Non-Parametric Bayesian Areal Linguistics

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, UT 84112
me@hal3.name

Abstract

We describe a statistical model over linguistic areas and phylogeny. Our model recovers known areas and identifies a plausible hierarchy of areal features. The use of areas improves genetic reconstruction of languages both qualitatively and quantitatively according to a variety of metrics. We model linguistic areas by a Pitman-Yor process and linguistic phylogeny by Kingman's coalescent.

1 Introduction

Why are some languages more alike than others? This question is one of the most central issues in historical linguistics. Typically, one of three answers is given (Aikhenvald and Dixon, 2001; Campbell, 2006). First, the languages may be related "genetically." That is, they may have all derived from a common ancestor language. Second, the similarities may be due to chance. Some language properties are simply more common than others, which is often attributed to be mostly due to linguistic universals (Greenberg, 1963). Third, the languages may be related *areally*. Languages that occupy the same geographic area often exhibit similar characteristics, not due to genetic relatedness, but due to sharing. Regions (and the languages contained within them) that exhibit sharing are called *linguistic areas* and the features that are shared are called *areal features*.

Much is not understood or agreed upon in the field of areal linguistics. Different linguists favor different defintions of what it means to be a linguistic area (are two languages sufficient to describe an area or do you need three (Thomason, 2001; Katz, 1975)?), what areal features are (is there a linear ordering of "borrowability" (Katz, 1975; Curnow, 2001) or is that too prescriptive?), and what causes sharing to take place (does social status or number of speakers play a role (Thomason, 2001)?).

In this paper, we attempt to provide a *statistical* answer to some of these questions. In particular, we develop a Bayesian model of typology that allows for, but does not force, the existence of linguistic areas. Our model also allows for, but does not force, preference for some feature to be shared areally. When applied to a large typological database of linguistic features (Haspelmath et al., 2005), we find that it discovers linguistic areas that are well documented in the literature (see Campbell (2005) for an overview), and a small preference for certain features to be shared areally. This latter agrees, to a lesser degree, with some of the published hierarchies of borrowability (Curnow, 2001). Finally, we show that reconstructing language family trees is significantly aided by knowledge of areal features. We note that Warnow et al. (2005) have independently proposed a model for phonological change in Indo-European (based on the Dyen dataset (Dyen et al., 1992)) that includes notions of borrowing. Our model is different in that we (a) base our model on typological features rather than just lexical patterns and (b) we explicitly represent language areas, not just one-time borrowing phenomena.

2 Background

We describe (in Section 3) a non-parametric, hierarchical Bayesian model for finding linguistic areas and areal features. In this section, we provide necessary background—both linguistic and statistical—

for understanding our model.

2.1 Areal Linguistics

Areal effects on linguistic typology have been studied since, at least, the late 1920s by Trubetzkoy, though the idea of tracing family trees for languages goes back to the mid 1800s and the comparative study of historical linguistics dates back, perhaps to Giraldus Cambrenis in 1194 (Campbell, In press). A recent article provides a short introduction to both the issues that surround areal linguistics, as well as an enumeration of many of the known language areas (Campbell, 2005). A fairly wide, modern treatment of the issues surrounding areal diffusion is also given by essays in a recent book edited by Aikhenvald and Dixon (2001). The essays in this book provide a good introduction to the issues in the field. Campbell (2006) provides a critical survey of these and other hypotheses relating to areal linguistics.

There are several issues which are basic to the study of areal linguistics (these are copied almost directly from Campbell (2006)). Must a linguistic area comprise more than two languages? Must it comprise more than one language family? Is a single trait sufficient to define an area? How "nearby" must languages in an area be to one another? Are some feature more easily borrowed that others?

Despite these formal definitional issues of what constitutes a language area and areal features, most historical linguists seem to believe that areal effects play *some* role in the change of languages.

2.1.1 Established Linguistic Areas

Below, we list some of the well-known linguistic areas; Campbell (2005) provides are more complete listing together with example areal features for these areas. For each area, we list associated languages:

The Balkans: Albanian, Bulgarian, Greek, Macedonian, Rumanian and Serbo-Croatian. (*Sometimes:* Romani and Turkish)

South Asian: Languages belonging to the Dravidian, Indo-Aryan, Munda, Tibeto-Burman families.

Meso-America: Cuitlatec, Huave, Mayan, Mixe-Zoquean, Nahua, Otomanguean, Tarascan, Tequistlatecan, Totonacan and Xincan.

