Andre Lee
Typology Final Paper

Theory of Linguistic Areas: Variable-Defined Areas and Predictive Areality Theory

I. Introduction to Linguistic Areas

A linguistic area is generally defined as a geographical region in which languages from different families show a "significant" set of variables (i.e. linguistic features) in common which are not present in areas directly outside the specified region, and which arise due to historical contact between speakers. The defining features of a linguistic area can have strong implications in typology, identifying individual features that are more likely to diffuse (and conversely, least likely to be inherited) among the world's languages, as well as linguistic features that are more likely to diffuse together. These types of questions can be further explored when considering linguistic area "leakage"; Bickel and Nichols define this as small amounts of defining features which spread beyond the boundaries of a predetermined linguistic area, often historically motivated (Bickel & Nichols, 2006).[1] In determining the extent of a linguistic area and its defining features, we explore two approaches, variable-defined areas and predictive areality theory.

II. Variable-Defined Areas:

Variable-defined areas are linguistic areas hypothesized by first choosing a set of "diagnostic" linguistic features and defining an area based on the demarcations of isoglosses made using those features. The Balkan language area (or *sprachbund*) is cited by Bickel and Nichols as an example containing isogloss-bundled areal features (Bickel & Nichols, 2006). The Balkan linguistic region traditionally consists of Albanian, Greek, Balkan Romance and Balkan Slavic (Friedman, 2006). However, during the 20th century linguists have added Balkan dialects

[1] Bickel and Nichols refer to these features as "escaped", giving the example of Mesoamerican language features moving eastward towards the Caribbean potentially fueled by the spread of domestication.

of Turkish, Romani and Jewish languages (e.g. varieties of Judezmo, Judeo-Greek and Hebrew).[2] Years of study have produced a catalogue of defining features, so-called "Balkanisms". Clitic doubling for objects is an example used quite often, an extremely rare feature outside the Balkan region. Balkan languages are characterized by the usage of clitics to specify gender, number and case, often reduplicating according to the animacy hierarchy. Clitic doubling refers to clitic pronouns co-occuring in the same verb phrase with the noun phrases to which they refer.

The occurrence of possessive doubling as a feature is more restrictive, but has been observed, specifically with kinship terms in Macedonian, Aromanian and Albanian, and for emphasis with some others.

1. Greek (colloquial):

   to vivlio mou mena

   the book me.GEN me.GEN

   "**my**[3] book"

2. Romanian (colloquial):

   propria-mi mea semnãtura

   own.FEM-me.DAT my signature.DEF

   "my very own signature"

In the colloquial Greek example, pronominal doubling is used for emphasis; this contrasts with the older, more formal form of possession in the formal language which utilizes the dative clitic after the definite noun. This suggests that the possessive doubling in older forms of Greek

[2] Friedman notes that Balkan linguistics have a relatively old tradition, and that 19th century standards for categorizing them still exist heavily today, particularly among Bulgarian linguists (due to political conflicts in the region).

[3] Convention used by Friedman to signify emphasis.

to express emphasis might have also used clitic doubling. The colloquial Romanian example uses the clitic doubling form for possessive doubling (Friedman, 2006, pp. 662-663).

Despite general agreement among linguists on areal features for regions like the Balkans and Mesoamerica, Bickel and Nichols identify potential issues with the variable-definition method. The first is the lack of criteria for the original determination of diagnostic features; they argue that before identifying a potential language area, linguists must already have a statistical grasp of the frequency of linguistic features worldwide, as well as which ones are likely to be diffused/inherited. Even with a high level of rigor, using isogloss-bundled areal features would lead to defining features that would easily co-occur by chance.[4] The second is an issue with isoglossing itself; situations where isoglossing might become inaccurate or create obstacles within multiple regions are possible, e.g. relatively new immigrants engaging in bilingualism/ code-switching but are yet unaffected by areal features. Linguists must make a choice to ignore these populations when creating an isogloss or isolate them with a discontinuous isogloss (Bickel & Nichols, 2006). Among other complaints, overall Bickel and Nichols very strongly state that "the variable-defined approach is unlikely to be able to define large, old, or inactive areas with significant linguistic immigration very satisfactorily" (Bickel & Nichols, 2006, pp. 3-5). As an alternative, they propose predictive areality theory.

