A production bias model of the Constant Rate Effect

Henri Kauhanen & George Walkden

The University of Manchester

NWASV2
Ghent, May 19–20, 2016

Context

- How can we make inferences about I-language on the basis of corpora?
- And can (diachronic) corpora tell us things that we wouldn't otherwise be able to discover?
- We need a **linking theory** that explains how grammatical knowledge is reflected in production data.
- The Constant Rate Hypothesis (Kroch 1989) is a seminal – but partial – theory of this kind.

Introduction

Aims of this talk:

- Outline some problems with the Constant Rate Hypothesis (CRH) as standardly described
- Provide a mechanism for Constant Rate Effects using a modified version of the learner in Yang (2000)
- Show that this model fits the data as well as or better than the traditional approach
- Illustrate an interesting consequence: when biases are in play, the grammar which can parse the most sentence types does not automatically win!

Part I

The Constant Rate Hypothesis and Constant Rate Effects

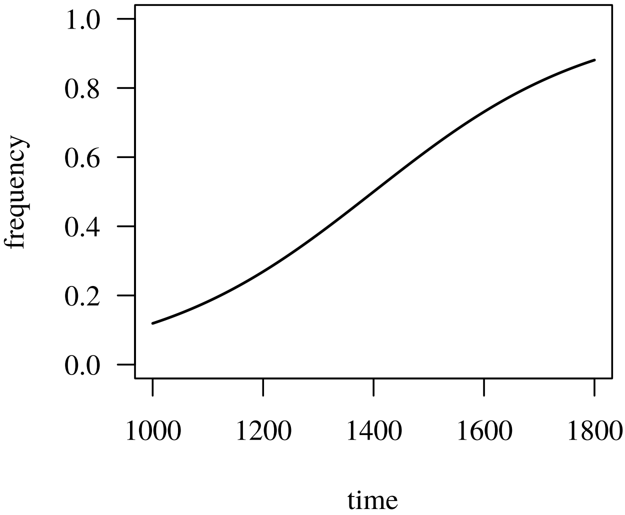
S-curves and the logistic equation

- Altmann et al. (1983), Kroch (1989): change is governed by the logistic equation.

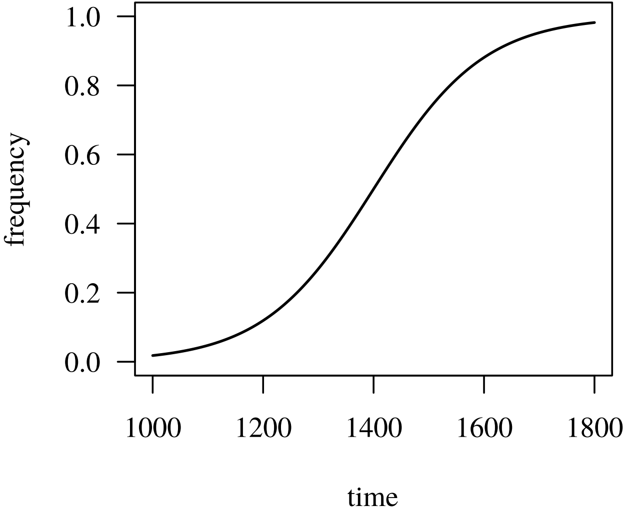
$$p(t) = \frac{1}{1 + e^{-s(t-k)}} \tag{1}$$

- The s parameter gives the overall steepness of the curve.
- The k parameter tells us where, in time, the curve is situated.

