

A Bayesian Model for Discovering Typological Implications

Hal Daumé III

School of Computing
University of Utah

me@hal3.name

Lyle Campbell

Department of Linguistics
University of Utah

lcampbel@hum.utah.edu

A “Typological *What?!*”

English:

I eat dinner in restaurants.

French:

je mange le diner dans les restaurants
I eat the dinner in the restaurants

Japanese:

boku -wabangohan -o resutoran -ni taberu
I -topic dinner -obj restaurants in eat

Hindi:

main raat ka khaana restra mein khaata hoon
I night-of-meal restaurants in eat am

VO --> PreP
PostP --> OV

Verb-Object (VO)
Prepositional (PreP)

Object-Verb (OV)
Postpositional
(PostP)

The Typologist's Life

	PreP	PostP
VO	16	0
OV	3	11

VO --> PreP (100%)
 OV --> PostP (79%)

(Greenberg, 1963) –
 Based on 30 diversely
 sampled languages

Now, repeat for lots of feature pairs

Difficulties with Typical Approach

**A --> B (99%) uninteresting if
A rare or B common**

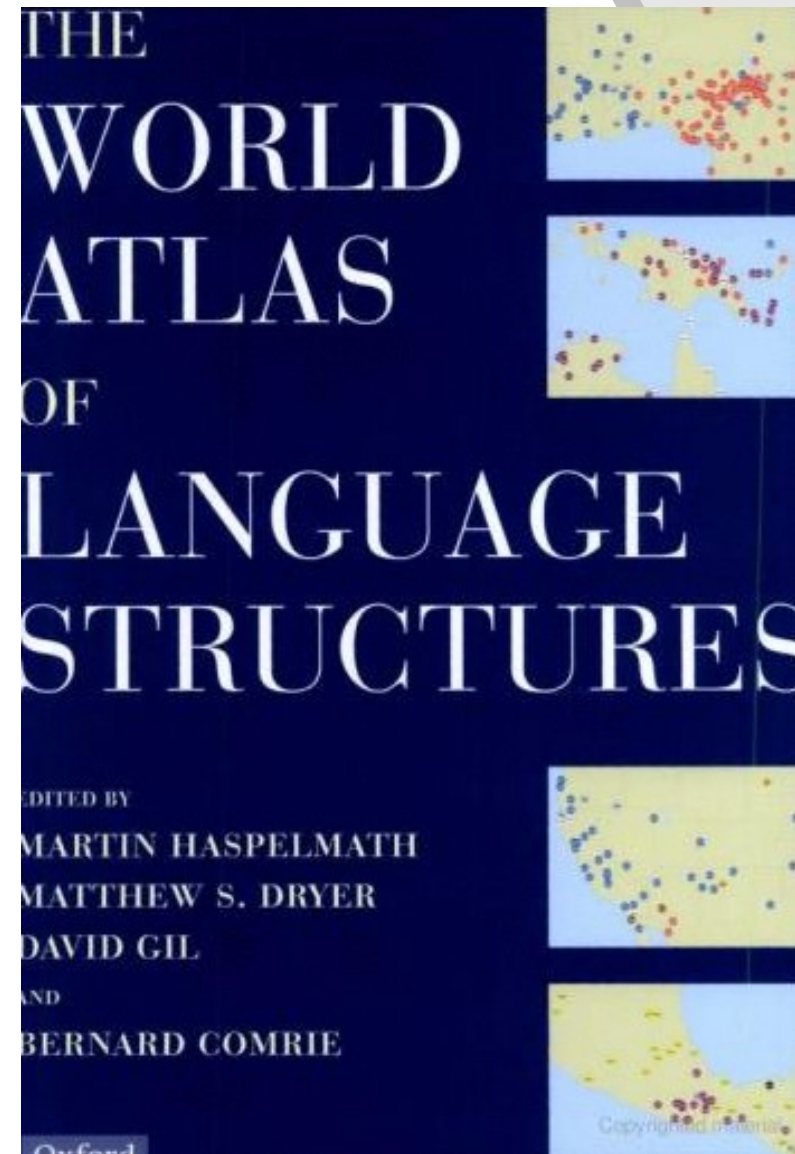
Search process tedious

**Sampling problem when
many languages considered**

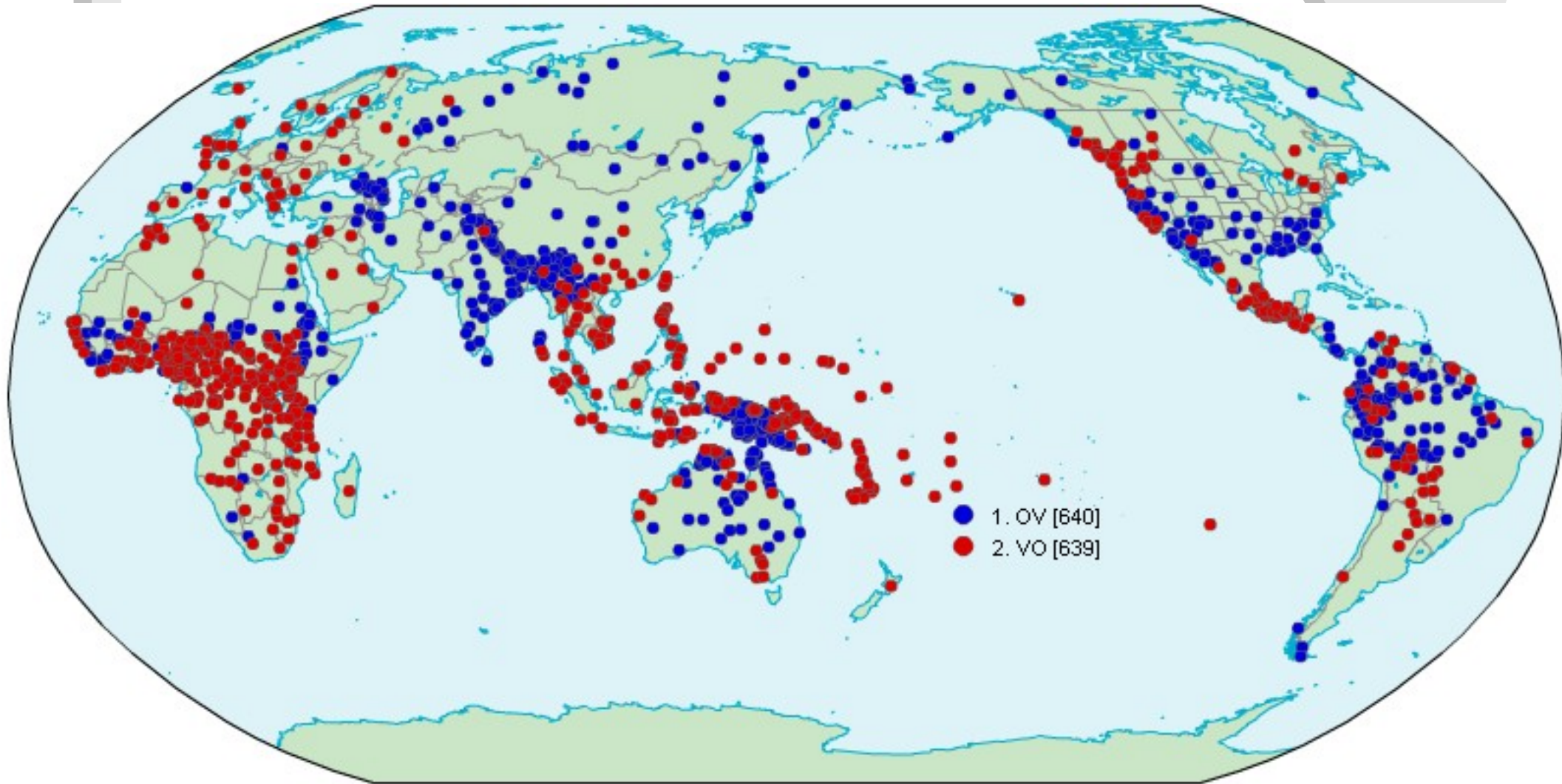
Process is inherently noisy

A Typological Database

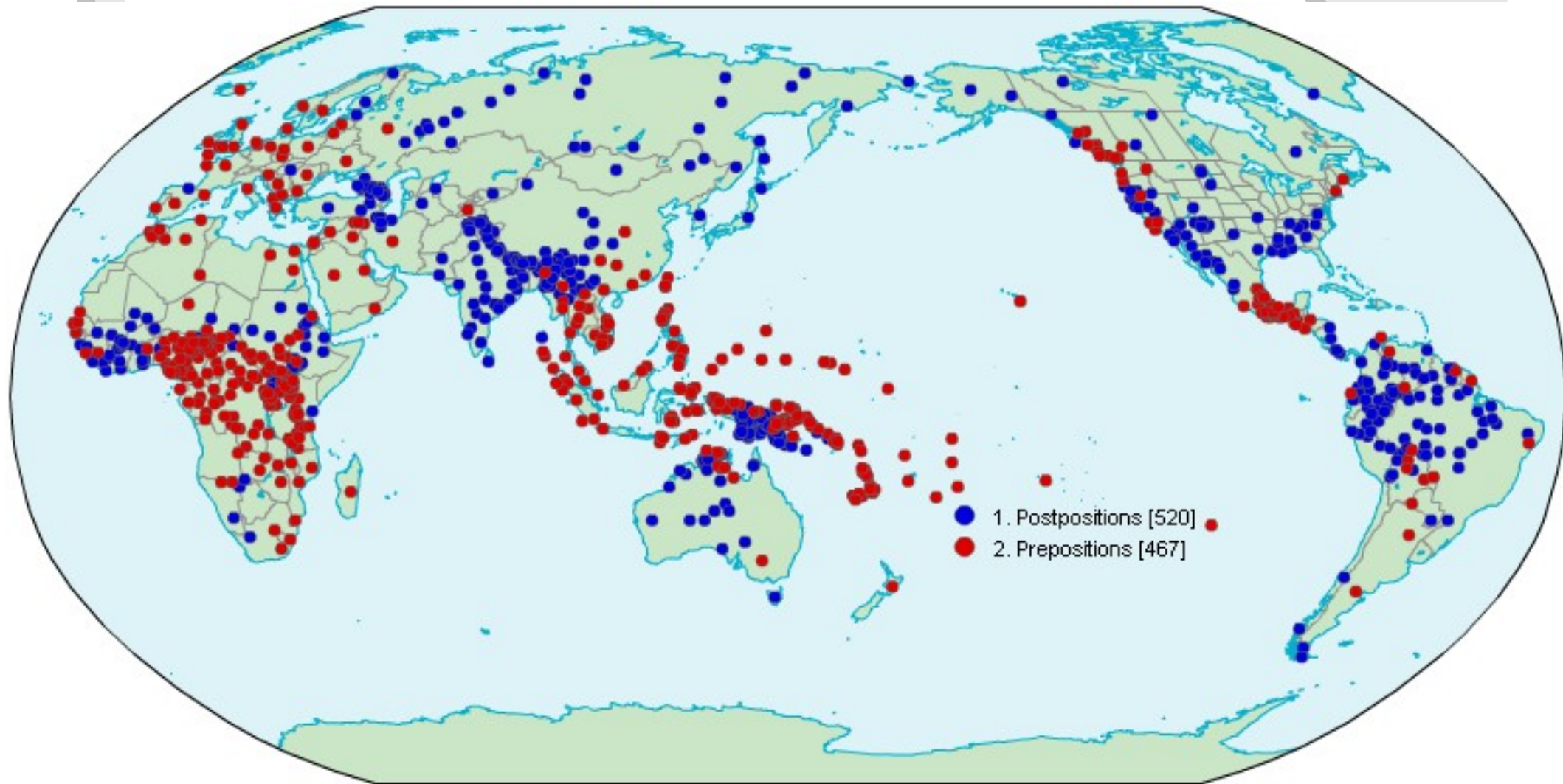
- 2150 Languages
 - 35 language families
 - 275 language geni
- 139 Features
 - 11 feature categories
- Sparsely sampled
 - 85% missing data