North-west America: Alsea, Chimakuan, Coosan, Eyak, Haida, Kalapuyan, Lower Chinook, Salishan, Takelman, Tlingit, Tsimshian and Wakashan.

The Baltic: Baltic languages, Baltic German, and Finnic languages (especially Estonian and Livonian). (Sometimes many more are included, such as: Belorussian, Lavian, Lithuanian, Norwegian, Old Prussian, Polish, Romani, Russian, Ukranian.)

Ethiopia: Afar, Amharic, Anyuak, Awngi, Beja, Ge'ez, Gumuz, Janjero, Kefa, Sidamo, Somali, Tigre, Tigrinya and Wellamo.

Needless to say, the exact definition and extent of the actual areas is up to significant debate. Moreover, claims have been made in favor of many linguistic areas not defined above. For instance, Dixon (2001) presents arguments for several Australian linguistic areas and Matisoff (2001) defines a South-East Asian language area. Finally, although "folk lore" is in favor of identifying a linguistic area including English, French and certain Norse languages (Norwegian, Swedish, Low Dutch, High German, etc.), there are counter-arguments to this position (Thomason, 2001) (see especially Case Study 9.8).

2.1.2 Linguistic Features

Identifying which linguistic features are most easily shared "areally" is a long standing problem in contact linguistics. Here we briefly review some of the major claims. Much of this overview is adoped from the summary given by Curnow (2001).

Haugen (1950) considers only borrowability as far as the lexicon is concerned. He provided evidence that nouns are the easiest, followed by verbs, adjectives, adverbs, prepositions, etc. Ross (1988) corroborates Haugen's analysis and deepens it to cover morphology, syntax and phonology. He proposes the following hierarchy of borrowability (easiest items coming first): nouns $>$ verbs $>$ adjectives $>$ syntax $>$ non-bound function words $>$ bound morphemes $>$ phonemes. Coming from a "constraints" perspective, Moravcsik (1978) suggests that: lexical items must be borrowed before lexical properties; inflected words before bound morphemes; verbal items can never be borrowed; etc.

Curnow (2001) argues that coming up with a reasonable hierarchy of borrowability is that "we may never be able to develop such constraints." Nevertheless, he divides the space of borrowable features into 15 categories and discusses the evidence supporting each of these categories, including: phonetics (rare), phonology (common), lexical (very common), interjections and discourse markers (com-

mon), free grammatical forms (occasional), bound grammatical forms (rare), position of morphology (rare), syntactic frames (rare), clause-internal syntax (common), between-clause syntax (occasional).

2.2 Non-parametric Bayesian Models

We treat the problem of understanding areal linguistics as a statistical question, based on a database of typological information. Due to the issues raised in the previous section, we do not want to commit to the existence of a particular number of linguistic areas, or particular sizes thereof. (Indeed, we do not even want to commit to the existence of *any* linguistic areas.) However, we will need to "unify" the languages that fall into a linguistic area (if such a thing exists) by means of some statistical parameter. Such problems have been studied under the name *non-parametric models.* The idea behind non-parametric models is that one does not commit *a priori* to a particularly number of parameters. Instead, we allow the data to dictate how many parameters there are. In Bayesian modeling, non-parametric distributions are typically used as *priors*; see Jordan (2005) or Ghahramani (2005) for overviews. In our model, we use two different non-parametric priors: the Pitman-Yor process (for modeling linguistic areas) and Kingman's coalescent (for modeling linguistic phylogeny), both described below.

2.2.1 The Pitman-Yor Process

One particular example of a non-parametric prior is the Pitman-Yor process (Pitman and Yor, 1997), which can be seen as an extension to the better-known Dirichlet process (Ferguson, 1974). The Pitman-Yor process can be understood as a particular example of a Chinese Restaurant process (CRP) (Pitman, 2002). The idea in all CRPs is that there exists a restaurant with an infinite number of tables. Customers come into the restaurant and have to choose a table at which to sit.

The Pitman-Yor process is described by three parameters: a base rate α, a discount parameter d and a mean distribution G_0. These combine to describe a process denoted by $\mathcal{PY}(\alpha, d, G_0)$. The parameters α and d must satisfy: $0 \leq d < 1$ and $\alpha > -d$. In the CRP analogy, the model works as follows. The first customer comes in and sits at any table. After N customers have come in and seated themselves (at a total of K tables), the Nth customer arrives. In the Pitman-Yor process, the Nth customer sits at a new table with probability proportional to $\alpha + Kd$ and sits at a previously occupied table k with probability proportional to $\#_k - d$, where $\#_k$ is the number of customers already seated at table k. Finally, with each table k we associate a parameter θ_k, with each θ_k drawn independently from G_0. An important property of the Pitman-Yor process is that draws from it are *exchangable*: perhaps counterintuitively, the distribution does not care about customer order.

