INDIVIDUAL AND POPULATION IN LANGUAGE CHANGE

MATHEMATICAL EXPLORATIONS OF THE ACQUISITION–USE LOOP

A thesis submitted to The University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Humanities

2018

Henri A Kauhanen

School of Arts, Languages and Cultures
# Contents

List of publications 5  
List of abbreviations 6  
List of Figures 7  
List of Tables 10  
Abstract 11  
Declaration 12  
Copyright statement 13  
Acknowledgements 14  

## Introduction 19  

1  Order and disorder in language 20  
   1.1  Order .......................... 20  
   1.2  Disorder .......................... 22  
   1.3  Plan .......................... 24  

2  Acquisition 26  
   2.1  The Z of language change .......................... 26  
   2.2  Variational Learning .......................... 30  

3  The diachronic interplay of acquisition and use 37  
   3.1  The Constant Rate Effect .......................... 37  
   3.2  Production biases .......................... 41  

4  Acquisition and change in higher-dimensional spaces 49  
   4.1  Multidimensional competition .......................... 49  
   4.2  Stable variation .......................... 56  

5  Networks and (near-)neutrality 62  
   5.1  Networks, differentials and order .......................... 62  
   5.2  Neutral change .......................... 64
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3 Why does it happen? Prolegomenon to a study of the $\sigma = 1$ case</td>
<td>67</td>
</tr>
<tr>
<td>5.4 A note on neutral and near-neutral models</td>
<td>73</td>
</tr>
<tr>
<td>6 Change and stability over large timescales</td>
<td>76</td>
</tr>
<tr>
<td>6.1 Typology across time and space</td>
<td>76</td>
</tr>
<tr>
<td>6.2 The inference of diachrony from synchrony</td>
<td>82</td>
</tr>
<tr>
<td>6.3 Coda: founder effects vs. ergodicity</td>
<td>86</td>
</tr>
</tbody>
</table>

## Original publications

<table>
<thead>
<tr>
<th>Publication</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Deriving the Constant Rate Effect</td>
<td>90</td>
</tr>
<tr>
<td><em>Henri Kauhanen &amp; George Walkden</em></td>
<td></td>
</tr>
<tr>
<td>I.1 Introduction</td>
<td>91</td>
</tr>
<tr>
<td>I.2 Grammar competition and production biases</td>
<td>97</td>
</tr>
<tr>
<td>I.3 Dynamics of the extended model</td>
<td>107</td>
</tr>
<tr>
<td>I.4 Evaluation</td>
<td>111</td>
</tr>
<tr>
<td>I.5 Discussion</td>
<td>124</td>
</tr>
<tr>
<td>I.6 Conclusion</td>
<td>127</td>
</tr>
<tr>
<td>I.7 Appendix: Derivations</td>
<td>128</td>
</tr>
<tr>
<td>II Stable variation in multidimensional competition</td>
<td>134</td>
</tr>
<tr>
<td><em>Henri Kauhanen</em></td>
<td></td>
</tr>
<tr>
<td>II.1 Variation, learning and diachronic stability</td>
<td>134</td>
</tr>
<tr>
<td>II.2 Two grammars</td>
<td>138</td>
</tr>
<tr>
<td>II.3 Advantage matrices and the cyclical balance criterion</td>
<td>141</td>
</tr>
<tr>
<td>II.4 Dynamics: general results</td>
<td>144</td>
</tr>
<tr>
<td>II.5 Babelian systems</td>
<td>147</td>
</tr>
<tr>
<td>II.6 Symmetric systems</td>
<td>149</td>
</tr>
<tr>
<td>II.7 Quasi-Babelian systems</td>
<td>151</td>
</tr>
<tr>
<td>II.8 Naive learning</td>
<td>153</td>
</tr>
<tr>
<td>II.9 Conclusions and conjectures</td>
<td>159</td>
</tr>
<tr>
<td>III Neutral change</td>
<td>163</td>
</tr>
<tr>
<td><em>Henri Kauhanen</em></td>
<td></td>
</tr>
<tr>
<td>III.1 Introduction</td>
<td>164</td>
</tr>
<tr>
<td>III.2 Neutrality</td>
<td>166</td>
</tr>
<tr>
<td>III.3 Model</td>
<td>170</td>
</tr>
</tbody>
</table>
List of publications

This thesis is composed of the following original publications, which will be referred to in the body of the work by their Roman numerals.


Copyright to Papers II and III is held by John Benjamins Publishing Company and Cambridge University Press, respectively. The publishers should be contacted for permission to reuse or reprint the material in any form.

For the coauthored papers, the responsibilities divided as follows.

- **Paper I.** Designed the study: HK & GW. Analysed the data: HK & GW. Proposed the mathematical model: HK. Carried out mathematical derivations and proofs: HK. Wrote the paper: HK & GW. Wrote the simulation and data processing code: HK.

- **Paper IV.** Designed the study: HK, DG, TG and RB-O. Analysed the data: HK, DG, TG and RB-O. Solved the mathematical model: HK, DG and TG, with the crucial steps taken by TG. Wrote the paper: HK, DG, TG and RB-O. Wrote visualization routines: HK and DG. Wrote the data analysis and simulation code: HK.
### List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRE</td>
<td>Constant Rate Effect</td>
</tr>
<tr>
<td>DN</td>
<td>Deductive–Nomological (Explanation)</td>
</tr>
<tr>
<td>DT</td>
<td>Dynamic Typology</td>
</tr>
<tr>
<td>L1</td>
<td>First Language</td>
</tr>
<tr>
<td>L2</td>
<td>Second Language</td>
</tr>
<tr>
<td>LRP</td>
<td>Linear Reward–Penalty (Learning)</td>
</tr>
<tr>
<td>PLD</td>
<td>Primary Linguistic Data</td>
</tr>
<tr>
<td>TLA</td>
<td>Triggering Learning Algorithm</td>
</tr>
<tr>
<td>UG</td>
<td>Universal Grammar</td>
</tr>
<tr>
<td>USM</td>
<td>Utterance Selection Model</td>
</tr>
<tr>
<td>V2</td>
<td>Verb-Second</td>
</tr>
<tr>
<td>VL</td>
<td>Variational Learning</td>
</tr>
<tr>
<td>WALS</td>
<td>World Atlas of Language Structures (Dryer &amp; Haspelmath, 2013)</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 The S-curve. ............................................................. 21
1.2 Macro- and micro-levels in thermodynamics and language change. . 22

2.1 Z-model of language change. ........................................ 27
2.2 Triggering Learning Algorithm (TLA). .............................. 29
2.3 Intrasentential syntactic codeswitching. ............................ 31
2.4 LRP learning of one Parameter...................................... 33
2.5 Parsing advantage diagram for one Parameter. .................... 34
2.6 Learning trajectories of reliable and unreliable LRP learners. ....... 35

3.1 Periphrastic *do* in early Modern English. ......................... 39
3.2 Standard operationalization of the CRE. ............................ 40
3.3 Unconstrained variation in intercept parameters. .................. 41
3.4 Family of context-bound propagation curves. ...................... 43
3.5 Inter-generational change with an acquisition–use loop .......... 44
3.6 CREs produced by the production bias model. ..................... 45
3.7 Fit of the production bias model to English periphrastic *do*. ..... 46
3.8 Maximally separated contextual curves for different slopes. ...... 48

4.1 LRP learning of $n$ grammars. ..................................... 51
4.2 Parsing advantage diagram for three grammars. .................. 52
4.3 Learning trajectories of reliable and unreliable LRP learners ($n = 3$). 55
4.4 Phase space plot of Babelian systems ($n = 3$). ................... 58
4.5 Phase space plot of symmetric systems ($n = 3$). ................. 59
4.6 Phase space plot of quasi-Babelian systems ($n = 3$). ............ 60

5.1 Different network topologies. ...................................... 65
5.2 Neutral evolution on clustered and unclustered networks. ......... 66
5.3 Shifting, dominance and monotonicity in different networks. ...... 68
5.4 Probability of successful nuclear mutant propagation. ............ 73

6.1 Global frequency distribution of linguistic features. ............... 77
6.2 Markov dynamics of linguistic features. .......................... 78
6.3 Geographic distribution of tone and indefinite article. .......... 79
6.4 Von Neumann neighbours on a toroidal lattice. .................. 81
6.5 Flowchart for the model put forward in Paper IV. ................ 82
6.6 Isogloss density as a function of feature frequency in WALS. .... 83
6.7 Isogloss density $\sigma$ as a function of $\rho$ and $\tau$. ........................................ 84
6.8 No correlation between feature frequency $\rho$ and temperature $\tau$. ... 86
6.9 Differences in $\rho$ and $\sigma$ between Old World and New World. .... 87

I.1 *Do*-support in Early Modern English. ................................. 92
I.2 Logistic curves. ................................................................. 93
I.3 Parsing advantage diagram for two grammars. ......................... 98
I.4 Contextual partition of a parsing advantage diagram. ................. 101
I.5 The biasing functions $F$ and $G$. ........................................ 103
I.6 Inter-generational change with and without production biases. ....... 105
I.7 Model runs with different sets of production biases. .................. 106
I.8 The dynamic interaction of parsing advantage and production biases. 110
I.9 Logistic approximation of the production bias model. ................ 112
I.10 Fit of the production bias model to English *do*-support. .......... 115
I.11 Fit of the production bias model to the English Jespersen Cycle. ... 117
I.12 Fit of the production bias model to final fortition in Early New High German. ......................................................... 119
I.13 Difference in error between standard operationalization and the production bias model. ......................................................... 121
I.14 An Anglo-Bavarian pseudo-CRE. ........................................ 122
I.15 Error comparison with pseudo-CRE included. ......................... 123

II.1 Pendula. ........................................................................ 137
II.2 Parsing advantage diagram for two grammars. ......................... 140
II.3 Sample evolution of a two-grammar system. ............................ 141
II.4 Parsing advantage diagram for three grammars. ....................... 142
II.5 A barycentric ternary plot. ................................................ 146
II.6 Phase space plot of Babelian 3-grammar systems. .................... 150
II.7 Sample trajectory in a 3-grammar system. .............................. 150
II.8 Phase space plot of a symmetric 3-grammar system. ................. 152
II.9 Phase space plots of canonical quasi-Babelian 3-grammar systems. 154
II.10 Orbit diagram of quasi-Babelian 3-grammar systems. ............... 155
II.11 Hierarchy of systems. ....................................................... 155
II.12 2-parameter Naive Parameter Learner at a vertex. ................. 158
II.13 2-parameter Naive Parameter Learner at an interior point. ...... 159
II.14 Naive Parameter Learners: diachrony. ................................ 160

III.1 Networks. ...................................................................... 172
III.2 Ill-behaved neutral trajectory. ........................................... 175
III.3 Well-behaved neutral trajectory. ........................................ 175
III.4 Illustration of monotonicity with four histories.  
III.5 Shifting: simulation results.  
III.6 Dominance and monotonicity: simulation results.  
III.7 Well-behavedness: simulation results.  
III.8 Effect of number of variants on shifting.  
III.9 Effect of number of variants on dominance and monotonicity.  
III.10 Shifting: simulation results with a less stringent threshold.  
III.11 Dominance and monotonicity: simulation results with a less stringent threshold.  
III.12 Effect of rewiring dynamics on well-behavedness.  
III.13 Time-to-dominance as a function of $K$ and $\sigma$.  
III.14 The probability of a central speaker distributing their variant to at least one other speaker before being removed from the social network.  
IV.1 Markov chain dynamics of a linguistic feature.  
IV.2 Illustration of feature frequency and isogloss density.  
IV.3 State of the lattice at the stationary distribution.  
IV.4 Empirical measurements of feature frequency and isogloss density.  
IV.5 Empirical and simulated geospatial distributions of tone and indefinite article.  
IV.6 Regression of temperature $\tau$ against Dediu’s PC1.  
7.1 The quasispecies and quasigrammar equations compared.  
8.1 Language change.  
8.2 Language change (production biases).  
8.3 Language change (acquisition biases).  
8.4 Language change (varied sources of PLD).  
8.5 VL: stochastic simulations and deterministic theory compared.
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>A non-exhaustive list of works investigating the CRE.</td>
<td>38</td>
</tr>
<tr>
<td>6.1</td>
<td>Distributional predictions of Greenberg’s Dynamic Typology.</td>
<td>80</td>
</tr>
<tr>
<td>I.1</td>
<td>English <em>do</em>-support.</td>
<td>115</td>
</tr>
<tr>
<td>I.2</td>
<td>English Jespersen Cycle.</td>
<td>116</td>
</tr>
<tr>
<td>I.3</td>
<td>Final fortition in Early New High German.</td>
<td>118</td>
</tr>
<tr>
<td>IV.1</td>
<td>Seven least stable and most stable features.</td>
<td>206</td>
</tr>
<tr>
<td>IV.2</td>
<td>35 binary linguistic features.</td>
<td>212</td>
</tr>
<tr>
<td>IV.3</td>
<td>Feature frequency, isogloss density and temperature for 35 features.</td>
<td>213</td>
</tr>
<tr>
<td>7.1</td>
<td>Biological evolution and language change.</td>
<td>236</td>
</tr>
<tr>
<td>8.1</td>
<td>Production biases.</td>
<td>243</td>
</tr>
</tbody>
</table>
Abstract

Language change is driven by a constellation of acquisition and usage factors operating at two ontological levels: the level of the individual and the level of the population. This thesis proposes that an improved understanding of processes of language change can be obtained through the use of mathematical models that incorporate detailed mechanisms of language acquisition and use and derive diachronic predictions as mathematical theorems of those mechanisms. Four aspects of linguistic diachrony are singled out for detailed study in the four original publications constituting the core of this journal-format dissertation: the Constant Rate Effect, stable variation in multidimensional grammatical competition, effects of social network topology and rewiring, and the relationship between processes of change and synchronic frequency and spatial distributions of linguistic traits.

Through mathematical analysis and computer simulations, the dissertation suggests (i) that diachronic patterns such as the Constant Rate Effect arise through an interaction of acquisition and usage effects, the latter modelled as probabilistic post-acquisition production biases that filter the underlying grammatical state; (ii) that the fundamental result on competition between two grammatical options that outlaws diachronically stable variation does not hold of multidimensional competition; (iii) that finite-size effects such as network topology and dynamic network rewiring may give rise to orderly phenomena such as S-curves even in the absence of traditional biasing factors; and (iv) that typological distributions of linguistic features arise through a dynamic interplay of faithful transmission and mutation, subject to both local and areal effects.

As a general conclusion, I propose that language change can be understood as the product of neither acquisition nor usage factors alone, but that both types of factor need to be incorporated in mechanistic models which make the interrelationships between these factors explicit. Acquisition and use constitute a diachronic feedback loop leading to nonlinear equations of linguistic change; it is in many cases a consequence of these nonlinearities that bifurcations arise which may throw change trajectories onto qualitatively unexpected courses simply as a response to minute variation in relevant control parameters. This sheds new light on explanation in diachronic linguistics: to explain a process of change is to model it with mechanisms which derive facets of that process as empirical predictions. Mathematical models are particularly well suited to this task, as they force the modeller to be explicit and exact, both qualitatively and quantitatively, about the mechanisms involved; explanations of this kind, moreover, can be either traditionally deductive–nomological, or probabilistic.
Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on Presentation of Theses.
Acknowledgements

Siempre mantén la calma.

(Santiago Metro)

It is a central claim of this thesis that not everything in linguistic diachrony is determined: various kinds of random events shape the histories (or futures, I should say) of languages. Something similar must have happened when, as an undergraduate research assistant in Professor Ismo T. Koponen’s group at the University of Helsinki many years ago, I stumbled upon copies of Hofbauer and Sigmund (1998) and Nowak (2006) on Ismo’s bookshelf. As will become apparent to any reader of what follows, the influence of these two books — not necessarily their content, for content-wise I have not always agreed, but certainly in terms of their style of argument, their flavour of epistemology, their Galilean insistence on not only the usefulness but the very necessity of mathematization — is felt in almost everything I have done or attempted to do. One thing led to another, and I soon found myself in Dr Tadeas Priklopil’s Introduction to Bifurcation Theory, finally putting matrix algebra, as it seemed, to good use.

Fast forward, and we meet HK in the Division of Linguistics and English Language at Manchester. It goes without saying that this thesis would have turned out a very different animal were it not for the excellent guidance and collaboration I enjoyed there. It is perhaps extraordinary for a graduate student to acknowledge the support of no fewer than four supervisors and five panel members along the way, but, given a number of random events again, I am privileged to do so. First and foremost, I am thankful to Professor (now Emeritus) David Denison and Dr (now Herr Prof Dr) George Walkden for believing in my project in the first place and making sure it had a smooth start. Dr Ricardo Bermúdez-Otero and Dr Tobias Galla joined in at a later juncture in a move that proved to have far-reaching, and entirely positive, consequences. I have not only greatly benefitted from all our discussions during the past four years, I have enjoyed them immensely, and thank all four of you: David, for showing me what it means to be a historical linguist with an acute eye for detail; George, for convincing me that theories of diachrony and synchrony must, and can, live in symbiosis; Ricardo, for being my sharpest critic, a well of inspiration, and a Renaissance man in a postmodern world; Tobias, for teaching me how to think like a physicist, a trait that has found surprising application in a domain traditionally (but falsely) considered diametrically opposite. George, who wore the main supervisor’s hat the longest, deserves a special mention for being an excellent mentor not just in
terms of academic substance, but in all related, less substantial yet no less important, extra, quasi and pan-academic matters.

Dr Laurel MacKenzie, Professor Alan McKane and Professor John Payne, as well as Ricardo and Tobias, each acted as independent reviewer for one or two of the six biannual panels. I thank you for perusing the panel papers with a critical eye and for providing a number of comments which I feel have greatly improved the final product. I also wish to thank Professor Eva Schultze-Berndt, as well as Andrew Fairhurst and Julie Fiwka of the Graduate School Office, for making sure the practical side of things ran smoothly. Many other people at Manchester and elsewhere have had an influence on the evolution of this thesis and the way I have come to think about language, language change and complex systems; while they are far too numerous to mention here, many are acknowledged in the individual Papers that make up the thesis.

I am extremely grateful to Dr Richard A. Blythe, of The University of Edinburgh, and Dr Joel C. Wallenberg, of Newcastle University, for agreeing to examine this thesis.