Typological Map: VO

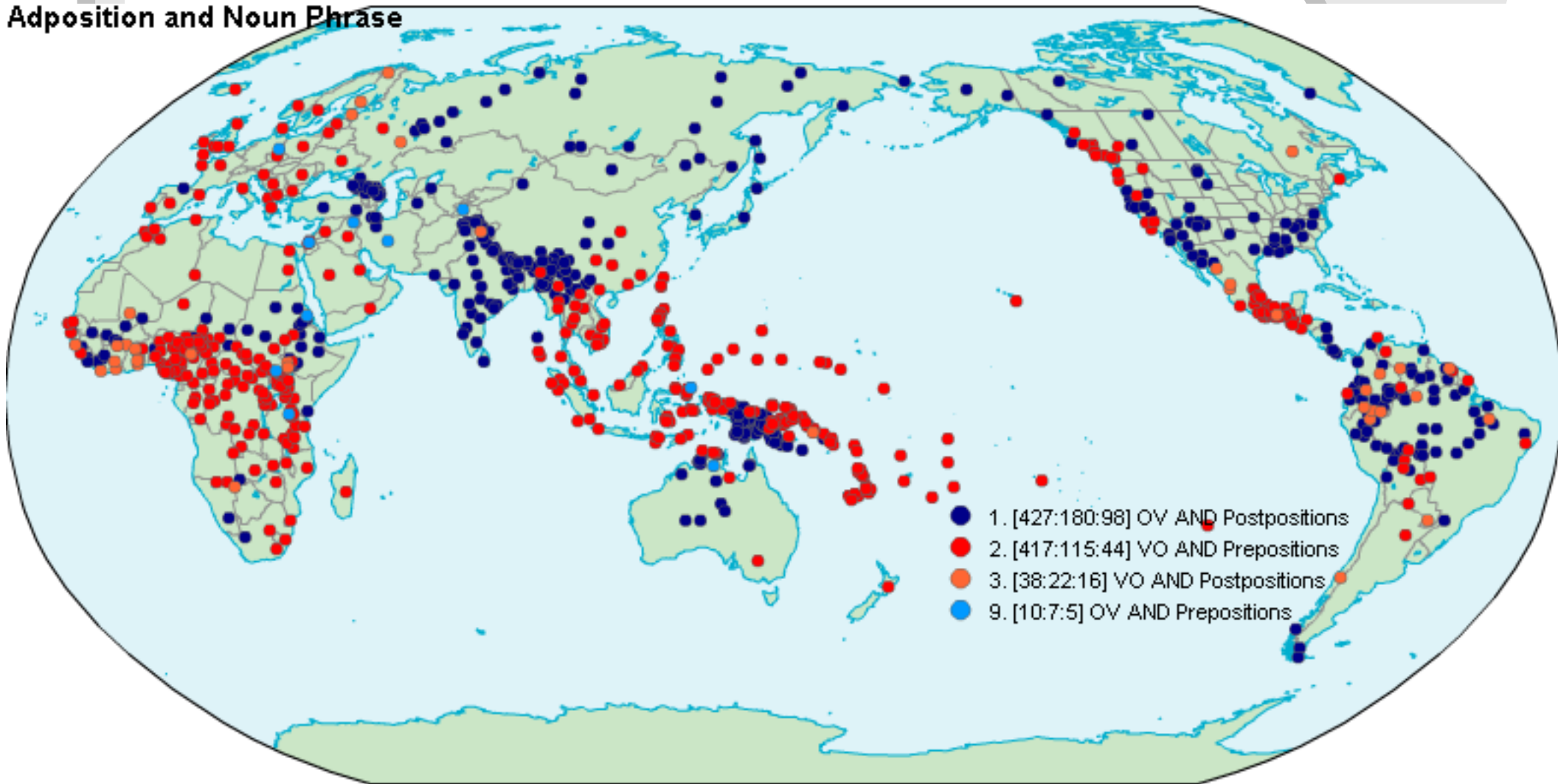


Typological Map: PreP



Typological Map: VO and PreP

Adposition and Noun Phrase



An Initial Model

- Consider two features --> 2xN matrix
- First, generate first column with prior probability π_1
- Next, decide if the implication holds
- Finally, generate the second column:
 - With probability π_2 if feature 1 is not “+” or if the implication doesn't hold
 - Forced to be “+” otherwise

VO	PreP
+	+
+	?
-	+
+	-
+	+
?	+
+	+
?	-
?	-
+	?
-	-
+	+
?	-
+	+
-	+

An Initial Model

- Consider two features --> 2xN matrix
- First, generate first column with prior probability π_1
- Next, decide if the implication holds
- Finally, generate the second column:
 - With probability π if feature 1 is not “+”