The Pitman-Yor process induces a power-law distribution on the number of singleton tables (i.e., the number of tables that have only one customer). This can be seen by noticing two things. In general, the number of singleton tables grows as $\mathcal{O}(\alpha N^d)$. When $d = 0$, we obtain a Dirichlet process with the number of singleton tables growing as $\mathcal{O}(\alpha \log N)$.

2.2.2 Kingman's Coalescent

Kingman's coalescent is a standard model in population genetics describing the common genealogy (ancestral tree) of a set of individuals (Kingman, 1982b; Kingman, 1982a). In its full form it is a distribution over the genealogy of a countable set.

Consider the genealogy of n individuals alive at the present time $t = 0$. We can trace their ancestry backwards in time to the distant past $t = -\infty$. Assume each individual has one parent (in genetics, *haploid* organisms), and therefore genealogies of $[n] = \{1, \ldots, n\}$ form a *directed forest*. Kingman's n-coalescent is simply a distribution over genealogies of n individuals. To describe the Markov process in its entirety, it is sufficient to describe the jump process (i.e. the embedded, discrete-time, Markov chain over partitions) and the distribution over coalescent times. In the n-coalescent, every pair of lineages merges independently with rate 1, with parents chosen uniformly at random from the set of possible parents at the previous time step.

The n-coalescent has some interesting statistical properties (Kingman, 1982b; Kingman, 1982a). The marginal distribution over tree topologies is uniform and independent of the coalescent times. Secondly, it is infinitely exchangeable: given a genealogy drawn from an n-coalescent, the genealogy of any m contemporary individuals alive at time $t \leq 0$ embedded within the genealogy is a draw from the m-coalescent. Thus, taking $n \to \infty$, there is a distri-

bution over genealogies of a countably infinite population for which the marginal distribution of the genealogy of any n individuals gives the n-coalescent. Kingman called this *the coalescent*.

Teh et al. (2007) recently described efficient inference algorithms for Kingman's coalescent. They applied the coalescent to the problem of recovering linguistic phylogenies. The application was largely successful—at least in comparison to alternative algorithms that use the same data-. Unfortunately, even in the results they present, one can see significant areal effects. For instance, in their Figure(3a), Romanian is very near Albanian and Bulgarian. This is likely an areal effect: specifically, an effect due to the Balkan langauge area. We will revisit this issue in our own experiments.

3 A Bayesian Model for Areal Linguistics

We will consider a data set consisting of N languages and F typological features. We denote the value of feature f in language n as $X_{n,f}$. For simplicity of exposition, we will assume two things: (1) there is no unobserved data and (2) all features are binary. In practice, for the data we use (described in Section 4), neither of these is true. However, both extensions are straightforward.

When we construct our model, we attempt to be as neutral to the "areal linguistics" questions defined in Section 2.1 as possible. We allow areas with only two languages (though for brevity we do not present them in the results). We allow areas with only one family (though, again, do not present them). We are generous with our notion of locality, allowing a radius of 1000 kilometers (though see Section 5.4 for an analysis of the effect of radius).[1] And we allow, but do not enforce trait weights. All of this is accomplished through the construction of the model and the choice of the model hyperparameters.

At a high-level, our model works as follows. Values $X_{n,f}$ appear for one of two reasons: they are either areally derived or genetically derived. A latent variable $Z_{n,f}$ determines this. If it is derived areally, then the value $X_{n,f}$ is drawn from a latent variable corresponding to the value preferences in the langauge area to which language n belongs. If it is derived genetically, then $X_{n,f}$ is drawn from a variable corresponding to value preferences for the genetic substrate to which language n belongs. The set of areas, and the area to which a language belongs are given by yet more latent variables. It is this aspect of the model for which we use the Pitman-Yor process: languages are customers, areas are tables and area value preferences are the parameters of the tables.

3.1 The formal model

We assume that the value a feature takes for a particular language (i.e., the value of $X_{n,f}$) can be explained *either* genetically or areally.[2] We denote this by a binary indicator variable $Z_{n,f}$, where a value 1 means "areal" and a value 0 means "genetic." We assume that each $Z_{n,f}$ is drawn from a feature-specific binomial parameter π_f. By having the parameter feature-specific, we express the fact that some features may be more or less likely to be shared than others. In other words, a high value of π_f would mean that feature f is easily shared areally, while a low value would mean that feature f is hard to share. Each language n has a known latitude/longitude ℓ_n.