During my time in Manchester I had the fortune to be part of a wonderful, bright and enormously fun postgraduate community. I would like to thank my friends for countless intellectual discussions, lunches, pub quizzes, walks in parks and impromptu ping pong matches in the Phonetics Lab, and generally for making my years in Manchester some of the happiest of my life: Dr Laura Arman, George Bailey, Fernanda Barrientos, Mary Begley, Deepthi Gopal, Dr Míša Hejná, Fang Jackson-Yang, Sarah Mahmood, Stephen Nichols, Jane Scanlon. During my visits to dear old Helsinki I have enjoyed the company and hospitality of Pipsa Enqvist, Tommi Gröndahl, Dr Otto Lappi and Sanna Tiirikainen, Dr Jussi Palomäki and Dr (the useful kind) Eeva Palomäki, and Ville A. Saarinen. Thanks to my onetime flatmates Kevin Sunny and Louis Zhang, for being such excellent company; to Chris and the rest of ManCoCo, for the espressi; to Voces Bibliothecae, for the polyphony; and, from a very special, beaty corner of my heart, to the people of Manchester Lindy, for all the dances.

The final stages of the writing up of this thesis coincided with a “post”-doctoral post at the University of Konstanz, on the beautiful Bodensee. I would like to thank the Department of Linguistics for making a wet-behind-the-ears natural philosopher feel welcome, in particular Dr Carmen Kelling and Professor Regine Eckardt, as well as the all important Welcome Centre team. I am particularly grateful for the opportunity I was given to teach classes on variation, change and typology, something that has deepened my thinking on these issues and affected this thesis directly.

On a more fiscal note, the work undertaken for this dissertation was funded by the University’s School of Arts, Languages and Cultures in its first year, by Emil Aaltonen Foundation in years two and three, and by The Ella and Georg Ehrnrooth Foundation during the submission pending period; for this financial patronage, without which
none of the work would have been possible, I extend my heartfelt thanks. I gratefully acknowledge conference travel funding from the School, the Hulme Hall Trust Fund, the Linguistics Association of Great Britain, the Department of Linguistics at the University of Konstanz, and the 39th Generative Linguistics in the Old World conference and University of Göttingen.

The thesis was typeset using \LaTeX, PGF/Ti\textit{k}Z, \texttt{knitr} and \texttt{pdftk} on a variety of GNU/Linux systems; all data analysis and simulation code was written in R and Fortran using Vim, and Maxima provided some handy symbolic calculation when algebra proved overbearing. Since these people are not acknowledged often enough, I think, I would like to thank those involved in the free software movement for making available these tools without which my workload would have obviously multiplied. (A collective thank-you to all contributors to Stack Overflow and \LaTeX Stack Exchange, too!) A very special thanks must go to the Faculty of Engineering and Physical Sciences at Manchester for allowing me the use of their HTCondor cluster for high-throughput parallel computing, as well as to the original Condor developers at the University of Wisconsin–Madison; without this, most of the computer simulations reported in this thesis would have been impossible in practice, if not in principle. Thanks are also due in this connection to Dr Michael Croucher for crucial advice on random number generation for parallel streams.

I turn, finally, to those closest to me. Feñi — thank you for your love and support; for picking me up from the floor every time I failed to heed the Metro’s advice; and quite simply for tolerating me, particularly during the latter stages of writing up. Isä, äiti, Anssi — kiitos, yksinkertaisesti!
To all Mancunians, in geography or in spirit

He remembered the greyness and griminess of Manchester, its appalling fogs, the coldness of houses, the warmth of the people [...].

(Hans Bethe obituary, The Guardian, 8 March 2005)
The most important act of theoretical integration that could be performed in historical linguistics is somehow establishing (if it is possible in principle) a clear and intelligible nexus between short-term individual behaviour and long-term linguistic evolution.

(Lass, 1997, 363, n. 28)

The most vitally characteristic fact about mathematics is, in my opinion, its quite peculiar relationship to the natural sciences, or, more generally, to any science which interprets experience on a higher than purely descriptive level.

(von Neumann, 1947/1961, 1)
INTRODUCTION
1 Order and disorder in language

Complex systems are characterized by several interacting levels; consequently the explanation of some observed orderly characteristic at one level may have to refer to processes, either orderly or disorderly, occurring at another level. Language is a complex system whose two main ontological levels are those of the individual and the population. This thesis explores how an explanatory connection between these two levels may be established, and what role mathematical formalism plays in this endeavour.

1.1 Order

The S-curve (Figure 1.1) is often said to be emblematic of language change: once a new linguistic variant is on the rise, its adoption follows a slow–fast–slow pattern (e.g. Bailey, 1973; Kroch, 1989; Croft, 2000). While the ubiquity of these sigmoidal curves has been questioned (Denison, 2003; Ghanbarnejad, Gerlach, Miotto & Altmann, 2014), they have been shown to occur in language change at all levels of the linguistic system from phonetics to syntax and semantics (see Blythe & Croft, 2012), and some mechanisms that generate them are now understood in detail, as we will have ample time to discuss in later sections of this thesis. The S-curve has then, not undeservedly, attracted much attention from diachronic linguists of diverse theoretical (as well as atheoretical) persuasions, so much so that it has come to be “established as a kind of template for change” (Chambers, 2002, 361).

The S-curve is but one manifestation of order in a complex system — in this case, in language, understood in an extended sense covering both the structural, synchronic language system and its space-time realization, the language community. That such orderly phenomena characterize the dynamics of language should not come as a surprise: after all, it is typical of complex systems to display orderly behaviour when viewed at a suitable level of description. For instance, a thermodynamic system in
equilibrium obeys the four fundamental laws of thermodynamics and consequently has predictable behaviour. This orderliness occurs despite the fact that at a lower, microscopic level of description, the system is characterized by properties of a very different kind, in fact ones not amenable to explanation within classical thermodynamics itself.

In physics, the theory of statistical mechanics is used to derive macro-level properties of equilibrated thermodynamic systems from the microscopic level, that is to say from a description of the system’s particles and their interactions.\(^1\) Similarly in language, one might expect macro-level properties of linguistic systems to result, in some yet to be explicated fashion, from micro-level constituents and their interactions. And just as it is a fundamental ontological commitment of any serious theory of synchronic linguistics that the observed properties of languages result from the operation of cognitive systems variously constrained, so must any serious theory of diachronic linguistics be able to reduce observed patterns of linguistic variation and language change to the operation of these systems \emph{and the interactions between several such systems}. It follows that a successful “statistical mechanics of language variation and change” must derive phenomenological properties of language from at least two different sources: the cognitive mechanism that accounts for the existence of languages (the language faculty) and population dynamics (the interaction, be it via

---

\(^1\)I pursue the analogy between physics and linguistics with the understanding that the reader has a basic understanding of the \emph{nature} of physical theory (though no knowledge of its \emph{substance} is assumed). For a particularly illuminating and accessible treatment of the link between the micro-level and the macro-level in thermal physics, see Schroeder (2000).
In physics, statistical mechanics (SM) serves as a bridging theory between a phenomenological description of the macroscopic level in terms of classical thermodynamics, and a description of interacting particle systems at the microscopic level. Similarly in language change, observed orderly macro-level properties such as the S-curve arise from interactions of individual speakers and language communities at the micro-level, though no established bridging theory exists so far.

![Diagram](image)

**Figure 1.2**
In physics, statistical mechanics (SM) serves as a bridging theory between a phenomenological description of the macroscopic level in terms of classical thermodynamics, and a description of interacting particle systems at the microscopic level. Similarly in language change, observed orderly macro-level properties such as the S-curve arise from interactions of individual speakers and language communities at the micro-level, though no established bridging theory exists so far.

language learning, language use or some other mechanism, of several such faculties). It is the purpose of this dissertation to defend a particular way of doing so, a particular manner of establishing a nexus between the macro-level of language change and the micro-level of speakers and speaker interactions (Figure 1.2).

While famous, the S-curve is not the only manifestation of order in language. Examples are easily multiplied: pathways of change, particularly in processes of grammaticalization (Hopper & Traugott, 1993) and phonologization (Bermúdez-Otero, 2015), typological universals, whether absolute or statistical (Comrie, 1981), and the Constant Rate Effect (Kroch, 1989) are all examples of macro-level order in the relevant sense. The latter two will receive extensive treatment in this study within the remit of Papers IV and I, respectively, whilst Paper III is devoted to an examination of S-curve-like patterns of change and Paper II to a hitherto little studied manifestation of order, stable variation. The four Papers are united in methodology — each proposes and explores the properties of a mathematical model of some aspect of language variation and change — and constitute an epistemic progression from the level of the individual (Papers I and II) through the level of language communities (Paper III) to the high-order level of language typology and the interactions between language communities on large timescales (Paper IV).

### 1.2 Disorder

With order comes disorder. Proceeding with our physical analogy, it might be tempting to conclude that this disorder resides entirely within the micro-level, such as when a particle is transported randomly in a gas by Brownian motion. Order could then be
seen as an emergent macro-level property of systems which are underlyingly disor-
derly. This picture must be true to some extent. For example, population displace-
ments in space (whether individuals commuting or moving house, or entire commu-
nities migrating) are part and parcel of human population dynamics and as such may
act as (partial) causes of language change, and these displacements are random when
viewed from the vantage point of linguistics proper. Equating the micro-level with
disorder and the macro-level with order would, however, hide a number of interesting
(and real) empirical phenomena from view.

The parallelism breaks twice, in fact, since there is both disorder at the macro-level
and order at the micro-level. More importantly, orderly phenomena at the macro-
level may be caused by either orderly or disorderly phenomena at a lower level: it
is the purpose of Paper III of this study to show that S-like curves may in some cir-
cumstances arise from random (though not “entirely” random — the issue turns out
to be convoluted) population dynamics, whilst Paper I constitutes an argument to
the effect that Kroch’s (1989) Constant Rate Effect arises from one specific orderly,
universal characteristic of language acquisition and use. To insist on a strict correla-
tion of order and level of description would also occlude important interaction effects
between orderly and disorderly phenomena, as Paper IV will argue.

It will, already at this early juncture, be useful to be a bit more specific about the
two notions order and disorder. By order, I understand determinism: a system is or-
derly insofar as it follows a deterministic trajectory in the sense that later states of
the system are determined by (and thereby predictable from) earlier states. In con-
trast to this, I will consider a system disorderly to the extent that it is stochastic,
that is to say random, non-deterministic and not amenable to prediction other than
in a probabilistic sense. Related but underlyingly deterministic notions such as the
dynamical-systems-theoretic notion of chaos — catastrophic sensitivity to variation
in initial conditions (e.g. Hirsch, Smale & Devaney, 2004) — fall outside this definition
of “disorder”. The ontological question whether nature is fundamentally determinis-
tic or non-deterministic also has no bearing on the concerns of the diachronic linguist.
Even if phenomena which we label stochastic turn out to be deterministic on a finer
analysis, they may still appear stochastic from the point of view of linguistic theory,
and may accordingly be approached using stochastic formalisms. By the same token,
if phenomena which we customarily label deterministic turn out to be somehow on-
tologically random, they may still be treated on deterministic terms if their effects are
sufficiently non-random.
1.3 Plan

This journal-format thesis consists of four original research articles flanked by the prefatory Chapters 1–6 and the concluding Chapters 7–8. Paper I derives the Constant Rate Effect (Kroch, 1989) by extending Yang’s (2000) model of Variational Learning (VL) with context-specific production biases, and compares the predictions of this model against three historical data sets. Paper II, similarly an extension of the VL framework, examines competition between multiple (more than two) grammars and finds that many such systems of grammar competition have attracting states which are states of stable variation. This invites experimentalists to look for evidence of stable variation in a specific kind of competition setting – situations of complex language contact. Paper III, a computational study at the community level, asks whether change can be neutral (non-biased) yet “well-behaved” in the sense that S-like curves of propagation are observed to punctuate quasi-equilibria. The answer, arrived at through a dynamical network model of the language community, is a qualified “yes”: well-behaved neutral trajectories emerge when networks are strongly clustered. Finally, Paper IV investigates the geospatial clustering of linguistic features within the framework of Greenberg’s (1978) Dynamic Typology (DT), with data mined from the World Atlas of Language Structures (Dryer & Haspelmath, 2013). A simple model is found to account for these data, but only if DT is extended with a theory of stochastic spatial interactions. A welcome consequence of this extension is that the proposed model makes possible the estimation of the “temperatures” of typological traits, a measure of their (inherent) stability or instability, in a manner which I will argue is superior to traditional accounts based on markedness theory and genealogical relatedness.

The four Papers are introduced and set against the backdrop of existing literature in Chapters 2–6. Chapter 7 revisits the model put forward in Paper II and shows that this model opens up a novel analogy between language change and biological evolution. Having explored the merits of this analogy, Chapter 8 then proceeds to consider its implications within the broader remit of the rest of the dissertation. The diagnosis is that much of the challenge in the current state of language change research turns, firstly, on (i) our incomplete understanding of the process of language acquisition, and secondly, on (ii) the insufficient attention that previous work has put on interactions between language acquisition and other facets of language dynamics, such as post-acquisition usage effects. The treatment offered in Chapter 8 must be considered experimental — as much as I would like to put forward a detailed research programme, the time is not quite ripe for such an undertaking. Some suggestions

---

For full citations of the original articles and a description of co-authors’ contributions, see the List of publications on p. 5.
for future work are outlined, however, including the sketch of a theory of probabilistic production biases designed to tackle problem (ii), and a “quasigrmam” equation intended to help establish a link between the individual and the population.

It is easy to underestimate the amount of computer programming that goes into computational work; in my experience, the lines of dirty, uncommented and variously macgyvered research code written for a typical computational article — simulations, data analysis, plotting routines, exploratory dead ends of all kinds — are numbered in the thousands. Of this amount, a few hundred lines typically survive in the form of the final product, but code is rarely included as part of published papers. To remedy this situation somewhat, the Appendix at the end of this dissertation contains Fortran code written for the simulations of Paper IV.
2

Acquisition

If processes of language change at the population level reduce to the level of the individ-
ual, we first need to ask how language acquisition works. Currently available evidence
suggests exploring the probabilistic framework of Variational Learning in more detail.
Assuming that learners are on the whole reliable in their process of language acquisition,
equations may be derived which relate the state of the population at time $t + 1$ to that at
time $t$. This forms the basis for a number of extensions studied in the following chapters.

2.1 The Z of language change

The reductionist philosophy sketched in Chapter 1 insists that properties of language
change be derived from properties of individuals and interactions between individuals
—in particular, this philosophy denies any explanation of change in terms of language
in the “reified sense”, as an entity existing independently of knowledge structures and
states internalized by speakers. Logically, a reductive explanation of language change
may look to at least two different sources of change: the way individuals acquire
language, and the way individuals use language. It is also possible that to explain
some phenomena of variation and change, recourse to neither source is sufficient on
its own — this theme, in fact, will form the central tenet of Paper I and will be revisited
a number of times in what follows.

The acquisition–use divide runs along well-established trench lines, with (dia-
chronically-minded) generative linguists often settling on the former side, and people
working in exemplar theory or other “cognitive”\(^1\) (Bybee, 2015) frameworks on the
latter. Usage-based models of change will receive substantial treatment in Chapter
5. In the present chapter, my focus will be on acquisition as the motor of change,

\(^1\)This is abuse of nomenclature, to some extent. A run-of-the-mill generative account of acquisition
and change, for example, is of course just as committed to the existence and operation of cognitive
systems as are frameworks based on exemplar theory, statistical learning, and so on.
The “Z-model” of language change (adapted from Andersen, 1973).

although as already suggested, Paper I will add a substantial usage component to the picture.

The idea that language change reduces to language acquisition goes back a long time: it was, in fact, already part and parcel of the theorizing of some 19th-century Neogrammarians (Lightfoot, 1999, 18–19). An oft-cited early exemplar in modern linguistics is Andersen (1973), in which change is conceptualized in terms of a “Z-model” (Figure 2.1).² A generation of speakers $t$ has internalized grammar $G_t$ — an I-language, to use Chomsky’s (1986) term. This grammar generates linguistic output (E-language) which, in turn, acts as input (primary linguistic data; PLD) to the language acquisition process of the following generation, $t+1$. Insofar as this acquisition process is reliable, the new generation acquires the same grammar, $G_{t+1} = G_t$, and there is no change. Due to a number of reasons, however, such as an external trigger that shifts the distribution of input received by the $(t+1)$th generation, this generation may set one or more grammatical Parameters³ differently from the parent generation. Change ensues: $G_{t+1} \neq G_t$.

In diachronic syntax, an influential early model of change along the above lines was put forward by Niyogi and Berwick (1997; also see Niyogi & Berwick, 1996, 2009; Niyogi, 2002, 2006). This model reduces syntactic change to syntactic acquisition by assuming that language learners set the values of syntactic Parameters along the

---

²Though Andersen (1973) does not use the term “Z-model”, which is a more recent label possibly originating in lecture notes circulated at the University of Cambridge in 2011 (George Walkden, p.c.). Since it is useful for useful things to have names, I will stick to speaking of the Z-model throughout this thesis. It should also be noted that adopting the Z-model does not amount to a denial of usage effects, either for Andersen in particular, or in linguistic theory more generally: language use of course factors into the PLD a generation of acquirers receive, and this may serve to trigger change. This theme will be developed in more detail throughout the thesis.

³In order to avoid confusion with 'parameter' in the dynamical-systems sense of model parameter, I will capitalize linguistic Parameters throughout this Introduction as well as in the Conclusion. Usage in the four original Papers is variable.
lines of the Triggering Learning Algorithm (TLA) of Gibson and Wexler (1994). This algorithm runs as follows (cf. Figure 2.2):

**Algorithm 2.1 (TLA; Gibson & Wexler, 1994).**

Suppose each possible grammar of human language is a vector \( G = (g_1, \ldots, g_n) \) of \( n \) binary Parameters, \( g_i \in \{0, 1\} \). Assume the learner starts with a randomly chosen grammar vector \( G \).

1. Learner receives an input sentence \( x \).
2. If the learner’s current grammar \( G \) parses \( x \), goto step 1; otherwise continue.
3. If the learner’s current grammar \( G \) fails to parse \( x \), the learner executes single-step hill-climbing:
   (a) The learner picks one Parameter \( g_k \) at random and flips its value (0 to 1; 1 to 0). Write \( G' \) for the grammar with the flipped Parameter value.
   (b) If \( G' \) parses \( x \), set \( G' \) as the learner’s current grammar and goto step 1; otherwise, retain the unflipped \( G \) and goto step 1.