VO	PreP
+	+
+	?
-	+
+	-
+	+
?	+
+	+
?	-
?	-
+	?
-	-
+	+
?	-
+	+
-	+

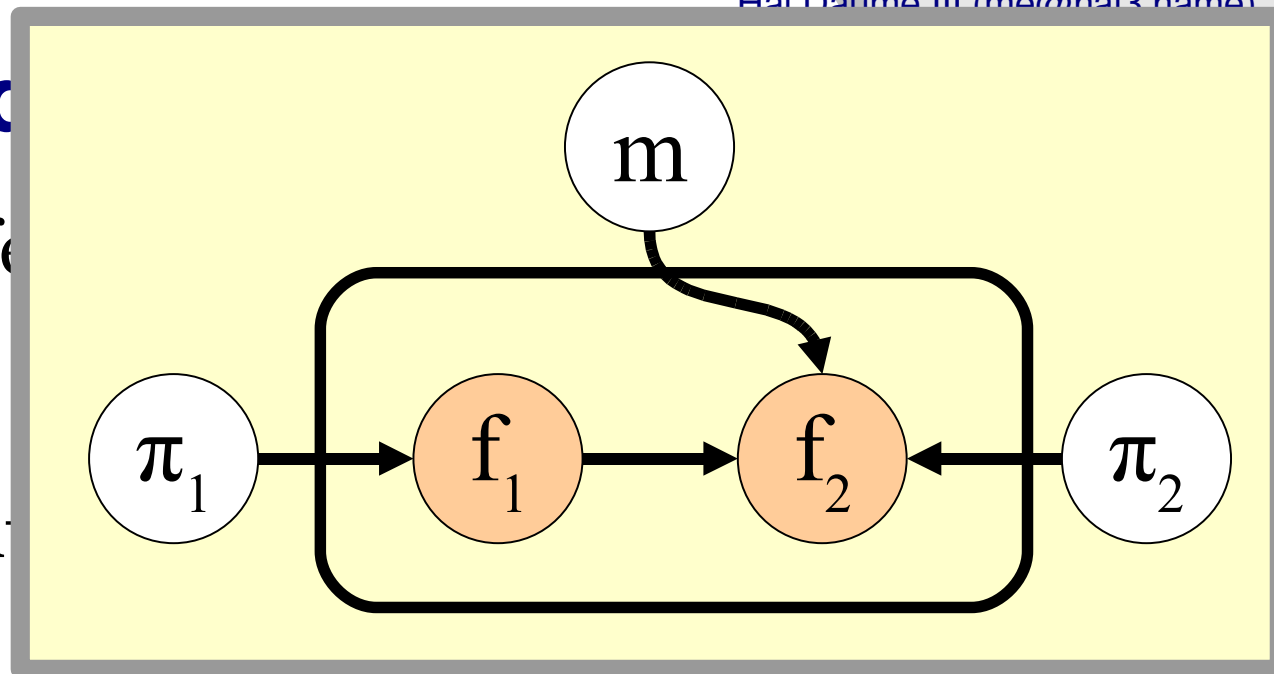
Problems:

Cannot handle noisy data

Doesn't address sampling problem

An Initial Model

- Consider two features
- First, generate prior probabilities



- Next, decide if the implication holds
- Finally, generate the second column:
 - With probability π_1 if feature 1 is not “+”

