Geospatial Modelling and Linguistics: What Can Maps Tell Us About Language?

Ceci G. Williamson*

Swarthmore College, December 2021

*I would like to thank Professors Ted Fernald and Brook Lillehaugen, for their thoughtful comments and feedback throughout the writing process. Thank you to Professor Emily Gasser for her guidance throughout my time in the Pacific Languages Lab and generosity with her expertise and data, and Professor Jake Grossman for his reality checks about ecological modelling. Special thanks also to my peer reviewers, Martín Rakowszczyk and Anatole Shukla, for their insight and encouragement.

Abstract

Linguistic geography has been relegated to the periphery of both linguistics and geography, with scholars citing the lack of a codified methodology for interpreting linguistic data and a dearth of geospatial linguistic data itself as reasons for this marginalization (Ambrose & Williams 1992, Luebbering 2011, Haynie & Gavin 2019, Haynie 2014:344). While most published grammars include an area map for illustrative purposes, linguistic map-making has no overarching standards for replicability, transparency, or overall methodology, and the analytical and explicative possibilities of geospatial techniques remain untapped (Ambrose & Williams 1992, Luebbering 2011, Haynie & Gavin 2019). Even as analytical tools like GIS software are becoming more easily accessible, "geolinguistics" has not made the same strides forward as disciplines like ecology or other social sciences.

The twin limitations of *lack of data* and *lack of methodology* combine to restrict the use, and therefore the wider understanding and further development, of geospatial techniques in linguistics. Dialect studies, historical linguistics, and language diversity studies, for instance, can all utilize the

illustrative and analytical capabilities of maps. In this paper, we will first consider the issue of geospatially relevant linguistic data – both how it's encoded and some inherent challenges – before addressing a series of case studies that do apply geospatial techniques to linguistic fields. I will highlight a few approaches across subdisciplines: isogloss and dialect maps, loanword visualization, and statistical modelling of language diversity. These cases provide an overview of the potential of geospatial techniques as well as highlighting the importance of transparent methodologies, data availability, and theoretical backing for any analysis undertaken.

Introduction

Linguistic geography has been relegated to the periphery of both linguistics and geography for the majority of the existence of both these fields, with scholars citing the lack of a codified methodology for interpreting linguistic data and a dearth of geospatial linguistic data itself as reasons for this marginalization (Ambrose & Williams 1992, Luebbering 2011, Haynie & Gavin 2019, Haynie 2014:344). Most published grammars include an area map for illustrative purposes, grounding their work in the physical location of the language community, and general-purpose linguistic sources like Ethnologue include maps in their descriptions of languages (Eberhard et al. 2021). Linguistic map-making, however, has no overarching standards for replicability, transparency, or overall methodology, and the analytical and explicative possibilities of geospatial techniques remain untapped (Ambrose & Williams 1992, Luebbering 2011, Haynie & Gavin 2019). Even as analytical tools like GIS software are becoming more easily accessible, "geolinguistics" has not made the same strides forward seen in disciplines like ecology or other social sciences.

The twin limitations of *lack of data* and *lack of methodology* combine to restrict the use, and therefore the wider understanding and further development, of geospatial techniques in linguistics. Dialect studies, historical linguistics, and language diversity studies all stand to benefit from the illustrative and analytical capabilities of maps. In this paper, we will first consider the issue of *geospatial linguistic data*. Any linguistic data that can be meaningfully connected to a physical location could be considered "geospatially relevant" – this could take the form of speaker locations or

hometowns in the case of lexical or phonetic surveys, the number of languages spoken in a given region in the case of language diversity studies, or even census data cataloguing a population's self-reported language use in sociolinguistics. How data is encoded and representing language spatially present some inherent challenges to geolinguistics, and this will be discussed in the first section. I will next address a series of case studies that do apply geospatial techniques to linguistic disciplines. While developing a complete methodology for linguistic mapping is beyond the scope of this paper, I will highlight a few approaches. Teerarojanarat and Tingsabadh (2011a) update the technique of isogloss mapping using digital tools, providing transparent methodology for work with lexical survey data and generating meaningful & illustrative results . My own work digitizing language area maps in West Papua combines loanword data and language areas to visually represent loan patterns across the Bird's Head region, expanding on the work of Gasser (2020). Finally, two studies, Pacheco Coelho et al. (2019) and Antunes et al. (2020), apply statistical modelling to language diversity studies, with varying levels of insights resulting. These cases provide an overview of the potential of geospatial techniques as well as highlighting the importance of transparent methodologies, data availability, and theoretical backing for any chosen analytical technique.

Why choose a geographic approach?

Not every discipline in linguistics is the best fit for geospatial tools. Geospatial analysis can shed light on the history of language interactions and origins, meaning historical linguistics and dialectology are good candidates (e.g. Haynie et al. 2014, Gasser 2020, Bowern et al. 2014). While the scope and type of insight that geospatial tools can provide depends on the available data, the relationship between time and physical distribution provides a theoretical basis for these approaches: namely, that people and languages disperse across space over time. Tobler's First Law of Geography (1970) states that *spatial autocorrelation* is the "tendency for spatially near values or language varieties to be more similar than spatially distant ones" (Haynie 2014:344). Languages near to one another are

more likely to share a common ancestor, and phylogenetically unrelated languages in close proximity to one another are more likely to trade features than distant languages that might never come into direct contact. Haynie typifies these interactions as "diffusion, divergence, and accommodation or convergence," any or all of which may take precedence in language interactions (2014:345).

Diffusion refers to the horizontal transfer of traits across languages, as with loanwords, but also including grammatical structures and phonological traits (Campbell & Mixco 2007:45). *Divergence* (also *diversification*) refers to the process by which a parent, or proto-language, splits into dialects and eventually discreet languages (Campbell & Mixco 2007:48-9). In contrast to divergence, *convergence* refers to diffusion in practice in an area: unrelated languages can become more similar over time as they continue to be in contact, potentially leading to the formation of a Sprachbund (or "diffusion area") in which "languages of a region come to share certain structural features" (Campbell & Mixco 2007:106). At the same time, *accommodation* (also *naturalization*) shifts the structure of borrowed terms and structures to better fit the recipient language (Campbell & Mixco 2007:134). These routes of language development can all be acting on a language or language family at the same time, and geographic spread can help establish relationships between languages, as all these processes are facilitated or inhibited by spatial proximity or *distalness*.

In both linguistics and biological evolution, "spatial scales of patterns typically correlate with the time depths of the associated phenomena" (Haynie 2014:345). Generally speaking, patterns across smaller regions reveal recent changes – dialect divergence or lexical loans – while large-scale, potentially continent-wide patterns prove more informative about language families and more distant proto-languages (Haynie 2014:345, Pacheco Coelho et al. 2019). The geospatial relationships between languages can serve as powerful explanatory factors, especially at large spatial scales, time depths, or in areas like Sprachbunds with high rates of borrowing and non-inherited linguistic similarity (Pacheco Coelho et al. 2019). The island of Papua (both Indonesian Papua and Papua New

Guinea) is an area with a high density of linguistic endemism, high rates of borrowing, and some ambiguity in language phylogenies, which we will return to in Case Study 2: West New Guinean Loanwords (Usher & Schapper 2018).

Section 1: Data

What does "Geolinguistic" Data Look Like?

Before approaching case studies or analytical techniques based in geospatial data, we first need to understand how this data is generated and stored. *Geographic data* refers to any piece of information that is linked to a specific location. For example, a wordlist generated by interviewing a speaker of a particular language is linguistic data, and if the interviewer writes down the town that speaker lives in, then the wordlist could also be geo-linguistic data. Geographic Information System (GIS) software has several ways of encoding this spatial information. Data can be *raster* or *vector* based. Rasters are akin to pixelated images: great at representing data densely sampled and without clear delineation, like rainfall across an area or percent vegetation cover. Raster data has limited resolution, though, based on the initial sampling density, similar to zooming in on a pixelated image. Vectors allow for discrete boundaries and are generally one of three types: point, line, or polygon. Lines and polygons are defined by points, which in turn are sets of coordinates on the earth's surface. This definition makes vectors theoretically useful at any spatial scale – zooming in past a certain point does not make the data less clear, since vector lines are defined by the relationships between fixed points rather than information gathered at a particular granularity. Each of these types of vectors, again, is suited to representing a different type of spatial information. Polygons represent areas; lines represent boundaries, paths or rivers; and points correspond with individual instances of items in the real world, like trees or sightings of a famous person. The example of

eliciting a wordlist given above might be encoded as a point – the coordinates of the center of town, for example, or even the coordinates of the speaker's house or where they grew up.