If the input sentences \( x \) are generated by a unique target grammar, it is easily seen that a TLA learner converges to that target grammar in the limit; in other words, as the number of learning iterations tends to infinity, the probability of the learner choosing the target grammar tends to 1. Diachronically interesting behaviour occurs when (i) learning is halted after a finite time, or when (ii) input to the learner is generated by at least two different parental grammars with non-identical extensions (weak generative capacities). In both of these situations, the learner may end up with a grammar different from that of the parent generation, a case of inter-generational change.

As a memoryless algorithm, the TLA yields to an exact analysis in terms of Markov chains, and Niyogi and Berwick (1997) carry out a systematic investigation of their diachronic model in some of its special cases. Notably, they show that one kind of S-curve (the logistic function) arises when TLA learners are arranged in a sequence of non-overlapping generations — though the mathematical argument only covers the trivial case of a one-Parameter system and a critical period of two (sic!) input sentences. In more realistic simulations of a three-Parameter system with varying sizes of critical period, Niyogi and Berwick (1997) model the evolution of the verb-second (V2) Parameter, specifically the disappearance of V2 from languages such as French and English (cf. R. Clark & Roberts, 1993). Here the modelling has a surprising outcome: it is found that, given the Gibson–Wexler Parameterization (Gibson & Wexler, 1994) and assuming TLA, V2 languages are found to be diachronically stable, i.e. loss of V2 is predicted to be impossible, contra empirical facts (Niyogi & Berwick, 1997).
Given the TLA, then, V2 grammars are local optima in the learner’s search space and consequently become sinks for the diachronic evolution of the entire population.\(^4\) As both Yang (2002b, 18–20) and Mitchener (2006) stress, the learner’s transition to the optimum may also be inordinately abrupt: in extreme cases, hearing just one sentence which is unambiguous evidence for V2 can lead the learner to fix onto a V2 grammar forever, irrespective of the preceding learning trajectory. Clearly, a human learner does more than just randomly flip Parameter values upon parsing failure, and the TLA can thus be criticized both on the grounds that it gives the wrong diachronic predictions (stability of V2) and, relatedly but independently, on grounds of lack of psycholinguistic realism.

For these kinds of reasons, researchers have turned to alternative ways of mod-

\(^4\)Niyogi and Berwick (1997) also explore a number of variations of the TLA as well as another set of Parameters, but the outcome is always that V2 languages are diachronically more stable than non-V2 languages.
elling acquisition and change. Mitchener (2006), for example, proposes a mathematical formalization of Lightfoot’s (1999) cue-based learner in an explicit attempt to rectify the TLA’s inability to explain phenomena such as the loss of V2. The cue-based learner differs from the TLA learner in having a memory: cues are understood as elements of I-language which the learner scans his or her PLD evidence for. Cues are assumed to have the ability to fix the learner onto a unique grammar, but only if the evidence for them exceeds a certain threshold. This, then, requires the learner to keep track of batches of input sentences. Mitchener (2006) shows that a dynamical system composed of iterated cue-based learners can correctly model the loss of V2, in the sense that the system features a bifurcation which reverses the stability of V2 grammars under suitable model parameter variation, for example due to dialect contact, as is thought to have happened when V2 was lost in Middle English (Lightfoot, 1999, 151–158; for an alternative explanation grounded in morphological erosion, see Fischer, van Kemenade, Koopman & van der Wurff, 2000, ch. 4).

2.2 Variational Learning

Both the TLA and the cue-based learner are firmly entrenched within the classical generative tradition which views grammatical competence as monolithic (Chomsky, 1986). Under this view, the purpose of the process of language acquisition is to consolidate a grammar in the mind of the speaker which then, for example, allows the speaker to make grammaticality judgements based on native intuition. Whatever variability is observed in the actual use of language by speakers (or whatever uncertainty speakers may evince when making grammaticality judgements) is relegated to the realm of performance factors, which are only thought to filter the output of the cognitive module that is responsible for grammatical competence.

One may wonder how useful this relegation of variation to a supposedly language-external performance is: multilingualism, diglossia and codeswitching are all widespread empirical phenomena which need to be accounted for in both their synchronic and diachronic manifestations (Figure 2.3). Given that under the standard Z-model (Figure 2.1) language change is an interplay of I-language and E-language, it would appear important for linguistic theory to explain how the aforementioned performance effects act on both acquisition and production. While nothing in principle prevents one from developing such an account within the monolithic tradition, to my knowledge this possibility has not been explored and certainly not formalized in terms of a model that made the requisite interrelations explicit.

An alternative is to maintain that language acquisition (as well as language use) is intrinsically variable and probabilistic. This is the line taken in the Variational Learning (VL) framework, introduced by Charles Yang in a sequence of publications (Yang,
“While the goats are being milked, and such other refreshments are preparing for us as the place affords”

(Landor, *Imaginary Conversations*, 1829, quoted in Denison, 1999, 153)

Intrasentential syntactic codeswitching in historical English: an example of the coexistence of both the archaic passival (*are preparing*) and the modern progressive passive (*are being milked*) in one sentence.

The VL framework takes issue with the monolithic tradition by allowing language learners to operate with probability distributions over the set of human grammars (which set, we should note, is still constrained by Universal Grammar; thus the variation space over which the relevant probability distributions are defined is fixed by innate factors). Variability is not seen as a performance issue, then, but rather as something endemic to acquisition itself.

VL could in principle be implemented in various different ways, but in practice most work in this domain has assumed that language learners arrive at grammar probabilities via a specific type of reinforcement learning known as *linear reward–penalty learning*, henceforth LRP learning (Bush & Mosteller, 1955). This is a domain-neutral algorithm for a cognitive agent (or a learning automaton; Narendra & Thathachar, 1989) to decide which actions to choose in response to stimuli, given a history of operant conditioning in the Skinnerian sense. Positive feedback tends to increase the probability of an action’s being chosen, while negative feedback decreases that probability. In language acquisition, the stimulus consists of relevant PLD (for the acquisition of syntax, it is usually taken to consist of unembedded sentences generated by one or more parental grammars); the actions of the learning agent correspond to the selection of a grammar among the set of possible grammars; positive feedback corresponds to successful parsing of an input token; negative feedback represents parsing failure. The learner starts with a neutral hypothesis (the probability that any given Parameter is set on is 0.5), receives input, and adjusts Parameter probabilities in response to parsing success and parsing failure. These adjustments are modulated by a *learning rate parameter* $\gamma$ which controls the speed and sensitivity of learning, with low values of $\gamma$ corresponding to small adjustments and hence slow but smooth learning trajectories, high values corresponding to large adjustments and more jagged learning paths.

Since most work on VL has focussed on the one-Parameter (two-grammar) special case, I will here limit my discussion to that case and set the general case aside until
Chapter 4. The algorithm then reads (cf. Figure 2.4):

**Algorithm 2.2 (LRP, 1 Parameter; Narendra & Thathachar, 1989, 110).**
Let $\pi$ denote the probability ($0 \leq \pi \leq 1$) with which the Parameter is set on. Let $\pi = 0.5$ initially, and assume a fixed learning rate $0 < \gamma < 1$.

1. Receive input sentence $x$.
2. Learner sets the Parameter on with probability $\pi$.
   
   (a) Assume Parameter was set on.
   
   i. If the resulting grammar parses $x$, the probability $\pi$ is increased: $\pi \leftarrow \pi + \gamma (1 - \pi)$.
   
   ii. If the resulting grammar fails to parse $x$, the probability $\pi$ is decreased: $\pi \leftarrow (1 - \gamma) \pi$.

   (b) Assume Parameter was set off.
   
   i. If the resulting grammar parses $x$, the probability $\pi$ is decreased: $\pi \leftarrow (1 - \gamma) \pi$.
   
   ii. If the resulting grammar fails to parse $x$, the probability $\pi$ is increased: $\pi \leftarrow \pi + \gamma (1 - \pi)$.
3. Repeat steps 1–2.

A central result of reinforcement learning theory is that a two-action (or, one-Parameter) agent following LRP learning converges towards an expected value of the probability $\pi$. This expected value can be expressed in terms of the penalty probabilities associated with each action, or each setting of the binary Parameter. The penalty probability for the “Parameter on” setting, $c_+$, is the probability of the learner encountering a sentence which that setting cannot parse:

$$c_+ = \text{Prob}\{x : x \text{ is not parsed when Parameter on}\}. \quad (2.1)$$

Symmetrically, the penalty probability for the “Parameter off” setting is

$$c_- = \text{Prob}\{x : x \text{ is not parsed when Parameter off}\}. \quad (2.2)$$

It can be shown that the expected value of $\pi$, the value towards which the learning agent tends if supplied with a small learning rate and allowed infinite time to learn, is given by the ratio $c_-/(c_+ + c_-)$:

**Theorem 2.1 (Narendra & Thathachar, 1989, sec. 5.3).**

Let $T$ stand for the number of input tokens heard by a one-Parameter LRP learner. Then the probability $\pi$ tends to a normal distribution with mean

$$\hat{\pi} = \frac{c_-}{c_+ + c_-} \quad (2.3)$$
initialize: \( \pi = 0.5 \)

receive input sentence \( x \)

set Parameter on with probability \( \pi \)

set \( \pi \leftarrow \pi + \gamma (1 - \pi) \)

set \( \pi \leftarrow (1 - \gamma)\pi \)

x parsed?

Parameter on?

set \( \pi \leftarrow (1 - \gamma)\pi \)

set \( \pi \leftarrow \pi + \gamma (1 - \pi) \)

no

no

yes

yes

\( x \) parsed?

\( x \) parsed?

\( x \) parsed?

yes

yes

no

no

\textbf{Figure 2.4}

LRP learning of one Parameter.

\textit{and zero variance, as} \( \gamma T \to \infty \) \textit{such that} \( \gamma \to 0 \).

Thus with long learning episodes and modest learning rates, the final value of the probability \( \pi \) attained by the learner is well approximated by \( c_-/(c_+ + c_-) \). In what follows, I will call such a learner \textit{reliable}.

Strictly speaking, this requires the penalty probabilities to be constant, corresponding to what in learning theory is referred to as a \textit{stationary environment}. Assuming such a stationary environment, we may outline some behaviours of LRP learning when applied to the problem of language acquisition. First, assume there is a unique target grammar — without loss of generality, assume that the “Parameter on” setting corresponds to this unique target. Then \( c_+ = 0 \). Consequently Theorem 2.1 implies that the learner tends towards the value \( \hat{\pi} = c_-/(c_+ + c_-) = 1 \) for the Parameter probability. Hence unique targets are, just as with the TLA (Section 2.1), learnable in the limit.

Again, more interesting behaviour ensues when the learner is exposed to a mix of grammars in his environment. Limiting our discussion to the one-Parameter case for the moment, assume the Parameter probability in the \textit{parent} generation (and thus in the learner’s environment) is \( p \). Assume, moreover, that some proportion \( a_+ \) of
sentences in the joint extension of the two grammars are uniquely parsed by the on setting, and some proportion $a_-$ uniquely parsed by the off setting (Figure 2.5). Then the probability of the learner encountering a sentence which is uniquely parsed by the on setting is $pa_+$, so that $c_- = pa_+$. Similarly, the penalty for the on setting is found to be $c_+ = (1 - p)a_-$. It follows that a reliable learner tends towards the limiting probability

$$\hat{p} = \frac{c_-}{c_+ + c_-} = \frac{pa_+}{(1 - p)a_- + pa_+}.$$  \hfill (2.4)

Figure 2.6 illustrates, comparing the behaviour of a reliable learner with that of an unreliable one.

If we now set a sequence of reliable learners in a diachronic progression of non-overlapping generations, generation $t$ serving as the environment for generation $t + 1$, the result is a dynamical system whose evolution is governed by the nonlinear difference equation

$$p(t + 1) = \frac{p(t)a_+}{(1 - p(t))a_- + p(t)a_+}.$$  \hfill (2.5)

It is not difficult to show that with this, the state $p = 1$ is attracting if $a_+ > a_-$, and that $p = 0$ is attracting if $a_+ < a_-$:

**Theorem 2.2 (adapted from Yang, 2000, 239).**

In LRP learning of one Parameter, the state $p = 1$ is an asymptotically stable rest point and the state $p = 0$ an unstable one (and hence the grammar with the Parameter set on wins out in diachrony) if $a_+ > a_-$. The state $p = 0$ is an asymptotically stable rest point and the state $p = 1$ an unstable one (and the grammar with the Parameter set off wins) if $a_+ < a_-.$
Learning trajectories ($T = 10^6$ learning steps) of a reliable ($\gamma = 0.0001$; blue) and an unreliable ($\gamma = 0.01$; red) LRP learner. For both learners, the advantage parameters are set at $a_+ = 0.2$ and $a_- = 0.1$. The blue learner operates in the environment $p = 0.9$, the red learner in the environment $p = 0.1$. The former thus tends towards the limiting Parameter probability $\hat{\pi} = (0.9 \cdot 0.2)/(0.1 \cdot 0.1 + 0.9 \cdot 0.2) \approx 0.95$, the latter towards the probability $\hat{\pi} = (0.1 \cdot 0.2)/(0.9 \cdot 0.1 + 0.1 \cdot 0.2) \approx 0.18$, illustrated here by the dashed lines (Theorem 2.1). The limiting variance, however, depends on the size of the learning rate $\gamma$, the red learner displaying high variance (hence, unreliable). The consequence is that a diachronic sequence of reliable learners may be approximated by a deterministic difference equation, whereas such an approximation would not be valid for a sequence of unreliable learners. (The horizontal axis is logarithmic, so as to better display the initial phase of convergence during the first $10^5$ learning steps.)
Importantly in light of the discussion to follow in Chapter 4, this result rules out the phenomenon of *stable variation*, where a state $0 < p < 1$ is asymptotically stable. Moreover, it can be shown that the evolution of (2.5) is logistic. Thus, Variational Learning is another mechanism giving rise to S-curves, alongside Niyogi & Berwick’s (1997) TLA-based model (see Section 2.1, above).

Having these desirable properties, the VL framework and LRP learning are increasingly being adopted as a mathematical model of change by variationists working under the umbrella of “competing grammars” (Kroch, 1994): for recent empirical applications of VL, see Ingason, Legate and Yang (2013) and Heycock and Wallenberg (2013). In the following chapter, I show how this framework may be extended to derive Kroch’s (1989) Constant Rate Effect, yielding additional support for Variational Learning.

\[\text{(2.5)}\]

3

The diachronic interplay of acquisition and use

Language acquisition and language use form a dynamical system over generations of speakers — a diachronic loop — in which one constantly feeds into the other. Extending the Variational Learning theory of acquisition with an account of post-acquisition production biases allows the derivation of the Constant Rate Effect, the observation that the contextual reflexes of an underlying Parametric change grow in parallel. This illustrates not only how diachronic phenomena may be derived from first principles, but also how mathematical modelling may be used to chart the possible behaviours of a linguistic system.

3.1 The Constant Rate Effect

The Constant Rate Effect (CRE) is the observation that when a linguistic option replaces another in a number of linguistic contexts, the rate of replacement is the same in each context. Formulated originally as a hypothesis by Kroch (1989), the CRE has since been demonstrated in a number of changes in a number of languages (see Table 3.1) and is thus a prime candidate for an orderly property in the sense outlined in Chapter 1. Although traditionally the CRE has been studied in the context of diachronic syntax, phonologists are also increasingly turning to it in an attempt to understand the diachronic predictions of competing theories of phonological representation and processing (Fruehwald, Gress-Wright & Wallenberg, 2009; Bermúdez-Otero, Baranowski, Bailey & Turton, 2015).

To take an example, Figure 3.1 illustrates the rise of periphrastic do in early Modern English, i.e. the increasing adoption of forms such as (3.1) over forms such as (3.2):

(3.1) I do not drink tea.
(3.2) I drink not tea.

This change affected a number of linguistic contexts, including negative declarative sentences and various types of interrogative sentence. However, as the data in Figure 3.1 preliminarily suggest, and as Kroch (1989) using statistical methods demonstrated, the adoption of periphrastic *do* proceeds at similar rates in all these contexts. This constant rate occurs despite the fact that, at any given moment in time, each context is located at a slightly different stage of the overall change, reflected as a temporal translation of the contextual curves. Kroch (1989) took these facts as diachronic support for the view that the underlying cause of the change was a Parameter switch (loss of V-to-T movement), and that the variability observed across contexts “reflects functional effects, discourse and processing, […] the strength of these effects [remaining] constant as the change proceeds” (Kroch, 1989, 238). The details of this development — which is in fact complicated by a number of factors\(^1\) — need not concern us here. What matters is the broad strokes of epistemology: diachronic patterns such as the CRE may be used to argue about theories of synchrony insofar as different synchronic theories make different diachronic predictions. For instance, the competing view that English periphrastic *do* adapted functionally to the different contexts in different ways (Bailey, 1973) is now largely discredited because of this evidence.

Kroch (1989) adopts the logistic function as a model of historical changes due to

---

\(^1\)Most famously, there is the complication of the positive declarative context (clauses such as “I do drink tea”), which I have glossed over in the above discussion. In this context, periphrastic *do* initially gains ground just as in the other contexts, but then declines abruptly in late 16th century, so that today clauses such as “I do drink tea” are rare and always carry an emphatic meaning. See Kroch (1989) and Postma (2010) for different explanations of this development.
the fact that it has the desired shape of an S. Importantly, the logistic function has two free parameters, $s$ and $k$:

$$ p(t) = \frac{1}{1 + e^{-s(t-k)}}. \quad (3.3) $$

Here, $p(t)$ is the frequency of either the innovative or the receding grammatical option (such as periphrastic $do$) as a function of time $t$. The parameter $s$ may be varied to increase or decrease the rate of the change; this parameter is then normally referred to as the *slope* of the logistic. The parameter $k$, in turn, controls the placement of the curve in the time domain: positive values of $k$ shift the curve towards positive time, negative values towards negative time; this parameter is known as the *intercept* of the logistic. The crucial insight is that, by varying $k$ but keeping $s$ fixed, a set of logistic curves can be made to model a CRE-like situation (see Figure 3.2, comparing with Figure 3.1). The standard operationalization of the CRE, then, is the following: if a set of contexts is well described by a set of logistics agreeing in their $s$ parameters but possibly varying in their $k$ parameters, what we have at hand is a CRE. This is the operationalization assumed in most work on the CRE up to date, though the precise manner of fitting a family of logistics to data varies from study to study.