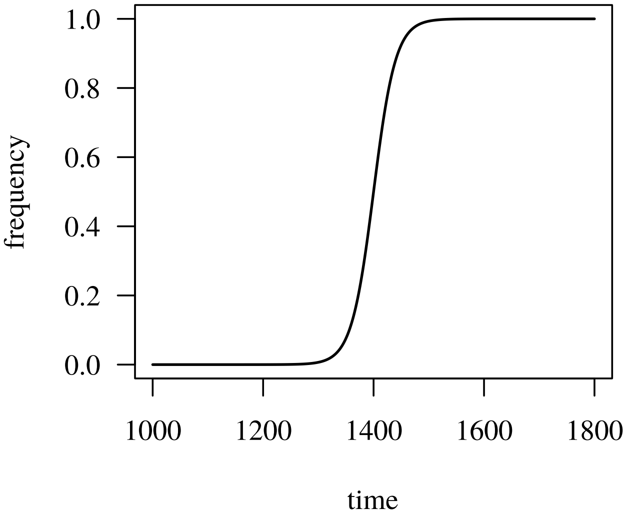
Varying the s parameter



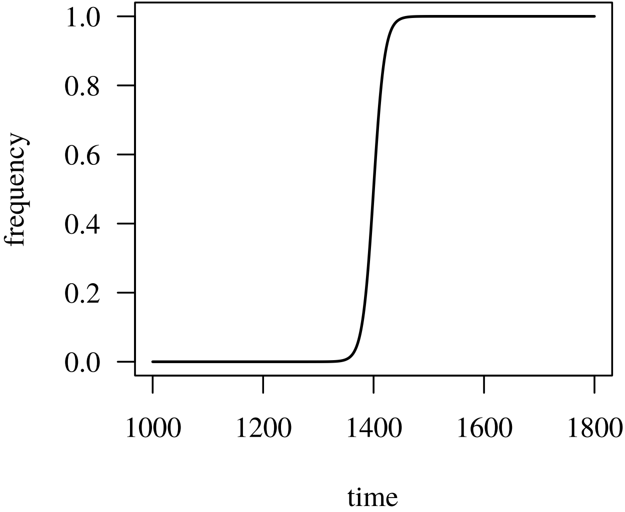
Varying the *s* parameter



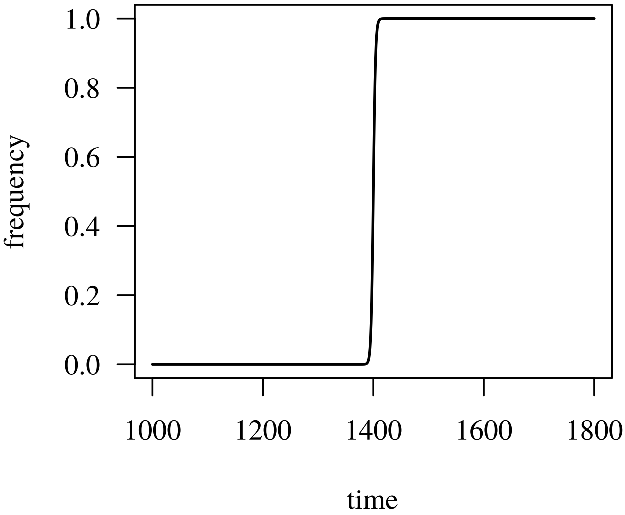
Varying the *s* parameter



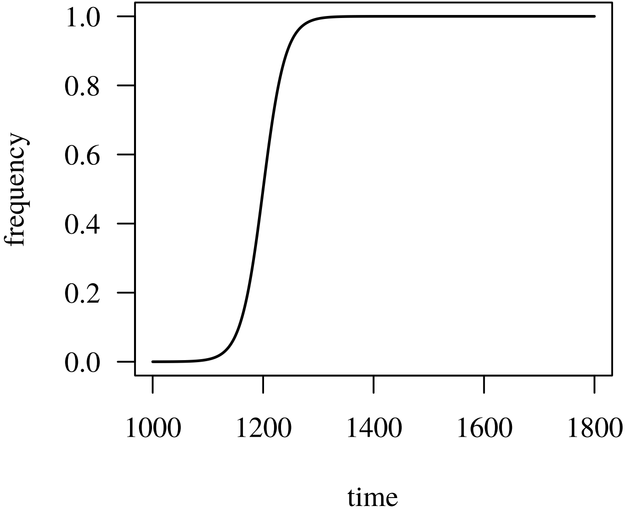
Varying the s parameter



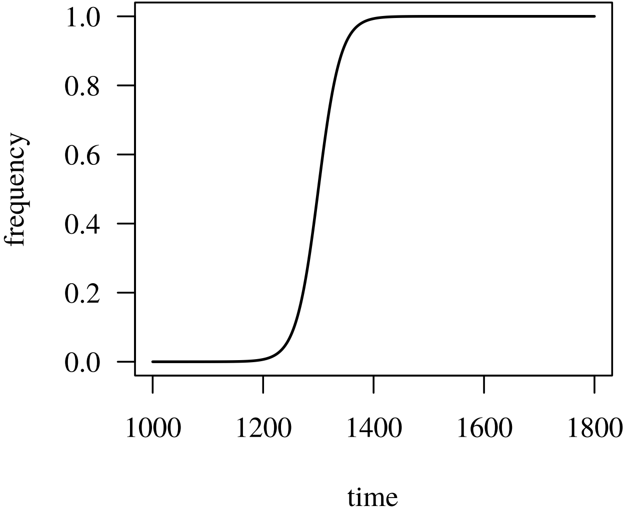
Varying the s parameter



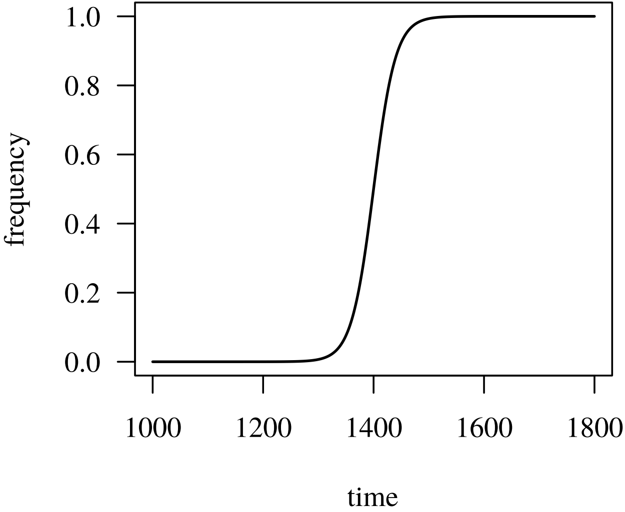
Varying the k parameter



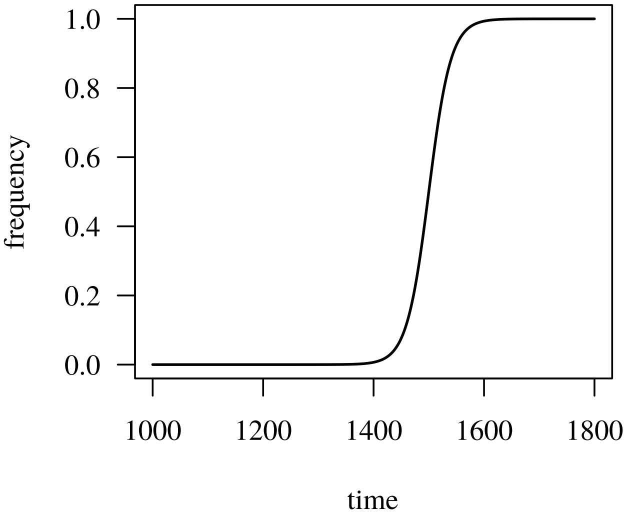
Varying the k parameter



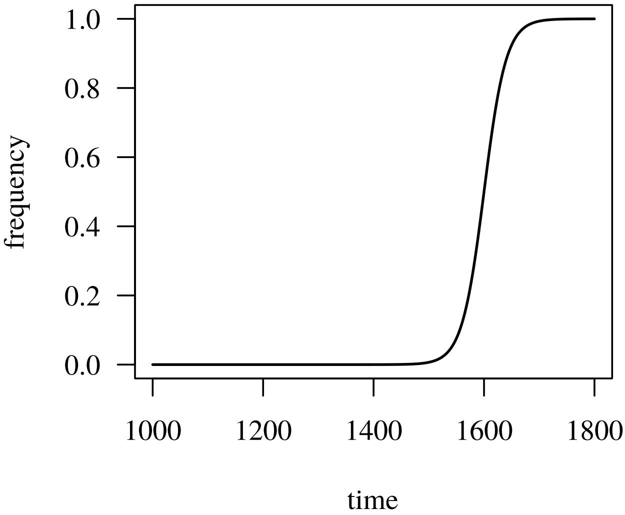
Varying the k parameter



Varying the k parameter



Varying the k parameter

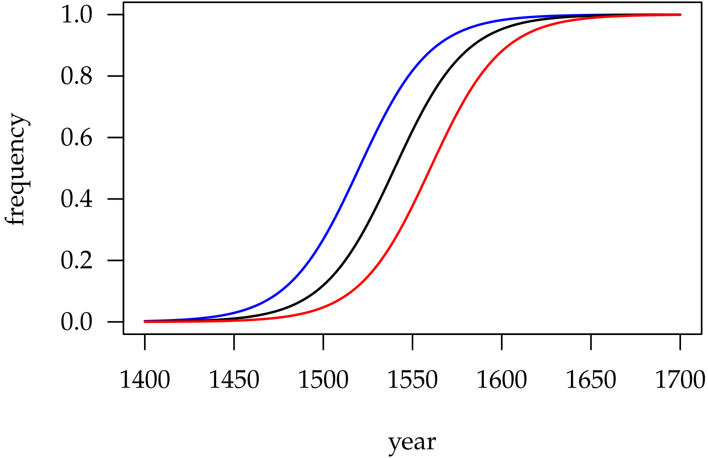


Constant Rate Hypothesis (CRH)

"[W]hen 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 terms of the logistic equation: the s values are the same, but the k values may differ.
- Since Kroch (1989), many empirical studies have found this sort of situation (henceforth Constant Rate Effects, CREs).
- Interpretation (Kroch 1989): a single syntactic parameter governs all contexts. Some contexts may favour the new form, others disfavour it, but those (dis)favouring effects are **constant**.
- Relies on individual speakers having access to multiple, probabilistically-weighted **competing grammars**.

Different linguistic contexts

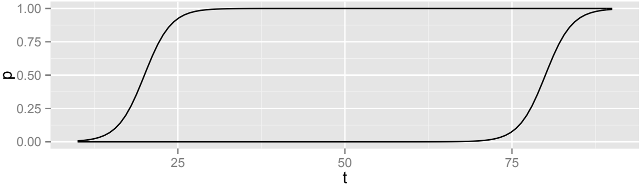


Problem 1: no mechanism

- Kroch (1989, 204) adopts the logistic equation because of its use in biology, but admits that it does not follow from anything.
- Since then, Yang (2000) has derived the logistic equation from his model of syntactic acquisition and change (on which more soon).
- But no mechanism has so far been presented **to derive the CRH**: how exactly are the contexts tied to an underlying change?

Problem 2: time separations

- For a CRE, the *s* values must not differ significantly. But what about large differences in *k*?