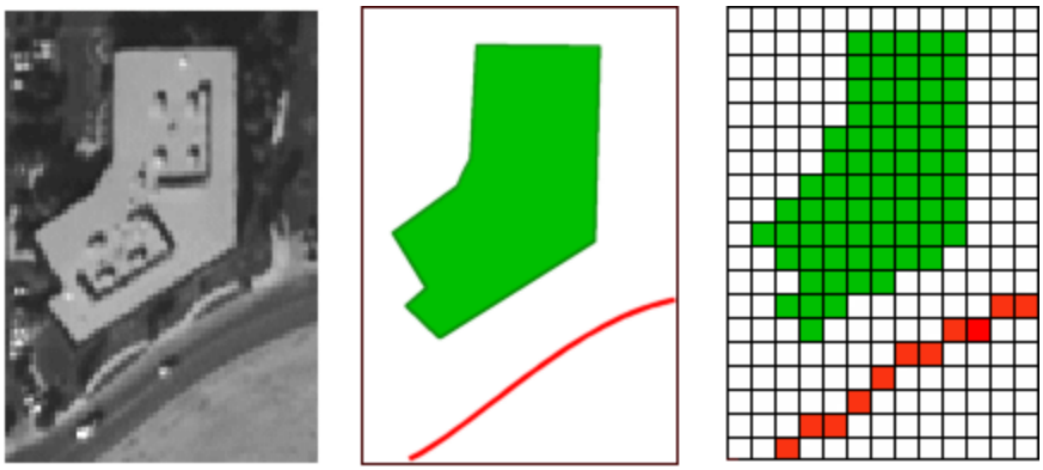

Figure 1. From left to right: aerial imagery, a vector representation of that image, and a rasterized representation of that image. Source: https://storymaps.arcgis.com/stories/187a2dab68f646a38d410a297b911348

Figure 2. Point, line, and polygon data. Polygon vector or raster representations are most commonly used for language areas, and point features are the most useful in geospatial analyses.

Languages aren't easily described by a point, line, or polygon. The distinctions between dialects are often a continuous gradient, not suited to the delineations made by vector lines (Teerarojanarat & Tingsabadh 2011a:56, Luebbering 2011:9, Stone 2018: 42). Migration and language change ensure that any map is a 'snapshot' in time, not a definitive description, and multilingualism refutes the clean distinctions of abutting polygons in favor of overlapping and intertwined areas of speaker populations (Luebbering 2011:9). Candace Luebbering remarks in her dissertation that mapping individual speakers would provide the most accurate units of linguistic data, but as this approach is infeasible in terms of both scale and ethics, sociopolitical borders are used as convenient abstractions (2011:9, Teerarojanarat and Tingsabadh 2011a: 61). Sociopolitical

borders are built into mapping software like QGIS, which makes them convenient and a useful reference point for the viewer, but alternative approaches like Thiessen polygons might be more true to the data actually collected (Teerarojanarat and Tingsabadh 2011a: 72).

In order to minimize encoding assumptions or generalizations in the geospatial data itself, data like that from a language survey can be encoded as point (incidence) data. One point might be a village or a recording location, and the data associated with that point could be a wordlist, a single utterance, a story, etc. While the survey data or historical resources that make up a point data set are as subject to biases as any nongeographic data set, this type of encoding at least deemphasizes the broad generalizations present in maps of language areas – especially those that rely on sociopolitical borders. Much like a scatterplot, point data alone does not allow for useful conclusions to be drawn, but it provides the first step towards more meaningful mapping techniques, including isogloss mapping, dialect boundaries, and statistical modelling, all of which will be explored in more depth in Section 2: Approach.

Point-based incidence data facilitates a variety of geospatial applications. One of the oldest of these is isogloss mapping (Haynie 2014). Isogloss mapping involves determining dialect boundaries based on bundles of features in surveyed data – the line between two regions falls where the features (mostly) change over from dialect to dialect, allowing for geographic delineations of dialect boundaries. (Teerarojanarat and Tingsabadh 2011b). One of Teerarojanarat and Tingsabadh's studies on Thai dialects successfully collated two non-coordinated surveys separated by 50 years to show the changes in dialect distributions over time (2011b:365).

In some instances, point data isn't available, or not enough of it is available to be useful. In the Case Study 2, in New Guinea, the finest resolution of language data available is in the form of languages area maps in prior publications (Figure 7) (Emily Gasser, pers. comm.). It's possible to

work with language areas like these using mathematical techniques, like calculating the centroid of a given area, or by choosing types of analysis compatible with areas rather than points.

Limitations of Geolinguistic Data

In addition to the difficulty of encoding language as points, lines, and polygons, the overall availability of detailed, georeferenced language data limits the employment of geospatial analytical techniques. First, data is often made available as images of maps included in language documentation like grammars or ethnographies (Haynie & Gavin 2019). These resources require extensive manpower to digitize and additional research to determine suitability for further geospatial analysis; moreover, they can vary widely in quality, accuracy, attribution, and level of detail (Haynie & Gavin 2019, Ambrose & Williams 1992: 309). Modern databases attempting to provide geographic language data still suffer from a high level of abstraction, often basing areas or parts of areas on geopolitical boundaries that may or may not be relevant to the phenomena under study (Lubbering 2011:2, Ambrose & Williams 1992:299). Even where more specific data is available, the coverage necessary to accurately map a dialect across an area is much greater than that needed for most purely linguistic, non-spatial studies (Ambrose & Williams 1992: 305). The researcher attempting to accurately map the extent of a linguistic phenomenon is left either conducting their own surveys at scale or working with abstracted areas from prior work, often in need of digitization.

The Theoretical Basis of Mapmaking

The foundation of a geospatial-linguistic methodology is to establish what a map, or tools associated with maps, can and cannot do. Before that, though, it must be understood that maps are, at their core, intended to tell a story or make a case (Krygier & Wood 2011:xiii). While "where a phenomenon is located" might be an objective statement, the choice to include that information in a map is a strategic decision, and every choice – from the colors used to differentiate language families to the designation of "language" versus "dialect" – changes the story that map conveys (Ambrose &

Williams 1992: 311). If a map is a subjective representation, and each step of its creation is a methodological choice, then each application geospatial tools needs to be tailored to the goal of the specific project. Works like Haynie & Gavin (2019) provide a blueprint for demystifying that methodology – in developing a continent-scale map of Indigenous languages at time of European contact, they include explicit priority rankings for conflicting language areas from a variety of sources and provide extensive metadata containing both attribution and acceptable use cases for each language area they include.

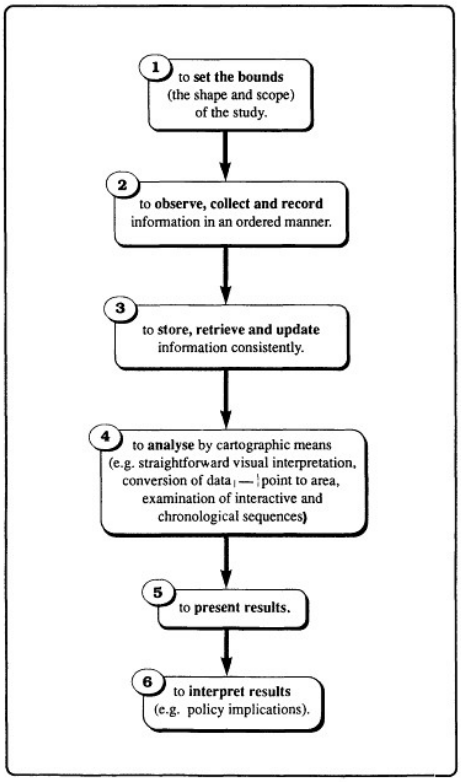

Figure 3. From Ambrose & Williams 1992. "The function of maps in geolinguistics."