An immediate problem with this operationalization has to do with specifying what, exactly, it means for some data to be “well described” by a set of curves and
In the standard operationalization, the CRE is modelled with a family of logistics (3.3) agreeing in their $s$ parameters but freely varying in their $k$ parameters (here $k = 4$ for the red curve, $k = 5$ for the black curve, $k = 6$ for the blue curve).

**Figure 3.2**

In the standard operationalization, the CRE is modelled with a family of logistics (3.3) agreeing in their $s$ parameters but freely varying in their $k$ parameters (here $k = 4$ for the red curve, $k = 5$ for the black curve, $k = 6$ for the blue curve).

when, exactly, a number of $s$ parameters are so close to each other as to “agree”. Paolillo (2011) argues that the variationist literature has not been able to put forward a satisfactory solution to this problem; consequently, he recommends forgoing the CRE altogether (also see Corley, 2014, for a number of further problems of the standard operationalization). Given the number of studies evincing the general pattern of the CRE (Table 3.1), however, it would seem premature to abandon study of the CRE just because a satisfactory solution to a *statistical* problem does not yet exist.

I submit that the standard operationalization involves another, more serious problem — one that can be overcome, however, as Paper I intends to show. Briefly, the problem is the following: under Kroch’s (1989) interpretation, a CRE is evidence of a single, unified underlying cause such as a Parameter switch filtered by extragrammatical contextual factors. However, the standard operationalization employs a family of logistic curves which are not bound to each other mathematically. Importantly, *arbitrary* variation in the $k$ parameters is allowed, in principle, leading to the absurd conclusion that, in extreme cases, one and the same change can go to completion in one context before it even takes off in another (Figure 3.3). Less grotesquely, but perhaps more importantly, there exists at the moment no theory of how and why a single Parameter switch, coupled with a number of contextual factors, should give rise to a family of logistics agreeing in their $s$ parameters but differing in their $k$ parameters. The standard operationalization is simply being used as a *descriptive proxy* to the data: it can be made to fit data, but this fact, crucially, does not
One problem of the standard operationalization of the CRE concerns unconstrained variation in \( k \) parameters. Thus, putative CREs include cases where, absurdly, the change goes to completion in one context even before it starts off in another.

suffice to establish a link between the empirical observation (the CRE as observed in diachronic data sets) and its intended explanation in terms of Parameter switches filtered by constant extra-grammatical factors: the CRE is never strictly deduced from an underlying mechanism (rather, a mechanism is suggested but the connection is never fleshed out in detail) and therefore not explained. This problem, which Paper I dubs the non-linking problem, then exposes a fundamental shortcoming in the epistemology of all CRE testing conducted under the standard operationalization.

The solution to this puzzle presented in Paper I builds upon Kroch’s (1989) original intuition — that a CRE arises whenever an underlying Parameter switch is modulated by constant extra-grammatical contextual factors — but supplies this intuition with a mathematical model of language acquisition and use that makes the relevant processes explicit. This model, then, links the purported mechanism and its diachronic outcome, the CRE, and presents one possible solution to the non-linking problem. What follows is a summary description of the model and its application to some real-world data; for full details of model definition and evaluation, see Paper I.

### 3.2 Production biases

In the VL framework, the \( t \)th generation internalizes a Parameter probability \( p(t) \) such that, with this probability \( p(t) \), a sentence is generated by these speakers with the “Parameter on” grammar (see Chapter 2). The acquisition process of a speaker
in the following, \((t + 1)\)th generation, then comes to depend on the value of \(p(t)\). As explained above (Section 2.2), the evolution of such a system is logistic with the “Parameter on” setting winning out in diachrony in the long term if and only if its parsing advantage is greater than that of the “Parameter off” setting: \(a_+ > a_-\). The reverse obtains when \(a_+ < a_-\).

This mechanism, in and of itself, cannot make sense of contextual effects as the notion of linguistic context is not built into the mechanism. An intuitive way of extending the model is to assume that the learner’s PLD is partitioned (in the usual set-theoretic sense) by a number of subsets \(C_1, \ldots, C_K\), each representing one linguistic context. Suppose, moreover, that speakers have differential preferences for using the novel Parameter setting in these contexts: some contexts favour the new setting, some contexts favour the old setting, and some contexts, possibly, are neutral with respect to the two settings. Formally, we associate with each context \(C_i\) a production bias \(b_i\), a constant real number, which filters the expression of the underlying Parameter in this context. Without loss of generality, I will here assume that positive values of \(b_i\) correspond to a favour for the on setting and negative values to a favour for the off setting, with \(b_i = 0\) indicating neutrality with respect to the underlying Parameter probability. Paper I defines this filtering as follows for context \(C_i\):

\[
p^{(i)}(t) = p(t) - b_i p(t) (1 - p(t))
\]  

(3.4)

where \(-1 \leq b_i \leq 1\); \(p^{(i)}(t)\) thus gives the probability of the “Parameter on” setting, not globally, but specifically in context \(C_i\). This definition is the most parsimonious one among a set of potential formulations satisfying the \(a priori\) requirement that the quantity \(p^{(i)}(t)\) remain, at all times, a probability (see Paper I, Theorem 2, for the technical details).

Now, suppose for the moment that \(p(t)\) in (3.4) is a logistic curve tracking the increasing use of the “Parameter on” setting and that application of the production biases \(b_i\) in generation \(t\) does not affect the value of \(p(t + 1)\). The biases \(b_i\) in equation (3.4) then parameterize a family of curves of the desired shape (Figure 3.4): the variable \(p\) tracks the “underlying” value of the Parameter probability, while positive biases \(b_i > 0\) produce “higher” contextual curves and negative biases \(b_i < 0\) “lower” ones. It should be noted that these curves, while in some sense “parallel”, are no longer logistic (except in the degenerate case \(b_i = 0\)); consequently, while the production bias hypothesis derives a CRE qualitatively, it is, strictly speaking, at odds with the standard operationalization via sets of logistics. More discussion of this fact — which affects the way in which the production bias model can be fit to data — follows below.

Although Figure 3.4 suggests that filtering an underlying Parameter probability through a set of biases does produce something like a CRE, the assumption that appli-
Assuming that application of production biases $b_i$ does not change the value of the underlying Parametric probability $p$, and given that $p$ has a definite form (such as the logistic function here), a family of contextual curves is generated, parameterized by $b_i$ (from equation 3.4). This "phenomenological CRE" is the first indication that the production bias mechanism may be able to derive the effect from first principles.

For this, we need to figure out exactly how application of the biases $b_i$ in generation $t$ affects the penalty probabilities in generation $t + 1$. Assuming that the proportion of input falling in context $C_i$ is $\lambda_i$, and that this proportion stays constant over time, the penalty probabilities $c_+$ and $c_-$ for the following generation can be calculated from (3.4). When this is done (see Paper I), it is found that as long as learners are reliable (see Section 2.1), the diachrony of the extended model is described by the

2Thus assuming, for instance, that the probability of a question rather than a declarative or imperative sentence being uttered is the same in 2017 as it was in 1917. This much is probably uncontroversial — but future work could explore if partitions of the PLD into contexts which do not stay constant over relevant timescales exist, and if so, figure out how this affects the modelling.
Inter-generational change with an acquisition–use loop: the interaction of Variational Learning and production biases. The $t$th generation internalizes a Parameter probability $p(t)$ during acquisition. Sentences are then drawn as PLD for the following generation, but filtered through a set of linguistic contexts in which the value of $p(t)$ is modulated by production biases; here, the PLD forks into three contexts, represented by the three lines flowing from one generation to the next.

**Figure 3.5**

Inter-generational change with an acquisition–use loop: the interaction of Variational Learning and production biases. The $t$th generation internalizes a Parameter probability $p(t)$ during acquisition. Sentences are then drawn as PLD for the following generation, but filtered through a set of linguistic contexts in which the value of $p(t)$ is modulated by production biases; here, the PLD forks into three contexts, represented by the three lines flowing from one generation to the next.

![Diagram of inter-generational change](image)

inter-generational difference equation\(^3\)

\[
p(t + 1) = \left[1 + \frac{a_- (1 - p(t))(1 - Bp(t))}{a_+ p(t)(1 + B(1 - p(t)))}\right]^{-1}
\]  

(3.5)

where $B = \sum_{i=1}^{K} \lambda_i b_i$ and we assume, without loss of generality, that $0 < p(t) < 1$. This inter-generational dynamics can be represented as an adaptation of the original Z-model of change (Figure 3.5).

Paper I provides a full analysis of equation (3.5) and uncovers, among other things, a bifurcation which extends the result of Theorem 2.2. Briefly put, whereas in Yang’s (2000) original model the parsing advantages $a_+$ and $a_-$ are all that determine the diachronic outcome, in the extended model this outcome depends on both the ratio of the parsing advantages $a_-/a_+$ and the factor $B = \sum_{i=1}^{K} \lambda_i b_i$, which Paper I dubs net bias. With $B = 0$, the dynamics of Yang’s (2000) model are recovered; a negative net bias corresponds to a net favour for the “Parameter off” setting, while a positive net bias corresponds to a net favour for the “Parameter on” setting. With non-zero net biases, one of the following novel outcomes is possible:

---

\(^3\)Equation (14) of Paper I. Note that the $q_t$ of Paper I corresponds to $p(t)$ here.
Behaviour of the production bias model for two hypothetical cases of Parametric change (black dots) in three contexts $C_1$ (red circles), $C_2$ (blue crosses) and $C_3$ (green squares). A CRE is observed in both cases, but in (b) the change is slowed down as the negatively biased contexts outweigh the one positively biased context. The parameter values are as follows: in both cases, $\lambda_1 = 0.2$, $\lambda_2 = 0.4$ and $\lambda_3 = 0.4$. In (a), the biases are $b_1 = 1$, $b_2 = -1$ and $b_3 = 0.5$; the net bias is then $B = 0$. In (b), the biases are $b_1 = 1$, $b_2 = -1$ and $b_3 = -0.5$; the net bias consequently has a value of $B = -0.4$. (Paper I, Figure 7.)

1. The on setting wins out in diachrony even though $a_- > a_+$. 

2. The off setting wins out in diachrony even though $a_+ > a_-$. 

See Paper I, Theorem 3, for a full statement of this result, which has implications for hybrid accounts of change that bring together acquisition and usage effects; further discussion of this point will follow in Chapter 8.

Turning now to the Constant Rate Effect, Figure 3.6 illustrates the behaviour of the full production bias model in two cases of a hypothetical historical change involving three linguistic contexts. In both cases, we start with an initial state where speakers have the off setting for a Parameter, so that $p(0) = 0$. An external trigger actuates a change, shifting this probability to some $p(1) = \varepsilon > 0$. Thanks to the combination of the parsing advantages $a_+$ and $a_-$ and the net bias $B$, the “Parameter on” setting wins out in diachrony. Having furnished the inter-generational acquisition–use loop in a mathematically rigorous way, a CRE is still observed (cf. Figure 3.4): we find visible contextual separation, the amount of this separation depending on the values of the production bias parameters $b_i$. The two trajectories differ, however, in the overall speed of the change. In Figure 3.6b, negative biases (biases that favour the archaic “Parameter off” setting) outweigh the positive ones, and change is slowed down.
These two hypothetical changes — as well as cases where a negative net bias serves to block a change entirely (see Paper I, Theorem 3) — illustrate a general challenge faced by increasingly complex models of language change. When an increasing number of model parameters implies an increasing range of theoretically possible trajectories of change, how best to test the model against empirical (e.g. historical) data? In the present model, the concrete problem is that the complexity of the diachronic acquisition–use loop implies that the evolution of the underlying Parameter probability $p$ may not be logistic, or indeed may not be describable by any kind of closed-form curve. In Paper I, this problem is solved by an examination of the model’s behaviour under varying combinations of bias parameters and advantage ratios; it is found that the evolution of $p$ is still reasonably well approximated by a logistic function, except in a subset of the model parameter space corresponding to very negative net biases (see Paper I, sec. 3.2). Paper I thus assumes a logistic approximation $\tilde{p}(t)$ for the underlying Parametric change, and then derives contextual curves $\tilde{p}^{(i)}(t)$ from this underlying approximation using equation (3.4). Such sets of contextual curves may be fit to historical data using nonlinear least squares optimization, and Paper I presents three case studies (two syntactic, one phonological) in which the model is fit to data which previous literature has identified as attesting CREs. Figure 3.7 illustrates the fit to the case of periphrastic do in English.
The production bias model presented in Paper I thus represents one solution to the non-linking problem, by making Kroch’s (1989) intuition about contextual factors mathematically precise and by demonstrating that the CRE indeed follows from these first principles. Crucially, in this model the contextual curves \( \tilde{p}^{(i)}(t) \) are now linked to an underlying logistic \( \tilde{p}(t) \). As a consequence, the model places a strict upper bound on the extent of contextual variability that is possible in principle: in brief, it can be shown that the maximal time separation between two contexts of one underlying change is roughly 1.76 times the reciprocal of \(|s|\), where \( s \) is the slope of the underlying logistic \( \tilde{p}(t) \).

**Theorem 3.1 (Time Separation Theorem; Paper I, Theorem 4).**

The maximal time separation between any two contexts of one underlying logistic change proceeding at rate \( s \), quantified as the difference between the points in time at which these contexts reach frequency 0.5, is

\[
\Delta(s) = \frac{2}{|s|} \log \left( \frac{1}{\sqrt{2} - 1} \right) \approx 1.76 \frac{1}{|s|}.
\]

(3.6)

The model thus predicts that CREs must occur within a limited time window whose extent is dictated by the rate at which the change occurs overall; in particular, the sort of scenario depicted in Figure 3.3 in which a change goes to completion in one context before even beginning in another, is outlawed (Figure 3.8). The model can then be used to disprove putative CREs in which contextual time separations are too large (see Paper I, sec. 4.6, for an example).

More importantly, the strong, quantifiable predictions of the production bias model mean that it is now possible to test the viability of the Constant Rate Hypothesis in a manner which improves upon the standard operationalization (via sets of independent logistics) in two respects. Firstly, the production bias model provides a mechanism for the derivation of contextual curves from an underlying Parametric change, thus dissolving at least some of the mystery surrounding the original hypothesis (Kroch, 1989) that constant rates are to be expected whenever changes are unified by an underlying Parameter switch. A fit between empirical data and the production bias model thus constitutes stronger evidence than a fit conducted with the standard operationalization, since in the former case, but not in the latter, a specific mechanism is postulated. Secondly, the Time Separation Theorem provides an entirely new kind of test of the Constant Rate Hypothesis. If a given change were to attest time separations longer than those licensed by the theorem, our belief in the model (and possibly in the Constant Rate Hypothesis, too) would be diminished (as long, at least, as the recalcitrant observation cannot be explained away otherwise, e.g. by reference to poor data resolution). On the other hand, the more datasets continue to conform to
the predictions of the Time Separation Theorem, the more the production bias model, and by extension the Constant Rate Hypothesis, is corroborated. An immediate direction for future research, then, is to test the model on more datasets than the three considered in Paper I.

Finally, it should be noted that while Paper I has explored the production bias hypothesis within the remit of the VL framework, the production bias mechanism could in principle be attached to any model of change that provides an underlying probability as the relevant dynamic variable. The diachronic feedback loop between the production bias mechanism and the underlying change mechanism — of which Paper I has given a full account in the case of VL — will have to be studied separately for each different combination of mechanisms. The central intuition behind the production bias hypothesis, however, remains: Constant Rate Effects arise whenever constant, context-bound production biases impinge on an underlying change.
4
Acquisition and change in higher-dimensional spaces

Explaining established empirical phenomena of linguistic diachrony is not the only purpose of mathematical modelling: generalizations of existing models and the pure mathematical analysis of these generalizations can chart the epistemic boundaries of modelling and make available predictions for future empirical work to refute or corroborate. Diachronic stability of variation is shown to be one such prediction of multidimensional linear reward–penalty learning.

4.1 Multidimensional competition

Our discussion of Variational Learning so far has assumed change in just one Parameter — as is the case with most literature on the subject. While it appears to be the case that many linguistic changes involve the resetting of no more than one Parameter at a time, the fact remains that each learner still has the entire set of humanly possible grammars as search space during acquisition, and the one-Parameter LRP procedure (Algorithm 2.2 and Figure 2.4) remains a simplification of the real-life facts of language acquisition. Fortunately, VL can be generalized for multiple Parameters in multiple ways, the simplest of which is to employ the multiple-action LRP learning procedure.\(^1\) The generalization is straightforward. Instead of two actions for the learning agent — one Parameter which to set, therefore two grammars which to choose between — we assume \(n\) actions instead. The learner thus attaches a probability to each of \(n\) grammars \(G_1, \ldots, G_n\). At every iteration of learning, the learner

\(^1\)Yang (2000) does not explore this generalization, but rather settles for the one-Parameter special case as a without-loss-of-generality approximation. Below, I show that a loss of generality is, in fact, implied. Yang (2002b) explores a different kind of generalization entirely, the so-called Naive Parameter Learner — for some discussion of this approach, see Section 8 of Paper II.
picks one of the \( n \) grammars and attempts to parse an input token. With successful parsing, the probability of the selected grammar is increased as in the one-Parameter version; with parsing failure, the probability of each of the remaining \( n - 1 \) grammars is increased instead (Figure 4.1):

**Algorithm 4.1 (LRP, Narendra & Thathachar, 1989, 116–117).**

Let \( \pi_i \) denote the probability \((0 \leq \pi_i \leq 1)\) of grammar \( G_i \) \((i = 1, \ldots, n)\). Let \( \pi_i = 1/n \) initially, and assume a fixed learning rate \( 0 < \gamma < 1 \).

1. Receive input sentence \( x \).
2. Learner picks grammar \( G_i \) with probability \( \pi_i \). Suppose \( G_k \) was picked.
   
   (a) If \( G_k \) parses \( x \), the value of \( \pi_k \) is increased and the value of \( \pi_j, j \neq k \), decreased: \( \pi_k \leftarrow \pi_k + \gamma(1 - \pi_k) \) and \( \pi_j \leftarrow (1 - \gamma)\pi_j \).
   
   (b) If \( G_k \) does not parse \( x \), the value of \( \pi_k \) is decreased and the value of \( \pi_j, j \neq k \), increased: \( \pi_k \leftarrow (1 - \gamma)\pi_k \) and \( \pi_j \leftarrow \frac{\gamma}{n-1} + (1 - \gamma)\pi_j \).

