Multilingual Learning with Parameter Co-occurence Clustering

Jason Riggle, Max Bane, James Kirby, John Sylak
University of Chicago

The Problem of Multilingual Discrimination

If learners in multilingual environments are given samples from a mixture of grammars, how well can they distinguish between individual grammars using information contained in those samples? For example, if a learner is exposed to grammars G1 and G2, where G1 epenthesizes onsets and G2 deletes codas, how can a learner avoid generalizing to a new grammar G3, which does both?

a.	G1: /VC/ → [CVC]
b.	G2: /VC/ → [V]
c.	*G3: /VC/ → [CV]

Figure 1

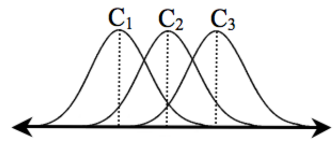

Figure 2

There are Optimality Theoretic models that account for free variation as a kind of multilingualism (Antilla, 1997; Boersma & Hayes, 2001), but these accounts assume that learners acquire unions of grammars. For a more general model of multilingualism, learners must also be able to acquire *disjunctions* of grammars. This rules out a Boersma-Hayes style model (as in Fig. 2) because variation between the rankings $C_1 \gg C_2 \gg C_3$ and $C_3 \gg C_2 \gg C_1$ implies non-zero probability for each of the six possible rankings of $\{C_1, C_2, C_3\}$.

Discrimination by Parameter Co-occurrence Clustering

We propose a heuristic for learning and distinguishing languages in multilingual scenarios by clustering grammatical parameters according to their co-occurrence in utterances. We assume a supervised learning scenario with input-output pairs drawn from multiple OT grammars, parameterized as sets of Elementary Ranking Conditions (ERCs; Prince 2002). Assuming the learner knows the constraints, it can infer the set of ERCs consistent with an observed utterance (I-O mapping). To perform clustering, our algorithm establishes a list of pairs of ERCs that co-occur in the ERC set of at least one utterance in the training sample. These pairs define a network of n ERCs whose dense regions correspond to collections of parameters associated with each other in the samples (see Fig. 3). We use these regions of strongly associated parameters as the basis for distinguishing grammars.

The algorithm proceeds as follows: for each connected component of ERCs in the network, if the set of ERCs in that component is *consistent* (i.e. free of internal contradictions), that ERC set becomes one of the learner's grammar hypotheses. If the connected component contains internal contradictions, the inconsistency is resolved by separating it into subcomponents that are free of contradiction. The points of separation are the ERCs with the highest *betweenness centrality* (Brandes, 2001) in the co-occurrence network. These are the nodes that lie on the highest number of shortest paths between nodes in the network (see Fig. 3). This algorithm is guaranteed to converge after $O(n)$ iterations to some number of discrete, internally consistent ERC sets; these sets are mutually exclusive hypothesis-grammars.

Results

To evaluate the algorithm, the learner received 2-30 training samples from each of 1-5 teachers speaking randomly generated 10-constraint syllable structure grammars. These were rated in terms of the number of hypotheses, the *overgeneralization ratio* (i.e. the fraction of the lexicon for which the learner 'mixed' grammars like G3 in Fig. 1), and *expected agreement displacement*, a measure of how different the set of hypothesis-grammars was from the training grammars. The average results of 100 trials for each pairing of s samples with n teachers are shown in Fig. 4 (14,000 trials total). For small numbers of languages (here, teachers), the algorithm is quite successful, but deteriorates quickly with additional languages. However, the strategy of clustering on parameter co-occurrence is extremely general and can incorporate other parameters. For instance, Fig. 5 illustrates improved discrimination if learners incorporate "speaker parameters" that index utterances by teachers. This suggests that discrimination could be further improved by adding more types of parameters, indexing properties like lexical content, sociolinguistic, syntactic, or semantic features, or anything else that could systematically co-occur with a language, dialect, or register.