Ambrose & Williams present one way of conceptualizing the function of maps in Figure 3. They contend that a map can address any subset of the boxes laid out in Figure 3, though likely not all of them at once. The success of the later initiatives – analysis, presentation, interpretation – are predicated on the successful completion of the prior boxes, or from drawing on sources that do

address those more fundamental concerns. Projects like OpenStreetMap, which provides open-source shapefiles for the entire globe (https://www.openstreetmap.org/), establish geopolitical borders and the shapes of landmasses & other physical features, making them both a useful resource in their own right and a foundation to build on. These projects fall into box 1. Raw survey data of the type collected by Teerarojanarat and Tingsabadh – that is, points indicating a speaker's judgement at a particular time – would comprise box 2, to "observe, collect and record information" (2011a, 2011b). Box 3, "store, retrieve and update," is one made much easier by the advent of widespread GIS in the intervening years since Ambrose & Williams first developed this schema, and any map project that is based in the first half of this chart should endeavor to encompass box 3 as well. Towards the latter half of this chart, matching the type of analysis to the scope and type of the data it is built on becomes critical. The composite isogloss map created by Teerarojanarat and Tingsabadh (2011a) is a great example of boxes 4 and 5, with nods to 6, as is explicated in the later section focusing on that project (Case Study 1: Dialect Boundaries and the Power of GIS).

Section 2: Approach

In this section, we will address a few analytical and representational techniques through case studies. These are intended both to illustrate the utility of mapmaking and geospatial analysis in linguistics, and to highlight best practices when employing these analytical techniques. Transparency of methodology and a strong theoretical backing are critical to drawing meaningful conclusions.

Case Study 1: Dialect Boundaries and the Power of GIS

Isogloss mapping is a technique for grouping dialect features to establish dialect distributions and continua that considers the geographic distributions of those dialects. In their 2011 paper "A GIS-based approach for dialect boundary studies," Teerarojanarat and Tingsabadh update the technique of isogloss mapping using digital tools. Isogloss mapping, a technique common in

dialectic linguistics, consists of first placing survey data – often phonological features – on a map, then drawing "isoglosses," or boundaries around groups of features used in common across an area (Haynie 2014:345). Isogloss maps can then be compiled to form dialect boundary maps (Teerarojanarat and Tingsabadh 2011a:65). This approach is vulnerable to subjectivity on the part of the researcher, especially when features do not "bundle" neatly or a dialect region contains more complex relationships, as in *Sprachbund* areas, meaning they depend heavily on the linguist's expertise (Haynie 2014:346). Isogloss maps are traditionally generated by hand-compiling maps from other publications, which may differ in scale, projection, level of detail, and accuracy, or suffer from copying errors, all of which contributes to uncertainty in the final product (Teerarojanarat and Tingsabadh 2011a: 56-7).

Isogloss maps are nonetheless a useful tool in addressing contested dialect areas, as in Haynie's 2012 work on Miwok dialect areas in central California (19-28). Archival sources on Miwok dialects recorded speaker biographical data, including place a birth or "hometown," alongside transcribed phonetic data (Haynie 2012:18). Plotting this data on a map allowed for dialect boundaries to be drawn on the basis of phonetic shifts across the Miwok range, which revealed that older sources placed the Central Sierra Miwok-Southern Sierra Miwok boundary accurately, but that the Northern Sierra-Central Sierra delineation was less distinct than anticipated (Haynie 2012:34). Isogloss maps form only one small part of the studies on Miwok conducted by Haynie, but their inclusion as an analytical tool provides insight into the historical relationships between plains and Sierra Miwok groups (Haynie 2012:34).

Given the utility of isogloss maps in describing dialect boundaries, there is a benefit to making the process of generating these maps more transparent, replicable, and fast. Teerarojanarat and Tingsabadh replace aspects of the traditional technique with digital, GIS-based approaches, with the understanding this produces maps that are more accurate, that allow for automation across large

datasets, and that can represent gradations of features in addition to discrete areas (2011a:67). Survey data on 170 Thai lexical items forms the basis of their study, with responses from 88% of subdistricts in Thailand (Teerarojanarat and Tingsabadh 2011a:61).

Responses were classified as belonging to the Central Thai dialect or a Non-Central Thai dialect – one of Northern Thai, North-eastern Thai, or Southern Thai (Teerarojanarat and Tingsabadh 2011a:60). They first assume that responses from a subdistrict are uniform across that area, and generate an isogloss map for each lexical item surveyed – 170 maps in all (Teerarojanarat and Tingsabadh 2011a:61). These maps are then compiled, with the number of Central Thai vs non-Central Thai lexical items reported in each subdistrict included as metadata. The use of GIS allowed the composite maps to depict gradations, as in Figure 5, showing both the percent usage of Central Thai in each area and the subdistricts for which no survey data was collected. This "gradation" map demonstrates the shift in dialects across the landscape with more nuance than a standard isogloss map.

Figure 4, below, was included in Teerarojanarat and Tingsabadh (2011a) to illustrate the changes made to the mapmaking process to incorporate digital tools. This diagram makes clear that isogloss maps, typically depicting only one unit of linguistic data (eg, a lexical item), are the foundation for generating dialect boundary maps, which encompass many pieces of linguistic data. The *linguistics* column in Figure 4 refers to techniques classically employed in dialect studies, beginning with lexical classifications – in this case study, classification by dialect – and moves on to manually drawing isogloss maps for each lexical item. The lexical items used to characterize dialects here could just as easily be morphemes, phonemes, speaker attestations as to what dialect they speak, or other data used to delineate one language group from another, depending on the data available to the researcher and their area of interest.

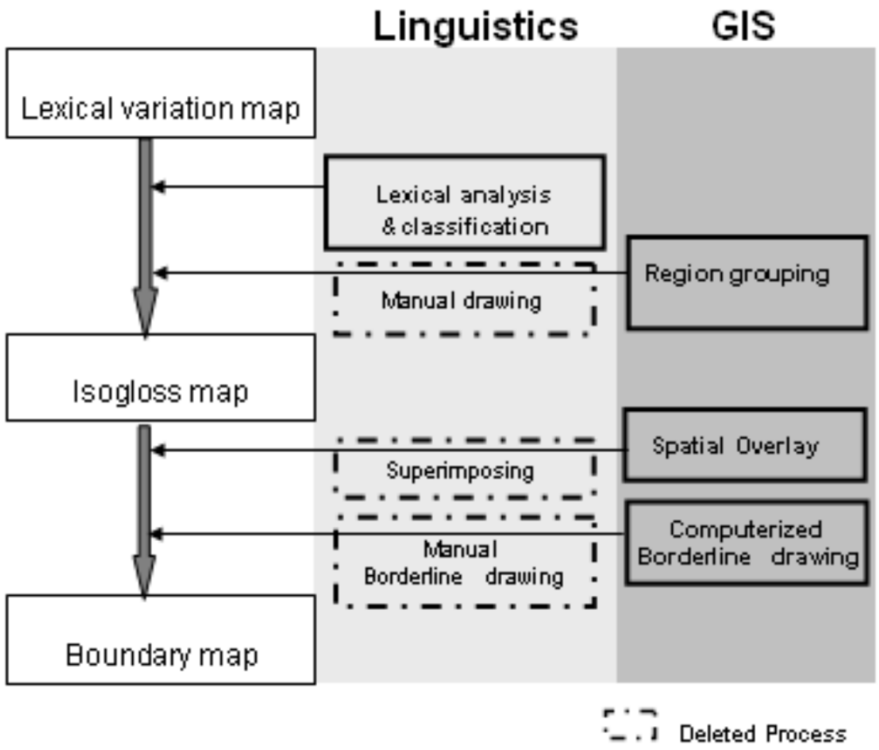

Figure 4: "Conceptual Methodology Diagram" from Teerarojanarat & Tingsabadh (2011a), noting changes between traditional and GIS-based mapping practices.

After the lexical analysis step, traditionally, the researcher uses their best judgement to place the boundaries when two dialects overlap, as in Haynie (2014). While Teerarojanarat and Tingsabadh classified their lexical items as per usual, they begin to employ digital tools at the isogloss drawing stage to make the process more objective and replicable – as shown in the *GIS* column. *Region grouping* and *spatial overlay* take the place of *manual drawing* and *superimposing* isogloss maps, different names for the same techniques but performed in GIS software.