Note that this algorithm reduces to one-Parameter LRP learning (Algorithm 2.2) in the \( n = 2 \) case, assuming (for example) that \( G_1 \) corresponds to the “Parameter on” grammar and \( G_2 \) to the “Parameter off” grammar.

The learner’s knowledge is now represented by a vector \( \pi = (\pi_1, \ldots, \pi_n) \) in the \( n \)-dimensional unit hypercube of grammar probabilities. Again, it is easily seen that if a target grammar exists that uniquely parses all input in the learner’s environment, then, given infinite time, the learner will tend to the basis vector at which the target grammar’s probability is 1. If the learner’s input is mixed, so that some part of it is uniquely parsed by some grammar(s) and some other part by some other grammar(s), the algorithm again has more interesting behaviour. In this chapter, I will summarize the results of Paper II, in which the diachronic dynamics of such multidimensional grammar competition is solved for a number of cases, assuming reliable learners.

Central in this is the following generalization of Theorem 2.1:

**Theorem 4.1 (Narendra & Thathachar, 1989, sec. 5.32).**

Let \( T \) be the number of input tokens encountered by an \( n \)-grammar LRP learner. Then \( \pi = (\pi_1, \ldots, \pi_n) \) tends towards a multivariate normal distribution with mean \( \hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n) \),

\[
\hat{\pi}_i = \frac{\prod_{j=1, j \neq i}^n c_j}{\sum_{j=1}^n \prod_{k=1, k \neq j}^n c_k}
\]  

(4.1)

for \( i = 1, \ldots, n \), and a null covariance matrix, as \( \gamma T \to \infty \) such that \( \gamma \to 0 \).
Thus, akin to the one-Parameter (two-grammar) case, a reliable learner will tend to
a known expected value of the grammar probability vector \( \pi = (\pi_1, \ldots, \pi_n) \). If we
knew the relevant penalty probabilities, we could then outline the diachronic, inter-
generational dynamics of the system, just as in the one-Parameter case.

Is it possible to determine penalty probabilities in higher dimensions? Figure 4.2
gives the relevant Venn diagram for the general three-dimensional case, illustrating
the added complexity that results even at the move from two to three grammars.
Using the notation of this diagram, it is easy to see that the penalty probability of
grammar \( G_1 \), for example, is

\[
c_1 = \alpha_{[2]} p_2 + \alpha_{[3]} p_3 + \alpha_{[2,3]} (p_2 + p_3) = (\alpha_{[2]} + \alpha_{[2,3]}) p_2 + (\alpha_{[3]} + \alpha_{[2,3]}) p_3.
\]

\[ (4.2) \]

Writing \( a_{12} = \alpha_{[2]} + \alpha_{[2,3]} \) and \( a_{13} = \alpha_{[3]} + \alpha_{[2,3]} \), we find that \( a_{12} \) gives the relative
advantage of \( G_2 \) over \( G_1 \) — the probability of a sentence parsed by \( G_2 \) but not by \( G_1 \) —

\[ \text{Figure 4.1} \]
LRP learning of \( n \) grammars.
and $a_{13}$ similarly gives the relative advantage of $G_3$ over $G_1$. Proceeding analogously for grammars $G_2$ and $G_3$, one finds the set of penalty probabilities

\[
\begin{align*}
    c_1 &= a_{12}p_2 + a_{13}p_3 \\
    c_2 &= a_{21}p_1 + a_{23}p_3 \\
    c_3 &= a_{31}p_1 + a_{32}p_2
\end{align*}
\]

where each $a_{ij}$ gives the relative advantage of $G_j$ over $G_i$, that is, the probability of the learner encountering a sentence which is parsed by $G_j$ but not by $G_i$.

It is possible (though somewhat mentally taxing) to extend this reasoning to $n$ dimensions — one has to consider all the possible intersections in the corresponding Venn diagram, the number of which grows superlinearly with $n$. We may consider the four-grammar case as an intermediate step. For $X \subseteq \{1, 2, 3, 4\}$, let $\alpha_X$ denote the probability of a sentence parsed by each grammar $G_i$ such that $i \in X$ but not parsed by any grammar $G_j$ such that $j \notin X$. Then, using the above reasoning, the penalty probability of grammar $G_1$ is found to be

\[
c_1 = \alpha_{\{2\}}p_2 + \alpha_{\{3\}}p_3 + \alpha_{\{4\}}p_4 + \\
     + \alpha_{\{2,3\}}(p_2 + p_3) + \alpha_{\{3,4\}}(p_3 + p_4) + \alpha_{\{2,4\}}(p_2 + p_4) + \\
     + \alpha_{\{2,3,4\}}(p_2 + p_3 + p_4)
\]

\[
= \alpha_{\{2\}} + \alpha_{\{2,3\}} + \alpha_{\{2,4\}} + \alpha_{\{2,3,4\}}p_2 + \\
    + \alpha_{\{3\}} + \alpha_{\{2,3\}} + \alpha_{\{3,4\}} + \alpha_{\{2,3,4\}}p_3 + \\
    + \alpha_{\{4\}} + \alpha_{\{3,4\}} + \alpha_{\{2,4\}} + \alpha_{\{2,3,4\}}p_4
\]

\[
= a_{12}p_2 + a_{13}p_3 + a_{14}p_4 \\
= a_{11}p_1 + a_{12}p_2 + a_{13}p_3 + a_{14}p_4,
\]

\[\text{(4.4)}\]
where the last equality follows from the fact that $a_{ii} = 0$ for any $i$ (logically, no grammar both parses and does not parse one and the same sentence). This suggests that in the general, $n$-grammar case, the penalty probabilities are given by

$$c_i = \sum_{j=1}^{n} a_{ij}p_j = \sum_{j=1}^{n} \sum_{X_{ij}} \alpha_{X_{ij}}p_j,$$

(4.5)

where $X_{ij}$ ranges over

$$\{Y \subseteq \{1, \ldots, n\} \setminus \{i\} : j \in X\},$$

(4.6)

the set of subsets of $\{1, \ldots, n\} \setminus \{i\}$ such that $j$ is a member of that subset.

The upshot is that, in a multidimensional system, the learner’s expected limiting value of the grammar probability vector $\pi$ is determined by two things: (i) the set of relative advantage quantities $a_{ij}$ and (ii) the probabilities $p_i$ with which the different grammars are used in the learner’s environment. It is then useful to collect the latter in a vector, $p = (p_1, \ldots, p_n)$, and the former in a matrix,

$$A = \begin{bmatrix}
0 & a_{12} & \cdots & a_{1n} \\
a_{21} & 0 & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \cdots & 0
\end{bmatrix}.$$

(4.7)

I shall refer to matrices of the form (4.7) as advantage matrices. In the general $n$-grammar case, an advantage matrix is thus an $n \times n$ square, zero-diagonal matrix of real numbers $0 \leq a_{ij} \leq 1$. Since the $\alpha_X$ quantities which make up the relative advantages $a_{ij}$ moreover have to sum to unity (cf. Figure 4.2), an additional well-definedness condition is imposed on advantage matrices; I will set this detail, which is discussed in Section 3 of Paper II, aside here. Having reanalysed the $p_i$ and $a_{ij}$ in terms of matrix algebra, and interpreting vectors as $n \times 1$ matrices as is customary, we may now note that the penalty $c_i$ of the $i$th grammar may actually be written as the $i$th row of the matrix product $Ap$:

$$c_i = \sum_{j=1}^{n} a_{ij}p_j = [Ap]_i = \begin{bmatrix}
a_{11}p_1 + a_{12}p_2 + \cdots + a_{1n}p_n \\
a_{21}p_1 + a_{22}p_2 + \cdots + a_{2n}p_n \\
\vdots \\
a_{n1}p_1 + a_{n2}p_2 + \cdots + a_{nn}p_n
\end{bmatrix}_i.$$

(4.8)

To summarize our discussion of multidimensional grammar competition so far, a learner following Algorithm 4.1 is exposed to PLD generated from an environment vector $p = (p_1, \ldots, p_n)$ of grammar probabilities, subject to an advantage matrix $A$ of relative parsing advantages between pairs of grammars. The latter is constant, re-
reflecting formal properties of the competing grammars (more specifically, intersection properties of the extensions, \textit{qua} weak generative capacities, of these grammars). The former, on the other hand, may change from generation to generation. From Theorem 4.1, a reliable learner will tend toward the limiting probability

\[ \hat{\pi}_i = \frac{\prod_{j \neq i} c_j}{\sum_j \prod_{k \neq j} c_k} \]  

(4.9)

for the \(i\)th grammar, and correspondingly toward the limiting vector \(\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)\) of grammar probabilities. This, then, becomes the \(p\) from which PLD to the following generation is drawn. The penalty probabilities \(c_i\) are linear combinations of \(p\) and \(A\):

\[ c_i = [A p]_i. \]  

(4.10)

Consequently, the inter-generational dynamics of a non-overlapping\(^3\) sequence of such learners depends on the composition of the advantage matrix \(A\) only, subject to the following difference equation:

\[ p_i(t + 1) = \frac{\prod_{j \neq i} [A p(t)]_j}{\sum_j \prod_{k \neq j} [A p(t)]_k} \quad (i = 1, \ldots, n). \]  

(4.11)

Writing \(D[p(t)] = \sum_j \prod_{k \neq j} [A p(t)]_k\) for the denominator (which is independent of \(i\)), and noting that an \(n\)-dimensional system of probabilities has \(n - 1\) degrees of freedom only, examination of the three-dimensional case, in particular, reduces to a study of the pair of equations

\[
\begin{align*}
\{ p_1(t + 1) &= [A p(t)]_2 [A p(t)]_3 / D[p(t)] \\
\{ p_2(t + 1) &= [A p(t)]_1 [A p(t)]_3 / D[p(t)] 
\end{align*}
\]  

(4.12)

with \(p_3 = 1 - p_1 - p_2\). Figure 4.3 gives three sample learning trajectories for varying learning rates \(\gamma\), illustrating the convergence of reliable learners to the deterministic limit (4.12) for one selection of advantage matrix \(A\).

In extreme cases, equation (4.11) may invite division by zero and thus be ill-defined; on the other hand, \(D[p] \neq 0\) whenever \(a_{ij} \neq 0\) for \(i \neq j\) (Paper II, Corollary to Theorem 3). Paper II thus limits discussion to the latter case, calling such advantage matrices \textit{proper}, and notes that improper matrices — corresponding to systems in which subset–superset grammar relations exist — need to be studied separately.

\(^3\)In line with the VL tradition, I assume non-overlapping generations here. Continuously overlapping generations can be explored by taking the continuous-time limit of the relevant difference equations and working with the resulting differential equations; I will, in fact, explore this option in Chapter 7 in an effort to shed light on the analogy between VL and certain models in evolutionary biology. As far as VL is concerned, the qualitative dynamics (equilibria and their stability) are the same for both the discrete and the continuous formulation, so that the choice of formulation is mostly a matter of taste.
Learning trajectories ($T = 10^6$ learning steps) of a reliable ($\gamma = 0.0001$; blue) and two unreliable learners (green and red; $\gamma = 0.001$ and $\gamma = 0.01$, respectively), illustrated on a ternary plot; for each learner, we set $a_{ij} = 0.1$ for $i \neq j$. Top: entire trajectories; bottom: final 100,000 learning steps. The blue learner operates in the environment $p = (0.9, 0.1, 0.0)$, the green learner in the environment $p = (0.1, 0.9, 0.0)$, and the red learner in the environment $p = (0.0, 0.1, 0.9)$. The three limiting vectors $\hat{\pi}$ (from Theorem 4.1) are shown by the three crosses. Increasing the learning rate $\gamma$ leads to more noise and more variance about $\hat{\pi}$ after initial transients have died out. Diachronic sequences of multidimensional LRP learners may thus be studied with difference equations in the deterministic limit (4.11), but only if learners are assumed to be reliable.
4.2 Stable variation

The state of a diachronic system of $n$-grammar LRP learners, $p$, is defined on the simplex

$$S_n = \{ p = (p_1, \ldots, p_n) : 0 \leq p_i \leq 1 \text{ and } \sum p_i = 1 \}.$$  \hspace{1cm} (4.13)

Above, we have seen that the evolution of such a system is parameterized by an advantage matrix $A$, an $n \times n$ square matrix that supplies relative pairwise parsing advantages. We may now ask: what kinds of trajectories are possible for $p$ on the simplex $S_n$? Where are the system’s rest points located, what is their stability, and how does all of this depend on the composition of the matrix $A$?\footnote{For a definition of the crucial dynamical-systems-theoretic notions of rest point and (asymptotic) stability, see Paper II, sec. 1.}

The work presented in Paper II points to a surprising conclusion: that in dimensions $n \geq 3$, all interior rest points are diachronically stable. Although all the special cases studied in Paper II conform to this hypothesis, its proof has been elusive in the general case, and I therefore put it forward here as a conjecture, to be confirmed by future work:

**Conjecture 4.1.**

For any $n$-dimensional diachronic sequence of LRP learners (i.e. for any $n \times n$ advantage matrix $A$):

1. The system has either $n$ or $n+1$ rest points. If $n$, then these are the vertices of the simplex $S_n$. If $n+1$, then these are the vertices plus one interior rest point $p^* \in \text{int}S_n = \{ p \in S_n : 0 < p_i < 1 \}$.

2. If the interior rest point exists, it is asymptotically stable.

It can, moreover, be shown that the antecedent of the implication in clause 2 of Conjecture 4.1 is true for several types of advantage matrix $A$ whenever $n \geq 3$. The upshot, then, is that some higher-dimensional systems of LRP learners attest stable variation, the phenomenon where a mixed grammar probability state $p \in \text{int}S_n$ is diachronically stable — in stark contrast to the two-dimensional case, in which such stable variation is ruled out by Theorem 2.2.

As already mentioned, Paper II assumes proper advantage matrices throughout. For these matrices, the following general result is established as a straightforward generalization of Theorems 4–5 of Paper II:

**Theorem 4.2.**

Assume $A$ is proper ($a_{ij} \neq 0$ whenever $i \neq j$). Then

1. the vertices $v_1 = (1, 0, \ldots, 0)$, $v_2 = (0, 1, 0, \ldots, 0)$, \ldots, $v_n = (0, \ldots, 0, 1)$ of the simplex are rest points;
2. no other point on the boundary \( \text{bd} S_n = S_n \setminus \text{int} S_n \) of the simplex is a rest point;

3. an interior point \( p = (p_1, \ldots, p_n) \in \text{int} S_n \) is a rest point if and only if \( c_1p_1 = c_2p_2 = \cdots = c_np_n \).

Lacking a solution to the (hopelessly nonlinear) general \( n \)-dimensional equation (4.11), however, it is difficult to establish further general results on the behaviour of these systems. In lieu of a general approach, one can then adopt two different strategies, or a combination of them: one can restrict attention to small \( n \), such as the case \( n = 3 \), or one can restrict attention to particular kinds of advantage matrix \( A \) with symmetry properties which reduce the number of free parameters of the system. Paper II adopts both strategies and proceeds to study three such classes of advantage matrix, starting from a consideration of systems in which all pairwise advantages \( a_{ij} \) are equal. These advantage matrices, then, are of the form

\[
A = \begin{bmatrix}
0 & a & \ldots & a \\
a & 0 & \ldots & a \\
\vdots & \vdots & \ddots & \vdots \\
a & a & \ldots & 0
\end{bmatrix}
\]  

for some fixed \( a \neq 0 \). I call them Babelian.

In Paper II (Theorem 6), Babelian systems are shown to have one and only one interior rest point, the maximum entropy state \( p = (1/n, \ldots, 1/n) \). Moreover, stability analysis via Jacobian matrices shows that in three dimensions, this interior rest point is asymptotically stable while the three vertices are all unstable (Paper II, Theorem 7). While I do not have a formal proof, this result intuitively carries over to \( n \) dimensions, and we may conclude that in Babelian systems, the diachronic tendency is always towards the maximum entropy state \( p = (1/n, \ldots, 1/n) \) in which each grammar is used with equal frequency, a state of stable variation. Figure 4.4 illustrates the \( n = 3 \) case.

Generalizing from Babelian systems, Paper II defines a symmetric system as one whose advantage matrix satisfies \( a_{ij} = a_{ji} \) for all \( i \neq j \). In three dimensions, in particular, one then has

\[
A = \begin{bmatrix}
0 & a & b \\
a & 0 & c \\
b & c & 0
\end{bmatrix}
\]  

with \( a, b, c \neq 0 \). Studying the three-dimensional case, Paper II finds that these systems,
too, have one and only one rest point in the simplex interior,
\[ \mathbf{p} = \begin{pmatrix} c/(a+b+c) & b/(a+b+c) & a/(a+b+c) \end{pmatrix}, \]  
and that this point is asymptotically stable. The vertices are unstable. The conclusion is that symmetric systems, too, tend diachronically toward states of stable variation, though now there is variation in the exact location of the rest point, parameterized by \( a, b \) and \( c \) (Figure 4.5).

A different sort of generalization results if one relaxes the strict symmetry requirement and considers advantage matrices with the following property: there exist constants \( a, b > 0 \) such that (i) \( a_{jj} = b \) for all \( j \neq 1 \), and (ii) for all \( j \neq 1 \), \( a_{kj} = a \) for all \( k \neq j \). In three dimensions,
\[ \mathbf{A} = \begin{bmatrix} 0 & a & a \\ b & 0 & a \\ b & a & 0 \end{bmatrix}. \]
Such advantage matrices I shall call \textit{quasi-Babelian}; they reduce to the Babelian case whenever \( a = b \).