We further assume that there are K linguistic areas, where K is treated non-parametrically by means of the Pitman-Yor process. Note that in our context, a linguistic area may contain *only one* language, which would technically not be allowed according to the linguistic definition. When a language belongs to a singleton area, we interpret this to mean that it does not belong to any language area.

Each language area k (including the singleton areas) has a set of F associated parameters $\phi_{k,f}$, where $\phi_{k,f}$ is the probability that feature f is "on" in area k. It also has a "central location" given by a longitude and latitude denoted c_k. We only allow languages to belong to areas that fall within a given radius R of them (distances computed according to geodesic distance). This accounts for the "geographical" constraints on language areas. We denote the area to which language n belongs as a_n.

We assume that each language belongs to a "family tree." We denote the parent of language n in the

[1] An reader might worry about exchangeability: Our method of making language centers and locations part of the Pitman-Yor distribution ensures this is not an issue. An alternative would be to use a location-sensitive process such as the kernel stick-breaking process (Dunson and Park, 2007), though we do not explore that here.

[2] As mentioned in the introduction, (at least) one more option is possible: chance. We treat "chance" as noise and model it in the data generation process, not as an alternative "source."

$X_{n,f} \sim \begin{cases} \mathcal{B}in(\theta_{p_n,f}) & \text{if } Z_{n,f}=0 \\ \mathcal{B}in(\phi_{a_n,f}) & \text{if } Z_{n,f}=1 \end{cases}$	feature values are derived genetically or areally
$Z_{n,f} \sim \mathcal{B}in(\pi_f)$	feature source is a biased coin, parameterized per feature
$\ell_n \sim \mathcal{B}all(c_{a_n}, R)$	language position is uniform within a ball around area center, radius R
$\pi_f \sim \mathcal{B}et(1,1)$	bias for a feature being genetic/areal is uniform
$(p,\theta) \sim \text{Coalescent}(\pi_0, m_0)$	language hierarchy and genetic traits are drawn from a Coalescent
$(a,\langle\phi,c\rangle) \sim \mathcal{PY}(\alpha_0, d_0, \mathcal{B}et(1,1)\times\mathcal{U}ni)$	area features are drawn Beta and centers Uniformly across the globe

Figure 1: Full hierarchical Areal model; see Section 3.1 for a complete description.

family tree by p_n. We associate with each node i in the family tree and each feature f a parameter $\theta_{i,f}$. As in the areal case, $\theta_{i,f}$ is the probability that feature f is on for languages that descend from node i in the family tree. We model genetic trees by Kingman's coalescent with binomial mutation.

Finally, we put non-informative priors on all the hyperparameters. Written hierarchically, our model has the following shown in Figure 1. There, by $(p,\theta) \sim \text{Coalescent}(\pi_0, m_0)$, we mean that the tree and parameters are given by a coalescent.

3.2 Inference

Inference in our model is mostly by Gibbs sampling. Most of the distributions used are conjugate, so Gibbs sampling can be implemented efficiently. The only exceptions are: (1) the coalescent for which we use the GreedyRate1 algorithm described by Teh et al. (2007); (2) the area centers c, for which we using a Metropolis-Hastings step. Our proposal distribution is a Gaussian centered at the previous center, with standard deviation of 5. Experimentally, this resulted in an acceptance rate of about 50%.

In our implementation, we analytically integrate out π and ϕ and sample only over Z, the coalescent tree, and the area assignments. In some of our experiments, we treat the family tree as given. In this case, we also analytically integrate out the θ parameters and sample only over Z and area assignments.

4 Typological Data

The database on which we perform our analysis is the *World Atlas of Language Structures* (henceforth, WALS) (Haspelmath et al., 2005). The database contains information about 2150 languages (sampled from across the world). There are 139 typological *features* in this database. The database is *sparse*: only 16% of the possible language/feature pairs are known. We use the version extracted and preprocessed by Daumé III and Campbell (2007).

In WALS, languages a grouped into 38 language families (including Indo-European, Afro-Asiatic, Austronesian, Niger-Congo, etc.). Each of these language families is grouped into a number of language geni. The Indo-European family includes ten geni, including: Germanic, Romance, Indic and Slavic. The Austronesian family includes seventeen geni, including: Borneo, Oceanic, Palauan and Sundic. Overall, there are 275 geni represented in WALS.