III. Predictive Areality Theory (PAT):

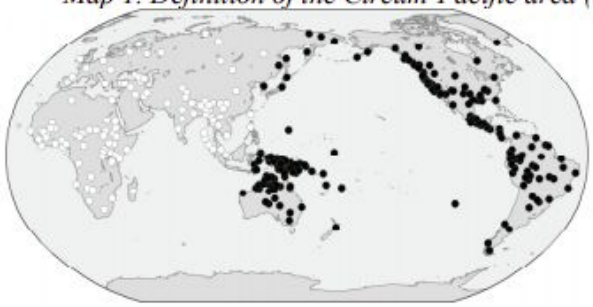
Predictive areality theory is a new approach to identifying potential language areas and their defining features. To form a hypothesis, we first define an area based on population history and archaeological theories. This hypothesis is "predictive" because it does not initially rely on

[4] Bickel and Nichols claim that out of 100 surveyed variables, half a dozen could easily co-occur if those variables were all relatively frequent in the world's languages.

overlapping distributions of linguistic features. Instead, by holding an initial assumption that such distributions are unclear or incomplete, historical data becomes the foundation of plausible theories of a language area's existence. Defining features are then chosen based on distributional overlap with the historical region, and tested statistically to make sure such overlap does not occur due to universals, shared genealogy or sheer chance. Interestingly, in this way Bickel and Nichols do not categorize areality as a typological property, but as its own predictor variable for observable typological distributions. As an example, they give an analysis of the Pacific Rim as a linguistic area (Bickel & Nichols, 2006).

The Pacific Rim region is defined as the strictly coastal regions of the larger Circum-Pacific region given below. The historical migrations of ancient Southeast Asian peoples, Austronesian expansion, and Bering Strait crossings were given as the rationale for the hypothesis.

Map 1. Definition of the Circum-Pacific area (black dots) in our sample



Using WALS to create a genealogically balanced sample, they then collected a sample of 75 relatively rare features that were also well-documented in the languages of the Pacific Rim (Bickel & Nichols, 2006). As an example, one of the extracted features was multiple possessive classes. From WALS, this is defined as multiple different possessive markings determined by lexical or semantic properties of the possessed noun. A glossed example from the Pacific Rim:

Warndarang (Maran; Northern Territory, Australia)

a. ng-baba
1-father
"my/our father"

b. wu-radburru ngini
NCM-country 1SG.GEN
"my country"

Nez Perce (Sahaptian; northwestern United States)

a. na'-tóot
1SG-father
"my father"

b. 'in'm-é:ks
1SG-man's.sister
"my sister" (man speaking)

c. 'íi-nim titóoqan
1SG-GEN people
"my people"

Here we see a coastal Australian and a Pacific Northwest language exhibit similar traits in regards to multiple possessive classes. Both make distinctions between kinship terms and common nouns, with Nez Perce having two classes for kinship terms based on the gender of the possessor (Dryer, 2005).

Of course, the predictive areality theory is not without its issues. Bickel and Nichols address possible concerns, as well as how they can be mitigated. One issue is variance, where Pacific Rim languages coexist quite regularly with languages that do not share areal features; this does not follow the level of consistency required by classical definitions of areality. Instead, that rigor of variance is substituted with additional statistical criteria that focuses on feature frequencies inside and outside the region. Another issue was mentioned previously, the issue of leakage. Bickel and Nichols note features like syntactic noun incorporation and ergativity that appear to spread inland within North America or Australia; they mitigate some of the error in the Western Hemisphere by using the larger Circum-Pacific region as a predictor to catch leakage. These and other issues raise historical questions about the preservation of Pacific Rim features

despite the disappearance of such features when reconstructing language families younger than the Pacific migration. It is hypothesized that areal pressure drives linguistic feature retention much more than inheritance (Bickel & Nichols, 2006).

IV. Experimentation: Neural Networks Trained on WALS as Tools for PAT

A. Introduction:

When discussing language areas, it is important to discuss distributions of language features. When talking about specific language features, we can computationally abstract a distribution for a specific feature as a function mapping a location to the presence of the feature.

Perceptrons and neural networks are computational models that run efficiently, able to learn patterns and approximate functions that are linearly separable. Using the WALS database as a training resource, we attempt to use a simple feed-forward neural network to predict the presence of a feature given locational inputs. A model is created for each feature in WALS and the most accurate models will be compared qualitatively against findings for the above mentioned proposed linguistic areas. We hope to test the viability of WALS as a machine learning resource the the expressiveness of simple locational-input models. As an application to PAT, we wish to evaluate the performance of neural-based models as a tool for validating the hypotheses surrounding predicted linguistic areas.