Quasi-Babelian systems exhibit more complicated behaviour since, essentially, they allow grammar \( G_1 \) to have either smaller or greater net advantage than its competitors (note that the \( i \)th column of an advantage matrix may be regarded as a kind of fitness vector for the \( i \)th grammar). Paper II examines the three-dimensional case in detail, and finds that these systems have either three or four rest points: in addition

\[ \text{Figure 4.4} \]
Phase space plot of Babelian systems (\( n = 3 \)). The asterisks give a sample trajectory from the initial state \( \mathbf{p}_0 = (0.1, 0.9, 0.0) \); the line segments show the system’s gradient at various points across the simplex. (Paper II, Figure 6.)
Figure 4.5
Phase space plot of symmetric systems \((n = 3)\). Here, \(a = 0.05\), \(b = 0.01\) and \(c = 0.02\). (Paper II, Figure 8.)

to the three simplex vertices \(v_1\), \(v_2\) and \(v_3\), an interior rest point exists at

\[
p = \left( \frac{1}{5 - 2\rho}, \frac{2 - \rho}{5 - 2\rho}, \frac{2 - \rho}{5 - 2\rho} \right)
\]
whenever \(0 < \rho < 2\), where \(\rho = b/a\). This point coalesces with the vertex \(v_1\) when \(\rho = 2\), and does not exist for \(\rho > 2\). It follows that \(\rho\) is a bifurcation parameter controlling a bifurcation which affects both the number and the stability of rest points in the system. For \(\rho > 2\), the three vertices are the only rest points; \(v_1\), corresponding to absolute use of \(G_1\), is asymptotically stable, while \(v_2\) and \(v_3\) are unstable. At the critical bifurcation value \(\rho = 2\), the vertex \(v_1\) loses stability to the interior rest point which now emerges and exists for \(0 < \rho < 2\).

**Theorem 4.3 (Paper II, Theorem 9).**

Assume a quasi-Babelian three-grammar system with advantage ratio \(\rho = b/a\). Then

1. the vertex rest points \(v_2 = (0, 1, 0)\) and \(v_3 = (0, 0, 1)\) are always unstable;
2. the vertex rest point \(v_1 = (1, 0, 0)\) is asymptotically stable for \(\rho \geq 2\) and unstable for \(0 < \rho < 2\);
3. the interior rest point \(p = \left( \frac{1}{5 - 2\rho}, \frac{2 - \rho}{5 - 2\rho}, \frac{2 - \rho}{5 - 2\rho} \right)\) is asymptotically stable whenever it exists, i.e. for \(0 < \rho < 2\).

Figure 4.6 illustrates the bifurcation.
Figure 4.6
Phase space plot of quasi-Babelian systems ($n = 3$): variation in $\rho = b/a$. (Paper II, Figure 9.)
To summarize, Paper II investigates the $n$-grammar extension of LRP learning (Algorithm 4.1) by introducing the notion of an advantage matrix and studying the diachronic, inter-generational dynamics of non-overlapping sequences of such $n$-grammar LRP learners. In addition to establishing the general result that the vertices of the simplex $S_n$ are always rest points of such systems (Theorem 4.2), the Paper examines three classes of advantage matrices with certain symmetry properties, in the $n = 3$ case for added mathematical tractability. In each case, it is found that these systems exhibit the phenomenon of diachronically stable variation — a rest point in the simplex interior which is asymptotically stable — either always (Babelian and symmetric systems), or in a subset of the system parameter space (quasi-Babelian systems). Moreover, this interior rest point is the only rest point in the interior; further rest points on the boundary are also excluded (Theorem 4.2). These observations point to the conclusion, formulated above as Conjecture 4.1, that whenever an interior rest point (a state of variation) exists in these systems, it is necessarily asymptotically stable (a state of stable variation). Future work will need to look for ways of establishing Conjecture 4.1 in the general, $n$-dimensional case, in which the Jacobian method employed in Paper II becomes infeasible. One possible line of attack is to look for suitable Lyapunov functions for the system. Another is to look for more general stability results in related fields such as mathematical biology.
5
Networks and (near-)neutrality

The relationship between the individual and population levels is not, however, reduc-
tive in any simple sense; in fact, the two levels interact, with influences flowing in both
directions. Notably, events at the population level may affect the process of language ac-
quisition at the level of the individual. Consideration of the connectivity patterns and the
dynamics of human social networks, in particular, leads to a re-evaluation of traditional
claims about the non-neutrality of linguistic change.

5.1 Networks, differentials and order

As discussed in Chapter 2, acquisition-based accounts of language dynamics reduce
language change in the diachronic domain to the process of language acquisition op-
erating at the level of the individual. Although the Z-model (Figure 2.1) acknowledges
that noise in the PLD a generation receives may have an effect — sometimes decisive
— on the evolution of the grammatical composition of a population, it has very little to
say either about that noise or about population composition. Explanation of language
acquisition has taken centre stage, with correspondingly less time and effort spent on
understanding how interactions between individuals affect language dynamics.

The usage-based tradition — which sometimes styles itself as diametrically op-
posed to the acquisition-based one (Beckner et al., 2009), though by the end of this
dissertation I hope to have convincingly argued that the two traditions can and need
to be reconciled — has explored these factors in much greater detail. Focussing on
utterances instead of Parameters as the basic dynamic variable (see e.g. Croft, 2003;
Baxter, Blythe, Croft & McKane, 2006), this tradition has placed heavy emphasis on
different types of interactions between speakers, on different connectivity patterns
of speakers in different sorts of linguistic communities and, pushing the limits of the
strong critical period assumption (Lenneberg, 1967), on change across the lifespan
(Sankoff & Blondeau, 2007). As far as computational modelling is concerned, perhaps the greatest contribution of the usage-based tradition so far is represented by various results on the respective roles of prestige differentials and social networks in linguistic change.

Ever since its original formulation in Baxter et al. (2006), and particularly after its more linguistic applications in Baxter, Blythe, Croft and McKane (2009), Blythe and Croft (2012) and Baxter and Croft (2016), the Utterance Selection Model (USM) can be considered the flagship of the computational usage-based tradition. Building on Croft’s (2000) evolutionary account of change, the USM focusses on interactions between individuals as the locus of change, utterances being conceptualized as the replicators in this evolutionary process. The model is stochastic in the sense that in each linguistic interaction only a small number of utterances are sampled from the space of all possible utterances; accordingly, its behaviour may be analysed using standard tools from statistical physics, such as the master equation (see Baxter et al., 2006). Yet in many circumstances the model produces near-predictable, that is to say orderly, behaviour. Importantly, Blythe and Croft (2012) show that when a Labovian prestige differential obtains between two competing linguistic variants, the USM predicts an S-curve; furthermore, lack of prestige predicts the absence of S-curves in their model, suggesting that a differential (originating in prestige or some other biasing mechanism) is necessary for S-curves.

Other computational studies have reached similar conclusions. Thus Ke, Gong and Wang (2008), modelling the propagation of variants on different network topologies (regular, random, small-world and scale-free), find that successful propagation of an innovatory variant is only guaranteed if adoption of that variant is heavily favoured (sometimes twenty-fold) over the adoption of the conventional one. Fagyal, Swarup, Escobar, Gasser and Lakkaraju (2010) similarly find that S-like patterns of propagation are only observed in their computational model if individuals are biased to weight interactions with prestigious (“socially desirable”) individuals more heavily than interactions with ordinary individuals. The overall conclusion from these studies would appear to be that, when change is modelled as a sociolinguistic interaction in a community (network) of speakers, a differential of some sort (usually prestige, attached either to speakers or to the competing linguistic variants themselves) is required to guarantee orderly behaviour. This is perhaps not surprising, considering that in acquisition-based models, too, change and S-curves are only observed when a differential (a parsing advantage differential) is in place (see Chapter 2, especially Theorem 2.2).
5.2 Neutral change

The above-mentioned models neglect, however, one central aspect of human social networks. This is the fact that social networks have a dynamic, evolving character, in contrast to the static networks assumed in modelling; even though previous work has explored different network topologies, these topologies have been assumed to be constant, in the sense that an individual’s neighbourhood is fixed for the duration of the computational simulation. In an effort to remedy this omission, Paper III defines a network-rewiring algorithm that incorporates a mechanism for removing old speakers from the network and adding new ones thereto. At each iteration of the model, an individual is chosen at random for removal from the network and replaced by a new individual who receives a set of individuals as his or her neighbourhood, as follows. The individuals in the network are first rank-ordered into a queue $Q$ in decreasing order of degree (the number of connections the individual has in the network). Then, with probability $\sigma$, a link is formed between the new individual and the individual having the greatest degree, and the latter individual is removed from $Q$.\(^1\) With the remaining probability mass $1 - \sigma$, the new individual is connected to a randomly chosen individual instead. This process is repeated for a predefined number $K$ of iterations, so that each new individual receives $K$ connections at the outset.

The rewiring algorithm then defines evolving network structures, with two free parameters $\sigma$ and $K$ — a further parameter $N$, held constant in the simulations reported in Paper III, gives the total number of individuals and thus defines the size of the network. Although no particular sociological realism is claimed for this algorithm at the moment,\(^2\) it can be noted that variation in the parameters $\sigma$ and $K$ allows for the modelling of different evolving network topologies, ranging from fully connected networks for $K = N$, to random networks for $\sigma = 0$ (the dynamic counterpart of static Erdős–Rényi graphs; see e.g. Newman, 2010), and highly star-like and clustered networks for $0 \ll \sigma \leq 1$ and $1 \leq K \ll N$ (Figure 5.1).

The surprising conclusion to follow from the modelling in Paper III is that for sufficiently clustered networks defined by this algorithm, prestigeless propagation of an innovative linguistic variant is possible. That is to say, if the probability of an individual adopting a variant $V$ is equal to the relative frequency of $V$ in this individual’s neighbourhood — if adoption is neutral — then an innovative “mutant” variant may percolate through the entire population, if the network is clustered enough.\(^3\)

\(^1\)Strictly speaking, the highest-degree individual may not be unique, as more than one individual may share that highest degree. In such a case, the new individual is connected to one of the highest-degree individuals chosen uniformly at random.

\(^2\)The definition of the rewiring algorithm received its initial inspiration from the Einsteinian dictum “make it as simple as possible, but not simpler”. On reflection, it is probably simpler, and future work is needed to introduce, and study the properties of, more realistic dynamic network algorithms.

\(^3\)The emergence of innovations in this model is accounted for by a mutation parameter $\mu$, which
Different settings of the model parameters $\sigma$ and $K$ yield different network topologies. Highly clustered networks are observed for high $\sigma$ and low $K$. The total number of individuals in the networks pictured is $N = 50$. (Paper III, Figure 1.)

is more, this propagation may be “smooth” in the sense that it displays a high degree of monotonicity, as illustrated by one of the simulation runs in Figure 5.2.

A natural response is to suggest that such an outcome is a statistical fluke.

gives the probability of the learner adopting a randomly sampled variant from the space of all (biologically, cognitively) possible variants. Thus the probability of the learner adopting a given variant $V$ is, under the assumption of neutral acquisition,

$$p_V = \mu \frac{1}{C} + (1 - \mu) f_V,$$

where $C$ is the size of the variation space (number of all possible variants) and $f_V$ gives the frequency of $V$ in the set of individuals to whom the learner is connected. In the simulations of Paper III, it is found that for coherent behaviour to emerge, $\mu$ typically has to have a small value on the order of $0.01 \leq \mu \leq 0.1$. Note that with this definition, both the innovation (mutation) of a new variant and its subsequent propagation through the network are neutral.
Neutral evolution of three competing variants (red, green, blue) on two different networks: a highly clustered network with $\sigma = 1$ and $K = 10$ (top) and an unclustered network with $\sigma = 0$ and $K = 10$ (bottom). Both networks have a total size of $N = 100$. Intuitively, the top trajectory is "well-behaved", while the bottom one is erratic. (Paper III, Figures 3 & 2.)

To examine this possibility, Paper III introduces a quantitative notion of the well-behavedness of a language community. This notion is composed of three measures, the values of which may be estimated numerically across batches of simulation runs for increased statistical reliability:

1. **Dominance.** — The dominance in a language community is measured as the proportion of time that at least $100 \cdot (1 - \delta)\%$ of the community uses some common variant, where $\delta > 0$ is a small real number.

2. **Shifting.** — Shifting is measured as the number of times the community tra-
verses from a state of dominance by one linguistic variant into a state of dominance by another variant in some given time period, e.g. during the entire length of a simulation run.

3. *Monotonicity.* — Language dynamics in a community is monotonic to the extent that each shifting event proceeds along a monotonic curve from one dominance state to the other. Quantitatively, monotonicity is measured as the number of trend-reversals in time windows of a specific length overlaid on the variant trajectories (see Paper III, sec. 4 and app. B, for the technical details).

Performing dozens of repetitions of simulations with varying values of $\sigma$ and $K$, Paper III shows that trajectories such as the one displayed in Figure 5.2 (top) are a reliable feature of the model in the regime of the model parameter space in which $\sigma$ is high and $K$ small (Figure 5.3). The model thus predicts that *well-behaved neutral change* may occur in highly clusterized communities, with innovations spreading along reasonably monotonic, S-like trajectories. Moreover, additional simulations on static networks show that the likelihood of these neutral shifting events is greatly reduced if the network is not rewired over time (Paper III, sec. 5.4). Thus the dynamics of the competition between linguistic variants and the dynamics of the underlying social network may have interactions which make a qualitative, and not just a quantitative, difference to the diachronic outcome. Paper III concludes that future work ought to look for more realistic dynamic network models in order to better understand this interaction in the domain of linguistics; this work can be expected to benefit from closely related work in the field of adaptive networks, which has explored similar interactions in epidemiology and ecology (Gross & Sayama, 2009).

5.3 *Why does it happen? Prolegomenon to a study of the $\sigma = 1$ case*

I next turn to a brief technical exploration of the question *why* neutral change is observed in the model put forward in Paper III. Computational simulations are notoriously poor at furnishing any understanding of the behaviours they produce; consequently, one turns to more informative, yet more difficult, analytical approaches. Although currently no analytical solution exists for the model of Paper III in the general case, some progress has been made in the limiting case $\sigma = 1$, that is to say, in the special case where every new individual is attached to the rank-ordered queue $Q$, with no connections established at random.

---

4By ‘static’ here I refer to networks that are not subject to the above-explained dynamic rewiring during the actual linguistic simulation. For these simulations, the network-rewiring algorithm was first run until the network’s degree distribution settled. At this point, the rewiring algorithm was turned off and the linguistic variant dynamics run on the resulting static network: at each time step, a random speaker was selected for update, and his or her variant was reset in proportion to the frequencies of the different variants in the speaker’s neighbourhood (i.e., neutrally), as outlined above.
Figure 5.3
Top: Number of shifting events in a simulation of $5 \times 10^4$ iterations for different values of $\sigma$ and $K$, on a network of $N = 100$ individuals and $C = 3$ competing linguistic variants; averages over 50 simulation runs. A notable increase in shifting is observed for large values of $\sigma$ and small values of $K$, i.e. for clustered networks. — Middle: Dominance (bottom surface) and monotonicity (top surface) estimated from these simulations. Both measures decrease with decreasing $\sigma$ and increasing mutation rate $\mu$. — Bottom: Combined measure of well-behavedness (shifting multiplied by dominance multiplied by monotonicity). A peak in well-behavedness occurs for large $\sigma$, small $K$ and intermediate $\mu$. (Paper III, Figures 5–7.)
In particular, in this special case it is possible to calculate the probability of an innovation spreading from the central component of the network to an individual in another part of the network; it can be shown that this probability is always greater than $1 - e^{-1} \approx 0.63$, which explains in part why the spread of innovations is so reliable in strongly clustered networks produced by the network-rewiring algorithm.

More specifically, let $\Gamma_t$ denote the set of individuals in the network whose degree is strictly greater than $K$ at time $t$, where $K$ is the number of connections given to each new individual, and call this set the nucleus of the network. Suppose the network has $N$ individuals in total, $N - 1$ of whom entertain variant $V$, and suppose there is one innovator entertaining variant $U \neq V$; in what follows, I will make use of biological terminology and call the conventional variant the “wild type” and the innovative one a “mutant”. We can then ask: supposing the mutant $U$ resides in $\Gamma_t$, what, on average, is the probability of $U$ being spread to at least one other individual before the mutant individual is removed from the network by the network-rewiring process? To answer this, we make use of the following fact:

**Theorem 5.1 (The Nuclear Trapping Property).**

Let $\sigma = 1$. Then

(i) if $|\Gamma_0| \leq K$, then $|\Gamma_t| \leq K$ for all $t \geq t_0$;

(ii) $\text{Prob}(|\Gamma_t| \leq K) \to 1$ as $t \to \infty$.

In other words, $0 \leq |\Gamma_t| \leq K$ is a trapping region for $|\Gamma_t|$.

*Proof (sketch).* For (i), consider an arbitrary rewiring event: an individual $i$ is chosen, her connections are removed, and finally she is given exactly $K$ new connections. Let $v \leq K$ be the size of the nucleus after the original connections have been removed but before the new ones are given; then $v \leq |\Gamma_0|$, and $v < |\Gamma_0|$ if there was an individual in the nucleus with degree $K + 1$ connected to $i$ before rewiring, or several such speakers. By the rewiring algorithm in the $\sigma = 1$ special case, individual $i$ must now be connected to these $v$ nuclear individuals, so when those connections are made, $K - v$ connections remain to be made which can potentially increase the size of the nucleus. If all of these are made to speakers of degree $K$, these latter speakers become nuclear and the size of the nucleus becomes $|\Gamma_{t+1}| = v + (K - v) = K$. If some of the $K - v$ connections are made to speakers of degree less than $K$, however, the size of the new nucleus will be strictly less than $K$, depending on the number of such connections made. In either case, the new nucleus is of size at most $K$.

For (ii), consider a nucleus of size $K + 1$. Suppose that a nuclear individual comes to be selected for rewiring. Then the size of the new nucleus will be $K$, and the process is trapped in the $|\Gamma_t| \leq K$ region by (i). This occurs with a nonzero probability, in fact with probability $(K + 1)/N$, since individuals are selected for rewiring uniformly at random. An inductive argument generalizes this reasoning for nuclei of sizes $K + k$ with $k > 1$.  

69
This result has the important corollary that, in the $\sigma = 1$ case after possible initial transients have died out and the nucleus has come to have a size at most $K$, every individual newly introduced to the network is necessarily connected to each individual in the nucleus.