We further preprocess the data as follows. For the Indo-European subset (hence-forth, "IE"), we remove all languages with ≤ 10 known features and then remove all features that appear in at most $1/4$ of the languages. This leads to 73 languages and 87 features. For the whole-world subset, we remove languages with ≤ 25 known features and then features that appear in at most $1/10$ of the languages. This leads to 349 languages and 129 features.

5 Experiments

5.1 Identifying Language Areas

Our first experiment is aimed at discovering language areas. We first focus on the IE family, and then extend the analysis to all languages. In both cases, we use a known family tree (for the IE experiment, we use a tree given by the language genus structure; for the whole-world experiment, we use a tree given by the language family structure). We run each experiment with five random restarts and 2000 iterations. We select the MAP configuration from the combination of these runs.

In the IE experiment, the model identified the areas shown in Figure 5.1. The best area identified by our model is the second one listed, which clearly correlates highly with the Balkans. There are two areas identified by our model (the first and last) that include only Indic and Iranian languages. While we are not aware of previous studies of these as linguistic areas, they are not implausible given

(Indic) Bhojpuri, Darai, Gujarati, Hindi, Kalami, Kashmiri, Kumauni, Nepali, Panjabi, Shekhawati, Sindhi **(Iranian)** Ormuri, Pashto
(Albanian) Albanian **(Greek)** Greek (Modern) **(Indic)** Romani (Kalderash) **(Romance)** Romanian, Romansch (Scharans), Romansch (Sursilvan), Sardinian **(Slavic)** Bulgarian, Macedonian, Serbian-Croatian, Slovak, Slovene, Sorbian
(Baltic) Latvian, Lithuanian **(Germanic)** Danish, Swedish **(Slavic)** Polish, Russian
(Celtic) Irish **(Germanic)** English, German, Norwegian **(Romance)** French
(Indic) Prasuni, Urdu **(Iranian)** Persian, Tajik
Plus 46 non-areal languages

Figure 2: IE areas identified. Areas that consist of just one genus are not listed, nor are areas with two languages.

(Mayan) Huastec, Jakaltek, Mam, Tzutujil **(Mixe-Zoque)** Zoque (Copainalá) **(Oto-Manguean)** Mixtec (Chalcatongo), Otomí (Mezquital) **(Uto-Aztecan)** Nahualtl (Tetelcingo), Pipil
(Baltic) Latvian, Lithuanian **(Finnic)** Estonian, Finnish **(Slavic)** Polish, Russian, Ukranian
(Austro-Asiatic) Khasi **(Dravidian)** Telugu **(IE)** Bengali **(Sino-Tibetan)** Bawm, Garo, Newari (Kathmandu)

Figure 3: A small subset of the world areas identified.

the history of the region. The fourth area identified by our model corresponds roughly to the debated "English" area. Our area includes the requisite French/English/German/Norwegian group, as well as the somewhat surprising Irish. However, in addition to being intuitively plausible, it is not hard to find evidence in the literature for the contact relationship between English and Irish (Sommerfelt, 1960).

In the whole-world experiment, the model identified too many linguistic areas to fit (39 in total that contained at least two languages, and contained at least two language families). In Figure 5.1, we depict the areas found by our model that best correspond to the areas described in Section 2.1.1. We acknowledge that this gives a warped sense of the quality of our model. Nevertheless, our model *is* able to identify large parts of the the Meso-American area, the Baltic area and the South Asian area. (It also finds the Balkans, but since these languages are all IE, we do not consider it a linguistic area in this evaluation.) While our model does find areas that match Meso-American and North-west American areas, neither is represented in its entirety (according to the definition of these areas given in Section 2.1.1).

Model	Rand	F-Sc	Edit	NVI
K-means	0.9149	0.0735	0.1856	0.5889
Pitman-Yor	0.9637	0.1871	0.6364	0.7998
Areal model	**0.9825**	**0.2637**	**0.8295**	**0.9090**

Table 1: Area identification scores for two baseline algorithms (K-means and Pitman-Yor clustering) that do not use hierarchical structure, and for the Areal model we have presented. Higher is better and all differences are statistically significant at the 95% level.

Despite the difficulty humans have in assigning linguistic areas, In Table 1, we explicitly compare the quality of the areal clusters found on the IE subset. We compare against the most inclusive areal lists from Section 2.1.1 for IE: the Balkans and the Baltic. When there is overlap (eg., Romani appears in both lists), we assigned it to the Balkans.