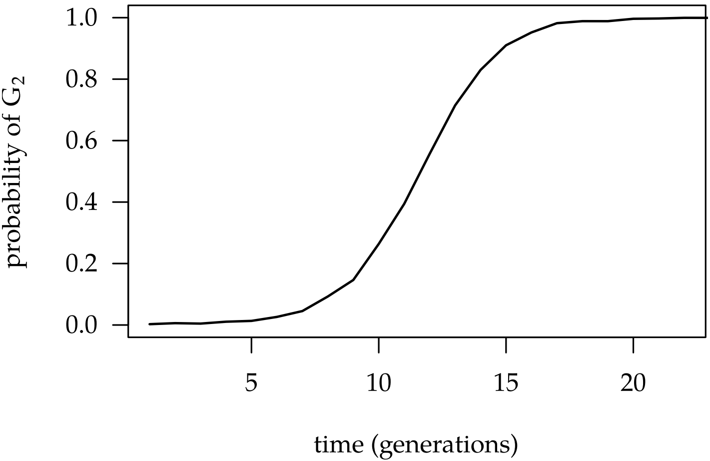
- If the change only gets off the ground in one context 50 years after it has run to completion in another context (or 500, or 5,000,000), how plausible is it that we are dealing with reflexes of a single parametric change? (Think of the acquirer!)
- But this would still qualify as a CRE.

Part II

The model

Change

- Assume two grammatical hypotheses G_1 and G_2 compete in a population of speakers, so that G_1 is used with probability p and G_2 with probability q ($p + q = 1$)
- How do $p = p(t)$ and $q = q(t)$ change as a function of time t?
- Safe assumption: change is logistic, with either $p(t)$ growing and $q(t)$ decreasing or $q(t)$ growing and $p(t)$ decreasing as $t \to \infty$
- A plausible mechanism for change of this kind has been given by Yang (2000)
  - In this model, change is driven by the fact that different grammars parse different proportions of sentences the learner encounters in its environment
  - The grammar with the greater parsing advantage wins



1.0
0.8
0.6
0.4
0.2
0.0
probability of G2
5
10
15
20
time (generations)

Production biases

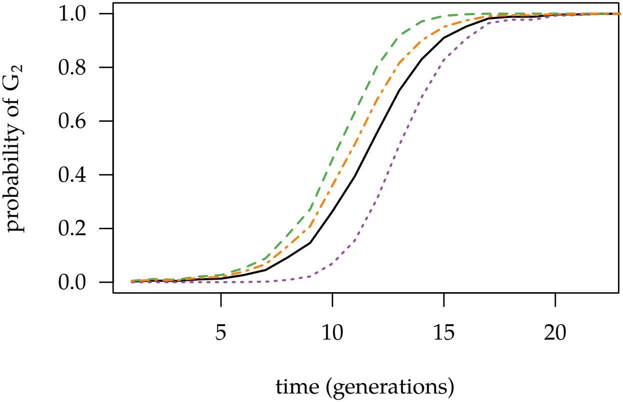
- What if there was an additional step in the process, where the learner's output is filtered by a **production bias**?
- I.e. learner emerges with a hypothesis (p, q) after learning, but in subsequent production filters this hypothesis according to some skewing mechanism that produces $(p - \epsilon, q + \epsilon)$ as the input for the following generation
- What if this biasing mechanism was sensitive to **linguistic contexts**?
  - E.g. different amount and/or direction of bias applied in questions and declaratives
- Intuitively, this should result in one of three outcomes:
  (1) Different contexts change at different rates
  (2) Different contexts begin to change at different times
  (3) Both (1) and (2)

Production biases

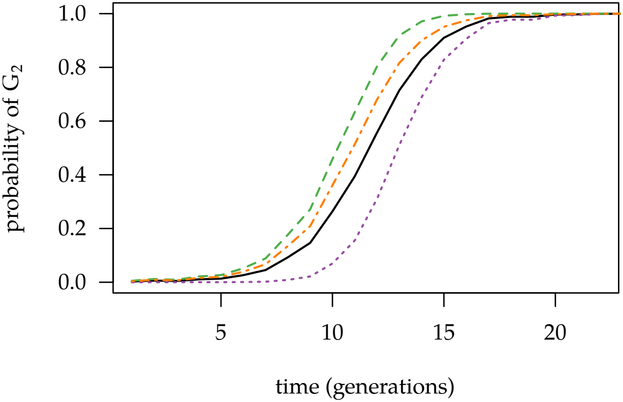
- Assume K linguistic contexts $C_1, \ldots, C_K$
- Assume that with each context C_i is associated a constant (time-invariant) production bias b_i with the speaker's actual production in context C_i given by

$$\begin{cases} p_i = p - b_i p(1-p) \\ q_i = q + b_i q(1-q) \end{cases} \qquad (2)$$

  with $-1 \leq b_i \leq 1$
- Intuition: degree of biasing for p_i has to depend on both p and $1-p$ (think of situations where $p = 0$ or $p = 1$)
- Notice:
  - if b_i is positive then context C_i favours G_2
  - if b_i is negative then context C_i favours G_1
  - if $b_i = 0$ there is no biasing in context C_i



1.0
0.8
0.6
0.4
0.2
0.0
probability of G_2
5
10
15
20
time (generations)



- Context curves not strictly logistic, but derived from one that is…
- …except in unusual circumstances (we'll return to this point below)