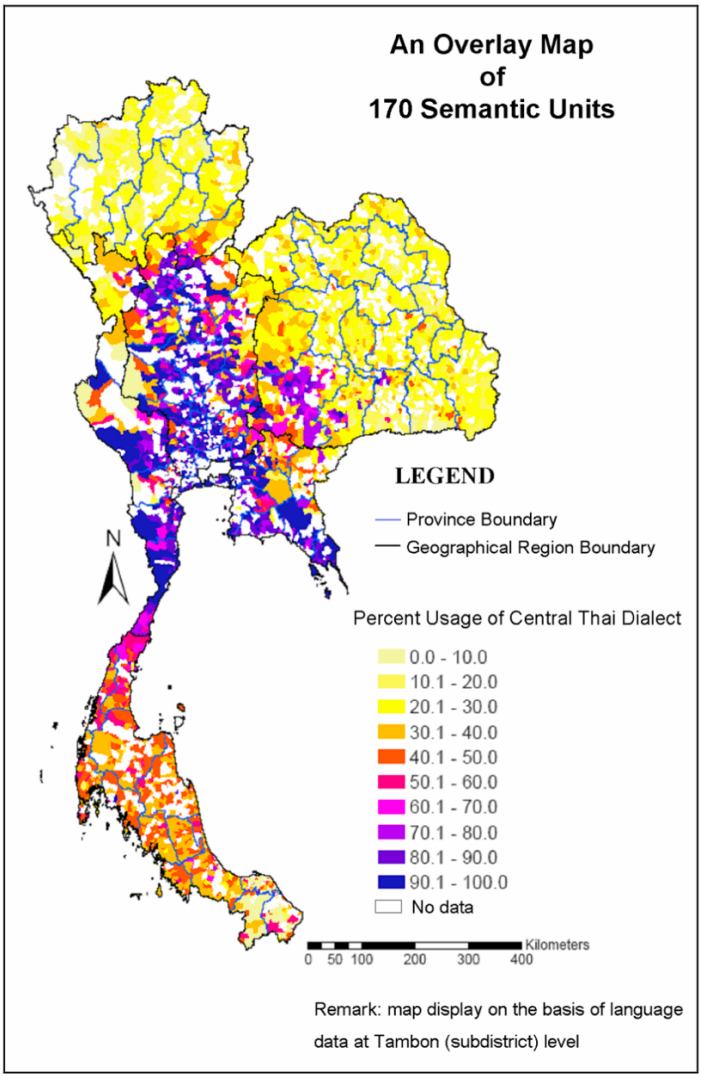

Figure 5. Gradiated isogloss map from Teerarojanarat & Tingsabadh (2011a).

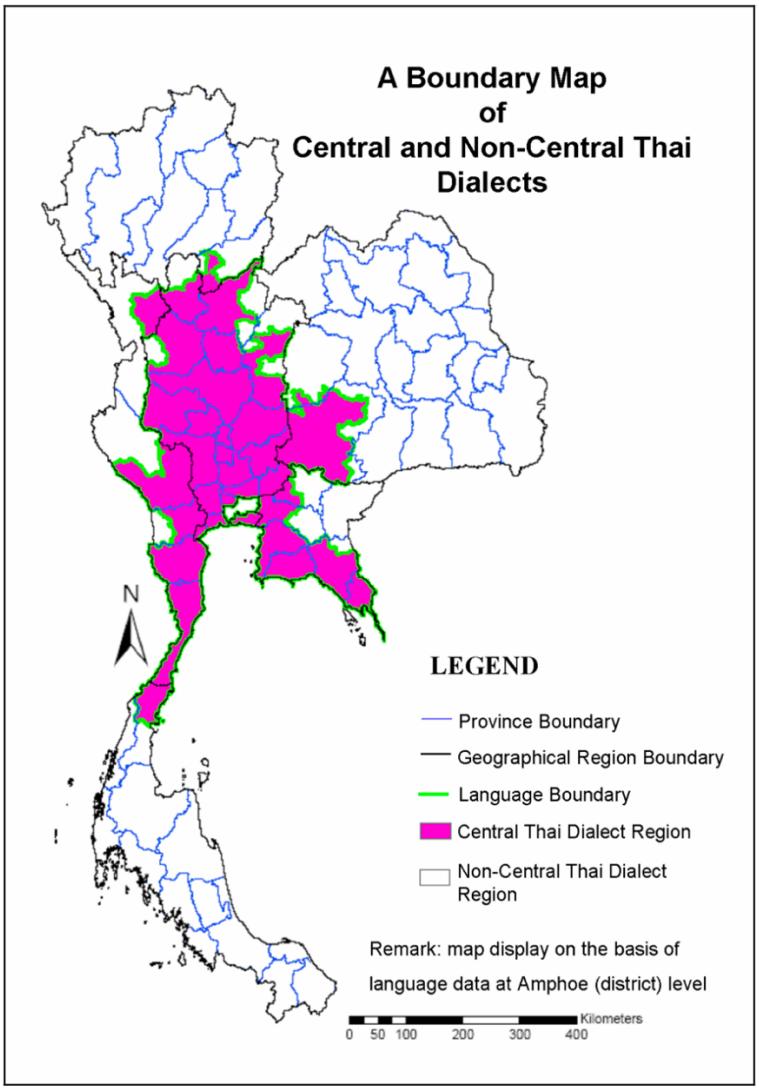

Figure 6. Dialect boundary based on 50% usage of the Central Thai dialect, from Teerarojanarat & Tingsabadh (2011a).

In the maps generated in Teerarojanarat and Tingsabadh 2011a, subdistrict level data coverage contained gaps from non-responding regions, so in order to generate a dialect boundary for Central Thai, data was aggregated at the district level (*region grouping* in Figure 4). In this case, after compiling all 170 isogloss maps (*spatial overlay*), areas where 50% or more of lexical items were attributable to Central Thai were included in the dialect boundary, and areas with 50% or more Non-Central Thai lexical items were excluded (Figure 6).

This approach results in a dialect boundary, much the same as traditional isogloss mapping, but with the important distinction that the process is both documented and replicable. The final border-drawing is still an imperfect representation of dialects that may be intermingled spatially or in actuality be a continuum, but GIS maps are more spatially accurate than hand-drawn ones, facilitate compiling larger numbers of isogloss maps easily, and allow the dialectologist tasked with drawing the final boundary to make their call with more confidence (Teerarojanarat and Tingsabadh 2011a:66). Thresholds like the 50% Central–Non-Central distinction, which appears to be an arbitrary choice, can be debated and improved upon, and additional data – for example, subdistricts that did not respond to the initial survey, additional lexical items, or other features – can be later added to the digital file to improve the accuracy of the boundary without re-building the entire map from scratch. This represents a sincere improvement on hand-complied maps, both in accuracy, clarity of methodology, and iterability.

This benefit, however, is predicated on the researchers' willingness to share and preserve digital versions of generated maps. A decade after its publication, the links in Teerarojanarat and Tingsabadh 2011a to online resources and nonpublished lexical item maps are dead (accessed 26 November 2021). While their approach is well-documented and clearly demonstrates the benefits of GIS integration with dialect studies, a researcher trying to pick up where Teerarojanarat and Tingsabadh left off might be set back to scanning and georeferencing the published images of maps in their paper. Digital archives and maintenance are critical to the long-term utility of digital mapping approaches – Ambrose & William's Step 3 (Figure 3).

Case Study 2: West New Guinean Loanwords

In addition to the process-oriented application of GIS mapping to dialect studies, as in Teerarojanarat and Tingsabadh (2011a), GIS mapping can illustrate spatial patterns in areas with high loanword incidences. Still focusing on lexical items as a form of language data, loanword maps

can use known language areas and recorded instances of borrowings across those languages to investigate potential spatial correlations. The example presented here focuses on the Bird's Head area of West Papua, a region of high language diversity and a dearth of documentation by linguists: 54% of West New Guinea languages are currently documented by a wordlist or less (Arnold, forthcoming:2).

The island of Papua was first settled by humans an estimated 47,000-51,000 years before the present day (Arnold, forthcoming:5). Small and mobile cultural groups across the island and surrounding archipelago developed extensive trade networks and a huge diversity of languages endemic to the island, the descendants of which are now classified as 'Papuan' languages though they are not necessarily related by heredity (Arnold, forthcoming:5, Gasser 2020:610). The phylogenetic relationships between Papuan languages are still being developed, and many are isolates or belong to extremely small families (Arnold, forthcoming:5, Gasser 2020:615). Loans between Papuan languages might have taken place at such a large time depth that sound change has rendered them undetectable, also making the linguistic phylogenies of the region difficult to characterize (Gasser 2020:629).