We compare our model with a flat Pitman-Yor model that does not use the hierarchy. We also compare to a baseline K-means algorithm. For K-means, we ran with $K \in \{5, 10, 15, \ldots, 80, 85\}$ and chose the value of K for each metric that did best (giving an unfair advantage). Clustering performance is measured on the Indo-European task according to the Rand Index, F-score, Normalized Edit Score (Pantel, 2003) and Normalized Variation of Information (Meila, 2003). In these results, we see that the Pitman-Yor process model dominates the K-means model and the Areal model dominates the Pitman-Yor model.

5.2 Identifying Areal Features

Our second experiment is an analysis of the *features* that tend to be shared areally (as opposed to genetically). For this experiment, we make use of the whole-world version of the data, again with known language family structure. We initialize a Gibbs sampler from the MAP configuration found in Section 5.1. We run the sampler for 1000 iterations and take samples every ten steps.

From one particular sample, we can estimate a posterior distribution over each π_f. Due to conjugacy, we obtain a posterior distribution of $\pi_f \sim \mathcal{B}et(1 + \sum_n Z_{n,f}, 1 + \sum_n [1 - Z_{n,f}])$. The 1s come from the prior. From this Beta distribution, we can ask the question: what is the probability that a value of π_f drawn from this distribution will have value < 0.5? If this value is high, then the feature is likely

p(gen)	#f	Feature Category
.00	1	Tea
.73	19	Phonology
.73	9	Lexicon
.74	4	Nominal Categories / Numerals
.79	5	Simple Clauses / Predication
.80	5	Verbal Categories / Tense and Aspect
.87	8	Nominal Syntax
.87	8	Simple Clauses / Simple Clauses
.91	12	Nominal Categories / Articles and Pronouns
.94	17	Word Order
.99	10	Morphology
.99	6	Simple Clauses / Valence and Voice
.99	7	Complex Sentences
.99	7	Nominal Categories / Gender and Number
.99	5	Simple Clauses / Negation and Questions
1.0	1	Other / Clicks
1.0	2	Verbal Categories / Suppletion
1.0	9	Verbal Categories / Modality
1.0	4	Nominal Categories / Case

Table 2: Average probability of genetic for each feature category and the number of features in that category.

to be a "genetic feature"; if it is low, then the feature is likely to be an "areal feature." We average these probabilities across all 100 samples.

The features that are *most* likely to be areal according to our model are summaries in Table 2. In this table, we list the *categories* to which each feature belongs, together with the number of features in that category, and the *average* probability that a feature in that category is genetically transmitted. Apparently, the vast majority of features are *not* areal.

We can treat the results presented in Table 2 as a hierarchy of borrowability. In doing so, we see that our hierarchy agrees to a large degree with the hierarchies summarized in Section 2.1.2. Indeed, (aside from "Tea", which we will ignore) the two most easily shared categories according to our model are phonology and the lexicon; this is in total agreement with the agreed state of affairs in linguistics.

Lower in our list, we see that noun-related categories tend to precede their verb-related counterparts (nominal categories before verbal categores, nominal syntax before complex sentences). According to Curnow (2001), the most difficult features to borrow are phonetics (for which we have no data), bound grammatical forms (which appear low on our list), morphology (which is 99% genetic, according to our model) and syntactic frames (which would roughly correspond to "complex sentences", another item which is 99% genetic in our model).

Indo-European

Model	Accuracy	Log Prob
Baseline	0.635 (± 0.007)	-0.583 (± 0.008)
Areal model	$\mathbf{0.689}$ (± 0.010)	$\mathbf{-0.526}$ (± 0.027)

World

Model	Accuracy	Log Prob
Baseline	0.628 (± 0.001)	-0.654 (± 0.003)
Areal model	$\mathbf{0.635}$ (± 0.002)	$\mathbf{-0.565}$ (± 0.011)

Table 3: Prediction accuracies and log probabilities for IE (top) and the world (bottom).

5.3 Genetic Reconstruction

In this section, we investigate whether the use of areal knowledge can improve the automatic reconstruction of language family trees. We use Kingman's coalescent (see Section 2.2.2) as a probabilistic model of trees, endowed with a binomial mutation process on the language features.

Our baseline model is to run the vanilla coalescent on the WALS data, effective reproducing the results presented by Teh et al. (2007). This method was already shown to outperform competing hierarchical clustering algorithms such as average-link agglomerative clustering (see, eg., Duda and Hart (1973)) and the Bayesian Hierarchical Clustering algorithm (Heller and Ghahramani, 2005).

