Power, linking and (the future of) hypothesis testing in variationist syntax

Over the past three decades, the *Constant Rate Hypothesis* (CRH; Kroch, 1989) has been widely used as a guiding principle in diachronic variationist syntax: according to the CRH, linguistic changes that are underlyingly connected by a single syntactic parameter have identical rates of change on the surface, resulting, in the usual case, in a diachronic pattern consisting of a set of parallel S-curves (a *Constant Rate Effect*, CRE). The empirical observation that several changes in several languages appear to follow such patterns has then been interpreted as support for the view that different surface contexts are not differently adaptive but rather stem from one unified change whose manifestation on the surface may only be modulated by factors that remain constant over time (Pintzuk, 2003). Diachrony recapitulates synchrony, and observations about the former may be used to reason about the properties of the latter.

Until recently, the variationist community had paid lip service to the sheer difficulty of the diachronic task of teasing constant rates of change apart from variable ones. Paolillo (2011) first pointed out that all work on the CRH necessarily casts the CRH as the null hypothesis of the statistical testing scenario, as what the CRH describes is the *absence* of an effect. This, however, has the unfortunate consequence of introducing type II errors into the process at an unknown rate: absence of evidence for an effect of surface context on rate of change is not evidence of the absence of such an effect, and the statistical power of current methods of CRE diagnosis remains unknown. This notwithstanding, work on the CRH continues without proper controls in place to mitigate this *Power Problem*, in syntax as well as other linguistic domains (for two recent examples, see Wallage, 2013 and Gardiner, 2015).

In Part One of this contribution, we compile a database of 50 changes reported in the variationist literature over the past 30 years and ask two related questions: (1) to what extent do these studies suffer from the Power Problem, and (2) what can be done to mitigate the problem on a more general level. Across our database, 8 different methods of CRE detection are used, 4 of which are quantitative in the sense that the decision is based on a statistical test and an objective criterion. To answer questions (1) and (2), we generated a large population of synthetic datasets with varying degrees of divergence from the complete identity of rates of change described by the CRH. The 4 methods were then applied to these synthetic datasets in a Monte Carlo experiment to compute their type II error rates as a function of the size of the effect on the one hand, and sample size on the other.

The results of our Monte Carlo power analysis show that reasonable type II error rates—defined such that the probability of incorrectly reporting a CRE is at most 0.05—are only attained for sample sizes on the order of tens of thousands of tokens, depending on the size of the potential confounding effect (Figure 1). These figures are well beyond the sample sizes typically attainable with historical corpora; indeed, only 2 datasets in our database of 50 have sample sizes in the tens of thousands. The answer to question (1), then, is pessimistic: most studies on the CRH to date lack the statistical power needed to support their conclusions. As an answer to question (2), we extract an equation from the Monte Carlo analysis that may be used to calculate the sample size required for the detection of diverging rates of change at a desired level of statistical power.

Quite orthogonally to the Power Problem, it has recently been argued that the CRE lacks a foundation, as it has never been rigorously derived from first principles (Kauhanen & Walkden, 2018). In other words the following argument, a rational reconstruction of Kroch (1989) and subsequent literature, is a *non sequitur* since the conclusion would not appear to follow from the premises:

> A grammatical parameter π undergoes change from one setting to another.
> Surface contexts $C_1,\ldots,C_n$ are all tied to π.
> Therefore: $C_1,\ldots,C_n$ change along equally sloped S-curves.

Dubbing this the *Non-Linking Problem*, and attempting to overcome it by reinterpreting the CRH within a mechanistic model, Kauhanen and Walkden (2018) equip Yang's (2002) variational learner with contextual biases that remain constant over time and iterate this learner–speaker over multiple successive generations. Surprisingly, the model predicts sets of curves which are not exactly equally sloped, i.e. the contexts change at slightly different rates even though they spring from one and the same underlying parametric change, modulated by a set of biases that remain constant over time, exactly in the spirit of the original CRH. Unfortunately, the quantitative objective methods with which CREs have been diagnosed under the classical interpretation cannot be employed with the bias model (Kauhanen & Walkden, 2018, sec. 4), and the empirical adequacy of the latter thus remains uncharted.