?	+
+	+
?	-
?	-
+	?
-	-
+	+
?	-
+	+
-	+

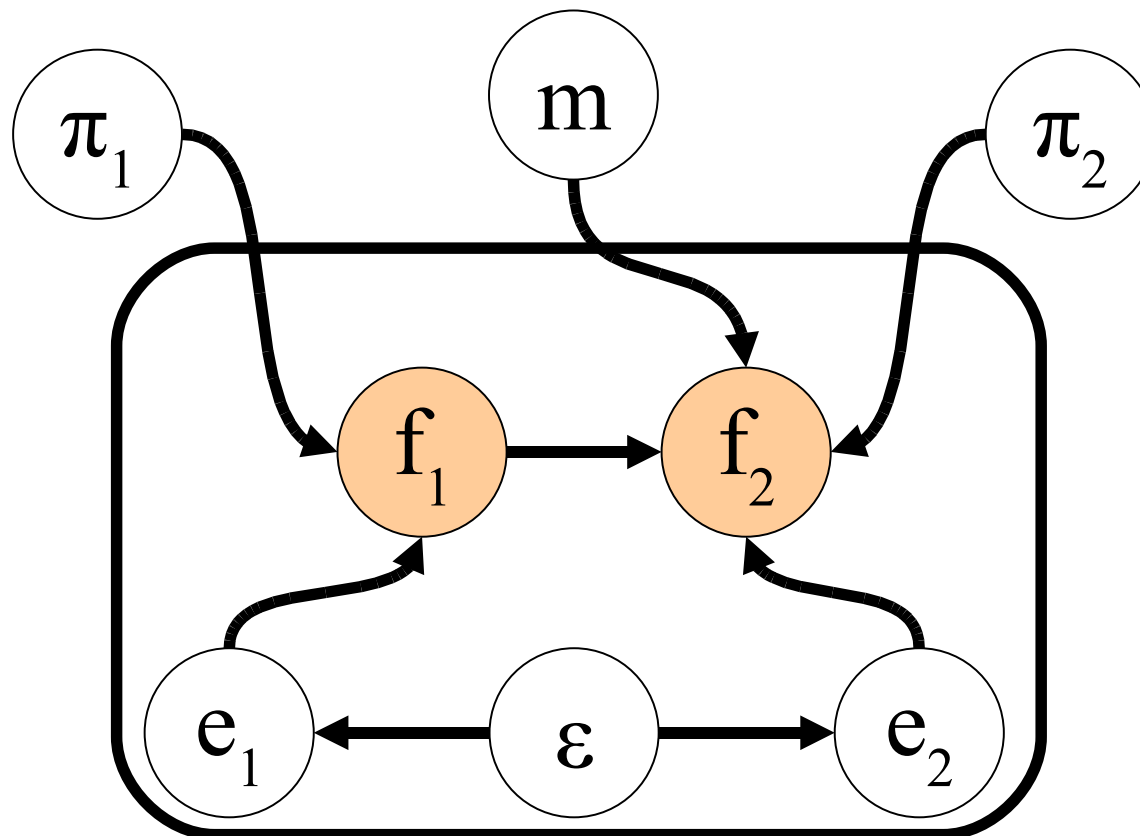
Problems:

