JHS Journal of Historical Syntax
Volume 7, Article 18: 1–31, 2023

GRAMMAR COMPETITION, SPEAKER MODELS AND RATES OF CHANGE: A CRITICAL REAPPRAISAL OF THE CONSTANT RATE HYPOTHESIS*

Henri Kauhanen
University of Konstanz

Abstract The notion that rates of replacement of an old grammatical option by a new one are identical across linguistic contexts during a period of change (the Constant Rate Hypothesis, CRH) has attracted considerable attention in the historical syntax literature. Here, I argue that any inferences made about change processes using models of the constancy or variability of rates of change must be conducted in a way that balances three considerations: (i) empirical fit, (ii) model complexity and (iii) ontological interpretability of model parameters. Five models involving constant or variable rates of change are examined with respect to three datasets with the help of the Akaike Information Criterion in an effort to explore how model selection can be carried out rigorously. Although this technique balances empirical fit and model complexity in a principled manner, thus offering an improvement over the statistical methods traditionally used in the examination of constancy of rates of change, it cannot weed out models which fail the criterion of ontological interpretability. I argue that such models should

* The work reported in this paper was begun during the author's ESRC Postdoctoral Fellowship at The University of Manchester Department of Linguistics and the Centre for Data Analytics and Society (Economic and Social Research Council grant no. ES/S011382/1); continued at the Zukunftskolleg of the University of Konstanz, financed by the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments; and finalized while the author was employed through funding provided by the European Research Council (ERC grant no. 851423, awarded to George Walkden). The Zukunftskolleg provided computer hardware which greatly facilitated initial prototyping of numerical optimization routines and executing the final analyses reported in the paper. I also wish to extend my gratitude to the audience at the workshop on Syntactic Change in Progress (SCiP) at DiGS 2021 for feedback on an earlier version of this paper read at that meeting; to three JHS reviewers for their critical feedback and numerous fruitful suggestions; to the editors for their guidance and patience; as well as to Ricardo Bermúdez-Otero, Gertjan Postma, George Walkden, Joel Wallenberg and Richard Zimmermann for discussions which have influenced my thinking on the Constant Rate Hypothesis and the nature of evidence and explanation in historical linguistics. It goes without saying that all errors are my own.

> be excluded from consideration on *a priori* grounds. What remains can be termed 'speaker models': mechanistic models whose components have interpretations in terms of the representation of knowledge of language and language use. Whether the CRH is such a model or can be derived from one remains an open question.

1 Motivation

There are arguably very few general, quantitative nomothetic statements in historical syntax; fewer still have been subjected to empirical scrutiny with the help of statistical modelling techniques. The Constant Rate Hypothesis (CRH) is one such statement:

> when one grammatical option replaces another with which it is in competition across a set of linguistic contexts, the rate of replacement, properly measured, is the same in all of them. (Kroch 1989: 200)

In the more than three decades since its introduction, the CRH has been explored, tested, broached, critiqued and reinterpreted in a number of studies (e.g. Santorini 1993, Ball 1994, Pintzuk 1995, Wagner 1996, Frisch 1997, Tagliamonte & Hudson 1999, Cukor-Avila 2002, Pintzuk & Taylor 2006, Sundquist 2006, Kallel 2007, Sundquist 2007, Tagliamonte & D'Arcy 2007, von Heusinger 2008, Wallage 2008, Postma 2010, Paolillo 2011, Durham, Haddican, Zweig, Johnson, Baker, Cockeram, Danks & Tyler 2012, Fruehwald, Gress-Wright & Wallenberg 2013, Wallage 2013, Corley 2014, Ecay 2015, Gardiner 2015, Bacovcin 2017, Postma 2017, Kauhanen & Walkden 2018, Simonenko, Crabbé & Prévost 2019, Wallenberg, Bailes, Cuskley & Ingason 2021, Zimmermann 2022).

The above quotation famously leaves open what it means to *properly* measure a rate of change in language. In practice, some operationalization of change in mathematical terms is necessary if the hypothesis is to be put to empirical test or application. Kroch (1989) recommended characterizing change using the logistic function, which allows the researcher to condense the notion of rate of change into a single number, namely the slope parameter of that function. This yields the following rational reconstruction of the CRH:

(1) If (P) a number of linguistic contexts $1, \ldots, m$ are observed to change,

and (Q) all those m developments occur because of a change in a single, underlying grammatical option,

> then (R) fitting a logistic model to the m contexts yields no evidence that the slope parameter is different from one context to another.

For ease of reference, let us refer to this conditional statement (P & $Q \rightarrow R$) as proposition C.

On the face of it, the logic here allows at least three different kinds of investigations. First, it is possible to assume that the conditional C holds and that the premise P is observationally true. Then, if sufficient quantitative evidence can be adduced to suggest that R is false, it follows that Q must be false, too. In other words, assuming that the CRH is true allows one to show that a set of observed changes are *not* tied to a single underlying grammatical option or parameter.[1] Secondly, if C and P are assumed to hold and sufficient evidence is adduced to suggest that R holds, it is possible to regard Q—although not as strictly true—as receiving a degree of 'corroboration' (Popper 1959) to the extent that repeated attempts at its falsification continue to fail. Finally, if the truth of both P and Q can be taken for granted in some particular instance of change, it is possible to attempt to corroborate or refute the CRH (the statement C) itself, by estimating the truth value of R.

In practice, serious difficulties of measurement and inference arise in any such undertaking. The majority of existing studies on the CRH have focused on relatively old changes, forcing them to rely on corpora in which problems of sample size and sample quality—for instance, uncertainty relating to the dating of texts, or manifest imbalances in terms of register or geographical area—are implicated. In other cases, problems may arise from the fact that the observations are not plausibly independent, as required by the usual statistical treatments. A serious problem of inference arises also from the fact that the commonly used statistical technique for evaluating the truth of proposition R takes (in fact, *has to* take) the CRH as the null hypothesis of the statistical inference situation. However, failure to reject a null is never conclusive evidence of the truth of that null, a problem that casts potentially serious doubt over many existing studies of the CRH (Paolillo 2011).

A further challenge concerns the precise status of the CRH as a theoretical statement. The hypothesis assumes underlying grammatical unity to be reflected as identity of rates of change—that P and Q imply R. Yet it is not entirely clear why this should be so: the conditional C would appear to present no *mechanism* whereby the identity of rates of change could be reduced back to first principles (cf. Kauhanen & Walkden 2018). The following is at least conceivable: perhaps the central intuition behind the CRH—that underlying

1 Strictly speaking, this assumes that no auxiliary assumptions with uncertain truth values are involved, a situation that rarely if ever obtains in a scientific investigation (the Duhem–Quine problem; see Harding 1976).

parametric unity ought to be reflected in *some* regular observable pattern in the diachronic record—is true, but identity of rates of change happens not to be that pattern (see also Kallel 2007).

Reflecting on these issues, in this paper I argue that diachronic linguistics needs to consider not just statistical models but also *speaker models*—mechanistic models of linguistic knowledge representation and language use. I also ask to what extent the CRH can be regarded as such a model. Concretely, I will consider three interpretations of the CRH, Kroch's (1989) original operationalization as a set of logistics, Kallel's (2007) extended model which includes an additional quadratic term to account for constant curvature, and Kauhanen & Walkden's (2018) model which attempts to derive the CRH from probabilistic production biases applied to underlying grammar probabilities learned in first-language acquisition; these models are further joined by two variable-rate models as implementations of the negation of the CRH. The models will be fit to three datasets from the existing literature; to avoid incurring type II errors at unknown rates (Paolillo 2011), an information-theoretic model selection procedure which balances model fit and model complexity (Burnham & Anderson 2002) is adopted instead of the standard null hypothesis testing framework. I will argue that the best model in any given case will tend to optimize three criteria simultaneously: (i) empirical fit, (ii) theoretical parsimony and (iii) ontological interpretability of key model parameters. An improved method for fitting the non-traditional model of Kauhanen & Walkden (2018) emerges as a side product of these investigations.

The models to be studied are introduced in §2; the problem of model selection is discussed in §3. The case studies from which the datasets of the empirical investigations are taken are summarized briefly in §4. Results are presented in §5 and discussed in §6–7. Technicalities are collected in the appendices A–C.

