Multilingual Learning with Parameter Co-occurrence Clustering

Jason Riggle, Max Bane*, James Kirby, John Sylak
*bane@uchicago.edu

Department of Linguistics
University of Chicago

39th Meeting of the North East Linguistic Society
November 7–9, 2008
This talk is about “multilingualism” in the broadest sense:
- The knowledge and use of many distinct linguistic systems by individual speakers (listeners).

*Multilingualism under this definition is pervasive.*

Examples along the language–dialect–register continuum:
- Native bilingualism and code-switching.
- “Multi-dialectism” (Clopper 2004).
- Register variation (e.g., Biber 1995).
On this view, any language user knows:

- A collection of potentially overlapping but different systems of communication (and when to use them).

⇒ If learners in multilingual environments are given samples from a mixture of languages, how can they distinguish between, and acquire, the individual component languages?

- We outline a strategy for accomplishing this within (almost) any parameterized linguistic theory.

- Case study: OT model of syllable structure phonotactics.
Plan of the Talk

1. The Problem of Multilingual Learning
2. A Strategy: Tracking Parameter Co-occurrence
3. Testing the Strategy: OT Syllable Grammars
4. Future Work
Formalized as mappings from some set of inputs (underlying representations) to a set of possible outputs (surface representations): \( L : I \rightarrow O \)

- **Syllabification**
  \[
  \text{Input} \quad \text{Output} \\
  CVCCV \quad \mapsto \quad .CV.CCV.
  \]

- **Morphology**
  \[
  \text{Input} \quad \text{Output} \\
  \text{PASTPART}(buy) \quad \mapsto \quad \text{bought (Most Englishes)} \\
  \text{PASTPART}(buy) \quad \mapsto \quad \text{boughten (Some Englishes)}
  \]

- **Syntax/Semantics**
  \[
  \exists t ( t < n \wedge \text{KISS}(\text{John})(\text{Mary})(t)) \quad \mapsto \quad \text{Mary kissed John.}
  \]
Learning as Parameter Estimation

$L$ is a potentially infinite mapping.

- Reduced to a finite set of parameters (a grammar) $g$ that define it according to some theory $\mathcal{G}$: $L = \mathcal{G}(g)$.

- Examples
  - If $\mathcal{G}$ is OT, then $g$ is a constraint ranking.
  - If $\mathcal{G}$ is minimalist syntax, then $g$ is a lexicon of attribute-value matrices.

The (supervised) language learning problem is then:

- Given a finite sample $S = \{(i_1, o_1), \ldots, (i_n, o_n)\}$ of observed input-output pairs, what parameter settings $g$ might define a mapping $L = \mathcal{G}(g)$ consistent with $S$?
A slight modification of the problem:

- Given a finite sample $S = \{(i_1, o_1), \ldots, (i_n, o_n)\}$ of observed input-output pairs drawn from multiple languages, what set of grammars $\{g_1, \ldots, g_k\}$ might define languages $\{L_1 = G(g_1), \ldots, L_k = G(g_k)\}$ consistent with $S$?
  - That is, for each observation $(i_i, o_i) \in S$, there must be some $j \in \{1, \ldots, k\}$ such that $(i_i, o_i) \in L_j = G(g_j)$.

- $k$ is not necessarily given.
Suppose the learner is exposed to mixed data from:

- $L_1$: /VC/ $\mapsto$ [CVC] (epenthesizes onsets)
- $L_2$: /VC/ $\mapsto$ [V] (deletes codas)

Learner must acquire $L_1$ and $L_2$ as distinct languages. How to avoid over-generalizing to $L_3$?

- $*L_3$: /VC/ $\mapsto$ [CV] (does both)
Learning a Union vs Learning a Disjunction

Some existing models treat free variation as a kind of multilingualism.

- Anttila 1997
- Boersma & Hayes 2001, Gradual Learning Algorithm

But they assume learner acquires a union of grammars. The general multilingual problem is to learn a disjunction of grammars.

E.g., under Boersma-Hayes style model, variation between rankings $C_1 \gg C_2 \gg C_3$ and $C_3 \gg C_2 \gg C_1$ gives non-zero probability to all 6 possible rankings of \{\(C_1, C_2, C_3\)\}. 

![Diagram showing three overlapping normal distributions](image-url)
In multilingual learning:

- Need to distinguish or separate the languages represented by the sample, but
- Need to accommodate possibility that target languages may be highly similar, and overlap significantly.
Plan of the Talk

1. The Problem of Multilingual Learning
2. A Strategy: Tracking Parameter Co-occurrence
3. Testing the Strategy: OT Syllable Grammars
4. Future Work
Our strategy will assume that:

- Given a single (input, output)-pair, it is possible to determine (efficiently) which parameter settings are consistent with that pair—i.e., which grammars define a language containing that pair.

This may not be trivial. Could imply monolingual learning problem is already solved.

- In OT, we can use Prince’s (2002) Elementary Ranking Conditions.
Elementary Ranking Conditions (ERCs)

ERCs encode which grammars (constraint orderings) are consistent with some observations. Statements in a 3-valued logic that express disjunctions of partial orderings.

The meaning of an ERC is that at least one of the elements marked with a \( W \) outranks all of the elements marked with \( L \)’s.