Cannot handle noisy data

Doesn't address sampling problem

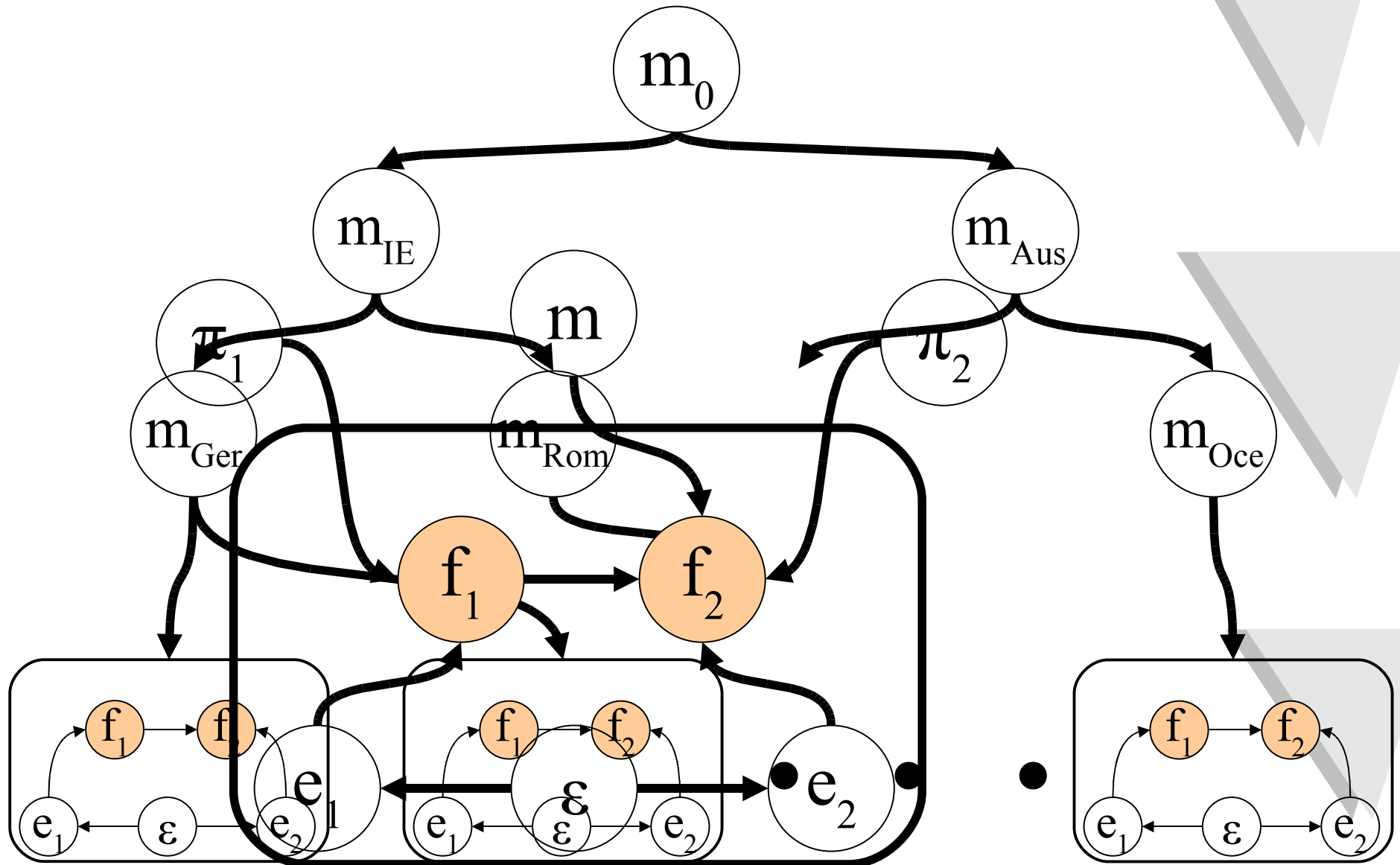
Fixing the Noise Problem

- Assume language-specific noise
- Model remains unchanged, except a new variable causes “f” to be flipped



Fixing the Sampling Problem

- Hierarchical Bayes prior...



Inference

- Binomials get Beta priors
 - $m \sim \text{Uniform}$
 - $\varepsilon \sim \text{Beta}$ with 5% mean, 0-10% with 50% probability
- Everything else gets uniform priors
- Inference by Gibbs sampling
 - Plus a rejection sampler subroutine