2 Speaker models

The classical formulation of the CRH characterizes linguistic changes using logistic curves: formally, the probability of one of two possible grammatical options involved in the change in context $c = 1, \ldots, m$ satisfies

$$p(c,t) = \frac{\exp(s_c t + k_c)}{1 + \exp(s_c t + k_c)} \tag{2}$$

where t is time, s_c is slope (rate of change) and k_c is the intercept (controlling temporal translation of the curve's tipping point, the value of the abscissa t at which the ordinate equals $p(c,t) = 0.5$). The CRH amounts to the statement that a unique slope s exists such that $s_c = s$ for all contexts c.

This is, furthermore, equivalent to carrying out a logistic regression with time, context and the interaction of time and context as explanatory variables: taking the logit transform of (2) yields

$$\log\left(\frac{p(c,t)}{1-p(c,t)}\right) = s_c t + k_c = st + s'_c t + k_c \tag{3}$$

where s is the rate of change common to all contexts and $s'_c = s_c - s$ gives the strength of the time–context interaction. The CRH amounts to the statement that $s'_c = 0$.

Interpreted as a statistical model, this is simply a statement about the likely values of a handful of parameters of a function used to describe a set of data. If the goal of language science is to explain and predict the behaviour of speakers—*speaker* here used as a convenient shorthand for 'cognitive agent with knowledge of language'—then a further step must be taken: either the parameters of the model must be given interpretations in terms of the ontology of the speaker, or else it must be shown how the statistical model can be obtained (ideally, deductively derived) from a substantive, mechanistic model.

If the first route is taken, then the implication is that speakers in fact implement quantities such as s, s_c and k_c. The CRH and its negation become statements such as (4) and (5), respectively:

(4) The speaker has a single s, applied to each context, and variable k_c

(5) The speaker has variable s_c and k_c

When constructing the probability of use of the grammatical option at time t in context c, in other words $p(c,t)$, the speaker makes use of these quantities, arriving at usage frequencies which, when collected over repeated measurements and across multiple speakers, yield the diachronic patterns which the statistical test taps into.

Alternatively, we may regard the logistic model as a pure statistical model whose various parameters do not necessarily have real-world interpretations, at least not on the level of individual speakers. In the ideal case, it should then be possible to show how the statistical model follows from some underlying model of the speaker. Some such intuition motivated the work in Kauhanen & Walkden (2018), which attempted to derive the CRH from an application of a set of biases to an underlying probability of employment of the competing grammatical options. Thus instead of assuming that speakers implement quantities such as s_c and k_c, this model assumes that the probability of the grammatical option in context c obeys the equation

$$p(c,t) = P(t) + b_c P(t)[1 - P(t)] \tag{6}$$

where $P(t)$ is an underlying, context-independent representation of the probability of the grammatical option, while b_c is a context-dependent but crucially time-independent bias (satisfying $-1 \leq b_c \leq 1$), the probability of either enhancing or suppressing expression of the grammatical option in this particular context. The idea is that speakers first acquire $P(t)$ as part of ordinary first-language acquisition and then 'filter' this knowledge through context-specific knowledge in actual language use with the help of the biases b_c. In Kauhanen & Walkden (2018), the dynamics of such a model were developed in the special case that $P(t)$ is a logistic, for instance a 'grammar weight' along the lines of the variational learning model of Yang (2002), which is known to predict logistic trajectories under fairly lenient assumptions. This yields contextual curves $p(c, t)$ which are similar, but not quite equivalent to a family of equally-sloped logistics.

It is important to highlight the difference between the two approaches, i.e. between viewing the mathematical equations as statistical descriptions and viewing them as mechanistic models. In the latter approach, the equations follow, by mathematical deduction, from a theory which is also expressed in mathematical terms. In Yang's (2002) model of variational learning, for example, the value of $P(t)$, the weight assigned by the speaker to a grammatical option, either increases or decreases over time. However, this is not all the model says, for it also supplies a mechanism whereby that weight either increases or decreases: the linear reward–penalty scheme of Bush & Mosteller (1955) dictates how individual speakers adjust their behaviour, and the population-level evolution follows from this via simple assumptions of intergenerational transmission. In other words, the model is not only a statistical description of observable longitudinal developments but also a mechanistic explanation of how speakers and populations adjust their grammar weights in response to external stimuli. The diachronic trajectories observable in corpora follow deductively from theoretical first principles, offering a particularly transparent way of testing the validity of those first principles.

In the particular case of Kauhanen & Walkden's (2018) model, the conceptual benefits of the move to a mechanistic description include the fact that a number of contexts are now mathematically tied to one underlying probabilistic representation, $P(t)$, and that the bias parameters b_c can be subsumed under the same probabilistic framework, as they are signed probabilities (positive if the context favours, and negative if it disfavours, the grammatical option). A number of challenges arise too, however, the most important of which is that it is no longer easy to fit the model to data. Whereas ordinary binomial regression can be used to fit the original CRH model, no off-the-shelf solution exists for its bias reinterpretation, which is nonlinear in its pa-

rameters but is not subsumed under generalized linear models or any similar family of models. In Kauhanen & Walkden (2018), an iterative least-squares algorithm was used; its use is somewhat unprincipled, however, particularly as the method yields no basis on which different models could be rigorously compared, hence offering no programmatic basis for model selection. In the following section, I will discuss how these problems can be overcome.

The bias model is not the only reinterpretation of the CRH to have appeared in the literature. In a study of the loss of negative concord in English, Kallel (2007) observes that a logistic regression model with an additional quadratic term improves the fit of the model for her dataset:

$$p(c,t) = \frac{\exp(q_c t^2 + s_c t + k_c)}{1 + \exp(q_c t^2 + s_c t + k_c)}. \tag{7}$$

Based on this, she proposes a 'context constancy principle' according to which what matters for underlying grammatical unity is that the different contexts show identical curvature at each time point. This occurs if q and s exist such that $q_c = q$ and $s_c = s$ for all c: in effect, a generalized form of the original CRH. Although the introduction of the quadratic term leads to improved fit in Kallel's (2007) case study, the additional term also introduces more model complexity. This aspect of the generalization is discussed in more detail below.

3 Model selection

A researcher interested in the constancy or variability of rates of change in a particular empirical application faces the following situation. Observed variation exists between two variants of a linguistic variable. A longitudinal dataset has been obtained consisting of tokens each of which evinces use of one or the other variant, across a number of contexts. We wish to model the probability of one of the variants (usually, the historically innovative one) appearing in context c at time t, $p(c,t)$, on the assumption that the speaker constitutes a Bernoulli random variable. As per the discussion in §2, five models are available, summarized in Table 1. Proceeding top-down, the first model represents the classical CRH of Kroch (1989); the second model allows variable slopes and will be called the Variable Rate Hypothesis (VRH) in what follows. The next model is Kauhanen & Walkden's (2018) bias model; this will also be referred to in the following as the Biased Rate Hypothesis (BRH). The remaining two models constitute Kallel's (2007) quadratic extensions of the original CRH and VRH and will be referred to as the qCRH and the qVRH, respectively. Table 1 also supplies K, the overall number of parameters for each model; this will be used as a measure of model complexity.

Model	Definition	Parameters	K
CRH (Kroch 1989)	$p(c,t) = \frac{\exp(st+k_c)}{1+\exp(st+k_c)}$	$s, k_1, \ldots, k_m$	$m+1$
VRH (Kroch 1989)	$p(c,t) = \frac{\exp(s_c t+k_c)}{1+\exp(s_c t+k_c)}$	$s_1, \ldots, s_m, k_1, \ldots, k_m$	$2m$
BRH (Kauhanen & Walkden 2018)	$p(c,t) = P(t) + b_c P(t)[1-P(t)]$ $P(t) = \frac{\exp(st+k)}{1+\exp(st+k)}$	$s, k, b_1, \ldots, b_m$ $(-1 \leq b_c \leq 1)$	$m+2$
qCRH (Kallel 2007)	$p(c,t) = \frac{\exp(qt^2+st+k_c)}{1+\exp(qt^2+st+k_c)}$	$q, s, k_1, \ldots, k_m$	$m+2$
qVRH (Kallel 2007)	$p(c,t) = \frac{\exp(q_c t^2+s_c t+k_c)}{1+\exp(q_c t^2+s_c t+k_c)}$	$q_1, \ldots, q_m, s_1, \ldots, s_m, k_1, \ldots k_m$	$3m$

Table 1 Five models of the probability $p(c,t)$ of expression of a grammatical option in context $c = 1, \ldots, m$ at time t. The rightmost column, K, gives the number of parameters and hence a measure of model complexity for each model.

