Computational methods are invaluable for typology, but the models must match the questions: Commentary on Dunn et al. (2011)

Roger Levy and Hal Daumé III

August 1, 2011

The primary goal of Dunn et al. is to understand the role that lineage plays in linguistic typology in general, and word order feature correlations in particular. The authors approach this goal through a combination of a large amount of linguistic data—lexical and typological-feature—and statistical analysis. Approaches based on the use of such data, especially as the complexity of available data grows, and analysis are likely to be central to the future of linguistic typology. They allow facts about diverse languages to speak directly to major theoretical questions about feature correlation and historical change, while accounting for how evidence is aggregated across languages known to be related by common ancestry. We applaud Dunn et al. for applying such methods to study word order universals. However, in order to get meaningful answers, one needs to ask the right statistical questions. For each of many feature pairs, the authors pose two central theoretical questions: (Q1) "is there evidence for correlation in the evolution of the features, taking into account common ancestry?" and (Q2) "is there evidence that feature correlation is lineage specific?". The evidence they adduce to answer these questions is based on statistical tests for correlation within each lineage, and thus bears fairly directly on (Q1). Additionally, Dunn et al. propose an affirmative answer to (Q2), noting that for no feature pair was evidence for correlated evolution found in every lineage. Unfortunately, this result does not necessarily support their affirmative answer to (Q2). In this commentary we explain in somewhat more detail the precise nature of the statistical tests conducted by Dunn et al. and why their results do not bear directly on (Q2). To foreshadow the conclusion, Dunn et al. have presented the *strength of evidence for* an effect but interpreted it as the *strength of the effect itself.* The difference between the two often boils down to sample size and sample diversity, as we hope to convey in the following discussion.

The statistical method used by Dunn et al., the *Bayesian hypothesis test*, is very simple at heart: it involves constructing two hypotheses—formally speaking, generative processes that might underlie the observed data—and comparing what likelihood the data would have if each hypothesis were true. To the extent that one hypothesis is more likely to generate the observed data than the other, it is favored by the Bayesian hypothesis test. The metric of *strength of evidence* for one hypothesis over the other is known as the *Bayes Factor*.[1] To give some basic intuitions for the strength of evidence denoted by a given Bayes Factor (BF), we use a simple example: assessing whether a coin is likely to be biased by flipping it repeatedly. In the Bayesian hypothesis test, Hypothesis 1 is that the coin is unbiased—each flip is 50% likely to come up heads—and Hypothesis 2 is that the coin is biased. Because Hypothesis 2 does not specify the *amount* of bias, we need to effectively consider *all* possible biases. To do so, we put a probability distribution on the coin's precise bias under Hypothesis 2: a common practice is to use the uniform distribution (all biases are *a priori* equally likely).[2] If the coin under scrutiny were flipped twice and came up heads both times, few observers would be inclined to conclude that it must be biased. Corresponding to this intuition, the Bayes Factor in favor of Hypothesis 2 is a paltry 0.58. Since this is greater than zero, it gives a mild indication in favor of this hypothesis.[3] But if the coin were flipped *ten* times and always came up heads, most observers would probably place their bets on the coin being weighted—and the Bayes Factor is a correspondingly higher 9.08. Yet in both scenarios the coin always came up heads; the magnitude of an observed trend (100% heads) is not the same thing as the degree of evidence against the trend being a coincidence. If we flip the coin a *hundred* times and obtain *seventy* heads, the evidence for bias is stronger (BF=12.1) than

[1] To be more mathematically precise, Dunn et al. define the Bayes Factor as twice the log of the ratio of the likelihood of the data under the two hypotheses. The Bayes Factor is not always expressed in log space as Dunn et al. have done, though it is common and desirable to do so. We thus stick with their convention in this commentary.

[2] This assumption can easily be changed if we know something about how biased coins are typically created, still within the Bayesian hypothesis testing framework. For the complete mathematical details underlying all examples in this commentary, please see the appendix.

[3] According to Jeffreys (1961), a BF in the range $[0, 2.2]$ is "barely worth mentioning"; in $[2.2, 4.6]$ is "substantial"; in $[4.6, 6.8]$ is "strong"; in $[6.8, 9.2]$ is "very strong"; and greater is "decisive."

in the ten-for-ten scenario, even though the effect itself (the size of the coin's departure from fairness) is weaker. The lineage/feature-pair combination for which Dunn et al. obtained the strongest evidence for correlated evolution is that of adjective-noun and genitive-noun word order in Indo-European; with a Bayes Factor of 21.23 (reported in the paper's online Supplementary Information) the strength of evidence in this case is about as great as that for bias in a coin which came up heads in 75 of 100 tosses, or in 585 of 1000 tosses.