The Time Separation Theorem

- This means we can derive each context curve $p_i(t)$ from an underlying logistic $p(t)$ as follows:

$$p_i(t) = p(t) + b_i p(t)(1 - p(t)) \tag{3}$$

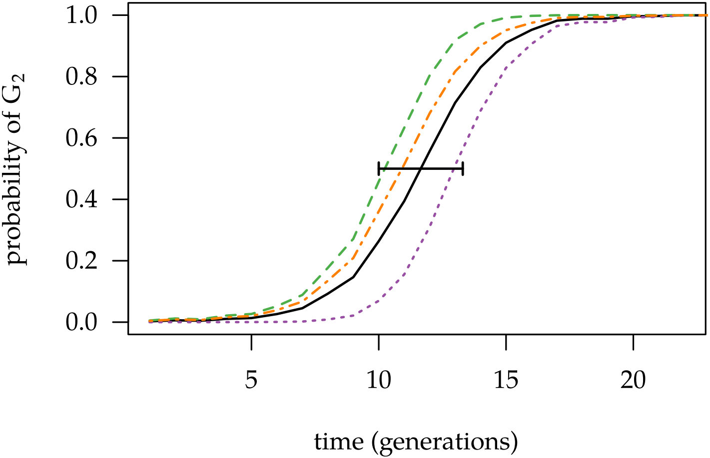
where $-1 \leq b_i \leq 1$.

- Importantly, this implies:

Theorem (The Time Separation Theorem)

For an underlying change $p(t)$ logistic with slope s, the maximal time separation between two contexts at the tipping points is

$$\Delta_{\max}(s) = \frac{2}{|s|} \log\left(\frac{1}{\sqrt{2} - 1}\right) \approx 1.76 \cdot \frac{1}{|s|}. \tag{4}$$

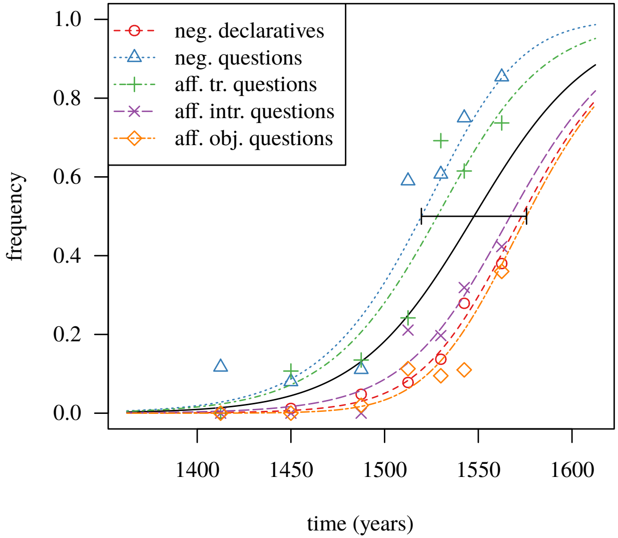


1.0
0.8
0.6
0.4
0.2
0.0
probability of G_2
5
10
15
20
time (generations)

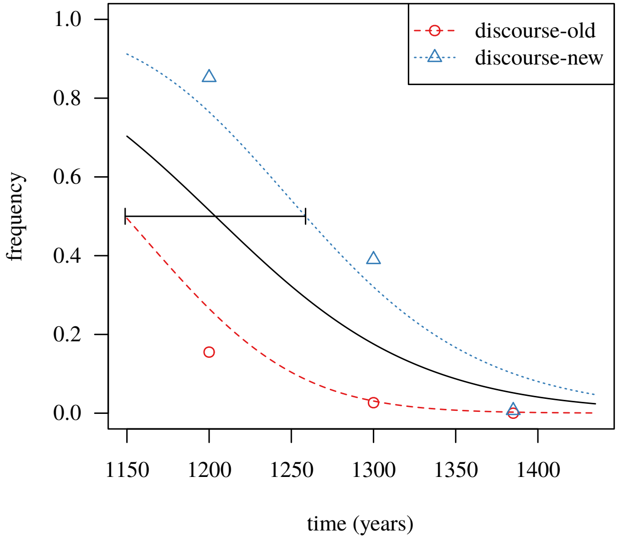
Part III

Evaluating the model: empirical fit

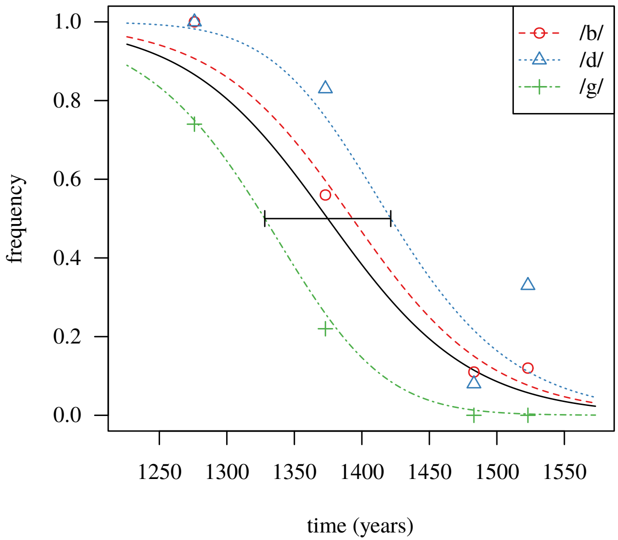
do-support in English (Kroch, 1989)