Any rigorous procedure for choosing between competing models of an empirical phenomenon must strike a balance between over- and underfitting: an overfitting model is a good fit to a particular dataset but fails to generalize to others, while an underfitting model fails to adequately capture underlying structure in the data (Burnham & Anderson 2002). Multiple ways of striking this balance exist; some of the conceptually simplest are based on likelihood maximization and information criteria (for earlier applications of information criteria to the problem of model selection in the specific context of the CRH, see Ecay 2015, Bacovcin 2017, Wallenberg et al. 2021). Briefly, the idea is to calculate the probability of obtaining the observed data assuming a model and a set of parameter values for that model; the likelihood function is a function of model parameters that supplies this probability. Maximizing the likelihood function means finding a combination of parameter values that maximizes that probability, effectively finding the best possible fit of that particular model to the data at hand.

Once a maximum of the likelihood function, $\hat{L}$, has been found, it is possible to compute the Akaike Information Criterion

$$\text{AIC} = 2K - 2\log(\hat{L}) \tag{8}$$

where K stands for the number of parameters of the model. The AIC balances model fit and model complexity and is an estimator, under fairly lenient assumptions, of the Kullback–Leibler divergence between the model and the true data-generating process (Burnham & Anderson 2002). It follows that, in a comparison between two or more candidate models, the one with the lowest AIC ought to be preferred, as that model incurs the least information loss when reality is represented using the model instead of the real data-generating process.

Obtaining the AIC for models such as (q)CRH and (q)VRH is standard practice; statistical packages will routinely provide these calculations. In Appendix A, it is shown how the maximum likelihood estimate can be found for the bias model BRH. This enables direct information-theoretic model comparison between all five candidate models, balancing model fit and model complexity, regardless of the fact that the BRH is not nested in any of the other models. In the remainder of the paper, this procedure will be applied to three datasets from two case studies from the existing literature, introduced in the following section. On the one hand, the goal of this exercise is simply to illustrate how an information-theoretic model selection procedure can be applied to problems of linguistic diachrony. On the other hand, the substantive question of which model is best supported in each empirical case is of theoretical interest, insofar as these models are speaker models in the sense sketched above and thus make ontological claims about the representation

and use of linguistic knowledge. In this, it is crucial to bear in mind that the five models have different complexities in the general case, and that in the model fit/model complexity tradeoff, relatively more complex models will have to fit the data better in order to be selected.

Once AIC values have been computed, model selection can be performed by comparing them across the model set: the notation $\Delta(M)$ will be employed to refer to the Akaike difference

$$\Delta(M) = \mathrm{AIC}(M) - \min_{M' \in \mathcal{M}} \mathrm{AIC}(M') \tag{9}$$

between model M and the model that scores the lowest AIC within the entire model set $\mathcal{M}$. Note that these differences are relative to $\mathcal{M}$: if the composition of the model set changes (if models are added or removed for any reason), so may the Akaike differences, insofar as the identity of the AIC-minimizing model changes.

4 Case studies

4.1 Do-support

As the first case study to apply the above ideas to, I will be looking at one of the datasets in Kroch's (1989) original paper, originally from Ellegård (1953). This concerns the emergence of periphrastic *do* in English, i.e. the replacement of forms such as (10) by forms such as (11), both quoted in Kroch (1989: 216):

(10) *How great and greuous tribulations suffered the Holy Appostyls?*

(11) *Where doth the grene knyght holde hym?*

Kroch (1989) tracked the increase in the frequency of *do*-support in five linguistic contexts: four types of questions, as well as negative declaratives. These all move in concert up to about the year 1575, meaning that no evidence was found for a difference in the slopes between the contexts in a logistic regression. Kroch (1989) analyses the emergence of *do*-support as a reflex of the loss of V-to-I raising of main verbs; ultimately, it is this parametric resetting that explains the development and, under Kroch's (1989) original CRH, the prediction that the rates of change ought to be identical across contexts.[2]

2 Ellegård's (1953) study also contains data on affirmative declaratives and on the positioning of adverbs; both of these phenomena bear some form of relation to V-to-I raising, discussed at length in Kroch (1989) and subsequent studies. I exclude them from consideration here—the affirmative declaratives because they constitute a failed change and hence cannot be modelled with the usual means of a logistic model; the adverbs because the relevant diagnostics, the orders I-Adv-V and Adv-I-V, cannot be distinguished on the surface when the auxiliary (I)

A complication of this case study is the fact that the 'well-behaved' nature of the development is broken around 1575. If Ellegård's (1953) data are taken at face value, then from this point onward not only does the rate of change slow down in each of the contexts, it also varies from context to context. Kroch (1989: 232–237) suggested a theoretical analysis of this perturbation, relating it to a major restructuring of the grammar; Warner (2005), however, argued that the perturbation was due to sociolinguistic and register effects, and that the vernacular is likely to have followed a simple trajectory.[3] On either analysis, Ellegård's (1953) data after 1575 must be analysed separately from those before 1575. I thus follow Kroch (1989) in excluding the data after 1575. The set of data constituted by the pre-1575 tokens will be referred to as dataset *K1* in what follows.

At the same time, the dataset when taken in its entirety constitutes a useful negative test case. Since legitimate reasons (either grammatical or extragrammatical in nature) exist for not regarding the full data as reflecting a single and simple development, we would not expect the full dataset to conform to the predictions of the CRH. In other words, any model of constancy of rates of change should return this dataset as a true negative. I will refer to the full dataset as *K2* in the following.

4.2 Word order: from OV to VO

The second case study comes from Pintzuk & Taylor's (2006) analysis of the change from OV to VO order in Old and Middle English, the replacement of forms such as (12) by forms such as (13), both quoted in Pintzuk & Taylor (2006: 258):

(12) *ȝef ȝe habbeð ani god don*
if you have any good done
'if you have done any good'

(13) *fordon þe he scal aȝein ȝeuen awiht*
for that he shall again give something
'for he shall again give something'

This is one of just a handful of studies to have appeared in the literature on constant rates that uses the CRH as part of a logic of refutation: drawing on historical corpora, Pintzuk & Taylor (2006) produce evidence that the rate

is missing, and hence second-order estimates of underlying orders not deducible from the surface form are required to carry out the analysis (see Kroch 1989: 225–232).

3 For detailed discussion of the complexities involved here, see Ecay (2015: 88–93). For an alternative curve-fitting approach, see Vulanović & Baayen (2007).

	CRH	VRH	BRH	qCRH	qVRH
K1	0.00	7.52	8.29	1.19	7.56
K2	169.39	140.19	184.58	21.82	0.00
PT	113.21	75.07	45.16	97.42	0.00

Table 2 Akaike differences $\Delta(M)$ for five models fit to the datasets K1, K2 and PT. Models with $\Delta(M) < 2$ shaded.

of replacement of OV by VO is *not* identical in three contexts: phrases with positive objects, negative objects and quantified objects. The authors take the variable rates of change as evidence for the position that different syntactic derivations are responsible for observed surface word order in the three contexts; the details of this are discussed in greater length in §6. At the outset, the expectation then is that of the five models examined in the present paper, either the VRH or the qVRH should be favoured for this particular dataset. I refer to these data as dataset *PT* in what follows.

5 Results

Table 2 shows the results of applying the five models introduced in §2 to the three datasets introduced in §4.[4] Conventionally, a model M is thought to have substantial support if its Akaike difference $\Delta(M)$ to the best model is not greater than 2; some support if $2 < \Delta(M) < 10$; and little to no support if $\Delta(M) > 10$ (Burnham & Anderson 2002). Following these (somewhat arbitrary) guidelines, the picture that emerges from Table 2 is clear: the CRH and qCRH models far outperform the others with the *do*-support dataset K1, while the qVRH is by far the best candidate for datasets K2 and PT.

Interpretation of these results is challenging, however, because interpretation of the models themselves is problematic: as was emphasized in §2, it is not clear whether the CRH and VRH models can be taken ontologically as speaker models, and the same problem applies, but compounded by the presence of the additional quadratic term, to the qCRH and qVRH models. The slope coefficient of the logistic function has the relatively unproblematic meaning 'rate of change', and it may even be possible to argue that speakers

4 The raw AIC values, maximum likelihood estimates and estimated model parameters, along with all code necessary to replicate the analyses, are available to download from https://doi.org/10.5281/zenodo.7734357.