Linguistic and genetic phylogenies do indicate that a second wave of settlement by Austronesian groups originating in Taiwan took place 4,500-4,000 years ago, mostly taking up residence in lowland and coastal regions (Arnold, forthcoming:8). Languages descended from those brought by Austronesian groups are termed Austronesian languages, and conversely the term 'Papuan' is synonymous with 'non-Austronesian' although both non-Austronesian and Papuan languages are spoken on the island of Papua & surrounding areas (Gasser 2020: 610). The majority of Austronesian languages in West Papua belong to the South Halmahera-West New Guinea (SHWNG) family. For a thorough discussion of the languages spoken in the Bird's Head region, see Gasser (2020).

The similarities between Austronesian and Papuan languages in West New Guinea are extensive enough that some scholars have argued the area is best classified as a Sprachbund, an indication of the high level of lexical and grammatical borrowings occurring across the Papuan-Austronesian divide (Gasser 2020:611). Gasser (2020:636) reports that the "volume of loanwords [in West New Guinea] … suggests fairly intense, long-term interaction between language groups, likely including a fair level of bilingualism," but that it is often difficult to determine the source versus recipient of a loan. She recommends searching for "tell-tale distributions" of one language loaning with many others (that do not necessarily loan with each other) in order to identify *lingua francas* of the region (Gasser 2020:636). Of the languages that do follow this pattern, Gasser (2020) further justifies their potential *lingua franca* status using geospatial reasoning – namely, coastal regions of Cenderawasih Bay as facilitating trade & travel, or the central location of Irarutu to the Bird's Head peninsula and 'neck' region of Papua (Gasser 2020:636). Hence, the spatial patterning of languages in the Bird's Head region can inform us about their connections and relationships.

To generate the map in Figure 7, language regions were sourced from Gasser (2020) and Laura Arnold, pers. comm. Source map image files were transformed, georeferenced, and digitized using additional background maps to increase resolution. For example, coastlines were traced from publicly available OpenStreetMap files (https://osmdata.openstreetmap.de). In addition to increasing resolution, the language areas for West Papua are now updateable. Given that these regions are severely under-documented, as more accurate language range data becomes available, these map files can be improved without the need to start from scratch again. Next, the centroid of each language area was calculated using QGIS software, and language family and loanword relationships for the term 'bird' were added as metadata for each area. Language areas were then recolored to represent family relationships – for example, the dark grey in Figure 7 – and centroid points were colored & reshaped to represent loaned lexical items.

Loan relationships between terms for 'bird' were determined based both on semantic and phonetic similarity. Though it is possible some loans with lexical changes were overlooked, and independently arising phonetic similarities were included, these possibilities are minimized by the use of both criteria (E. Gasser, pers. comm.) In Figure 7, black circles indicate unique words for 'bird,' or similar terms within the same language family. Triangles indicate words for 'bird' deemed likely to be loaned, with at least one language in a different family sharing the term.

Green triangles represent words for 'bird' that sound like *man, mani,* or *mna* (from Biak, Ambel, and Umar respectively). The dark grey represents the language areas of Austronesian descent, both the SHWNG language family and non-SHWNG Austronesian languages, and light grey indicates Papuan language families in the region. Of note is the cooccurrence of *man*-like 'bird' words and the Austronesian families, with few green triangles falling outside of dark grey areas. This suggests that 'bird' terms are inherited across SHWNG and other Austronesian languages, remaining stable throughout the history of these families in New Guinea, rather than being loaned within the group. This theory is in line with reconstructions of Austronesian protolanguage Proto-Malayo-Polynesian *manuk for 'bird,' and the inclusion of the *man-* morpheme in *mankukei* 'chicken' (Ambai), *mangkokei* 'chicken' (Pom and Wooi) (Gasser 2020:623-4).

Red triangles indicate *du* or *ru* (Tause and Maybrat) -like 'bird' terms, possibly indicating some relationship facilitated this loan – if indeed these terms are loaned – beyond geographic proximity, given that Maybrat is landlocked in the center of the Bird's Head area. Gasser (2020) identifies Maybrat as a surprising case, sharing loans with distant Biakic, Yapen, and coastal Cenderawasih Bay languages with geographically intervening languages not attesting the same loaned lexical items, the same as we see with *du/ru* 'bird' (Gasser 2020:630). Gasser hypothesizes "direct

contact between the Maybrat and seagoing traders" could have facilitated these loans (2020:630). Tause, a coastal Lakes Plain language, could fit this description.

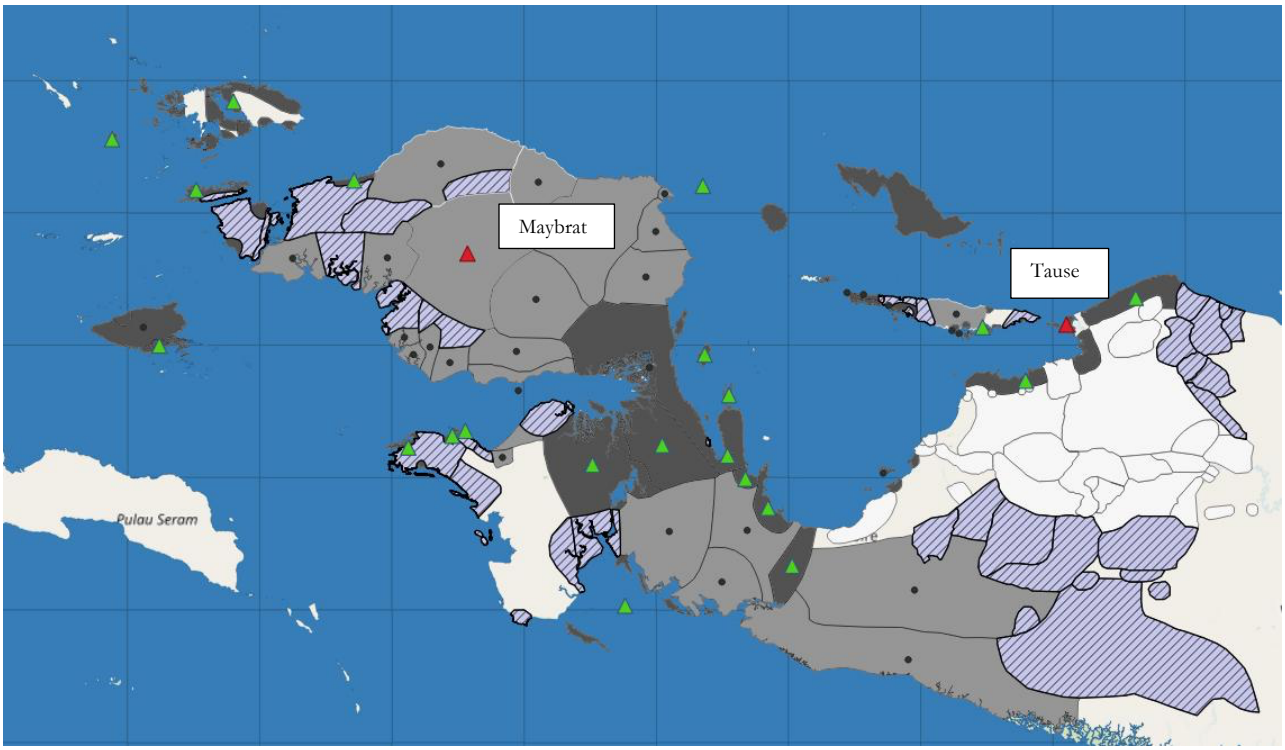

Figure 7. Western New Guinea language areas and representations of 'bird' terms. Light blue striped areas represent languages with no data, light grey regions are Papuan language areas, and dark grey regions are Austronesian language areas. Maybrat and Tause (red triangles) are labeled.

Clearly, one lexical item is not enough to draw any conclusions about loan patterns overall, though the relationships highlighted by this map do seem representative of previously established relationships. Further work with the loanword database developed by Dr. Emily Gasser could include composite maps with multiple loaned words encoded, similar to the methodology in Teerarojanarat and Tingsabadh 2011a as described in Case Study 1.