We run the same experiment both on the IE subset of data and on the whole-world subset. We evaluate the results qualitatively, by observing the trees found (on the IE subset) and quantitatively (below). For the qualitative analysis, we show the subset of IE that does not contain Indic languages or Iranian languages (just to keep the figures small). The tree derived from the original data is on the left in Figure 4, below:

The tree based on areal information is on the right in Figure 4, below. As we can see, the use of areal information qualitatively improves the structure of the tree. Where the original tree had a number of errors with respect to Romance and Germanic languages, these are sorted out in the areally-aware tree. Moreover, Greek now appears in a more appropriate part of the tree and English appears on a branch that is further out from the Norse languages.

We perform two varieties of quantitative analysis. In the first, we attempt to predict unknown feature values. In particular, we *hide* an addition 10% of the feature values in the WALS data and fit a model

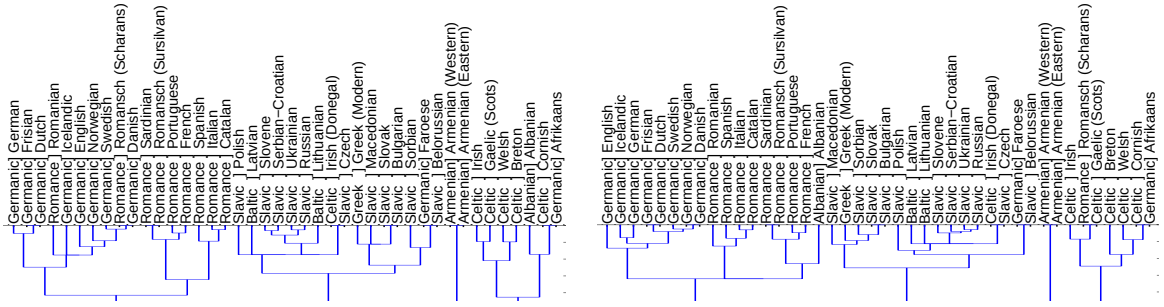

Figure 4: Genetic trees of IE languages. (Left) with no areal knowledge; (Right) with areal model.

Indo-European versus Genus

Model	Purity	Subtree	LOO Acc
Baseline	0.6078	0.5065	**0.3218**
Areal model	**0.6494**	**0.5455**	0.2528

World versus Genus

Model	Purity	Subtree	LOO Acc
Baseline	0.3599	0.2253	0.7747
Areal model	**0.4001**	**0.2450**	**0.7982**

World versus Family

Model	Purity	Subtree	LOO Acc
Baseline	0.4163	0.3280	0.4842
Areal model	**0.5143**	**0.3318**	**0.5198**

Table 4: Scores for IE as compared against genus (top); for world against genus (mid) and against family (low).

to the remaining 90%. We then use that model to predict the hidden 10%. The baseline model is to make predictions according to the family tree. The augmented model is to make predictions according to the family tree *for those features identified as genetic* and according to the linguistic area *for those features identified as areal.* For both settings, we compute both the absolute accuracy as well as the log probability of the hidden data under the model (the latter is less noisy). We repeat this experiment 10 times with a different random 10% hidden. The results are shown in Table 3, below. The differences are not large, but are outside one standard deviation.

For the second quantitative analysis, we use present purity scores (Heller and Ghahramani, 2005), subtree scores (the number of interior nodes with pure leaf labels, normalized) and leave-one-out log accuracies (all scores are between 0 and 1, and higher scores are better). These scores are computed against both language family and language genus as the "classes." The results are in Table 4, below. As we can see, the results are generally in favor of the Areal model (LOO Acc on IE versus Genus notwithstanding), depending on the evaluation metric.

Radius	Purity	Subtree	LOO Acc
125	0.6237	0.4855	0.2013
250	0.6457	0.5325	0.2299
500	0.6483	**0.5455**	0.2413
1000	**0.6494**	**0.5455**	0.2528
2000	0.6464	0.4935	0.3218
4000	0.6342	0.4156	**0.4138**

Table 5: Scores for IE vs genus at varying radii.

5.4 Effect of Radius

Finally, we evaluate the effect of the radius hyperparameter on performance. Table 5 shows performance for models built with varying radii. As can be seen by purity and subtree scores, there is a "sweet spot" around 500 to 1000 kilometers where the model seems optimal. LOO (strangely) seems to continue to improve as we allow areas to grow arbitrarily large. This is perhaps overfitting. Nevertheless, performance is robust for a range of radii.