	CRH	VRH	BRH	qCRH	qVRH
K1	0.00	7.60	8.29	1.91	12.68
K2	165.52	136.31	180.71	17.95	0.00
PT	113.21	75.07	45.16	97.42	0.00

Table 3 Akaike differences $\Delta(M)$ for five models fit to the datasets K1, K2 and PT, regressed on un-standardized time covariate. Models with $\Delta(M) < 2$ shaded.

have internalized representations of such quantities, as work on momentum-based change and age vectors has proposed (Labov 2001, 2010, Stanford, Severance & Baclawski Jr. 2014, Stadler, Blythe, Smith & Kirby 2016, Bermúdez-Otero 2020, Holmes-Elliott 2021). But it is far less clear what the quadratic terms in the qCRH and qVRH models might mean. It is unclear what the signal they represent may be, how speakers are able to tap into this signal, and how speakers are able to represent these quantities as part of their knowledge of language.

A significant conceptual challenge arises also from the fact that the quadratic t^2 is symmetric about the origin $t = 0$. If the time variable is standardized (z-scored) before the regression, as is customary practice and as also applies to the results in Table 2, both negative and positive times will occur in the covariate. This implies that the effect of the quadratic term will be not just nonlinear but also nonmonotonic: greater at the most negative times, then diminishing and passing through zero at the origin, and again increasing towards its maximum at the most positive times. It is unclear, on conceptual grounds, why this behaviour should be expected: nothing about either linguistic theory or our understanding of population dynamics predicts this kind of effect.[5]

Standardizing the time variable is not compulsory, of course. Table 3 displays the model comparison after regressions conducted on raw, un-standardized time (in the case of the specific datasets here, no negative times occur in the raw times). Interestingly, the broad shape of the results remains unchanged: models CRH and qCRH compete to be the best model of dataset

5 These remarks are not meant to imply that quadratic or other higher-order terms can never be used in regression models, but rather that they are subject to the same epistemological requirements as any other model terms: their inclusion must be justified by *a priori* theoretical reasons rather than for reasons of *ad hoc* improvements in model fit.

	CRH	VRH	BRH
K1	0.00	7.52	8.29
K2	29.20	0.00	44.39
PT	68.05	29.91	0.00

Table 4 Akaike differences $\Delta(M)$ when the model set consists of the three models CRH, VRH and BRH. Models with $\Delta(M) < 2$ shaded.

K1, while model qVRH is clearly the winner in the case of datasets K2 and PT. Thus, the quadratic models perform well whether the effect of the quadratic term is monotonic or not. This is arguably an unwelcome result, as it implies that the model does not make a specific prediction about the expected shape of that effect.

If the quadratic models are left out of the model comparison, the results in Table 4 emerge. Each dataset is favoured by a different model: the classical CRH model represents the best model of the *do*-support data K1, the VRH model is best for the extended *do*-support data K2, and the BRH model is selected for the word order dataset PT. In the following sections, I will attempt to argue that this outcome is the most consistent with our current understanding of the different models as well as with the specifics of these three datasets.

6 Discussion: particular

The above observations raise a number of points for discussion. I shall begin with remarks particular to the word order case study, attempting to explain the unexpected result that the BRH model is favoured over the VRH in this case. Remarks of a more general nature follow in §7.

Kauhanen & Walkden's (2018) BRH was intended as a mechanistic model of Kroch's (1989) original statement: something that derives the CRH, or something like it, from first principles. On the other hand, Pintzuk & Taylor (2006) show that, in the passage from OV to VO order over the course of the history of English, positive, negative and quantified objects attest non-identical rates of change. Why is the BRH model then selected by the information-theoretic model selection procedure over the VRH model in this case (Table 4)?

Although Pintzuk & Taylor (2006) show that the word order case study does not conform to the predictions of the CRH, they do not claim the deriva-

tion of surface word order in one context to be entirely independent of that in another. In fact, the change is ultimately explained by competition between the two options of a head-finality parameter in the base grammar. Paraphrasing Pintzuk & Taylor (2006: 264–265):

(14) a. There is underlying competition between head-final and head-initial VPs in the base component of the grammar. These are called the OV grammar and the VO grammar, respectively.

b. However, the OV grammar has an optional rule that postposes the object, moving it rightward of the verb.

c. Similarly, the VO grammar has an optional rule that preposes the object, moving it leftward of the verb.

d. Finally, different contexts allow postposing and preposing at different (but constant-in-time) rates.

Put together, these statements allow the derivation of the two surface orders (OV and VO) in two different ways—either the order is base-generated (15–16) or it follows from an application of preposing or postposing (17–18):

(15) S Aux $[_{\text{VP}}$ O V$]$

(16) S Aux $[_{\text{VP}}$ V O$]$

(17) S Aux O_i $[_{\text{VP}}$ V $t_i]$

(18) S Aux $[_{\text{VP}}$ t_i V$]$ O_i

Furthermore, since the preposing and postposing rules are assumed to apply in the different contexts (positive, negative and quantified objects) at different rates,[6] differences in the historical trajectories of surface word order in the three contexts are expected to arise.

On reflection, these assumptions fit the BRH model exactly. If $P(t)$ in equation (6) is taken to refer to the underlying probability of head-initial VPs (i.e. base VO grammar), acquired by speakers as part of the ordinary process of first-language acquisition for instance by way of a variational-learning-type mechanism (Yang 2002), then the biases b_c can be thought of as combined

6 Some of these rates may be zero: in fact, Pintzuk & Taylor (2006) argue that negative objects never postpose and that positive objects likely never prepose. This is immaterial to the general argument being made here.

preposing–postposing rates[7] for the three contexts, applied by speakers optionally and probabilistically: each bias parameter b_c provides the amount by which the underlying probability of VO gets either boosted or diminished in context c. If, moreover, these biases are constant over time (see Pintzuk & Taylor 2006: 262–263 for evidence that they are), we have the BRH model.

7 Discussion: general

In this paper, I have studied the CRH using an information-theoretic model selection technique (calculation of the AIC following likelihood maximization), with respect to three datasets. Linguistic considerations suggest that the CRH ought to be selected for one of these datasets only. This prediction was borne out: Kroch's (1989) original CRH turned out to be the best model for Ellegård's (1953) *do*-support data prior to 1575 and excluding the context of affirmative declaratives (dataset K1). When data points after 1575 are included (dataset K2), the CRH is no longer the best model. These conclusions are not surprising—they are perfectly in line with Kroch's (1989) original study. The model selection procedure here adopted has, however, the added benefits of not being vulnerable to earlier criticisms of the CRH as the null hypothesis of a statistical significance test (Paolillo 2011), and of taking both model fit and model complexity into account when competing models are compared.

The word order dataset PT is interesting for a number of reasons. Pintzuk & Taylor (2006) predicted the CRH not to hold for these data, on the hypothesis that word order for phrases containing positive, negative and quantified objects is, or can be, generated using different derivations. At the same time, they assume the underlying change to be explained by competition in a head-finality parameter. Above, it was shown that these assumptions form an instance of Kauhanen & Walkden's (2018) BRH, originally intended as a mechanistic model of the CRH. Empirically, the best model for the PT dataset (out of the five models here considered) is either the BRH or Kallel's (2007) qVRH with variable slopes and quadratic coefficients, depending on whether we believe the latter ought to be included in the candidate set (see below). The CRH model, however, finds little support here. This implies a puzzle: if the BRH was supposed to give a foundation for the CRH, why does it perform so well with dataset PT (where the CRH fares badly) and so badly with dataset K1 (where the CRH fares well)?

7 In the simplest case, each b_c would be a linear combination of a preposing rate α_c and a postposing rate ω_c; proper empirical exploration of this hypothesis is beyond the scope of the present paper.