Case Study 3: Niche modelling in North America and West New Guinea

In her 2014 overview of geospatial approaches in linguistics, Haynie dedicates a section to "Language Diversity and the Environment," or analyses that draw on both linguistic and

environmental data (Haynie 2014). One of these approaches is a relatively new technique borrowed from ecology, called, variously, *species distribution modelling* (SDM), *predictive habitat distribution modelling*, *niche modelling*, and, in conjunction with language area data, *eco-linguistic niche modelling* (ELNM), among others (Miller 2010, Antunes et al 2020). SDM typically uses spatial data like species incidence, biodiversity metrics, and ecological parameters like rainfall and temperature. While Haynie (2014) draws attention to the lack of current understanding of the mechanisms of language-environment interactions and cautions against overestimating the explanatory power of environmental variables on language distributions, the potential explanatory capability and versatility of niche modelling nonetheless calls for a critical look at how this practice could be successfully and meaningfully adapted from ecology to linguistics.

On the linguistic side of this equation, the geographic data needed for niche modelling is in many cases severely lacking (Haynie & Gavin 2019). Modern language areas reflect the relatively recent history of colonialism and displacement over the last few hundred years, so accounting for historical distribution – likely to have a much stronger relationship between environmental variables and language diversity and distribution – adds yet another layer of complexity and potentially requires researchers to have specialized historical knowledge in any area they hope to model (Haynie & Gavin 2019).

Background

To begin determine if a niche modelling approach might be useful to linguistic applications, we need to understand the underlying theory that gave rise to this approach in ecology. The basis of niche modelling is the *ecological niche*. Every living species is theorized to exist in its own niche, defined as the " 'n-dimensional hypervolume' in environmental space in which the species can exist indefinitely" (Miller 2010:491). The 'dimensions' of this volume are environmental gradients, like altitude, temperature, and food or light availability (Miller 2010:492). However, species do not

inhabit all the regions that they theoretically could, whether due to competition, lack of dispersal ability, or some other biotic or abiotic interaction. This leads to the distinction between *fundamental* and *realized* niche: a species' *fundamental* niche is the "range of environmental conditions a species is physiologically able to tolerate" while its *realized* niche is where the species is actually found, often influenced by dispersion capabilities and historical ranges as much as current conditions (Miller 2010:492-3). *Niche modelling* is the process of mapping the "distribution of suitable environmental factors" that describe the realized (or potential) niche of a given species via mathematical modelling (Miller 2010:491). The relationship between geographic area and species distribution is a descriptive or correlative one, meaning environmental *predictor variables* like light, water, or resource availability describe, with varying degrees of accuracy, the *response variable:* a species' geographic distribution (Miller 2010:492). Niche modelling is predicated on the idea that the environment meaningfully limits the distribution of the response variable. Generally, the predictor and response variables need to have a theoretical basis for the application of any model to provide interpretable results (Miller 2010:493).

When applied correctly, niche modelling can be a powerful tool. Species distribution models can characterize existing habitat areas for endangered species, for example, and subsequently be used to identify other areas in which that a species might thrive in even though it does not currently exist there (Miller 2010:491). Models can predict range shifts due to climate change, a critically important factor in current conservation efforts (Heikkinen et al. 2006 *via* Miller 2010). Given this tantalizing potential, can niche modelling offer the same sorts of insights to historical or areal linguistics?

Modelling North American Linguistic Diversity

Pacheco Coelho et al. use geospatial statistical models, similar to niche models, to investigate the spatial patterns of language diversity in North America (2019). "Language diversity," in this case,

refers to language richness, or the number of languages present in a given 300x300km area pre-colonization (Pacheco Coelho et al. 2019:3). Each of the eight environmental predictor variables they employ was chosen based on previous literature establishing a theoretical basis for its impact on language diversity (Pacheco Coelho et al. 2019:2-3). River density and topographic complexity, the first two variables, might increase the rate of language diversification if rivers or difficult terrain act as barriers between groups of people, but rivers are also hypothesized to decrease the rate of diversification if they are used for travel, thereby increasing language contact. Hence, the predictor "river density" might have a positive or negative relationship with language richness, but there are theories underlying these possible relationships. Biological diversity has been shown to be correlated with linguistic diversity (Pacheco Coelho et al. 2019, Antunes et al. 2020:1-2), so ecoregion richness, the third predictor, might correlate positively with language richness. Climatic variables, temperature & precipitation constancy and climate change velocity, the fourth through sixth predictors, have a number of possible relationships to language diversity. These include but are not limited to overall ecological risk to living in an area, agricultural availability, and the availability of migration routes to new areas (Pacheco Coelho et al. 2019:2). Population density, the seventh, both depends on the previous six variables and itself informs the carrying capacity of a region, the eighth predictor (Figure 8 a). For both these last two variables, the probability of "stochastic diversification events" increases with number of individuals present in an area, meaning higher populations and higher population densities are directly linked to greater language diversity (Pacheco Coelho et al. 2019:3). Group structures can also vary with the carrying capacity of a region, possibly complicating this relationship (Pacheco Coelho et al. 2019:3).

Critically, the relationships between predictor variables and the response variable – language richness – are not only justified by prior work but also built into the "niche" model Pacheco Coelho et al. implement. The explicitly causal relationship between population density, carrying capacity,

and language diversity is reflected in their model (Figure 8 a, Pacheco Coelho et al. 2019 Supplemental Materials (Text-Figures-Table):7). Their model is applied at a large spatial scale – across an entire continent – with comparatively few predictor variables, and predictor collinearity is accounted for (Pacheco Coelho et al. 2019 Supplemental Materials (Text-Figures-Table):7).

Pacheco Coelho et al.'s statistical analyses are informed by previous literature and a clear understanding of their data's limitations. The first statistical approach, Stationary Path Analysis, is based on prior work attempting to isolate a universal predictor of language diversity. It assumes that the mechanisms driving language diversity are constant across the entire region of study, the same assumption that led to other researchers using the same approach finding conflicting results in prior work across different regions (Pacheco Coelho et al. 2019:2). To assess the validity of the "universal predictor" assumption, a second model was implemented using the same data. The second approach, Geographically Weighted Path (GWPath) analysis, assumes each predictor varies in importance across space rather than having the same impact across the entire study area. They found that indeed, no one variable was a "universal predictor of language richness" (Pacheco Coelho 2019:5). The variable with the highest overall coefficient changed depending on the region (Figure 8 c). Their findings reframe prior studies on different regions that appeared to come to conflicting conclusions on predictors of language diversity – rather than conflicting, each might accurately depict the relevance of predictors for that specific area or set of languages. Pacheco Coelho et al. employ GWPath modelling on human diversity patterns for, as far as they are aware, the first time, and in doing so expand the toolbox of analytical techniques available to historical linguists and sociolinguists (Pacheco Coelho et al. 2019:5).

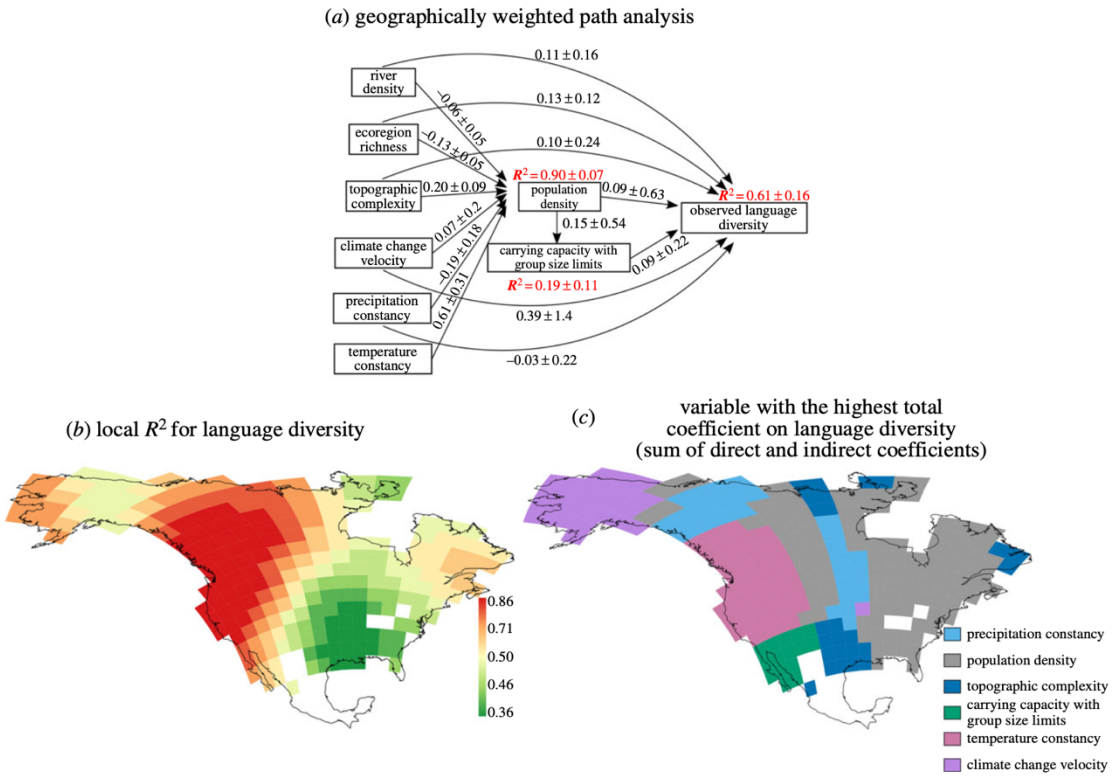

Figure 3. GWPath applied to North American linguistic diversity. (*a*) In the GWPath model, the standardized β coefficients of variables, as well as the R^2 for the direct relationships are represented by the average value over the continent, followed by its standard deviation. (*b*) Model fit varies over the geographical domain of North America. (*c*) Variables with the highest total coefficient (sum of direct and indirect effects) also vary across the continent. (Online version in colour.)