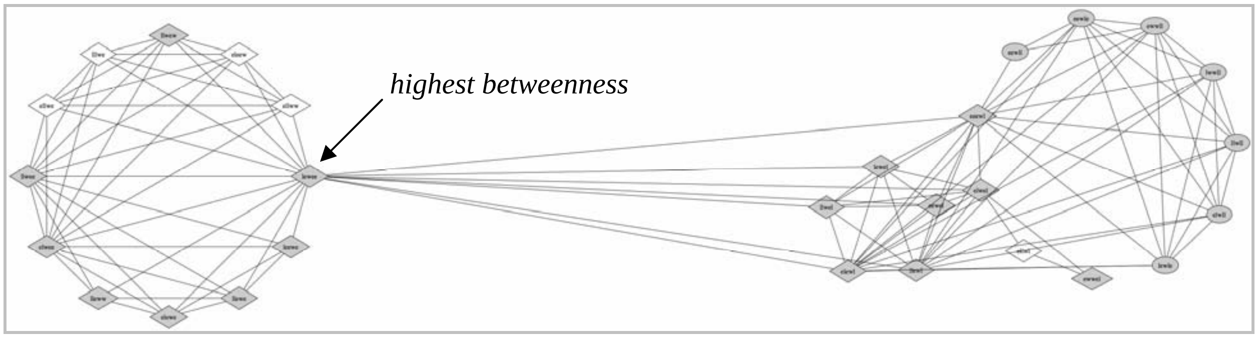

Figure 3: Three dense regions.

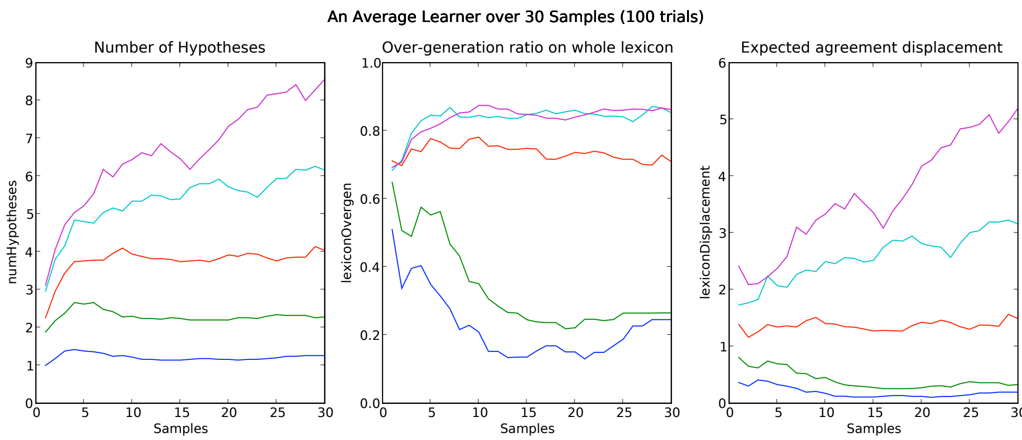

Figure 4: Lines represent different numbers of teachers. See legend in Fig. 5 (Purple = 5 teachers).

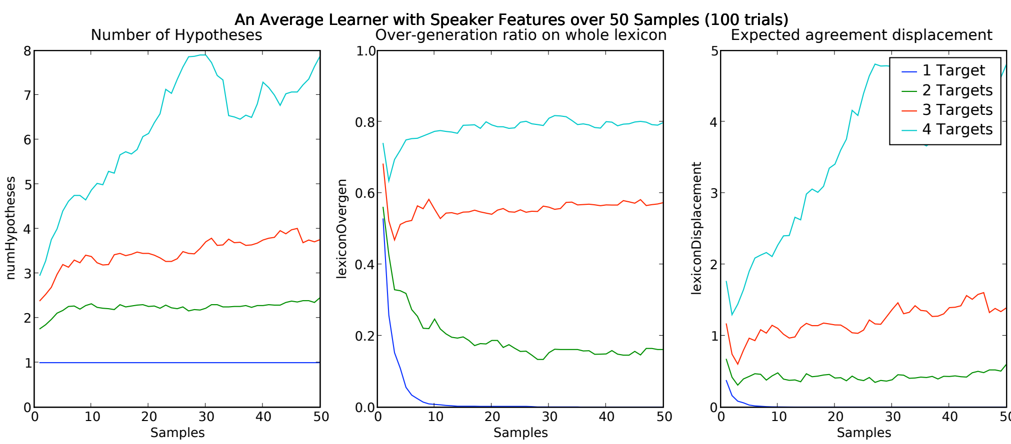

Figure 5: Discrimination results with speaker features.

References

Antilla, A. (1997). *Variation in Finnish Phonology and Morphology*. Ph.D. thesis, Stanford.

Boersma, P. & B. Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:1, pp. 45–86.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25:2, p. 163.

Prince, A. (2002). Entailed ranking arguments. *Rutgers Optimality Archive* ROA-500. `roa.rutgers.edu`.