B. Methodology:

A simple introduction to the feed-forward neural network: a neural network is a high-dimensional data structure which uses a series of large matrix operations to map between dimensionalities of one form to another. The "weights" which form the matrix values can be seen as weighted graph edges between vertices that capture and output intermediate values. The weights are adjusted to promote greater accuracy as the neural network is trained on a fixed set of correct input-output examples. An example of a neural network designed for a WALS language feature is detailed below[5]:

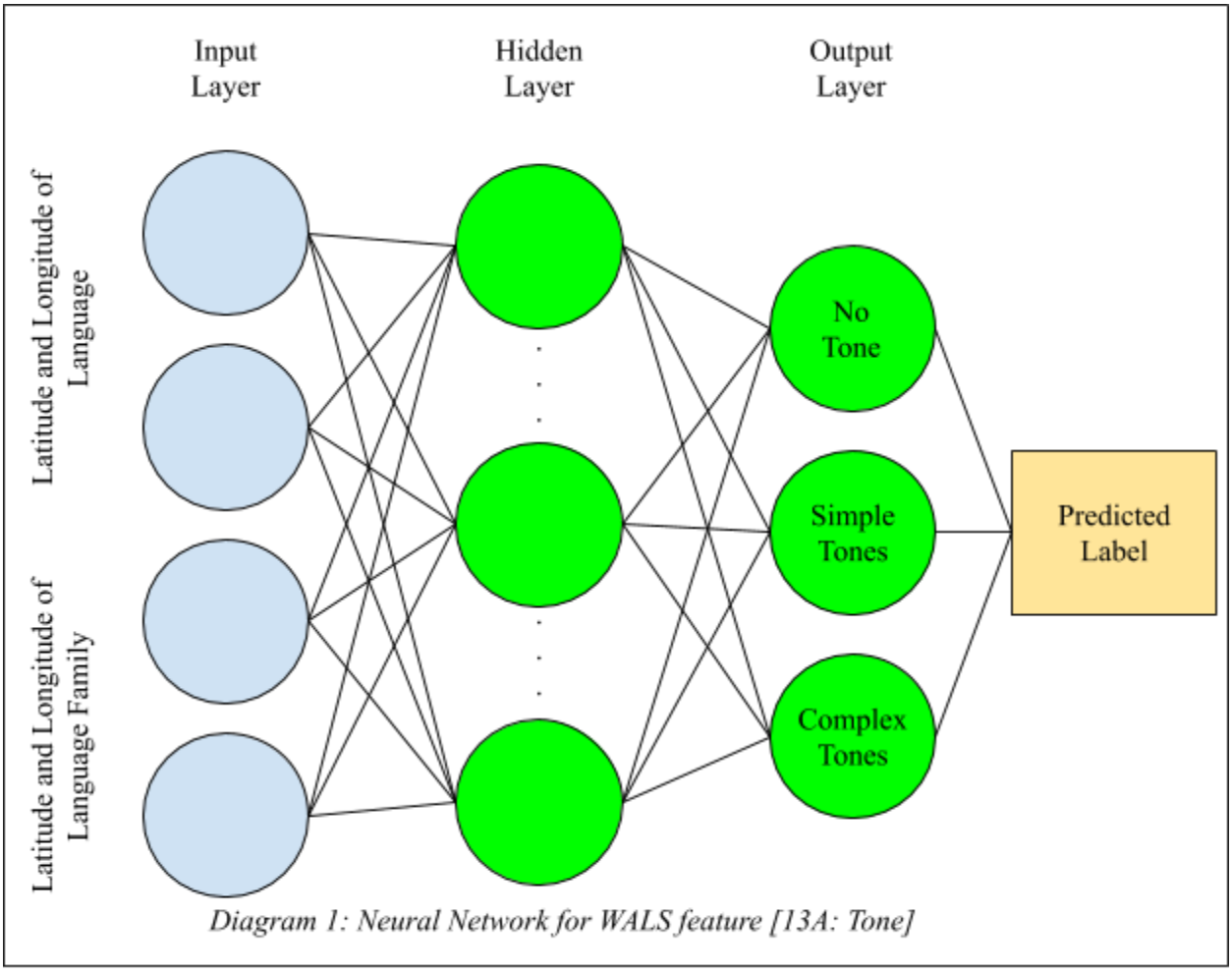


Diagram 1: Neural Network for WALS feature [13A: Tone]

[5] For more details on implementation specifics, the code repository is available here: https://github.com/andreezlee/ProgrammingProjects/tree/master/Predictive%20Areality%20Theory

The neural language uses a language's location as an input, and calculates the individual probabilities that the language contains each value of the feature. The highest probability output is the predicted feature for the language. Each input contains (a) the location of a language as documented by WALS, as well as (b) the location of the centroid of the language family (average location of all languages in the family, same as (a) if isolate). The reason for this input expansion is twofold: we want the neural network to capture patterns in regards to both language location and the family in which it is located, and we want a higher-dimensional input to increase the expressiveness of our design.