Focussing now on the above situation of one nuclear mutant with variant $U$ and $N - 1$ wild types with variant $V$, the probability of a new individual (newly inserted into the network by the rewiring process) adopting $U$ is given by

$$p_U = \mu \frac{1}{C} + (1 - \mu) \frac{1}{|\Gamma_i|},$$

(5.2)

where $C$ is the total number of possible variants accessible through the mutation parameter $\mu$ (see n. 3 on p. 64). In the worst case scenario for the mutant type, the size of the nucleus is equal to $K$, and consequently

$$p_U = \mu \frac{1}{C} + (1 - \mu) \frac{1}{K} = \frac{1}{K} - \mu \left( \frac{1}{K} - \frac{1}{C} \right).$$

(5.3)

Hence, the probability of the new individual adopting a variant that is not the mutant $U$ (i.e. is either $V$ or one of the remaining $C - 2$ possible variants) is given by

$$\bar{p}_U = 1 - p_U = 1 - \frac{1}{K} + \mu \left( \frac{1}{K} - \frac{1}{C} \right).$$

(5.4)

In a population of $N$ individuals from which individuals are removed uniformly at random, the nuclear mutant has an expected lifetime of $N$ time steps. Therefore, the probability of at least one new individual adopting the mutant variant during the mutant individual’s lifetime is estimated by

$$q = 1 - \bar{p}_U^N = 1 - \left[ 1 - \frac{1}{K} + \mu \left( \frac{1}{K} - \frac{1}{C} \right) \right]^N.$$

(5.5)

This probability is bounded from below:

**Theorem 5.2.**

Let $\sigma = 1$, $K \leq N - 1$ and $\mu \leq 1/K$. Then $q > 1 - e^{-1} \approx 0.63$.

To establish this result, an algebraic property is needed, whose proof (by induction) I omit:

**Lemma 5.1.**
For any positive integer $n$,

$$1 - \frac{1}{K^{2^n}} = \left(1 - \frac{1}{K}\right) \prod_{i=1}^{n-1} \left(1 + \frac{1}{K^{2^i}}\right). \tag{5.6}$$

**Proof (of Theorem 5.2).** Assume, without loss of generality, that $\mu > 1/C^\alpha$ and that $C^{\alpha+1} < K^\beta$ for some positive integer $\alpha$ and some positive integer $\beta$ such that $\beta - 2 = 2^\gamma$ for some positive integer $\gamma$. As we will see, $\alpha$ can be made arbitrarily large (i.e. $\mu$ can be made arbitrarily small), as can $\beta$ (i.e. the relationship of $C$ and $K$ can be made to vary freely), so these inequalities, which will be used in deriving the lower bound for $q$, do not affect the generality of the result. Then

$$\widetilde{p}_U = 1 - \frac{1}{K} + \frac{\mu}{K} - \frac{\mu}{C}$$
$$< 1 - \frac{1}{K} + \frac{1}{K^2} - \frac{1}{C^{\alpha+1}}$$
$$< 1 - \frac{1}{K} + \frac{1}{K^2} - \frac{1}{K^\beta}$$
$$= 1 - \frac{1}{K} + \frac{1}{K^2} \left(1 - \frac{1}{K^{\beta-2}}\right)$$
$$= 1 - \frac{1}{K} + \frac{1}{K^2} \left(1 - \frac{1}{K}\right) \prod_{i=1}^{\gamma-1} \left(1 + \frac{1}{K^{2^i}}\right)$$
$$= \left(1 - \frac{1}{K}\right) \left(1 + \frac{1}{K^2} \prod_{i=1}^{\gamma-1} \left(1 + \frac{1}{K^{2^i}}\right)\right),$$

where the second to last equality is an application of Lemma 5.1. Writing

$$\Lambda = 1 + \frac{1}{K^2} \prod_{i=1}^{\gamma-1} \left(1 + \frac{1}{K^{2^i}}\right) \tag{5.8}$$

we then have that

$$\widetilde{p}_U^N \leq \widetilde{p}_U \leq \left(1 - \frac{1}{K}\right)^K \Lambda^K =: L, \tag{5.9}$$

since $K < N$ and $0 \leq \widetilde{p}_U \leq 1$. Clearly, $\widetilde{p}_U$ is monotone increasing in $K$, so we wish to find if $L$ tends to a finite limit as $K \to \infty$. From the definition of Euler’s number,

$$\lim_{K \to \infty} \left(1 - \frac{1}{K}\right)^K = \frac{1}{e}, \tag{5.10}$$

so it suffices to investigate the limiting behaviour of $\Lambda^K$. For this, let $\lambda = \log \Lambda$ and $k = 1/K$. Then

$$\lim_{k \to 0} \lambda = \lim_{k \to 0} \frac{1}{k} \log \left(1 + k^2 \prod_{i=1}^{\gamma-1} \left(1 + k^{2^i}\right)\right). \tag{5.11}$$
Writing $F(k) = \log \left( 1 + k^2 \prod_{i=1}^{\gamma-1} (1 + k^{2i}) \right)$ and $G(k) = k$, we have $G'(k) = 1$ and

$$F'(k) = \frac{d}{dk} \left( \frac{1 + k^2 \prod_{i=1}^{\gamma-1} (1 + k^{2i})}{1 + k^2 \prod_{i=1}^{\gamma-1} (1 + k^{2i})} \right).$$

The denominator of $F'(k)$ clearly tends to $1$ as $k \to 0$. For the numerator, note that the argument to the differentiation operator is just a polynomial in $k$ with no degree-1 terms, so the derivative is a polynomial with no degree-0 terms and hence tends to $0$ as $k \to 0$. Therefore, $\lim_{k \to 0} F'(k) = 0$, and so

$$\lim_{k \to 0} \frac{F'(k)}{G'(k)} = 0.$$

By L’Hôpital’s Rule,

$$\lim_{k \to 0} \frac{F(k)}{G(k)} = 0,$$

hence $\lim_{K \to \infty} \lambda = 0$ so that $\lim_{K \to \infty} \Lambda = 1$. Thus, in the end,

$$\lim_{K \to \infty} L = \frac{1}{e} \cdot 1 = \frac{1}{e}.$$

Putting this together with (5.7), we have that

$$q = 1 - p^N_U \geq 1 - p^K_U > 1 - \frac{1}{e},$$

which is what we intended to show.

The above proof establishes the lower bound for any mutation rate satisfying $\mu \leq 1/K$, which is not an unreasonable assumption given the interpretation of $\mu$ and $K$ in the present model. In the case of very low mutation rates it is possible to approximate by substituting $\mu = 0$ in equation (5.5), whereby the proof will be greatly simplified. Moreover, examination of the form of (5.4) shows at once that $\tilde{p}_U \to 0$ as $K \to 1$ and $\mu \to 0$. Thus, in the limit of a minimal nucleus and zero mutation, the probability of successful propagation of a nuclear mutant variant approaches $1$, as is indeed intuitively sensible (Figure 5.4).

This result characterizes just one event in the overall dynamics of the model (and does so only in the special case $\sigma = 1$); the successful propagation of an innovatory variant across the entire network is a composite process depending on (i) the probability of innovation, (ii) the probability of a nuclear individual being replaced by the individual carrying the innovation,\(^5\) (iii) the probability of the innovator, once

---

\(^5\)The network-rewiring algorithm furnishes every new individual with exactly $K$ connections, and on the other hand every innovation or “mutation” event occurs in these new individuals (see above). Every innovator, then, resides outside the nucleus at the outset and must be “promoted” into it by some kind of process that increases the innovator’s degree. This can happen if (i) a nuclear individual comes to be removed and (ii) another newly added individual comes to be connected to the mutant

72
in the nucleus, spreading their variant to another individual in the network before being removed (the above-calculated probability \( q \)), (iv) the probability of this other individual spreading the mutant variant yet further, and (v) the cascading of all these processes over sufficiently many iterations so that the innovation may diffuse across the entire network. All of these things are, in principle, open to both computational and analytical exploration. The above argument — a first step towards a more comprehensive treatment — shows that in a strongly clustered network, nuclear mutants are more likely than not to spread their variant onwards. With more and more mutant individuals, the probability of reliable propagation of the mutant variant increases in a kind of snowball effect, allowing us a glimpse, at least, of an understanding of why well-behaved neutral change occurs in the model.

5.4 A note on neutral and near-neutral models

I have called the model of Paper III a ‘neutral’ model, since acquisition in this model is purely frequency-dependent: individuals are sensitive to the frequencies of the different variants in their social neighbourhood, without any biases promoting the adoption of a particular variant. The fact that with high values of the preferentiality parameter \( \sigma \) individuals connect preferentially to the highest-degree speakers in the

\( (\text{in addition to being connected to the } K - 1 \text{ nuclear individuals}), \) so that the mutant’s degree becomes \( K + 1 \). These promotion events obviously have a great effect on the spread of innovations in the current model, given the fact that every new individual is guaranteed to attach to every individual in the nucleus in the \( \sigma = 1 \) case. Here I have to set a more detailed exploration of promotion aside, but future work could attempt to calculate the probability of promotions, starting from the \( \sigma = 1 \) case.

Figure 5.4
Probability of successful mutant propagation from the nucleus as a function of \( K \) and \( N \) (equation 5.5), for mutation rate \( \mu = 0.01 \) and number of competing variants \( C = 30 \). (Paper III, Figure 14.)
social network, however, means that the model differs from classical random-copying models (Blythe, 2012) in an important respect: the population is structured, and indeed the results of Paper III suggest that well-behaved neutral change can only occur in such structured populations. Using the taxonomy proposed by Blythe and Croft (2012), what the model describes is neutral interactor selection rather than pure neutral evolution (drift), since interactions with speakers in a non-well-mixing population constitute differential replication (cf. Hull, 1988). Viewed from this vantage point, the type of model studied in Paper III is more profitably described as a near-neutral than a strictly neutral model.

Near-neutrality may arise from different sources, not just from non-random social network structures. Stadler, Blythe, Smith and Kirby (2016) introduce a model of change based on the idea (cf. Labov, 2001) that speakers are cognizant of trends in language dynamics: by tracking the recent history of the frequency trajectory of a variant, speakers are assumed to estimate whether a variant is on the increase or not, and to adjust their own behaviour in response to that estimation. Stadler et al. (2016) call this momentum (cf. Gureckis & Goldstone, 2009) and investigate the properties of an extension of the USM in which the bias consists of such a momentum term. The difference to a classical prestige-based explanation, for instance, is that momentum does not attach to any given variant specifically, but rather emerges from the vagaries of particular trajectories; a momentum-based model, then, is near-neutral. With computer simulations, Stadler et al. (2016) suggest that their model has two phases, depending on the values of a number of parameters controlling the estimation of momentum. In one phase, when momentum has a slight effect on an individual’s behaviour, the model predicts stability. In the other phase, however, individuals are more likely to boost their use of those variants which have seen increasing use in the recent past; here, the typical dynamic outcome is an S-curve from actuation to completion (Stadler et al., 2016).

It has been pointed out that neutral theory serves the important role of a null model in language dynamics (Blythe, 2012); specifically, a number of cases remain in which the postulation of a prestige or other differential, though conventional, is problematic (see Lass, 1997, for a number of examples). If such cases of change can be explained by a model that does away with differentials, all the better: just as it is beyond doubt that prestige is involved in many situations of variation and change, it is at least conceivable that some changes happen without its assistance. Mechanisms such as social network structure and rewiring (Paper III) and momentum (Stadler et al., 2016) may play a decisive role when replicator selection is absent: in the absence of a bias rooted in prestige, parsing advantage or some other such source that attaches to competing variants directly, near-neutral mechanisms come to the fore. Future work needs to address both formal and empirical issues: firstly, formal tools are needed to
better understand the properties of these essentially stochastic models in the different regimes of their parameter spaces, and secondly, empirical studies are needed to establish the extent to which the mechanisms postulated by near-neutral models (network structure and rewiring, momentum) actually play a role in real-world language dynamics.
6 Change and stability over large timescales

Although some of language change may be neutral, some of it must be non-neutral. This is shown, among other things, by a consideration of the overall synchronic frequency and spatial distributions of linguistic traits, which show significant deviations from the predictions of neutral theory. As an attempt at explanation of these deviations, a model of dynamic typology is explored. This makes possible the estimation of the “temperature” of individual linguistic features from their geospatial distributions, illustrating a particular mode of inference from synchrony to diachrony.

6.1 Typology across time and space

Even if Paper III suggests that previous research may have too rashly discounted the possibility of neutral linguistic change, ample evidence for non-neutral motivators of change exists; these motivators include purely physiological biases in perception or production (Ohala, 1989), prestige effects (Labov, 1972) and — although their role is less straightforward to pinpoint experimentally — effects of parsing advantage or economy (Yang, 2000; Roberts & Roussou, 2003). It may be fruitful to compare this situation with biology, where some mutations at the molecular level are argued to be neutral (Kimura, 1994) even though there is no doubt that non-neutral factors (Darwinian natural selection, favouring mutations that increase a phenotype’s fitness) account for the broad phylogenetic trajectories of species.

In biological evolution, phylogenies reflect the differential adaptation of organisms to environmental conditions over large timescales; consequently, the corresponding synchronic taxonomy or “typology” of organisms is non-neutral. In the domain of language, certain results on typological universals and tendencies point to the same conclusion (unless a cop-out is bought in the form of a very strong founder effect hypothesis, a position I will argue against in Section 6.3 below). The reasoning
The global frequency distribution of 35 binary linguistic features (see Paper IV for details) sampled from the World Atlas of Language Structures (Dryer & Haspelmath, 2013); \( \rho \) signifies the overall, global relative frequency of a feature.

Figure 6.1

The global frequency distribution of 35 binary linguistic features (see Paper IV for details) sampled from the World Atlas of Language Structures (Dryer & Haspelmath, 2013); \( \rho \) signifies the overall, global relative frequency of a feature.

goes as follows: in the absence of vigorous contact, the probability of finding any given variant in a randomly chosen language community at any given time ought to be equal, if all of language operated neutrally. Yet this is not what one observes at the cross-community, global, typological level: linguistic traits vary from the extremely uncommon to the extremely common, with every imaginable frequency occurring in between (Figure 6.1). This leads one to conclude — essentially correctly, although previous research has taken somewhat too simplistic a view on the matter, as I am about to argue — that some universal and non-neutral forces favouring certain linguistic features and disfavouring others are in operation.

Greenberg (1978) proposed to model these forces by treating the dynamics of individual linguistic features\(^1\) as Markov chains. Assuming binary features (each feature is either present in or absent from a language), the dynamics of a feature are then accounted for by two probability parameters, the probability of a language adopting this feature, and the probability of a language losing the feature; Bermúdez-Otero (1997) calls these the feature’s ingress and egress probability, respectively (Figure 6.2).

\(^1\)Where Paper III speaks of “variants”, here and in Paper IV linguistic variables are called “features”, in line with the typological tradition. Although ultimately a linguistic variable ought to be a primitive of synchronic theory (a Parameter in the sense of Chapter 2), this is not always the case with typological features, which may be composites. For present purposes, the distinction is mostly irrelevant, but more discussion follows below.
In Greenberg’s (1978) Markov chain model, the dynamics of a binary feature $F$ is accounted for by its ingress probability $p_I$ and egress probability $p_E$.

Given the Markov assumption, the stationary distribution of an individual feature can be calculated. Indeed, it is not difficult to show that for ingress probability $p_I$ and egress probability $p_E$, a system of this kind tends to a distribution such that the feature occurs with probability

$$\rho = \frac{p_I}{p_I + p_E}$$

in the limit of infinite time. To interpret this in more linguistically meaningful terms, the stationary distribution property means that one should expect any given linguistic community to exhibit the feature $100\rho\%$ of the time, if observed for a sufficiently long time and once transients (founder effects) have died out — or, looking at the matter from a comparative perspective, that $100\rho\%$ of a given sample of languages (or linguistic communities) should exhibit the feature, if the sample is large enough and the languages or communities evolve (sufficiently) independently of each other. Variation in the ingress and egress probabilities $p_I$ and $p_E$ among linguistic features could then be taken to explain the observed differences in the global frequencies of linguistic features (Figure 6.1), yielding a dynamicization of typology (Greenberg, 1978) or, simply, a dynamic typology (DT).

In addition to exhibiting differences in overall frequency, linguistic features differ in their distribution over geographical space. While this fact has long been appreciated — retroflexed consonants, for example, are a well-known areal feature in the sense of being restricted to certain Sprachbünde, mostly in (but not limited to) South Asia (Bhat, 1973) — a quantitative exploration of the spatial distributions of linguistic features is only now becoming a possibility, thanks to the recent compilation of digital typological atlases (Kortmann & Lunkenheimer, 2013; Donohue, Hetherington, McElvenny & Dawson, 2013; Dryer & Haspelmath, 2013; Moran, McCloy & Wright, 2014; Michael, Stark, Clem & Chang, 2015; Simons & Fennig, 2017). Figure 6.3 illustrates this variability in the spatial patterning of linguistic features by supplying the

---

2A simple way of seeing this is to note that in the stationary state, the flow out of the state in which the feature is attested must equal the flow into that state, yielding the stability condition $\rho p_E = (1-\rho) p_I$. Solving this for $\rho$ gives equation (6.1).
geographical distributions of the two features phonological tone and indefinite article, estimated from data mined from the World Atlas of Language Structures (Dryer & Haspelmath, 2013; henceforth WALS).

In the WALS data, tone and indefinite article have global frequencies of $\rho = 0.42$ and $\rho = 0.45$, respectively (see Paper IV, Supplementary information), yet the two features exhibit entirely different kinds of spatial distributions. Indeed, any feature of an intermediate frequency (i.e. not almost everywhere present and not entirely absent either) may distribute spatially along a continuum from a highly clustered or areal distribution (such as is found for tone) to a highly scattered or random one (such as is the case with the indefinite article). Greenberg (1978) hypothesized that such differences – as well as differences in the distribution of features among phylogenetic lineages, which matter I will set aside here — may be related to the features’ ingress and egress probabilities, in the following sense. Binning ingress and egress into two classes for ease of exposition, high and low, four combinations of the probabilities are possible. If a feature sports high ingress but low egress, we would expect this feature to be globally highly frequent. A feature with low ingress and high egress, on the other hand, should be practically non-existent. What to expect with the remaining two logical combinations is less clear: in Greenberg’s (1978) Markov chain model, both types of feature are predicted to appear with a similar frequency, since setting $p_I = p_E$ in equation (6.1) yields $\rho = 0.5$ regardless of the absolute value of $p_I = p_E$. Greenberg (1978) nevertheless put forward the prediction that features with high ingress and high egress ought to attest scattered spatial distributions, and features with low ingress and low egress clustered ones, yielding Table 6.1. This predic-
The distributional predictions of Greenberg’s (1978) Dynamic Typology, with example features.