6 Discussion

We presented a model that is able to recover well-known linguistic areas. Using this areas, we have shown improvement in the ability to recover phylogenetic trees of languages. It is important to note that despite our successes, there is much at our model does not account for: borrowing is known to be assymetric; contact is temporal; borrowing must obey univeral implications. Despite the failure of our model to account for these issues, however, it appears largely successful. Moreover, like any "data mining" expedition, our model suggests new linguistic areas (particularly in the "whole world" experiments) that deserve consideration.

Acknowledgments

Deep thanks to Lyle Campbell, Yee Whye Teh and Eric Xing for discussions; comments from the three anonymous reviewers were very helpful. This work was partially supported by NSF grant IIS0712764.

References

Alexandra Aikhenvald and R.M.W. Dixon, editors. 2001. *Areal diffusion and genetic inheritance: problems in comparative linguistics*. Oxford University Press.

Lyle Campbell. 2005. Areal linguistics. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*. Elsevier, 2 edition.

Lyle Campbell. 2006. Areal linguistics: the problem to the answer. In April McMahon, Nigel Vincent, and Yaron Matras, editors, *Language contact and areal linguistics*.

Lyle Campbell. In press. Why Sir William Jones got it all wrong, or Jones' role in how to establish language families. In Joseba Lakarra, editor, *Festschrift/Memorial volume for Larry Trask*.

Timothy Curnow. 2001. What language features can be "borrowed"? In Aikhenvald and Dixon, editors, *Areal diffusion and genetic inheritance: problems in comparative linguistics*, pages 412–436. Oxford University Press.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

R.M.W. Dixon. 2001. The Australian linguistic area. In Aikhenvald and Dixon, editors, *Areal diffusion and genetic inheritance: problems in comparative linguistics*, pages 64–104. Oxford University Press.

R. O. Duda and P. E. Hart. 1973. *Pattern Classification And Scene Analysis*. Wiley and Sons, New York.

David Dunson and Ju-Hyun Park. 2007. Kernel stick breaking processes. *Biometrika*, 95:307–323.

Isidore Dyen, Joseph Kurskal, and Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5). American Philosophical Society.

Thomas S. Ferguson. 1974. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629, July.

Zoubin Ghahramani. 2005. Nonparametric Bayesian methods. Tutorial presented at UAI conference.

Joseph Greenberg, editor. 1963. *Universals of Languages*. MIT Press.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press.

E. Haugen. 1950. The analysis of linguistic borrowing. *Language*, 26:210–231.

Katherine Heller and Zoubin Ghahramani. 2005. Bayesian hierarchical clustering. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 22.

Michael I. Jordan. 2005. Dirichlet processes, Chinese restaurant processes and all that. Tutorial presented at NIPS conference.

Harmut Katz. 1975. *Generative Phonologie und phonologische Sprachbünde des Ostjakischen un Samojedischen*. Wilhelm Fink.

J. F. C. Kingman. 1982a. The coalescent. *Stochastic Processes and their Applications*, 13:235–248.

J. F. C. Kingman. 1982b. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43. Essays in Statistical Science.

James Matisoff. 2001. Genetic versus contact relationship: prosodic diffusibility in South-East Asian languages. In Aikhenvald and Dixon, editors, *Areal diffusion and genetic inheritance: problems in comparative linguistics*, pages 291–327. Oxford University Press.

Marina Meila. 2003. Comparing clusterings. In *Proceedings of the Conference on Computational Learning Theory (COLT)*.

E. Moravcsik. 1978. Language contact. In J.H. Greenberg, C. Ferguson, and E. Moravcsik, editors, *Universals of Human Language*, volume 1; Method and Theory, pages 3–123. Stanford University Press.

Patrick Pantel. 2003. *Clustering by Committee*. Ph.D. thesis, University of Alberta.

J. Pitman and M. Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

Jim Pitman. 2002. Combinatorial stochastic processes. Technical Report 621, University of California at Berkeley. Lecture notes for St. Flour Summer School.

M.D. Ross. 1988. Proto Oceanic and the Austronesian languages of western melanesia. *Canberra: Pacific Linguitics, Australian National University*.

Alf Sommerfelt. 1960. External versus internal factors in the development of language. *Norsk Tidsskrift for Sprogvidenskap*, 19:296–315.

Yee Whye Teh, Hal Daumé III, and Daniel Roy. 2007. Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems (NIPS)*.

Sarah Thomason. 2001. *Language contact: an introduction*. Edinburgh University Press.

T. Warnow, S.N. Evans, D. Ringe, and L. Nakhleh. 2005. A stochastic model of language evolution that incorporates homoplasy and borrowing. In *Phylogenetic Methods and the Prehistory of Language*. Cambridge University Press. Invited paper.