Three Models

Flat – All languages independent

LingHier – Typological Hierarchy

DistHier – Obtained by clustering positionally

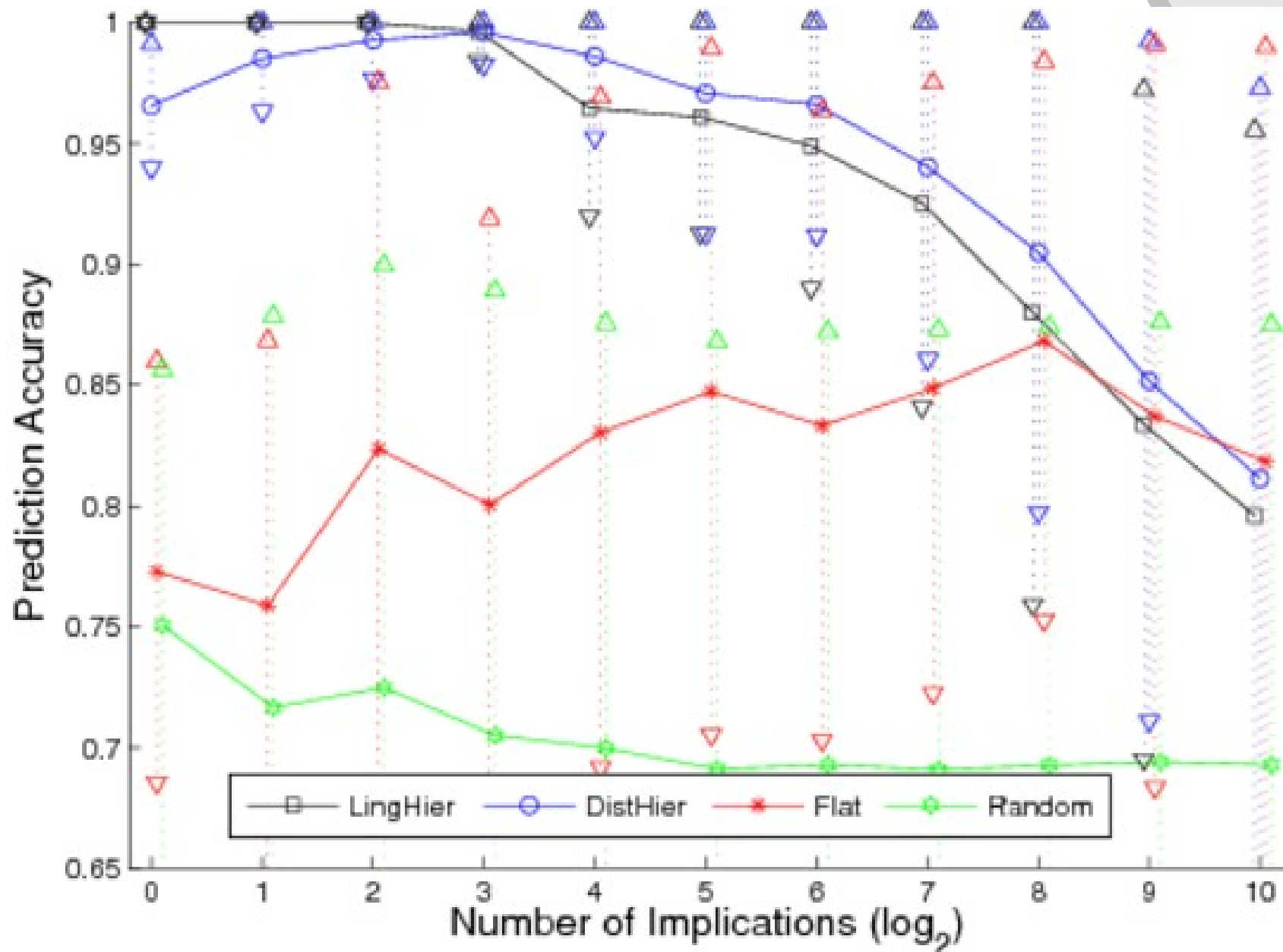
Automatically Extracting Implications

- Search only over pairs with:
 - 250 languages for which both features are known
 - 15 languages for which both hold simultaneously
 - When f_1 is true, f_2 is true with $>50\%$ probability
- Reduces space from 19,000 to 3442
- Sort by probability that m is true
- Evaluate:
 - Compare *restorative accuracy* versus each other
 - Compare against well-known implications

Restorative Accuracy

- Hide additional random data
- Sort implications by probability under each model
- Restore using top 1, 2, 4, 8, ... implications
- Compute accuracy as a function of ranking
 - (Will inevitably decline as count increases)

Restoration Accuracy by Model



Top Implications - If a language likes to have suffixes, it will probably have a suffix for tense/aspect.

Postpositions	Gen-N	
OV	PostP	
OV	Gen-N	Greenberg #4 + Greenberg #2a
Gen-Noun	PostP	Greenberg #2a (converse)
PostPositions	OV	Greenberg #2b (converse)
SV	Gen-N	???
Adj-N	Num-N	Greenberg #18
Suffixing	Tense Suf.	Clear explanation
VO	Noun-RelC	Lehmann
Intr. verb	No question prt.	Appeal to economy
Num-N	Dem-N	Hawkins XVI (for postpositional languages)
PreP	VO	Greenberg #3 (converse)
Adj-N	Dem-N	Greenberg #18
Noun-Adj	PostP	Lehmann

If languages have verb morphology to express “interrogativeness” they probably don't also need particles.

Top Implications – Both are rare consonants (LVs are /kp/ and /gb/; uvs are back-of throat sounds)

VO	PreP	
Init. Subord.	PreP	
Prefixing	PreP	Greenberg #27b
Little affixation	N-Adj	???
Labial-velars	No uvulars	See paper
Negative word	No pron poss afx	See paper
Strong prefixin	VO	Lehmann
Subord. Suffix	Suffixing	???
Final Sub. Wor	PostP	Operator-operand principle (Lehmann)
High+Mid F.V	Many vowels	See paper
Plural prefix	N-Gen	???
No fricatives	No tones	???
Oblig. subj. prc	No pron poss afx	See paper
Dem-N	Tense Suf	Operator-operand principle (Lehmann)
PreP	N-RelC	Lehmann, Hawkins

Front-rounded vowels are low on the list of a hierarchy of vowel types. In order to get them, you must have all other kinds.

Discussion

Model for automatic

**? Thanks!
Questions?**

Accounts for noise and sampling problem

**Different hierarchical models
quantitatively different**

**Discovered implications
correlated with known ones
(14 of top 21)**

Many worthy of further exploration:

<http://hal3.name/WALS>