I have argued that, in the ideal case, the models we use to explain processes and patterns of linguistic diachrony should be speaker models rather than just statistical models. By this I mean that these models need to have characteristics of what are sometimes termed mechanistic models in the philosophical and statistical literature (Lindsey 2001, Craver 2006, Baker, Peña, Jayamohan & Jérusalem 2018): they must describe entities, interactions and processes which explain, by deductive inference, the observed phenomena. In linguistics, such a model would normally have terms referring to the representation, processing and use of knowledge of language. I have argued that the BRH is interpretable in this way: the underlying probability $P(t)$ in equation (6) can be derived independently using a formalism such as variational learning (Yang 2002), while the biases b_c can be interpreted as constant signed probabilities which filter the speaker's underlying knowledge in language use, again derivable independently, at least in principle, in different ways depending on what the linguistic contexts involved in the change are (for details, see Kauhanen & Walkden 2018). It is less clear whether the terms involved in the other models (CRH, VRH, qCRH, qVRH) have similar interpretations. If not, then we ought to attempt to derive these models from some underlying mechanistic model, such as the BRH. The model comparisons here with respect to dataset K1 suggest, however, that the BRH, while offering a good account of the OV to VO change, may not be the best speaker model in every case in which constancy of rates of change is expected.

Much work on the CRH and other hypotheses in language variation and change has traditionally set out with the strategy of formulating two statistical models—one null and one alternative—and proceeded to demonstrate that one or another effect, covered by some term of the alternative model, is or is not real. To the extent that interpretability and parsimony are important goals, this framework is somewhat unsatisfactory. Particularly when only one dataset is being explored (and this is the normal situation in most studies), introduction of new model parameters and hence added model complexity carries a substantial danger of overfitting. Methods such as the information-theoretic model selection procedure here adopted can be used to safeguard against these problems to some extent. They are no panacea, however. Importantly, models with dubious interpretability should not be allowed to enter the set of candidate models in the first place—in the general case, nothing about the statistical methods *per se* can help to solve problems of the identification of candidate models, the interpretation of model parameters, and the threat of data dredging (for extended discussion of the challenges involved, see Burnham & Anderson 2002).

It is important to underscore the philosophical aspects of this way of approaching the explanation of empirical phenomena. Crucially, the use of an information-theoretic model comparison method such as the AIC does not tie the researcher to assuming that the set of candidate models contains the "true" model, or into expecting that application of the method will result in simple decisions about whether one or another model is "significant" in some sense. Rather than trying to reject a particular model cast as the null hypothesis of a statistical test, we are asking which model, out of a set of *theoretically motivated* candidate models, yields the best and most parsimonious description of the information contained in the data. One tangible benefit of this approach, in comparison to null hypothesis testing, is that the resulting measures (here, Akaike differences) quantify the relative merits of the competing models. By contrast, failure to reject the null hypothesis of a CRH in a classical null hypothesis statistical test gives us no idea of how much more confidence we ought to have in the CRH model relative to its competitor (the VRH).

These remarks notwithstanding, it must be acknowledged that measures such as the AIC do have their limitations. One of these concerns the way such measures equate model complexity with the number of model parameters. It seems clear that different parameters in different models may be doing different amounts of work, so that two models, while having identical degrees of freedom, may yet be rather unequally flexible in terms of the families of data patterns they predict. Ultimately, one would wish a measure of model complexity to somehow take the functional form of the model into account (Wagenmakers, Ratcliff, Gomez & Iverson 2004). It will be observed (see again Table 1) that the VRH, BRH and qCRH models, in particular, have identical Akaike complexities (identical numbers of parameters) if the number of contexts involved in the change is $m = 2$. Moreover, the BRH and qCRH models have identical Akaike complexities for any m. This means that teasing apart these models, especially when $m = 2$, will be difficult on a measure such as AIC. Alternative techniques for assessing model complexity and carrying out model selection, for instance by way of bootstrapping and cross-fitting (Wagenmakers et al. 2004) or by way of leave-one-out cross-validation (Vehtari, Gelman & Gabry 2017), are available. These methods incur greater computational challenges than the use of 'classical' information criteria. Nevertheless, their application to historical linguistic datasets ought to be explored in future work.

The case studies explored in this paper were selected on the basis that the historical developments they describe are relatively well-understood. The sort of approach I have sketched here should, naturally, be replicated using other datasets. Historical linguistic data are notoriously quirky and come

with a lot of uncertainties about the dating or geographical origin of texts, for instance. The scale of these problems is still essentially unknown, in the sense that we know little about how the statistical methods used to compare competing models are affected by data deficiencies of this kind. More simulation work is needed to address these questions, as is work extending the capabilities of models such as the BRH, for instance in the direction of inclusion of random effects, which would be useful in modelling certain kinds of variability and complex dependencies often attested in empirical datasets.[8] Another potentially useful way of handling the problem of idiosyncratic datasets would be to carry out a study similar to the one reported in this paper on a much larger database of case studies; this would further help to separate the signal from the noise.

Ultimately, more *conceptual* work is needed to understand why something like the CRH should hold, assuming it does. It has often been remarked (e.g. Postma 2017) that the CRH serves an important role as a linking device between I-language and E-language—that is to say, between mental representations on the level of the individual and language use on the level of the population, as recorded in corpora or other data sources. In fact, as stressed in §1, the CRH is one of only a handful of nomothetic statements in historical linguistics—in other words, a universal proposition that applies to a class of phenomena, rather than a descriptive statement about particular changes or historical contingencies. This is significant, *whether or not the CRH turns out to be true*: the crucial matter is simply that, as a nomothetic statement, the hypothesis makes predictions and thereby allows us to test not only the hypothesis itself but also a number of other, related predictions (for a recent example of this logic of inquiry, see Simonenko et al. 2019).

If the CRH is operationalized as a pure statistical model, as has been the practice in much work to date, then the connection between I-language and E-language necessarily remains somewhat nebulous. On the other hand, if a mechanistic model of the speaker is used to derive a population-level pattern, problems of a wholly different nature appear. Namely, it is at least an empirical possibility that whatever mechanism we suppose to exist in speakers is simply undetectable at population level, due to masking effects arising, firstly, from data deficiencies of the kind discussed above, and secondly, from the stochastic nature of population dynamics itself. One potential way out of this conundrum would be to show (mathematically) that such effects cancel out in the big picture, resulting in nothing but stochastic noise which can be

8 Inclusion of random effects is straightforward, in a technical sense, in the case of the other four models. However, the relevant variables (such as text identifier or author) are often missing from the data. This also applies to the versions of the datasets reported in this paper that were available to me.

teased apart from the underlying signal using statistical methods. But it is far too early to say with any confidence how the complex, interacting dynamics of language change will pan out in this regard. This demonstrates that ample room exists for continued investigation of the I-language–E-language connection, whether that be in the form of the classical CRH or some other characterization of that connection.

References

Bacovcin, Hezekiah Akiva. 2017. Modelling interactions between morphosyntactic changes. In Éric Mathieu & Robert Truswell (eds.), *Micro-change and macro-change in diachronic syntax*, 94–103. Oxford: Oxford University Press. doi:10.1093/oso/9780198747840.003.0007.

Baker, Ruth E., Jose-Maria Peña, Jayaratnam Jayamohan & Antoine Jérusalem. 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters* 14. 20170660. doi:10.1098/rsbl.2017.0660.

Ball, Catherine N. 1994. Relative pronouns in it-clefts: the last seven centuries. *Language Variation and Change* 6(2). 179–200. doi:10.1017/S0954394500001630.

Bermúdez-Otero, Ricardo. 2020. The initiation and incrementation of sound change: community-oriented momentum-sensitive learning. *Glossa* 5(1). 121. doi:10.5334/gjgl.627.

Burnham, Kenneth P. & David R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY: Springer 2nd edn. doi:10.1007/b97636.

Bush, Robert R. & Frederick Mosteller. 1955. *Stochastic models for learning*. New York, NY: Wiley.

Corley, Kerry. 2014. The constant rate hypothesis in syntactic change: empirical fact or "lies, damned lies, and statistics"? Bachelor's dissertation, University of Cambridge.

Craver, Carl F. 2006. When mechanistic models explain. *Synthese* 153. 355–376. doi:10.1007/s11229-006-9097-x.

Cukor-Avila, Patricia. 2002. She say, she go, she be like: verbs of quotation over time in African American Vernacular English. *American Speech* 77(1). 3–31. https://muse.jhu.edu/article/2842.

Durham, Mercedes, Bill Haddican, Eytan Zweig, Daniel Ezra Johnson, Zipporah Baker, David Cockeram, Esther Danks & Louise Tyler. 2012. Constant linguistic effects in the diffusion of be like. *Journal of English Linguistics* 40(4). 316–337. doi:10.1177/0075424211431266.