<table>
<thead>
<tr>
<th>input</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cand. a</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Cand. b</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cand. c</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ erc(a \succ b) = \langle W, L, e, L \rangle \]
\[ erc(a \succ c) = \langle W, W, L, L \rangle \]

Candidate \( a \) beats candidate \( c \) under any ranking in which either \( C_1 \) or \( C_2 \) outranks both \( C_3 \) and \( C_4 \).
A batch-learning strategy (three steps):

1. Begin with an empty, unweighted, undirected “co-occurrence graph.”

2. For each observation \((i, o)\) in the mixed sample:
   A. Make a list of the properties (parameter settings) a grammar must have in order to be consistent with the observation (in OT, a list/conjunction of ERCs).
   B. For each property in that list, add a node to the co-occurrence graph, and add an edge between each pair of those nodes.
So far the co-occurrence graph reflects which parameter settings were seen to be consistent with parts of the sample, and the edges indicate which were seen to be simultaneously required for a single observation.

**Hypothesis:** Intuitively, the “dense”—highly connected, mutually consistent—regions of the graph correspond to the grammars of the individual languages.
3. Apply some heuristic to identify the dense regions, or clusters, of the co-occurrence graph, and choose hypothesis grammars consistent with the parameter settings specified by those clusters.

Notes:

- Steps 1 and 2 just reduce the problem to a well-studied one: clustering.
- ...But with added requirements that the clusters might overlap, and should give mutually non-contradictory parameter settings.
- Many clustering heuristics do not presuppose the number of clusters (languages) as given.
Betweenness Centrality

We’ve experimented with a simple clustering heuristic based on the graph theoretic notion of “centrality”.

- The **betweenness centrality** (Freeman 1977) of a node $n$ in a graph $G$ is the proportion of shortest paths between pairs of nodes in $G$ that pass through $n$.

Can be used to identify clusters by locating the nodes on their edges—the nodes **between** dense areas.
Betweenness Centrality Clustering Heuristic

A possible implementation of Step 3:

A Begin with an empty set $H$ of hypothesis grammars.

B For each connected component $C$ of the co-occurrence graph $G$:
   i If $C$ is not internally contradictory, construct a hypothesis grammar consistent with $C$ and add it to $H$. Remove $C$ from $G$.
   ii Otherwise, find node $v \in C$ with greatest betweenness centrality. “Split” $v$ apart, distributing copies of it to each cluster that it lay between, possibly creating new connected components in $G$.

C Greedily merge any hypothesis grammars in $H$ that agree with each other on all inputs.
Betweenness Centrality Clustering Heuristic

By splitting nodes into copies, we allow hypothesis grammars to have overlapping properties/parameters.
Plan of the Talk

1. The Problem of Multilingual Learning
2. A Strategy: Tracking Parameter Co-occurrence
3. Testing the Strategy: OT Syllable Grammars
4. Future Work
Monte Carlo simulations of learning multiple syllable structure grammars.

- 12-constraint OT grammars, based on Prince and Smolensky’s (1993/2002) syllable theory.
- Implemented computationally with finite-state constraints (Riggle 2004).
- Alphabet \( \{C, V\} \), fixed lexicon of all length \( \leq 5 \) CV-strings (62 forms).
- Factorial typology of 679 possible syllable structure languages.
A Syllable Structure Case Study

- Each Monte Carlo trial, 1–5 target (i.e., “teacher’s”) grammars randomly generated (random constraint rankings).
- Learner receives mixed sample of target languages.
  - Learner must distinguish the languages.
- Only a portion of each target language’s lexicon is sampled.
  - Learner must generalize.
- Look at average results across 100–1,000 trials.
Quantifying the Learner’s Behavior

Mistakes learner can make:
- Undergeneralize, overgeneralize.
- Scrambled/mixed-up languages.
- Wrong number of hypotheses.

We measure:
- Number of hypotheses the learner makes (vs number of actual targets).
- “Over-generation ratio,” a metric of errors like $(L_1: \text{epenthesize}) + (L_2: \text{delete}) \mapsto (L_3: \text{do both}).$
- “Expected agreement displacement,” overall distance measure between set of hypotheses and set of target languages.
Tends to overestimate number of languages. Qualitative difference between bi- and trilingual+ cases?
Focus on Bilingual Learning

Does best with 10–20 samples. Sweet spot for balancing separation vs possibility of overlap?
Focus on Bilingual Learning

When hypotheses different than target languages are acquired, they nonetheless tend to be similar to the targets.
Incorporating Extralinguistic Information

So far, strategy is based solely on grammatical information implied by the sample.

- But we can also incorporate extra/socio-linguistic properties or parameters into the co-occurrence graph to aid in separating languages.
- We’ve experimented with “speaker features.”
  - In addition to ERCs, each utterance is associated with a feature/parameter indicating which of the “teachers” produced it.
  - Co-occurrence of these speaker features can aid significantly in separating otherwise confusable languages.
Speaker Features

Learning improves noticeably.

Readily extendable to other sorts of information:

- Co-occurrence with social context (parameterized somehow)
- Co-occurrence with grammatical context of other modules (syntactic, semantic, etc.)
Future Work

- Larger experiments testing the limits of the strategy.
- Weighted co-occurrence graphs (each edge counts the number of parameter co-occurrences).
- More sophisticated clustering techniques (spectral, Bayesian, etc.)
- Corpus analysis: e.g., distinguishing the registers present in a large body of transcribed telephone conversations.
Thank You!

Selected references on hand-out.