Do the Power and Non-Linking Problems, together with the difficulty of testing the bias model, imply that the CRH ought to be abandoned as a guiding principle in research on syntactic variation and change? It would be hasty to conclude so. In Part Two of this contribution, we show how an objective test for the bias model can be achieved through the use of the *Parametric Bootstrap Cross-fitting Method* (PBCM; Wagenmakers, Ratcliff, Gomez, and Iverson, 2004), a general-purpose procedure for binary classification. Conducting a second Monte Carlo power analysis, we additionally demonstrate how the type I and type II error rates of this model–method combination may be estimated. In general, we find that the bias model is less confusable with a variable rate model than is the original CRH, mitigating the Power Problem to a certain extent (Figure 2).

In Part Three, we ask whether and how a decision between the bias model and the original CRH may be made in an objective manner. Applying the PBCM now to a comparison between the two models of the CRE, we find that the bias model is preferred over the original CRH in 5 out of 6 datasets which are the most likely to contain a CRE on *either* model according to the analyses in Parts One and Two. We interpret this as evidence to the effect that the bias model, which derives a diachronic pattern from a closed loop of inter-generational variational learning and probabilistic production biases, may be better at recovering the original theoretical intuition behind the CRH than is the original CRH itself.

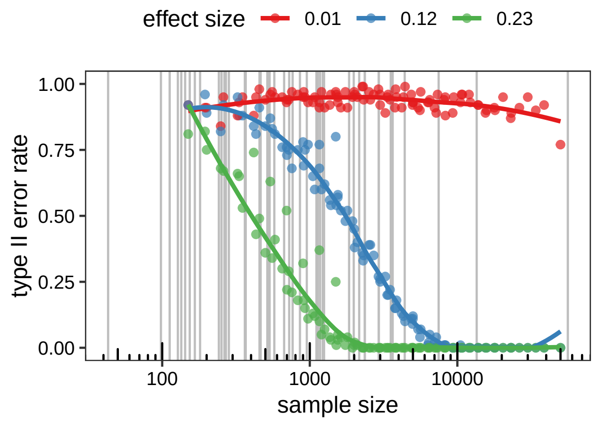

Figure 1. Type II error rate decreases with increasing sample size and increasing size of the confounding variable rate effect. Vertical lines indicate the sample sizes of the 50 datasets in our database. Curves are LOESS.

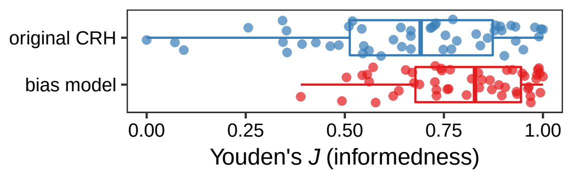

Figure 2. Probability of informed decision (Youden's *J* statistic) on the PBCM, using either the bias model or the original CRH, over the 50 datasets in our database.

Our results have consequences for variationist work trying to inform synchrony through meticulous analyses of diachronic patterns. Firstly, the data problem must be taken seriously; in particular, where the diagnosis of diachronic patterns rests on correctly accepting (rather than rejecting) a null hypothesis, a power analysis should be carried out; the methods here presented provide the means to do so. Secondly, mechanistic models that predict diachronic patterns as mathematical theorems of their first principles ought to be favoured over descriptive models in which the connection between pattern and principles remains tenuous. Thirdly, the analyses here reported testify to the suitability of the PBCM as a general-purpose model selection procedure; we advocate its wider adoption along with other related Monte Carlo methods in diachronic linguistics.

References

Gardiner, S. (2015). Taking possession of the constant rate hypothesis: Variation and change in Ancient Egyptian possessive constructions. *University of Pennsylvania Working Papers in Linguistics*, *21*(2), 69–78.

Kauhanen, H. & Walkden, G. (2018). Deriving the constant rate effect. *Natural Language & Linguistic Theory*, *36*(2), 483–521.

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, *1*(3), 199–244.

Paolillo, J. C. (2011). Independence claims in linguistics. *Language Variation and Change*, *23*, 257–274.

Pintzuk, S. (2003). Variationist approaches to syntactic change. In B. D. Joseph & R. D. Janda (Eds.), *The handbook of historical linguistics* (pp. 509–528). Oxford: Blackwell.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wallage, P. (2013). Functional differentiation and grammatical competition in the English Jespersen Cycle. *Journal of Historical Syntax*, *2*(1), 1–25.

Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.