For each network, the example set was formed with each language with the documented linguistic feature as a data point (Dryer & Haspelmath, 2014). The training set would be a randomized 90% of the example set, with the validation set being the other 10%[6]. Features were deemed unusable if there were less than 300 data points or if training the neural network led to accuracy reaching 0. Due to the dearth of data, accuracy for determining each neural network's performance was an average of the training and validation accuracies on the final training epoch. Fine-tuning on metaparameters were conducted to produce more accurate models.

C. Simplified Results:

Final Model Parameters:	
Highest-Performing Hidden Layer Size:	16 vertices

[6] This is standard practice in regards to machine learning and natural language processing. The network is trained using the training set a certain number of times (epochs), and its performance in general usability is measured using accuracy on the validation set.

Highest-Performing Training Time:	6 epochs
Avg. Model Accuracy:	49.25%
Highest Model Accuracy:	99.58%
Usable WALS Features:	68

After fine-tuning, we used only the top 50 features whose models had the highest accuracy (refer to Appendix A). Given a defined area on the global map, random points within that area were chosen, treated as language isolates, and the predictions were recorded. Predictions were evaluated on two fronts: how accurately the neural networks approximated WALS, and how accurately the neural networks approximated linguistic findings regarding areal features. For this we chose 50 random points in the Balkan region, and 100 random points in the Pacific Rim.

Overall, results were quite mixed. For almost all of the 50 features in the regions tested, the predictions mostly reflected the true distribution according to the WALS database. However, there were some clear issues that were made apparent given the regions tested. When examining the Balkans, it becomes apparent that using accuracy as the only performance measure created a bias for models to indiscriminately choose the most common value of a feature. Some examples would be features “4A-Voicing in Plosives and Fricatives” and “144I-SNegVO Order”. The neural networks labeled all sample points in the Balkans as having no voicing contrasts and no SNegVO word order, while in WALS almost all Balkan languages have voicing with plosives and fricatives, as well as SNegVO word order possible with a separate negative word. When

examining the Pacific Rim, there are many smaller regions where certain features go completely undocumented on WALS. For example, features like "78A-Coding of Evidentiality" and "91A-Order of Degree Word and Adjective" are very sparsely documented in the Eastern Russia/Bering Strait regions. Results from samples at the northernmost areas of the Pacific Rim are therefore not very accurate.

While so-called "Balkanisms" tend to be more specific than what is documented in WALS, a similarity detected by both a neural model and Friedman was the absence of front-rounded vowels (Friedman, 2006). Within the Pacific Rim, there are more features noticed by Bickel and Nichols matching with those documented by WALS (Bickel & Nichols, 2006). A good example is in the morphological coding of evidentiality. Evidentiality is an expression of the source a speaker gives for their statement, be it first-hand account (e.g. 'I saw X happen') or otherwise. This feature focuses on how evidentiality is expressed grammatically (de Haan, 2013). The neural network identified a large amount of the sample Pacific Rim points as under the category of using a verbal affix or clitic, corroborated in WALS as most of the languages with this feature also lie around the Pacific Rim.

1. Takelma (Takelman; Oregon):

   naga=ihiʔ

   say.AOR.3SG=QUOT

   'he said it is said'

2. Namibiquara (Namibiquara; Brazil):

   wakon=na=ra

   work=action.currently.observed.by.multiple.witnesses=PERFV

   'He is working. (He is being observed)'

In both examples, the evidential morpheme is a clitic. In the Takelman example, the clitic *ihiʔ* attaches to the end of the phrasal verb and functions as a quotative, used by a speaker to denote a statement whose source was someone else (de Haan, 2013). In the Namibiquara example, the enclitic *na* is used to denote multiple sources (Lowe, 1999, p. 275). While they denote different levels of evidentiality, in both examples the clitic attaches to the verb in similar ways.

V. Conclusion:

From the results of the experimentation, we believe it is safe to say that while neural networks might show promise as a tool for validating new language areas, there are still some issues that need to be addressed. The first would be the database itself. While an oft-cited resource for a variety of typological features, for many of them WALS is still too sparse for training the level of specificity necessary for such neural networks.