Figure 8. From Pacheco Coelho et al. 2019, results of GWPath modelling on North American language richness data.

Niche Modelling Papuan New Guinea Language Diversity

From the work of Pacheco Coelho et al., it is clear that statistical modelling has the potential to generate novel insights at the intersection of linguistic data, geography, and the environment, but what about niche modelling specifically? Antunes et al. (2020) apply niche modelling to language areas in New Guinea, attempting to determine the relationship between environmental factors and modern-day language family distributions as well as overall language diversity.

In their study, languages were grouped at the family level and divided into two clades: Austronesian and Trans-New Guinea (TNG). Languages falling outside these two groups (eg., "Other Austronesian" and "non-Austronesian NG") were discarded in their analysis (Antunes et al. 2020:18). Austronesian and TNG reflect two proposed phylogenetic groups discussed in Case Study

2: Austronesian languages in Papua descend from the Austronesian lineage originating in Taiwan, whose speakers migrated around Island Southeast Asia around 4000-5000 years before the present day (Arnold, forthcoming:8). TNG, on the other hand, is a hypothesized macro-family of Papuan languages across the island of Papua and surrounding archipelago, possibly originating from the central highlands. The boundaries and internal structure of TNG are still debated, though family groupings within the larger phylogenetic group are less controversial (Arnold, forthcoming:6-7). They also identified 29 language families, which served as the smallest units of analysis (Antunes et al. 2020:18).

Spatial data for the languages of the region was generated by first digitizing language areas across the island of Papua, then overlaying settlement/village data from DIVA-GIS (Antunes et al. 2020:18). Villages were assumed to speak the language of the language-area polygon they were located in. This approach allowed for more in-depth geospatial techniques to be applied to the region using point-based data at the sacrifice of an individual point's accuracy; languages (and subsequently their ranges) in Papua are frequently under- or ill-described, especially in the Bird's Head region (Arnold, forthcoming:2).

Antunes et al. (2020) used principal component analysis to characterize the environmental space of the island of Papua to determine the ecological niches available there. Using all 19 climatic and environmental variables available on WorldClim and ETOPOI, they conducted principle component analysis (PCA) in order to place each of the 29 language families in environmental 'space' (Antunes et al. 2020:20). Climate data was gathered from 1950-2000, notably two orders of magnitude shorter than the time depth at which the languages and language families of West Papua have been interacting, (Antunes et al. 2020:19). This leads to the unstated assumption of static climatic conditions across multiple millennia. "Eco-Linguistic Niches" were predicted for a given language group based on a "consensus" model – applying 10 algorithms designed to describe a

species' ecological niche, and using a consensus technique developed by Antunes in 2015 to reach a final model (Antunes et al. 2020 S2).

Highly colinear variables were not discarded on the basis that the consensus model was able to minimize model overfitting, and that micro-scale environmental variation is better captured using more predictor variables (Antunes et al. 2020, S2:6). *Colinearity* describes data that are not statistically independent predictors of a response variable: one common reason for collinearity is that two variables are different ways of measuring the same underlying phenomena. For Antunes et al., these cases are fairly obvious, like *mean annual temperature* and *maximum annual temperature;* or *slope azimuth* and *flow direction of water* (2020:20). Miller, in an overview of niche modelling techniques in ecology, writes that consensus approaches "increase accuracy while reducing uncertainty," though makes no mention of reduced overfitting (2010:501).

It is interesting to note that the environmental and linguistic data for the entirety of Papua is limited in resolution based on environmental sampling resolution. WorldClim relies heavily on interpolation and is the only climate dataset employed by Antunes et al. (2020:19). They reference the number of algorithms employed as a mitigating factor for the use of only one source of climate data (Antunes et al. 2020:19). Limited published linguistic area data further limits the resolution of Antunes et al.'s analysis, as digitized and georeferenced areas are not a high-resolution source for spatial data, bringing into question the ability of their models to account for micro-scale climatic variations if neither the predictor nor the response variable is measured at a fine resolution.

Armed with this understanding of Antunes et al.'s methodology, what about the theory that undergirds it? Ecology, like linguistics, relies on observational studies of phenomena that cannot easily be replicated in a laboratory setting, which might suggest ecological tools' possible relevance to geospatial linguistics. The first hurdle to applying niche modelling to linguistic data is applying the concept of a *niche*. While languages are often described as 'like species,' with the theory of language

family phylogenies drawing from evolutionary biology (Atkinson & Gray 2005:513), how far can this comparison go? In ecology, the relationship between predictive variables and the realized niche of a given species is backed by ecological theory and, in many cases, physiological data. For example, woody tree species are limited in their northern range by cold-hardiness, a trait determined by a combination of xylem physiology, temporal adaptations, and chemical processes (J. Grossman, pers. comm.)

Antunes et al. argue that a human group that shares a language acts in the same manner as a biological species with respect to its environment, shifting its range and existing in competition with other language groups (2020:2). According to this logic, culture, forms of agriculture, and environment-specific adaptations should be closely tied to language and presumably change at the same rate. For example, they posit that the dispersal of taro-growing horticulturalists speaking Trans-New Guinea (TNG) languages was influenced by the regions where taro grows well (Antunes et al. 2020:14). They do not, however, provide any data on current or historical taro distribution, nor do they explicate the taro-affiliated versus non-taro-affiliated languages at any level more specific than the TNG family. They also do not provide a basis for why TNG languages might be spoken by a taro-growing group rather than an Austronesian language, a phylogenetic distinction used in their models.

Antunes et al. further explain that cultural adaptations allowed languages groups to "exploit environmental components that previously were used only rarely or intermittently" (Antunes et al. 2020:16). As I understand it, their argument is that the niche of a given language has cultural and social dimensions in addition to environmental ones, and these dimensions can also be described, albeit indirectly, by the environmental variables they used. While plausible, this mechanism is far removed from the environmental data they use in their models, and would require significant ethnological backing to provide the theoretic basis Miller (2010) recommends for any application of

niche modelling. To use social dimensions in a model, either these dimensions need to be directly quantified and included as predictors, or some link between the environmental variables and the social ones should be made explicit in the model developed, as we see in the levels and dependencies included in Pacheco Coelho et al.'s modelling approach (2019, Figure 8).

However, a nonmechanistically described predictor can sometimes be used in modelling when mechanistic relationships are still being developed (Miller 2010). In some cases, niche modelling can inform the discovery of unanticipated or nonobvious links between a species' realized range and environmental conditions. In ecology, however, there is an underlying understanding that the physical adaptation of a species directly informs its potential niche, which is then limited by the environment to form its realized niche. Humans, as a species, share the same physiological potential niche, so Antunes et al.'s use of niche modelling must therefore apply to exclusively the realized niche or social dimensions of said niche, and that realized niche must somehow be reflected in the language people speak. This limitation is not addressed in the text.

Previous work relating language diversity to environmental conditions, like Pacheco Coelho et al. (2019), used some metric of diversity, often language richness, and interrogated the relationship of *diversity itself* with environmental variables. Language diversification occurs in a manner similar to species diversification, and studying overall diversity does not require tying a specific language to specific cultural or niche-defining behaviors in order to draw conclusions (Atkinson & Gray 2005). The basis for the relationship between language richness & the environment in which that richness is developed and maintained is well established (Pacheco Coelho et al. 2019:2-3, Antunes et al. 2020:1-2).

