC by the numbers. (<https://www.mcny.org/exhibition/who-we-are>) (Accessed 2021-05-19.)
- Musgrave, Simon & John Hajek. 2010. Sudanese languages in Melbourne: Linguistic demography and language maintenance. In Treis, Yvonne & Rik De Busser (eds.), *Selected papers from the 2009 conference of the Australian Linguistic Society*, 1–17. Australian Linguistic Society.
- Mutter, Christina & Florian Zacherl. 2019. Visualising language in space: New approaches in linguistic cartography. (Workshop on Visualization for Digital Humanities, Vancouver.)
- NYC Census 2020. n.d. Census. (<https://www1.nyc.gov/site/census/index.page>) (Accessed 2021-05-19.)




- NYC Department of City Planning. n.d. Metro Region Explorer. (Digital map.) (<https://metroexplorer.planning.nyc.gov>) (Accessed 2021-05-19.)
- NYC *Language Mapping*. n.d. (GitHub site.) (<https://github.com/Language-Mapping>) (Accessed 2021-05-19.)
- Pennycook, Alastair & Emi Otsuji. 2015. *Metrolingualism: Language in the city*. London: Routledge.
- Perlin, Ross. 2016. The race to save a dying language. *The Guardian*, Aug. 1. (<https://www.theguardian.com/news/2016/aug/10/race-to-save-hawaii-sign-language>) (Accessed 2021-05-24.)
- Perlin, Ross & Daniel Kaufman (eds.). 2020. *Languages of New York City*, 3rd edn. New York: Endangered Language Alliance. (Map.)
- Perlin, Ross, Daniel Kaufman, Jason Lampel, Maya Daurio, Mark Turin, & Sienna Craig (eds.). 2021. *Languages of New York City*. New York: Endangered Language Alliance. (Digital map.) (<https://languagemap.nyc/>) (Accessed 2021-05-25.)
- Quinto-Pozos, D. (2008). Sign language contact and interference: ASL and LSM. *Language in Society* 37(2). 161–189.
- Schillo, Julia & Mark Turin. 2020. Applications and Innovations in Typeface Design for North American Indigenous Languages. *Book 2.0* 10 (1). 71–98.
- Sharifian, Farzad & Simon Musgrave. 2013. Migration and multilingualism: Focus on Melbourne. *International Journal of Multilingualism* 10(4). 361–74.
- Smakman, Dick & Patrick Heinrich (eds.). 2017. *Urban sociolinguistics: The city as a linguistic process and experience*. London: Routledge.
- Sohnit, Rebecca & Joshua Jelly-Schapiro. 2016. *Nonstop metropolis: A New York City atlas*. Berkeley: University of California Press.
- Taylor, Charles. 1994. The politics of recognition. In Taylor, Charles, K. Anthony Appiah, Jürgen Habermas, Steven C. Rockefeller, Michael Walzer, & Susan Wolf (eds.), *Multiculturalism: Examining the politics of recognition*, 25–73. Princeton: Princeton University Press.
- Turin, Mark. 2021. The Digital Himalaya project: Collection, protection & connection. In *Visualizing objects, places, and spaces: A digital project handbook*. (Online handbook.) <https://doi.org/10.21428/51bee781.028ec770>
- Turin, Mark. 2014. Mother tongues and language competence: The shifting politics of linguistic belongings in the Himalayas. In Toffin, Gérard & Joanna Pfaff-Czarnecka (eds.), *Facing globalization in the Himalayas: Belonging and the politics of the self*, vol. 5: Governance, conflict and civic action, 372–396. New Delhi: SAGE Publications India.
- United Nations Population Fund. n.d. *Urbanization*. (<https://www.unfpa.org/urbanization>) (Accessed 2021-05-18.)
- United States Census Bureau. n.d. American Community Survey. (<https://www.census.gov/acs/www/about/why-we-ask-each-question/language/>) (Accessed 2021-05-18.)
- Vandenbroucke, Mieke. 2020. Mapping visible multilingualism in Brussels' linguistic landscapes. In Brunn, Stanley D. & Roland Kehrein (eds.), *Handbook of the changing world language map*, 1525–1543. Cham: Springer.

- Van der Merwe, I. J. 1993. The urban geolinguistics of Cape Town. *GeoJournal* 31(4). 409–417.
- Van Rees, Eric. 2015. Mapbox and the new age of mapping. *GeoInformatics* 18(5). 3.
- Vertovec, Steven. 2007. Super-diversity and its implications. *Ethnic and Racial Studies* 30(6). 1024–1054.
- Veselinova, Ljuba Nikolova & J. C. Booza. 2009. Studying the multilingual city: A GIS-based approach. *Journal of Multilingual and Multicultural Development* 30(2). 145–165.
- Wallace, Mike. 2017. *Greater Gotham: A history of New York City from 1898 to 1919*. New York: Oxford University Press.
- Weslager, C. A. 1999. *The Delaware Indians: A history*. New Brunswick: Rutgers University Press.
- Willis, Christina M. 2013. *The voices of Houston: A linguistic survey*. Houston: Rice University Kinder Institute for Urban Research. (Report.) <https://doi.org/10.25611/l3qn-pd1k>


Ross Perlin

[perlin@elalliance.org](mailto:perlin@elalliance.org)

 [orcid.org/0000-0003-1932-7057](https://orcid.org/0000-0003-1932-7057)


Daniel Kaufman

[kaufman@elalliance.org](mailto:kaufman@elalliance.org)

 [orcid.org/0000-0003-0971-8409](https://orcid.org/0000-0003-0971-8409)


Mark Turin

[mark.turin@ubc.ca](mailto:mark.turin@ubc.ca)

 [orcid.org/0000-0002-2262-0986](https://orcid.org/0000-0002-2262-0986)


Maya Daurio

[maya.daurio@ubc.ca](mailto:maya.daurio@ubc.ca)

 [orcid.org/0000-0002-5650-6604](https://orcid.org/0000-0002-5650-6604)


Sienna Craig

[sienna.r.craig@dartmouth.edu](mailto:sienna.r.craig@dartmouth.edu)

 [orcid.org/0000-0002-6760-762X](https://orcid.org/0000-0002-6760-762X)

Jason Lampel

[jason@abettermap.com](mailto:jason@abettermap.com)

 [orcid.org/0000-0002-2026-9272](https://orcid.org/0000-0002-2026-9272)