Probability Theory of Lexicostatistics

Michael Cysouw

November 8, 2010

Consider two Languages L_1 and L_2 with a common ancestor $\mathbf{L}$. For the investigation of the changes from $\mathbf{L}$ to these two languages, we take a list of meanings (e.g. the list as proposed by Swadesh 1950, 1952) and collect expressions of these meanings in the two languages ('word lists'). If we find that the two languages have different reflexes for a particular meaning, than something has happened.

A central assumption of lexicostatistics in the sense as proposed by Swadesh is that the probability of stasis (non-change) of the expression of a meaning in a particular time frame t (say, 1000 years) is constant r. This assumption that there actually is a constant is of course far from uncontroversial!

Swadesh proposed that there is the following relation between r ('retention rate'), t ('time') and c ('common vocabulary'), i.e. the fraction of all meanings in which nothing has changed. He then established a value for r by comparing various pairs of language in different point in time, thus using known dates t for various language pairs with known c.

$$\log r = \frac{\log c}{t} \tag{1}$$

Then, if we assume that we know r, then we can estimate the 'divergence time' d from any observed 'common vocabulary' c by

$$d = \frac{\log c}{2 \log r} \tag{2}$$

Note the appearance of a factor 2 here. This factor is necessary, because the divergence time is calculated between L_1 and L_2, which is actually the sum of the divergence times from the common ancestor $\mathbf{L}$ to L_1 and L_2:

$$d(L_1, L_2) = d(\mathbf{L}, L_1) + d(\mathbf{L}, L_2) \tag{3}$$

Assuming the divergence times $d(\mathbf{L}, L_1)$ and $d(\mathbf{L}, L_2)$ are the same (which only makes sense when L_1 and L_2 are both observed at the same point in time), then

$$d(\mathbf{L}, L_1) = d(\mathbf{L}, L_2) = \frac{d(L_1, L_2)}{2} \tag{4}$$

But what is the rationale behind this formula? Swadesh refers to radioactive decay as an inspiration, but why should radioactive decay be a good model for language change? Sankoff (1972) fleshed out the probabilistic details behind this formula.

The first assumption is that, given a meaning m, the probability that the expression of this meaning will change in a time interval t follows a *Poisson Distribution* with a parameter λ defining the possibility of change. Strictly speaking this assumption would imply that the probability of m changing k times in this interval is

$$\frac{e^{-\lambda t}(\lambda t)^k}{k!} \tag{5}$$

Now, assuming that there has been no change in the expression of m during the time frame t, then this reduces to the probability that m remains unchanged in the interval t is

$$\frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t} \tag{6}$$

The second assumption is that the probability of one meaning m_1 changing is independent of a meaning m_2 changing. Given that the set of meaning proposed by Swadesh is rather diverse, this seems to be a relatively unproblematic assumption, though there are still many scenarios possibly in which coupled changes occur.

If we have a set of unrelated events, the combined probability is given by the *Binomial Distribution*. Given a probability p for each individual event, the probability of observing exactly M events in a set of N cases is

$$\binom{N}{M} p^M (1-p)^{N-M} \tag{7}$$

In our case, the probability of an event was described by $e^{-\lambda t}$, so the combined probability of observing M non-changes in a word list of N meanings becomes

$$\binom{N}{M} (e^{-\lambda t})^M (1-e^{-\lambda t})^{N-M} \tag{8}$$

Now, given this binomial distribution, we can establish the *expected value* of the distribution, which is roughly speaking the average outcome of the process

$$\mathrm{E}[M] = \frac{1}{N} \sum_{i=1}^{N} \binom{N}{M} (e^{-\lambda t})^M (1-e^{-\lambda t})^{N-M} \tag{9}$$

which fortunately reduces nicely to

$$\mathrm{E}[M] = N e^{-\lambda t} \tag{10}$$

which can then be divided by N to get close to Swadesh' formula (really, we're almost there!):

$$\mathrm{E}[\frac{M}{N}] = e^{-\lambda t} \tag{11}$$

As you might have noticed, the probability of no change happening to a single meaning m is identical to the expected value of the proportion $\frac{M}{N}$ of words that did not change in a word list, which is an effect of the assumption of *Binomial Distribution*.

If we want to estimate the probability function p (here $e^{-\lambda t}$) in the *Binomial Distribution*, we can use a so-called *Maximum Likelihood Estimator*, which conveniently is the fraction of M and N, so we have again

$$\hat{p} = \frac{M}{N} = e^{-\lambda t} \tag{12}$$

Finally now, we can understand where the Swadesh formula comes from. Given a particular fraction of $c = \frac{M}{N}$ observed common vocabulary, our best guess (when we are pressed to make such a guess) at the amount of time t that might have passed to get to this fraction can be provided by a *Maximum Likelihood Estimator*, so

$$c = \frac{M}{N} = e^{-\lambda t} \tag{13}$$

which can be rewritten as

$$\log c = \log e^{-\lambda t} = -\lambda t \log e = -\lambda t \tag{14}$$

or

$$-\lambda = \frac{\log c}{t} \tag{15}$$

Note that the 'retention rate' r in the original Swadesh formula is a constant, and so the logarithm in $\log r$ is also just a constant. The logarithm is only used in the formula to bring the constant into the same interpretable dimension as c. However, because r is a fraction between 0 and 1, then the logarithm will always be a negative number. In the derivation as given above, λ is always a positive value, and the negative sign is explicitly added. Setting $r = e^{-\lambda}$ gives the original Swadesh formula:

$$\log r = \frac{\log c}{t} \tag{16}$$