Ecay, Aaron W. 2015. *A multi-step analysis of the evolution of English do-support*:

University of Pennsylvania dissertation. http://repository.upenn.edu/edissertations/1049/.

Ellegård, Alvar. 1953. *The auxiliary do: the establishment and regulation of its use in English*. Stockholm: Almqvist & Wiksell.

Frisch, Stefan. 1997. The change in negation in middle English: a NEGP licensing account. *Lingua* 101(1–2). 21–64. doi:10.1016/S0024-3841(96)00018-6.

Fruehwald, Josef, Jonathan Gress-Wright & Joel C. Wallenberg. 2013. Phonological change: the constant rate effect. In *Proceedings of NELS 40*, Amherst, MA: GLSA.

Gardiner, Shayna. 2015. Taking possession of the constant rate hypothesis: variation and change in Ancient Egyptian possessive constructions. *University of Pennsylvania Working Papers in Linguistics* 21. 69–78. https://repository.upenn.edu/pwpl/vol21/iss2/9.

Harding, Sandra. 1976. Introduction. In Sandra Harding (ed.), *Can theories be refuted? Essays on the Duhem–Quine thesis*, ix–xxi. Dordrecht: Reidel.

von Heusinger, Klaus. 2008. Verbal semantics and the diachronic development of DOM in Spanish. *Probus* 20(1). 1–31. doi:10.1515/PROBUS.2008.001.

Holmes-Elliott, Sophie. 2021. Calibrate to innovate: community age vectors and the real time incrementation of language change. *Language in Society* 50(3). 441–474. doi:10.1017/S0047404520000834.

Hosmer, David W., Jr., Stanley Lemeshow & Rodney X. Sturdivant. 2013. *Applied logistic regression*. Hoboken, NJ: Wiley 3rd edn. doi:10.1002/9781118548387.

Kallel, Amel. 2007. The loss of negative concord in Standard English: internal factors. *Language Variation and Change* 19(1). 27–49. doi:10.1017/S0954394507070019.

Kauhanen, Henri & George Walkden. 2018. Deriving the Constant Rate Effect. *Natural Language & Linguistic Theory* 36(2). 483–521. doi:10.1007/s11049-017-9380-1.

Kroch, Anthony S. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1(3). 199–244. doi:10.1017/S0954394500000168.

Labov, William. 2001. *Principles of linguistic change. Volume 2: Social factors*. Malden, MA: Blackwell.

Labov, William. 2010. *Principles of linguistic change. Volume 3: Cognitive and cultural factors*. Chichester: Wiley-Blackwell.

Lindsey, J. K. 2001. *Nonlinear models in medical statistics*. Oxford: Oxford University Press.

Paolillo, John C. 2011. Independence claims in linguistics. *Language Variation*

and Change 23. 257–274. doi:10.1017/S0954394511000081.

Pintzuk, Susan. 1995. Variation and change in Old English clause structure. *Language Variation and Change* 7(2). 229–260. doi:10.1017/S0954394500001009.

Pintzuk, Susan & Ann Taylor. 2006. The loss of OV order in the history of English. In Ans van Kemenade & Bettelou Los (eds.), *The handbook of the history of English*, 249–278. Malden, MA: Blackwell. doi:10.1002/9780470757048.ch11.

Popper, Karl. 1959. *The logic of scientific discovery*. London: Hutchinson.

Postma, Gertjan. 2010. The impact of failed changes. In Anne Breitbarth, Christopher Lucas, Sheila Watts & David Willis (eds.), *Continuity and change in grammar*, 269–302. Amsterdam: Benjamins. doi:10.1075/la.159.13pos.

Postma, Gertjan. 2017. Modelling transient states in language change. In Éric Mathieu & Robert Truswell (eds.), *Micro-change and macro-change in diachronic syntax*, 75–93. Oxford: Oxford University Press. doi:10.1093/oso/9780198747840.003.0006.

R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Santorini, Beatrice. 1993. The rate of phrase structure change in the history of Yiddish. *Language Variation and Change* 5(3). 257–283. doi:10.1017/S0954394500001502.

Simonenko, Alexandra, Benoit Crabbé & Sophie Prévost. 2019. Agreement syncretization and the loss of null subjects: quantificational models for Medieval French. *Language Variation and Change* 31(3). 275–301. doi:10.1017/S0954394519000188.

Stadler, Kevin, Richard A. Blythe, Kenny Smith & Simon Kirby. 2016. Momentum in language change: a model of self-actuating s-shaped curves. *Language Dynamics and Change* 6. 171–198. doi:10.1163/22105832-00602005.

Stanford, James N., Nathan A. Severance & Kenneth P. Baclawski Jr. 2014. Multiple vectors of unidirectional dialect change in eastern New England. *Language Variation and Change* 26(1). 103–140. doi:10.1017/S0954394513000227.

Sundquist, John D. 2006. Syntactic variation in the history of Norwegian and the decline of XV word order. *Diachronica* 23(1). 105–141. doi:10.1075/dia.23.1.06sun.

Sundquist, John D. 2007. Variable use of negation in Middle Low German. In Joseph C. Salmons & Shannon Dubenion-Smith (eds.), *Historical linguistics 2005*, 149–166. Amsterdam: John Benjamins. doi:10.1075/cilt.284.12sun.

Tagliamonte, Sali A. & Alexandra D'Arcy. 2007. Frequency and variation in the community grammar: tracking a new change through the generations. *Language Variation and Change* 19(2). 199–217. doi:10.1017/S095439450707007X.

Tagliamonte, Sali A. & Rachel Hudson. 1999. Be like et al. beyond America: the quotative system in British and Canadian youth. *Journal of Sociolinguistics* 3(2). 147–172. doi:10.1111/1467-9481.00070.

Varadhan, Ravi, Hans W. Borchers & Vincent Bechard. 2020. *dfoptim: Derivative-free optimization*. https://CRAN.R-project.org/package=dfoptim. R package version 2020.10-1.

Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27. 1413–1432. doi:10.1007/s11222-016-9696-4.

Vulanović, Relja & Harald Baayen. 2007. Fitting the development of periphrastic do in all sentence types. In Peter Grzybek & Reinhard Köhler (eds.), *Exact methods in the study of language and text*, 679–688. Berlin: Mouton de Gruyter. doi:10.1515/9783110894219.

Wagenmakers, Eric-Jan, Roger Ratcliff, Pablo Gomez & Geoffrey J. Iverson. 2004. Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology* 48. 28–50. doi:10.1016/j.jmp.2003.11.004.

Wagner, Laura. 1996. The transition from haver to ter in Portuguese. *University of Pennsylvania Working Papers in Linguistics* 3(2). 133–145. https://repository.upenn.edu/pwpl/vol3/iss2/11.

Wallage, Phillip. 2008. Jespersen's Cycle in Middle English: parametric variation and grammatical competition. *Lingua* 118(5). 643–674. doi:10.1016/j.lingua.2007.09.001.

Wallage, Phillip. 2013. Functional differentiation and grammatical competition in the English Jespersen Cycle. *Journal of Historical Syntax* 2(1). 1–25. doi:10.18148/hs/2013.v2i1.4.

Wallenberg, Joel C., Rachael Bailes, Christine Cuskley & Anton Karl Ingason. 2021. Smooth signals and syntactic change. *Languages* 6. 60. doi:10.3390/languages6020060.

Warner, Anthony. 2005. Why DO dove: evidence for register variation in Early Modern English negatives. *Language Variation and Change* 17. 257–280. doi:10.1017/S0954394505050106.

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

Zimmermann, Richard. 2022. An improved test of the constant rate hypothesis: late Modern American English possessive have. *Corpus Linguistics and Linguistic Theory* doi:https://doi.org/10.1515/cllt-2021-0038.

Henri Kauhanen
Department of Linguistics
University of Konstanz
Universitätsstraße 10
78464 Konstanz
Germany
henri.kauhanen@uni-konstanz.de

A Likelihood functions

Models are fit using the method of maximum likelihood, i.e. by finding (an estimate of) the vector of parameters that maximizes the probability of the observed data, given the model. This requires knowledge of the (log-)likelihood functions and their gradients (first partial derivatives).[9] In this appendix, I show how these can be computed for the five models considered in this paper. It is not possible to find the maxima analytically, as it is not possible to solve the roots of the relevant gradient equations analytically; the maxima must therefore be approximated numerically by making use of the form of the likelihood function and its gradient, starting from an initial parameter guess. The algorithmic details of this optimization are outlined in Appendix B.

