Entropic bases for artificial grammar learning and infant mispronounciation studies

Adamantios Gafos, Universität Potsdam

We present different ways to quantify how much information is gained by participants in artificial grammar learning (AGL) experiments and infant mispronounciation studies. We show that under conditions where participants learn a 'natural' but not an 'unatural' rule or where infants show sensitivity to phonemic change only in one direction, there are asymmetries in entropic quantities such as information uncertainty or information gain under the different conditions.

We begin with an AGL study. Martin and Peperkamp (2020), henceforth MP, exposed English participants to a HARMONY condition where stems with two front vowels took a front 'plural' suffix and stems with two back vowels took a back 'plural' suffix: /i-e/, /e-i/, /i-ɛ̃/, /ɛ̃-i/, /e-ɛ̃/, /ɛ̃-e/ (front stem vowel sequence plus front suffix +/tɛl/) and /o-u/, /u-ɔ̃/, /ɔ̃-u/, /o-ɔ̃/, /ɔ̃-o/ (back stem vowel sequence plus back suffix +/tɔl/). In the DISHARMONY condition, suffix choice was reversed. An impressive 173 participants took part. Results showed that harmony was learned but disharmony was not. One interpretation of this asymmetry (the one adopted by MP) is that learners come to the task with biases about natural (harmony) versus unnatural rules (disharmony).

In the trisyllabic words of this task, there are eight [±back] vowel combinations (two choices, [+back] or [−back], for each of the three vowel positions). Of these, two are harmonic (all three vowels back or all three front); the rest are disharmonic. We simulated the learning scenario by setting the initial probabilities of the disharmonic vowel sequences to about three times as probable as the harmonic ones, given that 'only around 35% of English polysyllabic words are harmonic', MP: 82). We then evolved the entropies of the distributions of vowel sequences under the two conditions (HARMONY, DISHARMONY) using Bayesian learning. Let j be an index for the possible vowel sequence and i an index the trial number in the training phase (36 trials as per the experimental study), and n_j^i be the number of occurrences of sequence j up to observation i. Consider the presentation of the next sequence j. This event will have some surprise associated with it, which in information theory (Shannon 1948) is given by the negative logarithm of its probability (the more rare the event, the more surprising it is): $\mathrm{I}(x_i) = -\log(p_j(x_i))$. Surprise is unique to a particular event and measures its improbability. We can estimate the probability of each j in the next observation by $p_j(x_i) = (n_j^i + 1) / \sum_k n_k^i + 1$, that is, increase the counter for sequence j by one and divide by the total occurrences of all sequences plus one. This estimate $p_j(x_i)$ is the mean value of a Dirichlet probability distribution updated using Bayesian learning (Bernardo & Smith 1994). Once some sequence is observed, its presentation will have some impact on the probability distribution of the trained alternatives. Thus, we can also now estimate the entropy of the training situation by $H = -\sum_j p_j * log p_j$. This is the usual sum of the probabilities of each vowel sequence multiplied by the logarithm of the probability of the sequence. Entropy measures the **uncertainty** over all (so far) presented sequences. Surprise measures the improbability of a specific sequence. Entropy is the running average of surprise. We then go to the next observation (present another sequence) and repeat till end of training. The probability distribution of the various vowel sequences changes progressively as a result of training. Figure 1 depicts the evolution of entropy throughout the training phase given the above methods.

At the end of training, the resulting entropy expresses the uncertainty of the learner about what is likely to happen (what vowel pattern is being established during training). Entropy (uncertainty) is less at the end of training for the HARMONY than for the DISHARMONY condition. At test, participants were 'asked to select which of the two plural forms they thought was correct' (MP:

72) with the unsuffixed stem form presented along with the two possible suffixed forms, CVCV-/tɛl/ and CVCV-/tɔl/. The ability to choose between the two different forms is a function of the entropy at the end of training. The higher the entropy, the less the certainty of the choice of the correct response. Another way of expressing the same is to compare, at the end of each training condition, the estimated surprise values of the two competing choices. Figure 2 does this. It can be seen that the two surprise values are not as different in the DISHARMONY condition as they are in the HARMONY condition. Failure to (confidently) choose correctly between the two options in the DISHARMONY condition is thus equivalent to learners being more uncertain about their choice. Hence, asymmetries in the degree of (un)certainty, accumulated over the training phase, about different patterns of vowel sequencing provide a plausible alternative basis for the results.