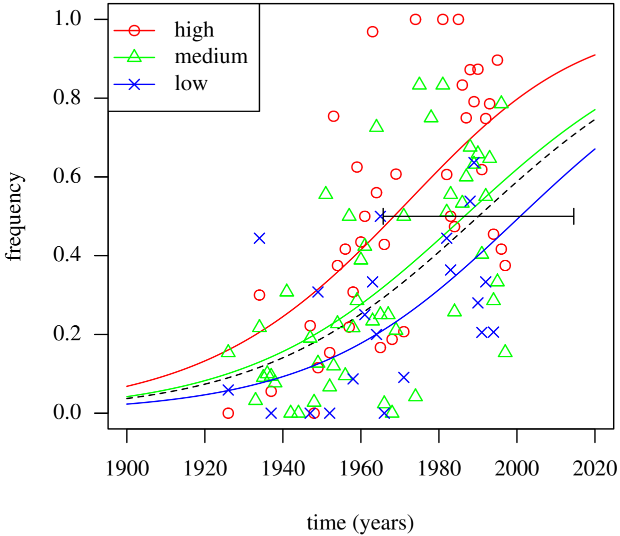
English Jespersen cycle (Wallage, 2013)



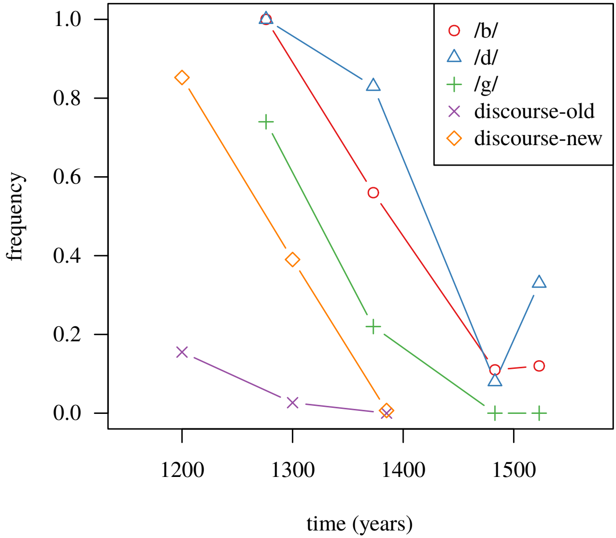
German final fortition (Fruehwald et al., 2009)



Manchester /t/-glottalling (Bermúdez-Otero et al., 2016)

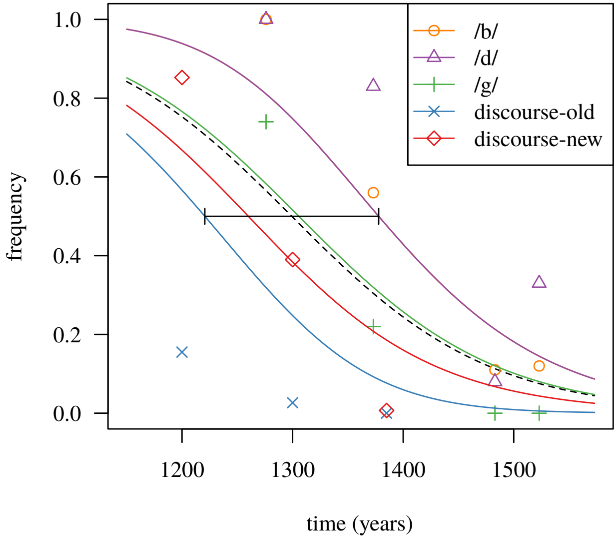


Pseudo-CRE



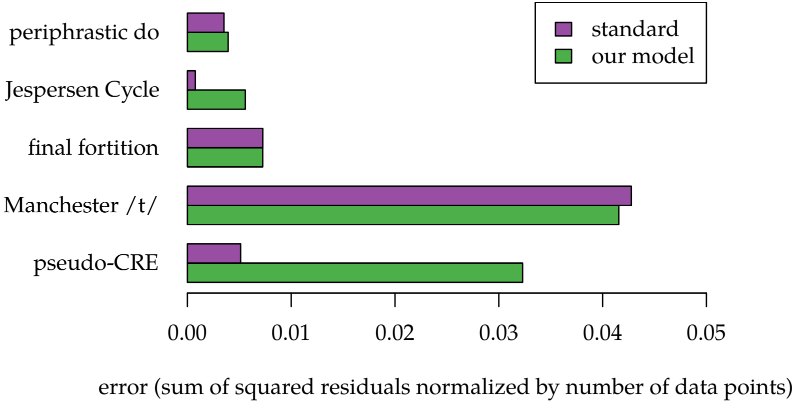
1.0
0.8
0.6
0.4
0.2
0.0
frequency
1200
1300
1400
1500
time (years)
/b/
/d/
/g/
discourse-old
discourse-new