A.1 General remarks

Each of the five models considered here assumes that nature generates data following a Bernoulli process with success probability p, the exact form of this probability function depending on the model. Accordingly, the likelihood of the data, given a model, is the following function of the model's parameter vector $\vec{\theta}$:

$$L(\vec{\theta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \tag{19}$$

where n is sample size, each data point is of the form (t_i, c_i, y_i), where t_i is time, c_i indexes context (from 1 to m, the total number of contexts), and y_i is

9 Although the gradients are not particularly difficult to derive for the models here considered, this task may be laborious or impossible in the case of other models. In such cases, the same general model-fitting framework can still be used, as long as a derivative-free optimization routine is used instead. In R, the *dfoptim* package (Varadhan, Borchers & Bechard 2020) provides a number of such routines.

the value of the dependent variable (1 or 0). The form of p_i depends on the model as follows:[10]

$$p_i = \frac{1}{1 + \exp(-(st_i + k_{c_i}))} \quad \text{(CRH)} \tag{20}$$

$$p_i = \frac{1}{1 + \exp(-(s_{c_i} t_i + k_{c_i}))} \quad \text{(VRH)} \tag{21}$$

$$p_i = \frac{1}{1 + \exp(-(qt_i^2 + st_i + k_{c_i}))} \quad \text{(qCRH)} \tag{22}$$

$$p_i = \frac{1}{1 + \exp(-(q_{c_i} t_i^2 + s_{c_i} t_i + k_{c_i}))} \quad \text{(qVRH)} \tag{23}$$

$$\begin{cases} p_i = P_i + \sigma_i P_i (1 - P_i) \\ P_i = \dfrac{1}{1 + \exp(-(st_i + k))} \\ \sigma_i = -1 + \dfrac{2}{1 + \exp(-b_{c_i})} \end{cases} \quad \text{(BRH)} \tag{24}$$

Since in the case of the BRH the bias parameters have the strict lower and upper bounds of -1 and 1 (Kauhanen & Walkden 2018), the sigmoid σ_i is here employed to map unbounded reals b_{c_i} to the interval $(-1, 1)$; this technicality is adopted so that the parameters can be optimized without specifying bounds for the optimization.

The parameter vector $\vec{\theta}$ is

i. $\vec{\theta} = (\theta_1, \ldots, \theta_{m+1}) = (s, k_1, \ldots, k_m)$ (CRH)

ii. $\vec{\theta} = (\theta_1, \ldots, \theta_{2m}) = (s_1, \ldots, s_m, k_1, \ldots, k_m)$ (VRH)

iii. $\vec{\theta} = (\theta_1, \ldots, \theta_{m+2}) = (q, s, k_1, \ldots, k_m)$ (qCRH)

iv. $\vec{\theta} = (\theta_1, \ldots, \theta_{3m}) = (q_1, \ldots, q_m, s_1, \ldots, s_m, k_1, \ldots, k_m)$ (qVRH)

v. $\vec{\theta} = (\theta_1, \ldots, \theta_{m+2}) = (s, k, b_1, \ldots, b_m)$ (BRH)

10 This p_i is of course a function of the parameters $\vec{\theta}$; I suppress the argument for reasons of indolence and legibility. The same remark applies to the functions P_i and σ_i in the case of the BRH.

where m denotes the total number of contexts.

As usual, it is easier (and more numerically stable) to work with the (natural) logarithm of the likelihood:

$$\begin{aligned}(25)\qquad \ell(\vec{\theta}) := \log L(\vec{\theta}) &= \log \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \sum_{i=1}^{n} \left[\log p_i^{y_i} + \log(1-p_i)^{1-y_i}\right] \\ &= \sum_{i=1}^{n} \left[y_i \log p_i + (1-y_i)\log(1-p_i)\right].\end{aligned}$$

For the gradient of the log-likelihood function,

$$(26)\qquad \nabla\ell(\vec{\theta}) = \left(\frac{\partial}{\partial\theta_1}\ell(\vec{\theta}), \dots, \frac{\partial}{\partial\theta_K}\ell(\vec{\theta})\right),$$

where K is the length of $\vec{\theta}$, we then have

$$\begin{aligned}(27)\qquad \frac{\partial}{\partial\theta_j}\ell(\vec{\theta}) &= \sum_{i=1}^{n}\left(y_i\frac{\partial}{\partial\theta_j}\log p_i + (1-y_i)\frac{\partial}{\partial\theta_j}\log(1-p_i)\right) \\ &= \sum_{i=1}^{n}\left(\frac{y_i}{p_i}\frac{\partial}{\partial\theta_j}p_i - \frac{1-y_i}{1-p_i}\frac{\partial}{\partial\theta_j}p_i\right) \\ &= \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\right)\frac{\partial}{\partial\theta_j}p_i,\end{aligned}$$

so computing the gradient in each case reduces to the problem of evaluating $\partial p_i/\partial\theta_j$.

In the calculations to follow, it will be useful to have a concise way of referring to contexts *other* than the one that applies to a given data point i. Recall that c_i will be used to refer to the context that does apply to the ith data point; I will then employ the notation c_{-i} to refer to any other context. Thus suppose there are $m = 3$ contexts in total, for example, and that $c_i = 1$. Then the symbols $k_{c_{-i}}$ and $s_{c_{-i}}$ refer to the intercept and slope of either the second or third context. Similarly, $b_{c_{-i}}$ would refer to the bias parameter in either the second or the third context in the case of the BRH.

A.2 (q)CRH and (q)VRH

The gradient of the simple logistic regression model is of course well-known (see e.g. Hosmer, Lemeshow & Sturdivant 2013). For the sake of completeness, I will nevertheless complete the derivation for the qVRH here. The results for the qCRH, CRH and VRH models follow as special cases.

Write $x_i = q_{c_i} t_i^2 + s_{c_i} t_i + k_{c_i}$ for brevity. Applying well-known properties of the differentiation of logarithms and exponentials, as well as the quotient rule, we have

$$\begin{aligned}\frac{\partial}{\partial\theta_j} p_i &= \frac{\partial}{\partial\theta_j} \frac{1}{1+\exp(-x_i)} \\ &= -\frac{1}{(1+\exp(-x_i))^2} \cdot \frac{\partial}{\partial\theta_j}[1+\exp(-x_i)] \\ &= -p_i^2 \cdot \frac{\partial}{\partial\theta_j}[1+\exp(-x_i)] \\ &= p_i^2 \exp(-x_i) \frac{\partial}{\partial\theta_j} x_i,\end{aligned} \tag{28}$$

i.e.

$$\begin{aligned}\frac{\partial}{\partial\theta_j} p_i &= p_i \cdot \frac{1}{1+\exp(-x_i)} \exp(-x_i) \frac{\partial}{\partial\theta_j} x_i \\ &= p_i \cdot \frac{\exp(-x_i)}{1+\exp(-x_i)} \cdot \frac{\partial}{\partial\theta_j} x_i \\ &= p_i \cdot \frac{1}{1+\exp(x_i)} \cdot \frac{\partial}{\partial\theta_j} x_i \\ &= p_i(1-p_i) \frac{\partial}{\partial\theta_j} x_i.\end{aligned} \tag{29}$$

Thus

$$\begin{aligned}\frac{\partial}{\partial\theta_j} \ell(\vec{\theta}) &= \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) \frac{\partial}{\partial\theta_j} p_i \\ &= \sum_{i=1}^{n} \frac{y_i(1-p_i) - (1-y_i)p_i}{p_i(1-p_i)} p_i(1-p_i) \frac{\partial}{\partial\theta_j} x_i \\ &= \sum_{i=1}^{n} [y_i(1-p_i) - (1-y_i)p_i] \frac{\partial}{\partial\theta_j} x_i\end{aligned} \tag{30}$$

with

$$\frac{\partial}{\partial\theta_j} x_i = \frac{\partial}{\partial\theta_j} (q_{c_i} t_i^2 + s_{c_i} t_i + k_{c_i}) = \begin{cases} t_i^2 & \text{if } \theta_j = q_{c_i}, \\ t_i & \text{if } \theta_j = s_{c_i}, \\ 1 & \text{if } \theta_j = k_{c_i}, \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

