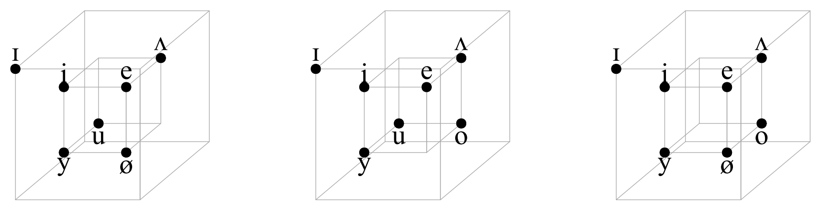
Dunbar, Ewan and Emmanuel Dupoux

## Evidence for two different kinds of feature economy effects on natural inventories

The speech sounds of human languages are thought to have an internal structure based on distinctive features. Previous work has suggested that this structure should influence the shape of inventories (Clements 2003). Intuitively, since [b] versus [p] makes use of the same featural distinction as [d] versus [t], we predict more languages with both pairs than with only one. Given a set of $S$ segments that can be specified using no more than $K$ distinctive features, (for some feature theory), their *feature economy* can be defined as $S/K$ (or, more coherently, as $S/2^K$, assuming a binary feature theory: Hall 2007). Empirical work has shown that natural inventories show greater economy than would be expected by chance (Mackie and Mielke 2011). Here, we dissociate two different tendencies that could both independently give rise to economy: a tendency for *clustering* and a tendency for *dispersion* (Flemming 2001). We show that, together, these two tendencies fully statistically account for the economy in natural inventories. We give two different analyses suggesting that clustering is the more important property of the two.




The above figure shows three vowel inventories that all need four features for their specification, and that all contain seven segments. Thus, they have equal levels of feature economy. However, the segments in the middle inventory are better dispersed than in the other two: they fill up the space more evenly (in particular, along the front–back dimension). The segments in the right-hand inventory are more clustered than in the other two: there are more pairs of minimally featurally different sounds. Forces that influence inventories to have greater clustering will necessarily drive up the economy of those inventories, and forces that lead to greater dispersion will also increase economy in most cases. We define measures for each of these tendencies (both of which are weakly correlated with economy, $S/2^K$). We quantify dispersion for a feature $f$ as $1 - \frac{I(f)}{S-1}$, where $I(f)$ is the absolute difference between $+_f$, the number of sounds specified $+$ for $f$, and $-_f$, the number specified $-$. (The maximum value of $I(f)$ is $S-1$, because if all sounds in the system had the same value for feature $f$, $f$ would not be considered: see below.) We quantify clustering for $f$ as $\frac{P(f)}{\min(-_f, +_f)}$, where $P(f)$ is the number of pairs of segments in the inventory differing only on $f$. We average over $f$ to get measures of dispersion and clustering for each inventory. (Clustering is also averaged over specifications, as it depends on which features are used: see below.)

We analyze the inventories of the languages in the P-Base database (Mielke et al. 2011); in addition, for each language, we extract the lists of consonants, of stops, and of vowels. We include a sound system (or sub-inventory) in our analysis if it can be analyzed exclusively using binary features, under the 23-feature representation used in Mackie and Mielke 2011 (adapted from Clements and Halle 1983). We thus analyze 235 full inventories, 354 consonant inventories, 456 stop inventories, and 480 vowel inventories. To determine whether these sound systems are more economical than would be expected by chance, we generate a random baseline. For each sound system, we draw a fake system of the same size randomly from the set of all attested segments (or consonants, or stops, or vowels). We draw segments in proportion to their frequency in the database, but we put no restrictions on how they combine (e.g., a full inventory could consist only of consonants).

Our measures require choosing a subset of the 23 features that is "necessary" for specifying a given system which is irreducible, in that, if any feature were removed, it would be impossible to specify that system. There may be more than one such set of features (for example, [i] versus [u] can be specified with [back] or [round] if there are no front rounded or back unrounded vowels). We perform an approximate search returning a large sample of the irreducible subsets for each inventory. ($K$ in the economy measure is the size of the smallest irreducible subset.)

For each class of segments, and for each measure, we test whether natural inventories show a tendency toward higher values than the baseline by computing the area under the curve (AUC), a statistic quantifying how well the status of an inventory (natural/random) could be predicted from the value of that measure. Values greater than 0.5 (maximum 1.0) indicate a tendency towards higher values in natural inventories. All classes of segments show tendencies towards economy (quantitative results are summarized in the table); stop and vowel systems show effects of dispersion; and all classes of segments show tendencies towards clustering.

Both dispersion and clustering are correlated with economy. The combination of both measures fully statistically explains the tendency towards economy (after orthogonalizing dispersion and clustering, we regress both out of economy; the residual value no longer differs between natural and random). However, for all classes of segments except vowels, clustering accounts for at least as much of the variance in economy as dispersion (regressing clustering out of economy leads to at least as much decrease in AUC as regressing out dispersion). Furthermore, an alternative random baseline, which, a priori, ought to give inventories *less* like natural ones, actually shows *greater* economy and dispersion than natural inventories (sampling feature matrices without respecting the attestedness or phonetic possibility of segments); however, this kind of inventory still shows less clustering, suggesting that clustering is also the least trivial of the three properties.

| | Full | Consonant | Stop | Vowel |
|---|---|---|---|---|
| Economy (mean nat/rand//AUC) | .10/.06//**.74** | .20/.13//**.72** | .62/.40//**.74** | .57/.42//**.70** |
| Dispersion | .56/.55//.56 | .50/.51//.47 | .70/.59//**.68** | .70/.60//**.68** |
| Clustering | .41/.32//**.80** | .59/.43//**.84** | .88/.67//**.77** | .90/.75//**.71** |
| Residual economy AUC | .51 | .54 | .55 | .50 |
| Decrease in AUC (Disp) | .08 | .02 | .09 | **.14** |
| Decrease in AUC (Clust) | .08 | **.07** | **.12** | .11 |
| Economy vs alt. random | .10/.47//.00 | .20/.56//.04 | .62/.77//.32 | .57/.71//.29 |
| Dispersion | .56/.86//.00 | .50/.83//.01 | .70/.70//.49 | .70/.70//.51 |
| Clustering | .41/.20//**.99** | .59/.26//**.99** | .88/.58//**.84** | .90/.52//**.91** |

Clements, G.N. 2003. Feature economy in sound systems. Phonology 20 287–333. Clements, G.N., and Halle, M. 1983. Problem Book in Phonology (MIT Press). Flemming, E. 2001. Scalar and categorical phenomena in a unified model of phonetics and phonology. Phonology, 18 7–44. Hall, D.C. 2007. The role and representation of contrast in phonological theory. PhD thesis, Toronto. Mackie, S., and Mielke, J. 2011. Feature economy in natural, random, and synthetic inventories. In *Where do phonological features come from?* (Benjamins). Mielke, J., Magloughlin, L., and Hume, E. 2011. Evaluating the effectiveness of unified feature theory and three other feature systems. In *Tones and features: Phonetic and phonological perspectives* (de Gruyter).