Pseudo-CRE

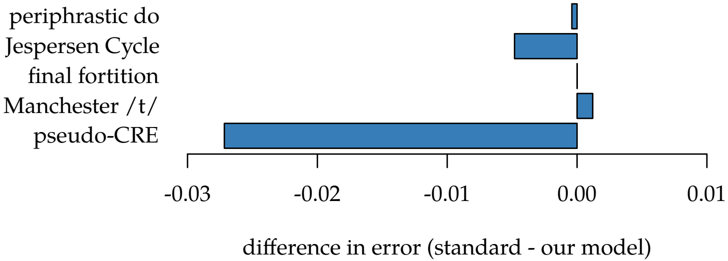


1.0
0.8
0.6
0.4
0.2
0.0
frequency
1200
1300
1400
1500
time (years)
/b/
/d/
/g/
discourse-old
discourse-new

Fit

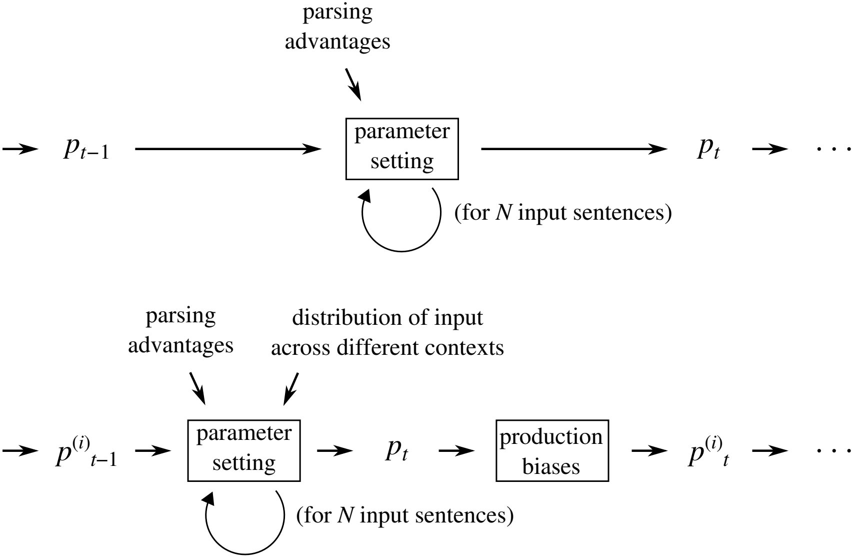

periphrastic do
Jespersen Cycle
final fortition
Manchester /t/
pseudo-CRE
standard
our model
0.00
0.01
0.02
0.03
0.04
0.05
error (sum of squared residuals normalized by number of data points)

Fit (relative)



Part IV

The non-logistic regime



parsing advantages
p_{t-1}
parameter setting
p_t
. . .
(for N input sentences)
parsing advantages
distribution of input across different contexts
$p^{(i)}{}_{t-1}$
parameter setting
p_t
production biases
$p^{(i)}{}_t$
. . .
(for N input sentences)

Yang's (2000) Fundamental Theorem

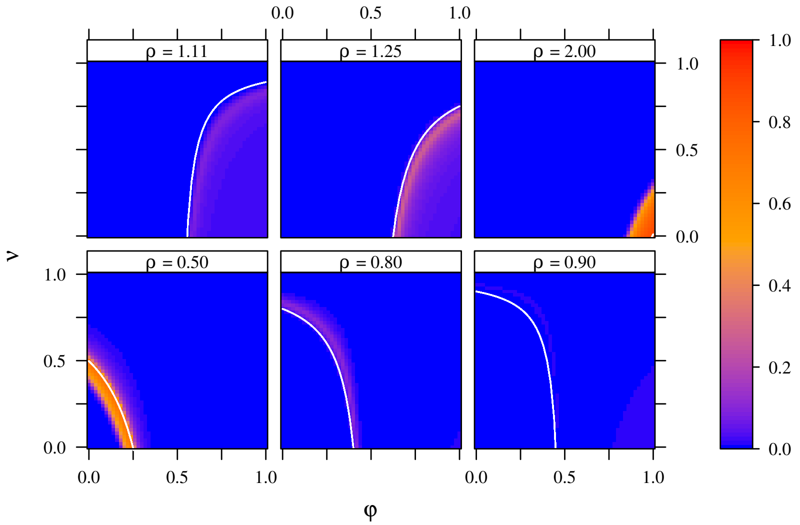
- There is a strong prediction here regarding the outcome of grammatical competition
- Let $\rho = \alpha/\beta$ be the ratio of the weak generative capacities of the competing grammars G_1 and G_2 (called "advantages" in Yang, 2000)
  - if $\rho > 1$, then the proportion of sentences parsed by G_1 exclusively is larger than the proportion of sentences parsed by G_2 exclusively
  - if $\rho < 1$, vice versa
  - if $\rho = 1$, grammars "equally powerful"
- It can be shown (Yang, 2000):

Theorem (The Fundamental Theorem of Language Change)

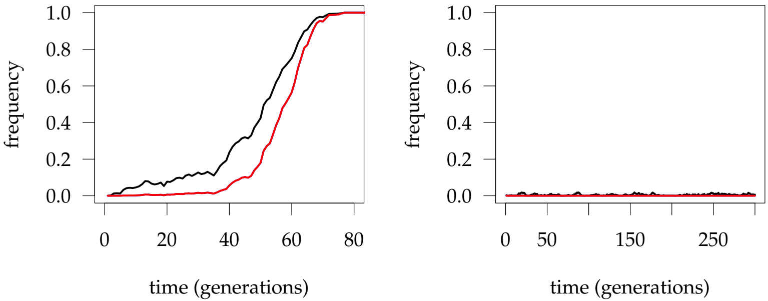
Suppose learners are reliable (get a large number of input sentences and apply a reasonable learning rate). Then grammar G_2 overtakes grammar G_1 if, and only if, $\rho < 1$.