The neural network training must also improve. While overfitting to the training data does not seem like an issue with the adjusted accuracy measure, using accuracy alone as a performance measure creates a statistical bias in trained models. In the future, perhaps it would be better to incorporate precision and recall into the metrics.

Also, when testing a hypothetical language area, to avoid issues brought up by the Balkan example, the results might improve if focus was placed more on larger regions. For inputs,

instead of treating each sample point as an isolate, the sample points would each be treated as part of a language family whose centroid is located outside the hypothetical region, in order to more clearly find features more likely to be diffused than inherited.

An interesting phenomenon observed was a group of features not useful for training models (i.e. models trained on the data of those features saw training and validation performance dropping to near zero as training continued). While in machine learning this is still unexplained by theory, investigations into this phenomenon might yield insights into features with a presence close to independent from location.

Overall, as a PAT tool, with models pre-trained on much larger amounts of data, neural methods of statistical validation might be useful for validating hypotheses. From fine-tuning we see these neural networks do not need much training time and are relatively small in size, leading to greater efficiency. The experiment also demonstrates the relatively new method of PAT as a more general approach for generating hypothetical linguistic areas compared to the variable-defined approach. By not strongly relying on existing data through isoglosses, PAT allows us to discuss areal features in a much broader context, and can incorporate new statistical methods for validating them. However, without set boundaries, issues of leakage still arise, and relying on anthropological or archaeological evidence might yield only less specific features shared among the languages of the region. As we explore new computational tools and gather more data, a hybrid approach might also emerge which can solve more of these issues.

VI. References:

Bickel, Balthasar, and Nichols, Johanna (2006) *Oceania, the Pacific Rim, and the Theory of Linguistic Areas*. Annual Meeting of the Berkeley Linguistics Society, vol. 32, no. 2, doi:10.3765/bls.v32i2.3488.

de Haan, Ferdinand (2013). *Coding of Evidentiality*. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. http://wals.info/chapter/78

Dryer, M. S. (2005) *Possessive Classes*. The World Atlas of Language Structures Online. https://wals.info/chapter/59

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2014. The World Atlas of Language Structures Online. Jena: Max Planck Institute for the Science of Human History. (Available online at https://wals.info)

Friedman V A (2006), Balkans as a Linguistic Area. In: Keith Brown, (Editor-in-Chief) Encyclopedia of Language & Linguistics, Second Edition, volume 1, pp. 657-672. Oxford: Elsevier.

Lowe, Ivan. "Nambiquara." *The Amazonian Languages*, edited by Alexandra Y. Aikhenwald and R. M. W. Dixon: Cambridge University Press (1999): 269-292.

Appendix A: Most Accurate Neural Net Features

2A	Vowel Quality Inventories
4A	Voicing in Plosives and Fricatives
5A	Voicing and Gaps in Plosive Systems
6A	Uvular Consonants
7A	Glottalized Consonants
8A	Lateral Consonants
9A	The Velar Nasal
11A	Front Rounded Vowels
12A	Syllable Structure
13A	Tone
15A	Weight-Sensitive Stress
16A	Weight Factors in Weight-Sensitive Stress
17A	Rhythm Types
18A	Absence of Common Consonants
19A	Presence of Uncommon Consonants
27A	Reduplication
44A	Gender Distinctions in Independent Personal Pronouns
46A	Indefinite Pronouns
52A	Comitatives and Instrumentals
55A	Numeral Classifiers
64A	Nominal and Verbal Conjunction
70A	The Morphological Imperative
72A	Imperative-Hortative Systems
73A	The Optative
78A	Coding of Evidentiality

90A	Order of Relative Clause and Noun
90C	Postnominal Relative Clauses
91A	Order of Degree Word and Adjective
94A	Order of Adverbial Subordinator and Clause
100A	Alignment of Verbal Person Marking
101A	Expression of Pronominal Subjects
104A	Order of Person Markers on the Verb
105A	Ditransitive Constructions: The Verb ‘Give’
107A	Passive Constructions
111A	Nonperiphrastic Causative Constructions
116A	Polar Questions
118A	Predicative Adjectives
119A	Nominal and Locational Predication
120A	Zero Copula for Predicate Nominals
129A	Hand and Arm
130A	Finger and Hand
143G	Minor Morphological Means of Signaling Negation
144B	Position of Negative Words Relative to Beginning and End of Clause and with Respect to Adjacency to Verb
144H	NegSVO Order
144I	SNegVO Order
144J	SVNegO Order
144K	SVONeg Order
144P	NegSOV Order
144Q	SNegOV Order
144R	SONegV Order