But the results of tests for feature correlation within each lineage do not necessarily bear on the question of whether the feature correlations across lineages are the same. To illustrate how orthogonal these issues can be, we continue with the coin-flipping analogy: now we have two coins, and are interested in whether each is biased; this is analogous to Dunn et al.'s (Q1). We are also interested whether, if they both are biased, their bias is the same; this is analogous to their (Q2). To illustrate how orthogonal these issues can be, consider the following experiment on our two coins, A and B. We flip coin A ten times, and obtain eight heads; we flip coin B fifty times, and obtain forty heads. Both came up heads on 80% of the flips. For each coin, we can compute the Bayes Factor for the hypothesis that the given coin is biased. In this case the evidence in favor of coin A is barely worth mentioning (BF=1.45), but is decisive for coin B (BF=15.35). Thus, the *evidence* for bias is *much stronger* in the case of coin B than coin A.

But regarding the question "do these coins have the *same* bias," your intuition no doubt tells you that it is fairly likely that they have the same bias. In fact, we can set up a *new*, simple Bayesian Hypothesis Test to *directly* test this question (details are in the Appendix). Indeed, the Bayes Factor for this question indicates substantial evidence (BF=2.22) for the same-bias hypothesis test. Importantly, note that for (Q1), when we independently tested each coin for a bias, we reached dramatically different conclusions for the two coins. But to conclude further that the biases for the two coins must be different would be gravely mistaken—there is substantial evidence that they have the same bias! Connecting this back to typology, Dunn et al. could not find a feature pair that had strong evidence of bias in multiple lineages, and concluded that these feature pairs had different biases in the different lineages. As we've just seen, however, such conclusions may be unwarranted.

In the foregoing discussion, it has no doubt become obvious that one source of discrepancy between the strength of an effect and the strength of evidence for it can be sample size (e.g., the number of times a coin is flipped). Sample size may well underlie some of the apparent lineage-specificities reported by Dunn et al.. In the case of AdjN/DemN, for example, Dunn et al. find evidence for correlation in Austronesian and Indo-European, but not in Bantu or Uto-Aztecan. Yet there are far more languages in the samples for the former two families (129 and 79) than in those for the latter two (48 and 26). Furthermore, in tests for correlation such as those by Dunn et al., there is a second source of discrepancy which is absent in the coin-flipping analogy: sample diversity. For AdjN/DemN, there is virtually no variability in DemN for Uto-Aztecan or for AdjN in Bantu, hence there is little opportunity to observe correlation between the two features in either lineage.

Both sample size and sample diversity limit the magnitude that the Bayes Factor for correlated between these traits in these families could possibly reach. In the case of AdjN/GenN, it is only in Indo-European that both traits have much variability. Hence it's no surprise that no conclusive evidence for correlated traits is found in any of the other three families. Note also that none of Dunn et al.'s analyses ever finds strong evidence *against* correlated traits (though to be fair it is typically very difficult in a Bayesian hypothesis test to find strong evidence in favor of a simpler hypothesis without very large sample sizes.)

Hence the bulk of Dunn et al.'s presented analyses do not in fact directly test their second central question (Q2) of whether word-order feature correlations may be lineage-specific. Their methods *do* provide a relatively direct test of their first question (Q1), whether there is evidence for feature correlation in the first place. However, even there we must observe that they did not aggregate evidence across families—e.g., there is weak evidence for correlation between AdjN and ObjV for each of the three families in which both features have meaningful variation, but we do not know how strong this evidence is in aggregate.[4] These are by no means failures of computational statistics in general for problems of language typology, however. Their analysis could be changed straightforwardly to aggregate evidence for correlation across families, simply by attaching the four families together at a global root node extremely (in principle, infinitely) far back in time and applying their model to test for correlated traits on the resulting dataset. And just as we formalized and tested the hypothesis that two coins may have the same bias (differently from testing whether each coin is fair), the hypothesis that the distribution of a given feature pair is governed in all lineages by the same correlated-evolution model (versus lineage-specific correlated-evolution models) could be formalized and tested as well. The more detailed results reported for AdpN/ObjV

[4] The strength of aggregated evidence cannot be measured by simply adding the reported Bayes Factors across the three families as Bayes Factors are not generally additive; in the coin-flipping example, two heads out of two flips is BF=0.58, but raising the count to four heads out of four flips brings us to BF=2.33.

in Austronesian and Indo-European, for example, where the best-fit correlated-evolution models look very different from one another, suggest that the evidence for lineage-specificity is likely to be strong for this feature pair; perhaps it is strong for other feature pairs as well. We hope that Dunn et al.'s paper will stimulate further research that uses the power of modern probabilistic methods to directly address the issue of lineage-specificity of feature correlations, as well many other issues central to the field of linguistic typology.

References

Askey, R. A., Roy, R., 2010. Beta function. In: Olver, F. W. J., Lozier, D. M., Boisvert, R. F. (Eds.), NIST Handbook of Mathematical Functions. Cambridge University Press.

Dunn, M., Greenhill, S. J., Levinson, S. C., Gray, R. D., 2011. Evolved structure of language shows lineage-specific trends in word-order universals. Nature 473, 79–82.

Jeffreys, H., 1961. Theory of Probability. Oxford University Press.