Antunes et al. move beyond the abstracted metric of language richness and attempt to describe the specific conditions in which language families, and by extension cultural groups, will survive. This logical leap from previous work poses an interesting question: will a certain aspect of

New Guinea language environments be able to predict which family exists in a given location? While interesting to investigate, ecological niche modelling may not be the way to proceed. Using ENM with linguistic distribution data assumes that "language" and "environmental adaptations" are inexorably tied – and yet Antunes et al. provide a counterexample in their own text, saying of an Austronesian group,"[t]he Mari kept their pottery tradition (a typical element of Austronesian cultures) but adopted the language of their TNG-speaking Gadsup neighbors" after conflict caused them to relocate (2020:16). They find that "Eco-Linguistic Niches" do not predict the distributions of individual language families, but do correspond broadly to the TNG/Austronesian distinction – provided "marginal" TNG languages are discarded, and any non-Austronesian non-TNG New Guinean languages are also ignored (Antunes et al. 2020:14).

Overall, I find Antunes et al.'s implementation of niche modelling to be in need of a stronger theoretical backing and in need of further support for their methodological choices beyond reliance on a consensus model to overcome limitations in their predictor data. Antunes et al. clearly see the potential in the application of GIS technology and niche modelling, and take an ambitious approach to geospatial linguistics, but they might benefit from a more careful understanding of what tools are being employed and why they are appropriate for a given application, as is advocated for by Miller (2010:493).

Conclusion

Maps have historically held a great deal of power, especially in the realm of endangered and Indigenous languages (Stone 2018:42). Stone writes that "[w]hoever controls depictions of a given geographical, political or linguistic territory has the means of shaping a society's thoughts regarding that territory," especially in the context of settler-colonialism and the enforcement of conceptions of European nation-states (2018:42). Conversely, maps can also be tools of resistance and

revitalization, lending credence and authority to marginalized groups (Haynie & Gavin 2019, First Peoples Map of British Colombia). Projects like the First Peoples Map of British Colombia blend community engagement, talking dictionary functionality, and an online map to create a showcase of language diversity and Indigenous art, culture, and heritage (https://maps.fpcc.ca/languages). A map is a form of visual storytelling, not an objective view of the world, and methodological choices need to reflect this understanding.

GIS software and mapmaking represent a powerful explanatory and analytical set of tools that linguists can and should employ to describe language data and reveal deeper spatial patterns. However, the nature of language data and the dearth of available and comprehensive data sets present challenges. As scholars are beginning to propose and move towards coherent methodologies for the use of geospatial techniques in dialect studies, historical linguistics, or language diversity studies (Stone 2018, Haynie & Gavin 2019), it is critical to understand the fundamentals of mapmaking and the basis on which these methodologies are developed, and to employ them in ways that are transparent and suitable for the data in question.

Works Cited

Ambrose, J E & C H Williams. 1991. Language Made Visible: Representation in Geolinguistics. In *Linguistic Minorities, Society and Territory* (Multilingual Matters), vol. 78, 17. Multilingual Matters Ltd.

Antunes, Nicolas, Wulf Schiefenhövel, Francesco d'Errico, William E. Banks & Marian Vanhaeren. 2020. Quantitative methods demonstrate that environment alone is an insufficient predictor of present-day language distributions in New Guinea. (Ed.) Richard A Blythe. *PLOS ONE* 15(10). e0239359. https://doi.org/10.1371/journal.pone.0239359.

Arnold, Laura (forthcoming). 'Linguistics as a lens into the prehistory of western New Guinea', in Dylan Gaffney and Marlin Tolla (eds), *Archaeology and Material Culture in Western New Guinea*. Canberra: Terra Australis, Australian National University Press.

Atkinson, Quentin D. & Russell D. Gray. 2005. Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics. (Ed.) Chris Simon. *Systematic Biology* 54(4). 513–526. https://doi.org/10.1080/10635150590950317.

Bowern, Claire, Hannah Haynie, Catherine Sheard, Barry Alpher, Patience Epps, Jane Hill & Patrick McConvell. 2014. Loan and Inheritance Patterns in Hunter-Gatherer Ethnobiological Systems. *Journal of Ethnobiology* 34(2). 195–227. https://doi.org/10.2993/0278-0771-34.2.195.

Campbell, Lyle & Mauricio J. Mixco. 2007. *A Glossary of Historical Linguistics* (Glossaries in Linguistics). Edinburgh: Edinburgh Univ. Press.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2021. *Ethnologue: Languages of the World*. Twenty-fourth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Gasser, Emily. 2020. Borrowed Color and Flora/Fauna Terminology in Northwest New Guinea. *Journal of Language Contact* 12(3). 609–659. https://doi.org/10.1163/19552629-01203003.

Haynie, Hannah, Claire Bowern, Patience Epps, Jane Hill & Patrick McConvell. 2014. Wanderwörter in languages of the Americas and Australia. *Ampersand* 1. 1–18. https://doi.org/10.1016/j.amper.2014.10.001.

Haynie, Hannah J. 2014. Geography and Spatial Analysis in Historical Linguistics: Geography and Spatial Analysis in Historical Linguistics. *Language and Linguistics Compass* 8(8). 344–357. https://doi.org/10.1111/lnc3.12087.

Haynie, Hannah J. & Michael C. Gavin. 2019. *Modern Language Range Mapping for the Study of Language Diversity*. Preprint. SocArXiv. https://doi.org/10.31235/osf.io/9fu7g. https://osf.io/9fu7g (17 September, 2021).

Haynie, Hannah Jane. 2012. *Studies in the History and Geography of California Languages*. ProQuest LLC.

Kauhanen, Henri, Deepthi Gopal, Tobias Galla & Ricardo Bermúdez-Otero. 2018. Geospatial distributions reflect rates of evolution of features of language. *arXiv:1801.09637 [cond-mat, physics:nlin, physics:physics]*. http://arxiv.org/abs/1801.09637 (8 October, 2021).

Lozier, J. D., P. Aniello & M. J. Hickerson. 2009. Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. *Journal of Biogeography* 36(9). 1623–1627. https://doi.org/10.1111/j.1365-2699.2009.02152.x.

Luebbering, Candice Rae. 2011. The Cartographic Representation of Language: Understanding language map construction and visualizing language diversity. Virginia Tech. https://vtechworks.lib.vt.edu/handle/10919/37543 (10 September, 2021).

Miller, Jennifer. 2010. Species Distribution Modeling: Species distribution modeling. *Geography Compass* 4(6). 490–509. https://doi.org/10.1111/j.1749-8198.2010.00351.x.

OpenStreetMap. 2021. https://www.openstreetmap.org/.

Pacheco Coelho, Marco Túlio, Elisa Barreto Pereira, Hannah J. Haynie, Thiago F. Rangel, Patrick Kavanagh, Kathryn R. Kirby, Simon J. Greenhill, et al. 2019. Drivers of geographical patterns of North American language diversity. *Proceedings of the Royal Society B: Biological Sciences* 286(1899). 20190242. https://doi.org/10.1098/rspb.2019.0242.

Peeters, Yvo J D & Colin H. Williams (eds.). 1992. The Cartographic Representation of Linguistic Data. In *Discussion Papers in Geolinguistics*, vol. 19–21, 104. Le Pailly, France.

Singh, D, M E Wagih & D Hunter. 2018. Genetic resources of taro (Colocasia esculenta (L.) Schott) in Papua New Guinea: A review. 19.

Teerarojanarat, Sirivilai & Kalaya Tingsabadh. 2011a. A GIS-Based Approach for Dialect Boundary Studies. *Dialectologia* 6. 55–75.

Teerarojanarat, Sirivilai & Kalaya Tingsabadh. 2011b. Using GIS for Linguistic Study: A Case of Dialect Change in the Northeastern Region of Thailand. *Procedia - Social and Behavioral Sciences* 21. 362–371. https://doi.org/10.1016/j.sbspro.2011.07.015.

Usher, Timothy & Antoinette Schapper. 2018. The Lexicons Of The Papuan Languages Of The Onin Peninsula And Their Influences. Zenodo. https://doi.org/10.5281/ZENODO.1451029. https://zenodo.org/record/1451029 (17 September, 2021).

First Peoples' Map of B.C. https://maps.fpcc.ca/ (3 December, 2021).