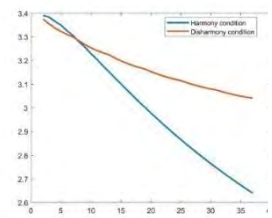

Figure 1: Entropy evolution during two conditions, HARMONY and DISHARMONY.

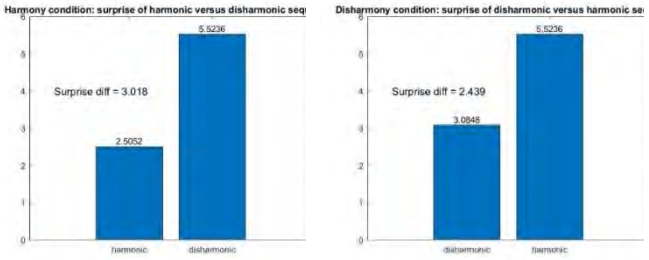

Figure 2: Bar plot comparison of surprise values of harmonic and disharmonic words in two training conditions, HARMONY (left) and DISHARMONY (right).

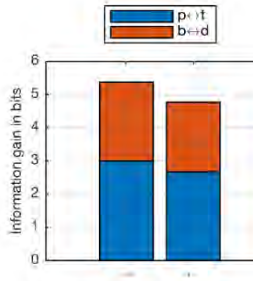

Figure 3: IG for labial to coronal (left column) vs. coronal to labial (right column) change.

The present approach extends to non-AGL settings. Here, we show how it provides a basis for asymmetries in infant word-learning tasks where responses to changes from labial to coronal are detectable but not vice versa. That is, infants perk up when the word /bal/ is pronounced as /dal/ but not vice versa (van der Feest 2007). Let $p(x)$ be the distribution of a correctly pronounced and $q(x)$ that of a mispronounced consonant. We conceive of consonants as distributions of categories. This is because the (in)correctly pronounced consonant is cashed in as a distribution of potential outcomes (on the perceiver's side). This is the data a confusion matrix provides us with: any row of such a matrix specifies the probability distribution of one category, say, /ba/, being perceived as one of several alternatives (/ba/, /pa/, /da/ and so on in the columns of the matrix). The relevant entropic measure quantifying infants' surprise is known as *information gain* (or *expected surprise*) between two distributions, $p(x)$ and $q(x)$: $D[q(x)||p(x)] = \sum_x q(x)\log[q(x)/p(x)]$. This quantifies the expected amount of surprise when perceiving $q(x)$ while expecting $p(x)$. The prediction is that a b→d has higher expected surprise than a d→b change, i.e. $D[d||b] > D[b||d]$. Figure 3 verifies this prediction based on (Dutch) confusion matrix data (Pols 1983): p, b → t, d has higher expected surprise than t, d → p, b. In the talk, we demonstrate how the same approach extends to other cases where such results have been demonstrated.

Selected References: Bernardo, J. M. & A. F. M. Smith. 1994. *Bayesian Theory*. John Wiley; Martin, A. & S. Peperkamp. 2020. Phonetically natural rules benefit from a learning bias. *Phonology* 37, 65-90. Pols, L. C. W. 1983. Three-mode principal component analysis of confusion matrices. *Speech Communication* 2, 275-293; Shannon, C. E. 1948. A mathematical theory of communication. *Bell Systems Technical Journal* 27, 379-423. van der Feest, S. 2007. *Building a phonological lexicon*. Ph.D. Dissertation, Radboud University.