Appendix

In Bayesian statistics, a two-hypothesis test between H_1 and H_2 involves using data D to infer *posterior probabilities* $P(H_1|D)$ and $P(H_2|D)$ for the two hypotheses using Bayes' Rule:[5]

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D)} \tag{1}$$

$$P(H_2|D) = \frac{P(D|H_2)P(H_2)}{P(D)} \tag{2}$$

where $P(D|H_i)$ is the *likelihood* of the data D being generated under hypothesis H_i, and $P(H_i)$ is the *prior probability* of H_i. We can proceed by dividing Equation 1 by Equation 2 giving us the *posterior odds* of H_2 versus H_1:

$$\frac{P(H_2|D)}{P(H_1|D)} = \frac{P(D|H_2)}{P(D|H_1)} \times \frac{P(H_2)}{P(H_1)}$$

[5] This can be generalized to tests among more than two hypotheses, though for simplicity we stick to the case of only two hypotheses here.

The term $\frac{P(D|H_2)}{P(D|H_1)}$ is the contribution of the data D in deciding between H_1 and H_1, and is often called the *Bayes Factor*, or BF. When converted into log space, BF > 0 indicates that the data favor H_2, BF < 0 that the data favor H_1. The larger the Bayes Factor, the more strongly the data speak. BF $= 0$ indicates that the data are truly neutral between the hypotheses. This commentary uses the same convention as Dunn et al. of expressing the Bayes Factor as $2 \log \frac{P(D|H_2)}{P(D|H_1)}$.

For the biased-coin examples given in the main text of this commentary, we need to specify the probability distribution $P(D|H)$ for each hypothesis H_i under consideration, where data D will consist of heads/tails counts. If a coin has bias p, meaning that a single toss has probability p of coming up heads, then the probability of any given sequence with h heads and t tails is $P(D|p) = p^h(1-p)^t$. Hence in the first example, where the hypothesis H_1 is that the coin is unbiased (the special case of $p = \frac{1}{2}$), we have $P(D|H_1) = \left(\frac{1}{2}\right)^{h+t}$.

For the hypothesis H_2 that the coin is not necessarily fair, we do not know what the bias might be. To account for this, we must put a probability distribution $P(p|H_2)$ over the bias p, which allows us to assess $P(D|H_2)$ by averaging over all possible values of p:

$$P(D|H_2) = \int_0^1 P(D|p)P(p|H_2)\,\mathrm{d}p$$

For simplicity, we assume that all biases are equally likely[6] (namely $P(p|H_2) = 1$), and find that:

$$P(D|H_2) = \int_0^1 p^h(1-p)^t\,\mathrm{d}p = \frac{h!\ t!}{(h+t+1)!} \tag{3}$$

As expected, this expression is maximized when either the number of heads h or the number of tails t is zero[7]. For the first example of whether a coin is fair given that it's been flipped twice and came up heads both times ($h = 2, y = 0$), we have $P(D|H_1) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$, $P(D|H_2) = \frac{2!\ 0!}{3!} = \frac{1}{3}$, $\frac{P(D|H_2)}{P(D|H_1)} = \frac{4}{3}$, and thus the Bayes Factor in favor of the coin being weighted is $2 \log \frac{4}{3} = 0.58$.

To apply this methodology to the question of whether two coins have the same bias—analogous to Dunn et al.'s central question of whether two lineages have the same underlying correlation structure between two linguistic features—we take our data D to be the heads/tails counts D_A and D_B for two different coins A and B. For coin A we obtained h_A

[6] This assumption could be changed based on our expectations about how unfairly weighted coins behave.

[7] For any integer $k \geq 1$, the value $k!$ ("k factorial") is $k \times (k-1) \times \ldots \times 2 \times 1$; for the special case of $k = 0$ we have $0! = 1$. The integral in Equation (3) is known as *Euler's Beta Integral* (Askey and Roy, 2010).

heads and t_A tails; for coin B we obtained h_B heads and t_B tails. On the first hypothesis H_{same} of identical bias, the two heads/tails sets were generated with the same bias p, so we have:

$$P(D|H_{\mathrm{same}}) = \frac{(h_A + h_B)!\ (t_A + t_B)!}{(h_A + t_A + h_B + t_B + 1)!}$$

To formalize the second hypothesis H_{diff} of different biases, we once again assume that for each coin all biases are equally likely a-priori; we also assume that the biases of the two coins are generated independently from one another. This gives:

$$P(D|H_{\mathrm{diff}}) = P(D_A|H_{\mathrm{diff}}) \times P(D_B|H_{\mathrm{diff}}) = \frac{h_A!\ t_A!}{(h_A + t_A + 1)!} \times \frac{h_B!\ t_B!}{(h_B + t_B + 1)!}$$

Thus the Bayes Factor for the two-coin example is:

$$\frac{P(D|H_{\mathrm{same}})}{P(D|H_{\mathrm{diff}})} = \frac{(h_A + h_B)!\ (t_A + t_B)!\ (h_A + t_A + 1)!\ (h_B + t_B + 1)!}{h_A!\ t_A!\ h_B!\ t_B!\ (h_A + t_A + h_B + t_B + 1)!}$$