The logistic language learning curve?

Hezekiah Akiva Bacovcin & Christopher Ahern

Introduction: The logistic function, which has long served as the basic model of discrete language change (cf. Altmann et al. 1983; Kroch 1989), has both deep connections to change in biological populations as well as a practical implementation in terms of logistic regression. However, no particular cognitive mechanism has been proposed to underlie the logistic (Kroch 1989, 4). Yang's (2002) variational learning model offers a cognitive basis for change insofar as its mean dynamics yield *S*-shaped curves. Indeed, using simulations, we show that given sufficiently large datasets, it is possible to gain insight into whether the learning or logistic model generated an empirical S-shaped curve.

Background: The *variational learning* model of acquisition (Yang, 2002) consists of a finite set of parametrically varying grammars, a probability distribution over these grammars that the learner tracks, and an update rule that governs how learners change the distribution over grammars given input from the environment (cf. Bush and Mosteller (1955)). When presented with variable evidence, individual learners converge in expectation to a set of weights over grammars that reflects the relative evidence for grammars in the linguistic environment. In the case of two grammars, we can calculate the expected change, or mean dynamics, in the distribution over grammars in a population of learners over time as shown on the left where s is the selection coefficient in favor of one grammar over another (Ingason et al., 2013):

$$\dot{p} = p(1-p)\frac{s}{1-s(1-p)} \qquad \dot{p} = p(1-p)s \tag{1}$$

While the dynamics of the learning model on the left and the logistic on the right are distinct, it is not necessarily the case that they will in practice always be distinguishable. We simulate data under a range of parameters to determine when we can distinguish the two kinds of models.

Design: Based on a set of empirical case studies in the history of English (e.g. rise of "do" with "have" in American English, Zimmermann (2015)), we found an average selection coefficient (s) of approximately 0.03 (0.01-0.05), which corresponds to a change on the scale of 2-8 centuries. Using this selection coefficient (0.03), we simulated datasets varying the number of tokens in each simulation (randomly distributed across a 500 year span) with 200 simulations for each size. We selected the number of samples such that the average number of tokens per year ranged from 1 to 1024. For each simulation, we generated two sets of responses: (1) using the learning model and (2) using the logistic model. We then fit both models to each set of responses in each simulation (4 models per simulation) in order to determine if the models could distinguish between the different underlying data-generating mechanisms.

Results: To compare model discrimination at various sample sizes, we calculated the log-likelihood of each of the four models (two models for data generated by the learning model; two models for data generated by the logistic) fit for each simulation. We converted the log-likelihoods into two signals, one for each data-generating mechanism. Each signal recorded which model's log-likelihood was higher. Using these signals, we calculated the KL-divergence statistic for each dataset size. The KL-divergence is given by the following equation, where the state t corresponds to the mechanism that actually generated the data and the signal m corresponds to observing which model had the higher log-likelihood given the data.

$$KL(m) = \sum_{t} P(t \mid m) log \left(\frac{P(t \mid m)}{P(t)} \right) \tag{2}$$

The KL divergence statistics for our simulations are given in Table 1 assuming no prior preference for either data-generating model, where higher scores indicate greater informativity regarding the data-generating mechanism.

Average number of tokens/year	KL (Learning Better)	KL (Logistic Better)
1	0.00	0.01
2	0.00	0.00
4	0.00	0.01
8	0.00	0.00
16	0.01	0.01
32	0.00	0.01
64	0.05	0.08
128	0.05	0.11
256	0.09	0.16
512	0.23	0.24
1024	0.42	0.42

Table 1: Simulation results with increasing sample sizes

Conclusion: We have demonstrated that it is possible to gain information about the underlying processes that lead to historical change. In particular, with plausibly sized datasets that have an average of 500–1000 tokens per year (e.g. COHA, Davies (2010-)), the results of comparing the fit of the learning and logistic models is informative about the underlying model that generates change. In our presentation, we will demonstrate how to fit these models to empirical datasets using R (R Core Team, 2015) and review our simulations and results.

References: Altmann, Gabriel, Haro von Buttlar, Walter Rott, and Udo Strauss. 1983. A law of change in language. In *Historical linguistics*, ed. Brian Brainard, 104–115. Bush, Robert, and Frederick Mosteller. 1955. *Stochastic models for learning*. Wiley. Davies, Mark. 2010-. The Corpus of Historical American English: 400 million words, 1810-2009. Ingason, Anton Karl, Julie Anne Legate, and Charles Yang. 2013. The evolutionary trajectory of the icelandic new passive. *University of Pennsylvania Working Papers in Linguistics* 19:11. Kroch, Anthony S. 1989. Reflexes of grammar in patterns of language change. *Language variation and change* 1:199–244. R Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. Yang, Charles. 2002. *Knowledge and learning in natural language*. Oxford University Press. Zimmermann, Richard. 2015. A syntactic change with lots of data: The rise of do-support with possessive 'have'in american english. University of Manchester Linguistics and English Language (LEL) Research Seminar. 3 November.