A.3 BRH

Let $\mathcal{P}_i = P_i(1 - P_i)$ and $x_i = st_i + k$. For the BRH we now have

$$\frac{\partial}{\partial \theta_j} p_i = \frac{\partial}{\partial \theta_j} P_i + \frac{\partial}{\partial \theta_j}[\sigma_i \mathcal{P}_i] = \frac{\partial}{\partial \theta_j} P_i + \sigma_i \frac{\partial}{\partial \theta_j} \mathcal{P}_i + \mathcal{P}_i \frac{\partial}{\partial \theta_j} \sigma_i \tag{32}$$

using the product rule. Now clearly

$$\frac{\partial}{\partial b_{c_i}} p_i = \mathcal{P}_i \frac{\partial}{\partial b_{c_i}} \sigma_i \quad \text{and} \quad \frac{\partial}{\partial b_{c_{-i}}} p_i = 0. \tag{33}$$

On the other hand,

$$\begin{aligned} \frac{\partial}{\partial b_{c_i}} \sigma_i &= \frac{\partial}{\partial b_{c_i}} \left[-1 + \frac{2}{1 + \exp(-b_{c_i})} \right] \\ &= 2 \frac{\partial}{\partial b_{c_i}} \frac{1}{1 + \exp(-b_{c_i})} \\ &= \frac{-2}{(1 + \exp(-b_{c_i}))^2} \cdot \frac{\partial}{\partial b_{c_i}} [1 + \exp(-b_{c_i})] \\ &= \frac{2 \exp(-b_{c_i})}{(1 + \exp(-b_{c_i}))^2}. \end{aligned} \tag{34}$$

Thus

$$\frac{\partial}{\partial b_{c_i}} p_i = \mathcal{P}_i \frac{2 \exp(-b_{c_i})}{(1 + \exp(-b_{c_i}))^2}. \tag{35}$$

Then let $\theta_j \in \{s, k\}$. Now

$$\frac{\partial}{\partial \theta_j} p_i = \frac{\partial}{\partial \theta_j} P_i + \sigma_i \frac{\partial}{\partial \theta_j} \mathcal{P}_i. \tag{36}$$

The first term was already computed in §A.2, so let us focus on the second. Since $\mathcal{P}_i = P_i(1 - P_i)$, we can again make use of the product rule:

$$\frac{\partial}{\partial \theta_j} \mathcal{P}_i = P_i \frac{\partial}{\partial \theta_j} (1 - P_i) + (1 - P_i) \frac{\partial}{\partial \theta_j} P_i. \tag{37}$$

Now

$$\frac{\partial}{\partial \theta_j} (1 - P_i) = -\frac{\partial}{\partial \theta_j} P_i, \tag{38}$$

so that

$$\frac{\partial}{\partial \theta_j} \mathcal{P}_i = -P_i \frac{\partial}{\partial \theta_j} P_i + (1 - P_i) \frac{\partial}{\partial \theta_j} P_i = (1 - 2P_i) \frac{\partial}{\partial \theta_j} P_i. \tag{39}$$

All in all,

$$\frac{\partial}{\partial\theta_j}p_i = \frac{\partial}{\partial\theta_j}P_i + \sigma_i(1-2P_i)\frac{\partial}{\partial\theta_j}P_i = [1+\sigma_i(1-2P_i)]\frac{\partial}{\partial\theta_j}P_i. \tag{40}$$

From §A.2,

$$\frac{\partial}{\partial\theta_j}P_i = P_i(1-P_i)\frac{\partial}{\partial\theta_j}x_i = \mathcal{D}_i\frac{\partial}{\partial\theta_j}x_i. \tag{41}$$

Hence

$$\frac{\partial}{\partial\theta_j}p_i = \mathcal{D}_i[1+\sigma_i(1-2P_i)]\frac{\partial}{\partial\theta_j}x_i. \tag{42}$$

Finally,

$$\frac{\partial}{\partial s}x_i = \frac{\partial}{\partial s}(st_i+k) = t_i \quad \text{and} \quad \frac{\partial}{\partial k}x_i = \frac{\partial}{\partial k}(st_i+k) = 1. \tag{43}$$

Therefore

$$\begin{cases} \dfrac{\partial}{\partial s}p_i = \mathcal{D}_i[1+\sigma_i(1-2P_i)]t_i \\ \dfrac{\partial}{\partial k}p_i = \mathcal{D}_i[1+\sigma_i(1-2P_i)]. \end{cases} \tag{44}$$

Putting everything together, we arrive at the gradient of the log-likelihood function for the BRH:

$$\frac{\partial}{\partial\theta_j}\ell(\vec{\theta}) = \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\right)P_i(1-P_i)B_i \tag{45}$$

with

$$B_i = \begin{cases} [1+\sigma_i(1-2P_i)]t_i & \text{if } \theta_j = s, \\ 1+\sigma_i(1-2P_i) & \text{if } \theta_j = k, \\ 2\exp(-b_{c_i})/(1+\exp(-b_{c_i}))^2 & \text{if } \theta_j = b_{c_i}, \\ 0 & \text{if } \theta_j = b_{c_{-i}}. \end{cases} \tag{46}$$

B Optimization

Once the likelihood function and its gradient are known, maxima can be found using a numerical optimization procedure. Here, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton method, implemented in the *optim* function of the R statistical computing environment, version 4.0.4 (R Core

Team 2021), was used. Initial values for the slope and intercept parameters were obtained for each model by first fitting a simple logistic using R's *glm* function. Fitting was repeated 10 times for each model–dataset combination in an effort to mitigate the potential problem of convergence to local optima, adding a small amount of noise to the initial values on each run. The initial values for the bias parameters in the BRH and the quadratic terms in the qCRH and qVRH models, however, were set to 0 as this turned out to best guarantee convergence of the optimization. The run that returned the maximal maximum likelihood estimate was retained for analysis. All optimizations reached numerical convergence.

For consistency, BFGS was used for each of the five models, but the CRH, qCRH, VRH and qVRH models were additionally also fit using R's default implementation for generalized linear models (*glm*) as a sanity check. The absolute value of the difference between the BFGS maximum likelihood estimate and the maximum likelihood estimate from *glm* was less than 10^{-6} in each case, i.e. identical up to the sixth decimal.[11] While this says nothing about whether BFGS tends to converge to global rather than merely local optima of the likelihood function in case of the BRH model, the high correlation does imply that use of this algorithm does not lead to less optimal estimates in the case of the models which can also be fit using generalized linear methods. Since the shape of the likelihood surface of the BRH model is essentially unknown, this is the best we can do at the present moment without launching into a lengthy technical treatise; future work should explore the likelihood surface in more mathematical detail.

All code required to replicate the analyses can be obtained from https://doi.org/10.5281/zenodo.7734357. Under the hood, the code calls routines provided by the *cre2* R package, version 0.1.1, which implements the above log-likelihood and gradient calculations. Development versions of this package are available at https://github.com/hkauhanen/cre2.

C Notes on data sources

Datasets K1 and K2 were taken from Table 3 in Kroch (1989), with percentages transformed into absolute frequencies (categorical responses), rounding when necessary. K1 contains only the subset of data up to 1575. The time

11 Regressing the log-likelihood values pooled across all models and datasets against each other using an ordinary linear model, the resulting slope coefficient estimate is $\beta = 1$, with a standard error of 2.3×10^{-11}, $t = 4.4 \times 10^{10}$, $p < 2.0 \times 10^{-16}$. In other words, the maximum likelihood estimates are identical for all intents and purposes, with the remaining difference probably explained by different convergence tolerances implemented in the optimization algorithms.

variable corresponds to the midpoint (arithmetic mean) of the two end points of each time period given in the source, z-score-standardized (except for the results in Table 3, which employed raw time midpoints).

Dataset PT was taken from Table 11.9 of Pintzuk & Taylor (2006). This study uses the Helsinki Corpus periodization for texts; since the earliest texts (those labelled 'OE1' in Pintzuk & Taylor 2006) cannot be dated with any accuracy, I have left them out of consideration here. Five periods remain for the regressions, corresponding to date ranges 950–1150, 1150–1250, 1250–1350, 1350–1420 and 1420–1500; these correspond to original rather than manuscript dates in the Helsinki Corpus. Again, date range midpoints were used for the time variable in the regressions, z-scored apart from the regressions reported in Table 3.