Production biases

- Crucially, in our model there is a **feedback loop** between the underlying parametric change and the production biases
- This leads to a more complicated dynamics and a correspondingly more complicated difference equation for the underlying change
- Numerical simulations suggest that in our model underlying change is still logistic in **most** cases


ρ = 1.11
ρ = 1.25
ρ = 2.00
ρ = 0.50
ρ = 0.80
ρ = 0.90
ν
φ
0.0
0.5
1.0
0.2
0.4
0.6
0.8

The non-logistic regime: advantage vs. bias



- Sufficiently negative production biases can, however,
  - slow down change and make it non-logistic (still S-shaped though!)
  - block change altogether

The non-logistic regime: advantage vs. bias

- So what, exactly, happens to the Fundamental Theorem?

Theorem (The Fundamental Theorem of Language Change)

Suppose learners are reliable (get a large number of input sentences and apply a reasonable learning rate). Then grammar G_2 overtakes grammar G_1 if, and only if, $\rho < 1$.

- Recall:
  - $\rho > 1$: the amount of data parsed by G_1 but not by G_2 is greater than the amount of data parsed by G_2 but not by G_1
  - $\rho < 1$: vice versa

The non-logistic regime: advantage vs. bias

Theorem (The Extended Fundamental Theorem of Language Change)

Assume reliable learners and write

$$B = \sum_{i=1}^{K} \lambda_i b_i \tag{5}$$

and

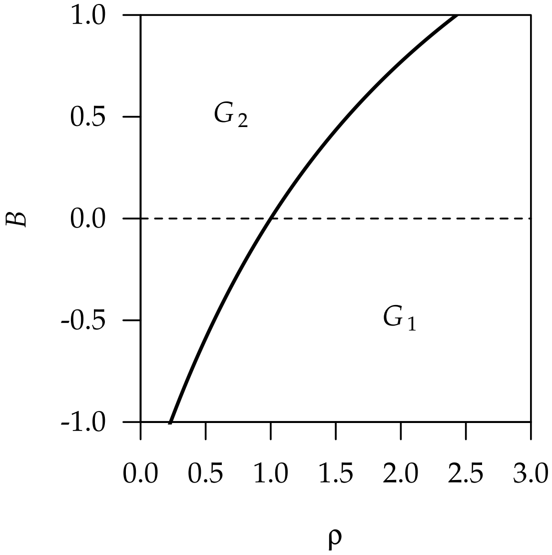
$$B_c = \frac{\rho - 1}{1 + q(0)(\rho - 1)}. \tag{6}$$

Then

1. $q(t) \to 1$ *as* $t \to \infty$, *if* $B > B_c$;
2. $q(t) = q(0)$ *for all* t, *if* $B = B_c$;
3. $q(t) \to 0$ *as* $t \to \infty$, *if* $B < B_c$.

In other words, G_2 *overtakes* G_1 *if, and only if,* $B > B_c$.

Bifurcation



Part V

Conclusions

Conclusions

- We have presented a **mechanism** for CREs, derived from simple ingredients:
  - The variational learner of Yang (2000)
  - Iterated learning (in an infinite, well-mixing population...)
  - Production biases with fixed weights
- A corollary of this is that an upper limit can be placed on the time separation of context curves (traditionally *k*).
- The model gives a good fit for a number of case studies.
- In some instances, production biases can outweigh the parsing 'advantage' of a grammar, so that it does not triumph in diachrony.

References

Altmann, G., von Buttlar, H., Rott, W. & Strauss, U. (1983). A law of change in language. In B. Brainerd (ed.), *Historical linguistics*, 104–115. Bochum: Brockmeyer.

Bermúdez-Otero, R., Baranowski, M., Bailey, G. & Turton, D. (2016). A constant rate effect in Manchester /t/-glottalling: high-frequency words are ahead of, but change at the same rate as, low-frequency words. Paper presented at OCP 13, Budapest, January 2016.

Fruehwald, J., Gress-Wright, J., and Wallenberg, J. (2009). Phonological change: the Constant Rate Effect. *Proceedings of NELS 40*. Amherst, MA: GLSA.

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

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

Yang, C. (2000). Internal and external forces in language change. *Language Variation and Change*, 12: 231–250.

Acknowledgements

Audiences in Manchester, Naples (ICHL22), Amsterdam (*Dealing with Bad Data in Linguistic Theory*) and Göttingen (GLOW39), and Ricardo Bermúdez-Otero, for discussions and for allowing the use of the data in Bermúdez-Otero et al. (2016)