<table>
<thead>
<tr>
<th>Ingress</th>
<th>Egress</th>
<th>Predicted Distribution</th>
<th>Example (from Greenberg, 1978)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>High</td>
<td>Low</td>
<td>Near-universal</td>
</tr>
<tr>
<td>II</td>
<td>High</td>
<td>High</td>
<td>Intermediate, scattered</td>
</tr>
<tr>
<td>III</td>
<td>Low</td>
<td>Low</td>
<td>Intermediate, clustered</td>
</tr>
<tr>
<td>IV</td>
<td>Low</td>
<td>High</td>
<td>Near-absent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vowel nasalization</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vowel harmony</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Velar implosives</td>
</tr>
</tbody>
</table>

...
The model assumes language communities to be situated on a two-dimensional regular lattice with periodic boundary conditions, wrapping around to form a torus. Each site (red) is affected by its four von Neumann neighbours (blue) as well as by a non-spatial ingress–egress dynamics (Figure 6.2).

Community $C$ and a feature $F$ are chosen at random. With a predefined probability $q$, a “voter” event occurs, and $C$ adopts the value of $F$ (either “present” or “absent”) in one of its four von Neumann neighbors $C'$ (Figure 6.4). With the remaining probability mass $1 - q$, a Greenbergian ingress–egress event occurs instead: if $F$ is absent from $C$, it is adopted with probability $p_I$, and if $F$ is present in $C$, it is lost with probability $p_E$. The whole process is summarized as a flowchart in Figure 6.5.

To test the predictions of this model against empirical data, Paper IV mines WALS for information on 35 binary linguistic features. Since visual inspection of maps such as those in Figure 6.3 is both infeasible when the number of features is large, and inexact no matter what the number of features, in Paper IV a more quantitative angle of attack is assumed. Following most work on voter-like models in the statistical physics literature (see e.g. Krapivsky, Redner & Ben-Naim, 2010), Paper IV adopts the density of reactive interfaces as a measure of spatial correlation. For our present purposes, this measure is defined as follows, for feature $F$: (i) count the number of geographical nearest-neighbour$^4$ language pairs such that one member of the pair has $F$ and the other lacks it; (ii) divide by the total number of language pairs in the sample. The resulting quantity, which Paper IV calls isogloss density and denotes by $\sigma$,$^5$ then has a theoretical range from 0 to 1 and estimates the probability of finding a reactive interface — an isogloss — between a pair of linguistic communities. For low values of $\sigma$, this probability is small, and consequently the feature in question tends to cluster in space. Features with high values of $\sigma$, on the other hand, sport scattered spatial distributions. The two features tone and indefinite article mapped in Figure 6.3, for example, have isogloss densities $\sigma = 0.17$ and $\sigma = 0.36$, respectively (Paper

---

$^4$Nearest neighbours were determined from WALS coordinate data by the great-circle distance on an ideally spherical Earth.

$^5$Not to be confused with the $\sigma$ of Paper III and Chapter 5 above!
Flowchart for the model put forward in Paper IV. $F(C)$ refers to the value of feature $F$ in community $C$, ranging over 1 (feature present) and −1 (feature absent); w.p.: ‘with probability’.  

IV, Table S2). Figure 6.6 provides an overall picture by plotting isogloss density $\sigma$ as a function of feature frequency $\rho$ for each of the 35 features considered in Paper IV.

### 6.2 The inference of diachrony from synchrony

What can be done to explain such an empirical observation? Turning now to the lattice model, it would be possible to run the model with varying values of the parameters $p_I$ (ingress), $p_E$ (egress) and $q$ (probability of a spatially extended interaction), and measure the feature frequency $\rho$ and isogloss density $\sigma$ for the lattice at some large number of model iterations. It turns out that simulations are unnecessary, however, in the sense that the model is simple enough to yield to solution by analytical means. In Paper IV (Supplementary information) it is shown that the value of $\rho$ tends to

$$\rho = \frac{p_I}{p_I + p_E},$$

exactly as in Greenberg’s (1978) original Markov chain model. The isogloss density $\sigma$ is also soluble, though the mathematics get much more complicated: $\sigma$ tends to the
Isogloss density $\sigma$ as a function of overall feature frequency $\rho$ for 35 features: data from WALS. The crosshairs give bootstrap confidence intervals as explained in Paper IV. (Paper IV, Figure 4.)

The crosshairs give bootstrap confidence intervals as explained in Paper IV. (Paper IV, Figure 4.)

The parameter $\tau$ gives the relative rate of the non-spatial ingress–egress process over the spatial process. Since the latter process copies features faithfully, while the former serves to flip feature values, $\tau$ is a temperature — it measures, roughly, the

$$\sigma = 2H(\tau)\rho(1 - \rho)$$  \hspace{1cm} (6.3)

with

$$\tau = \frac{(1 - q)(p_I + p_E)}{q}$$  \hspace{1cm} (6.4)

and

$$H(\tau) = \frac{\pi(1 + \tau)}{2K\left(\frac{1}{1+\tau}\right)} - \tau$$  \hspace{1cm} (6.5)

where $K(\cdot)$ is the complete elliptic integral of the first kind.\(^6\) Importantly, from equation (6.3) one finds that $\sigma$ is a parabolic function of $\rho$ with the factor $H(\tau)$ controlling the height of the parabola. Moreover, $H(\tau)$ is a monotonic increasing function of $\tau$, so that the value of $\sigma$ for any given value of $\rho$ is determined by the ratio $\tau$ (Figure 6.7).

The parameter $\tau$ gives the relative rate of the non-spatial ingress–egress process over the spatial process. Since the latter process copies features faithfully, while the former serves to flip feature values, $\tau$ is a temperature — it measures, roughly, the

\(^6\)The complete elliptic integral of the first kind is the function

$$K(z) = \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - z^2 \sin^2 \theta}}.$$  \hspace{1cm} (6.6)

It cannot be expressed in terms of elementary functions but its value can be efficiently computed with the arithmetic–geometric mean.
average kinetic energy (average tendency to change) in the system. What accounts for differences in spatial distributions such as those reported in the two maps in Figure 6.3, then, is not simply the ingress and egress rates of features, but the relation of these rates to an independent rate of faithful linguistic transmission. The more faithful interactions occur per unit time, the more clustered a feature comes to be in geographical space; conversely, the fewer such interactions, the more the feature is at the mercy of the ebb and flow of ingress and egress, and has a correspondingly scattered spatial distribution.\footnote{7}

Thanks to the lattice model, it is now possible to draw diachronic predictions from synchronic feature distributions: specifically, given empirical values of feature frequency $\rho$ and isogloss density $\sigma$ estimated from a database such as WALS, the temperatures of individual features may be inferred by inverting the function $H(\tau)$.\footnote{8} This has nontrivial implications for typological work attempting to infer predictions of stability from observations of synchronic distribution. Thus consider the traditional claim (e.g. Hawkins, 1983) that the typological rarity of a feature is a direct measure of that feature’s “markedness”, “unnaturalness” or “instability”. This claim runs

---

\footnote{7}{For a fuller discussion of the linguistic interpretation of the two subprocesses of the model, see Paper IV, sec. 2.}

\footnote{8}{As emphasized in n. 6 above, the elliptic integral does not have an analytical solution and $H(\tau)$ therefore cannot be inverted analytically. However, the inversion can be performed numerically, by computing values of $H(\tau)$ for a range of values of $\tau$ using the arithmetic–geometric mean, and then using this as a lookup table to infer in the direction from $H(\tau)$ to $\tau$. This is the procedure followed in Paper IV.}
into both conceptual and empirical problems: conceptually, there is nothing logically
contradictory about a rare but stable feature, and an empirical example is afforded
by retroflex stops, whose frequency is estimated at just 8–11% globally (Ladefoged
& Bhaskararao, 1983; Moran et al., 2014), despite the fact that they have apparently
never undergone egress since their onetime ingress in the Dravidian languages in
which they commonly occur (Bhat, 1973; Lass, 1975). The model put forward in Pa-
er IV can now deal with such empirical counterexamples, since, mathematically, a
feature can be rare (have a low value of $\rho$) yet at the same time be stable (have a
low temperature $\tau$): from equation (6.2), low $\rho$ implies low $p_I$ and high $p_E$, and from
equation (6.4), low $\tau$, low $p_I$ and high $p_E$ together imply high $q$.

Conversely, there is in principle nothing to outlaw a frequent (high $\rho$) but unsta-
ble (high $\tau$) feature. A candidate is verbal person marking, which in the data of Paper
IV has a frequency of $\rho = 0.78$ and an estimated temperature of $\tau = 0.51$; considering
that the range of $\tau$ estimates in Paper IV is from 0.00002 to 0.51 (i.e., verbal person
marking is the highest-temperature feature in our sample), this is a high-temperature,
or unstable, feature. Thus even though a given language is more likely than not to
display verbal person marking, there is a tendency for this feature to be lost and ac-
quired again, lost and acquired again, and so on, over time. Such a prediction makes
sense when considered from the point of view of language contact: for an L2 learner,
verbal inflection is difficult but highly dispensable (a “nuisance”, in fact), since the
referents of actions are usually deducible from context even without overt morpho-
logical marking. Features such as verbal person marking, then, are predicted to be
unstable not because they are infrequent (in fact, they are not infrequent), but be-
cause of independently credible facts of language learning in humans beyond the L1
and the critical period, reflected in high ingress and egress probabilities in the lattice
model (cf. Bentz & Winter, 2013, who show empirically that presence of L2 learners
correlates with absence of nominal case). More generally, there is no correlation be-
tween the overall global frequency $\rho$ of a feature and its estimated temperature $\tau$, or
between the distance $|0.5 - \rho|$ and $\tau$ (Figure 6.8), indicating that accounts of marked-
ness and stability based exclusively on synchronic frequency distributions neglect an
important contribution to stability arising, essentially, from the spatial domain.

Recent work in quantitative typology has attempted to refine the notion of
diachronic stability by employing genealogical methods of stability estimation
(Maslova, 2004; Parkvall, 2008; Dediu, 2011; Dediu & Cysouw, 2013). Briefly, the idea
is that stable linguistic features ought to attest little to no within-family variance,
while unstable features, by contrast, should exhibit variation even among closely re-
lated languages. Although details of implementation vary, all genealogical methods
assume a prior classification of the world’s languages into families, a procedure that
incurs at least two problems. Firstly, since within-family variation is proportional to
In a simple linear regression model, there is no correlation between feature frequency $\rho$ and feature temperature $\tau$ (left; Pearson’s $r = 0.12, p = 0.48$), or between the distance from half, $|0.5 - \rho|$, and $\tau$ (right; Pearson’s $r = 0.02, p = 0.92$). Data points from Paper IV.

Secondly, genealogical methods do not incorporate effects of horizontal transfer between lineages and may thus overestimate the stability of features subject to frequent modification in contact settings, such as nominal case (Bentz & Winter, 2013). These problems are discussed at length in Paper IV, which also compares temperature ($\tau$) estimates derived from the lattice model against those predicted by a genealogical method (Dediu, 2011). Paper IV argues that the genealogical method incurs a number of false positives resulting from the above two problems, and that stability estimates derived from the lattice model via the temperature quantity ($\tau$) are likely to be more reliable, given the fact that this procedure does not rely on genealogical classifications (and therefore avoids the problem of family time depths) and incorporates an explicit (if simple) model of horizontal interactions (in the form of the spatial “voter” process).

6.3 Coda: founder effects vs. ergodicity

It is a fundamental mathematical property of the model put forward in Paper IV that the state of the lattice tends, with increasing time, towards a stationary distribution. This fact, embodied in equations (6.2) and (6.3) above, means that any initial transients, such as an idiosyncratic initial distribution of features across the lattice, are
Differences in feature frequency $\rho$ and isogloss density $\sigma$ between the Old World (OW) and New World (NW) subsets of the WALS data. Both differences cluster around zero, meaning the distributions of both $\rho$ and $\sigma$ are largely identical in the two geographical subsets; this is evidence for ergodicity and against the founder effect hypothesis. Differences were calculated for only those features attested in both geographical subsets; this means 33 out of the 35 features overall.

The question then becomes: does the (empirical) distribution of features over human languages today carry a protolinguistic signal? Note that it is an essential prerequisite of the modelling in Paper IV, and eo ipso of the stability estimates derived from that modelling, that the answer to this question be negative. Although it is unclear how the question could be decisively answered, the ergodic hypothesis (no signal) derives some support from the great timescales involved: sober estimates put the emergence of human language (with syntax and, crucially, recursion) to at least 100,000 years ago (Bickerton, 2007; Tallerman, 2012; Pagel, 2017). Considering that some linguistic changes sweep through their populations in a matter of decades (e.g. Baxter et al., 2009), and that even the most drawn-out cases of slow syntactic change complete in a few centuries (e.g. Kroch, 1989; but see Wallenberg, 2016), it
would not seem out of place to suggest that the system (human language) has had enough time to mix so that initial transients have been lost.

A possibly more convincing argument can in fact be rooted in empirical data. If there were significant founder effects in language, different linguistic features should exhibit different distributions — either in plain frequency, or spatially, or both — in different parts of the globe. Rerunning the analysis of Paper IV, this time partitioning the Earth first into Old World (Africa, Europe and Asia) and New World (North and South America, Australia and Papunesia), does not support founder effects: both the frequency $\rho$ and isogloss density $\sigma$ are roughly the same in both subsets of the world for any feature consulted (Figure 6.9). Thus even though there is no question that linguistic bottlenecks and the like may sometimes throw the dynamics of a language community onto an essentially stochastic track (on this point see Lass, 1997, 382–383), in the greater scheme of things (on deep timescales, across the globe and, epistemically, speaking in statistical averages) the dynamics of language are guided by universal factors (translating into ingress and egress rates) which (instead of founder effects) leave a signal.

These signals may be extremely subtle and require sophisticated signal processing to unpack; in fact, the main methodological contribution of Paper IV may be in suggesting that to make sense of such essentially empirical matters as ingress and egress rates and founder effects, more theory (specifically, more theory of the mathematical modelling type) is needed.

The four original Papers follow in unedited pre/post-print form. For full citations, copyright statement, and an explanation of co-authors’ contributions, see the List of publications on p. 5.
ORIGINAL PUBLICATIONS
I

Deriving the Constant Rate Effect

Henri Kauhanen & George Walkden

Abstract

The Constant Rate Hypothesis (Kroch, 1989) states that when grammar competition leads to language change, the rate of replacement is the same in all contexts affected by the change (the Constant Rate Effect, or CRE). Despite nearly three decades of empirical work into this hypothesis, the theoretical foundations of the CRE remain problematic: it can be shown that the standard way of operationalizing the CRE via sets of independent logistic curves is neither sufficient nor necessary for assuming that a single change has occurred. To address this problem, we introduce a mathematical model of the CRE by augmenting Yang’s (2000) variational learner with production biases over an arbitrary number of linguistic contexts. We show that this model naturally gives rise to the CRE and prove that under our model the time separation possible between any two reflexes of a single underlying change necessarily has a finite upper bound, inversely proportional to the rate of the underlying change. Testing the predictions of this time separation theorem against three case studies, we find that our model gives fits which are no worse than regressions conducted using the standard operationalization of CREs. However, unlike the standard operationalization, our more constrained model can correctly differentiate between actual CREs and pseudo-CREs — patterns in usage data which are superficially connected by similar rates of change yet clearly not unified by a single underlying cause. More generally, we probe the effects of introducing context-specific production biases by con-

∗This paper is to appear in Natural Language & Linguistic Theory, DOI:10.1007/s11049-017-9380-1. The version included here is a post-print reflecting changes made in the peer reviewing process, but not the final publisher’s version. When citing, please use the page numbers given in the publisher’s version. © The Authors.
ducting a full bifurcation analysis of the proposed model. In particular, this analysis implies that a difference in the weak generative capacity of two competing grammars is neither a sufficient nor a necessary condition of language change when contextual effects are present.

**Keywords:** Constant Rate Effect; language change; dynamical systems; mathematical models; nonlinear regression

### I.1 Introduction

#### I.1.1 The Constant Rate Effect

In a seminal paper in historical syntax, Kroch (1989) proposed the Constant Rate Hypothesis:

> [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)

Initially (and still logically) a hypothesis, the notion of a constant rate has accumulated enough support over the last three decades for this to be referred to as the Constant Rate Effect, or CRE (see e.g. Pintzuk, 2003, 511).

The logic behind CREs is as follows: if a variant replaces another variant in two or more different contexts and the rate of change is the same in each of these contexts, then we should assume that only a single change has occurred. CREs have therefore been deployed to argue that two or more apparently unrelated surface changes are in fact manifestations of a single underlying change (Figure I.1). Unifying changes in this way provides strong support for approaches to language in which syntactic variation consists not primarily in lexical or contextual idiosyncrasies but in the values of a finite number of universal parameters, as in the classical Principles & Parameters approach (Chomsky, 1981; Chomsky & Lasnik, 1993). There is no necessary link between Principles & Parameters and CREs, as Pintzuk (2003, 511) emphasizes; the variationist approach within which the Constant Rate Hypothesis is couched is theory-neutral. However, CREs are a useful tool in the armoury of diachronic syntacticians who wish to argue for “the controlling effect of abstract grammatical analyses on patterns in usage data” (Kroch, 1989, 239).

CREs offer a fresh perspective on the causation of changes. Kroch (1989, 238) criticizes the approach to causation in which “the finding that a given context is most favorable to the use of an innovation is taken to show that the innovation is an accommodation to the linguistic functionality of that context”. Where there is a disparity
between contexts that share the same rate of change, this “reflects functional effects, discourse and processing, on the choices speakers make among the alternatives available to them in the language as they know it; and the strength of these effects remains constant as the change proceeds” (Kroch, 1989, 238). In other words: surface changes are to be thought of as reflexes of underlying grammatical changes; the discrepancies in frequencies seen at the surface level are due to extra-grammatical factors, or contextual effects, which are independent of the underlying change itself and constant across time.

The usual procedure for detecting a CRE in some diachronic data is to fit a logistic curve (I.1) to each of the contexts separately and then to compare the growth rates of these curves against each other.

\[
p_t = \frac{e^{s(t-k)}}{1 + e^{s(t-k)}} = \frac{1}{1 + e^{-s(t-k)}}
\]

Here, \(p_t\) is the frequency of either the innovatory or the receding variant (or parameter value) in a given context at time \(t\), and \(s\) is the (time-independent) rate of change in that context. The \(k\) parameter serves to translate the curve along the time axis, indicating the point of greatest growth, or the tipping point, of \(p_t\) (Figure I.2). With this
operationalization, we have the following procedure for establishing a CRE: a logistic curve of the form (I.1) is first fit to each of the contexts of interest separately and independently. Then, if variation among the $s$ or ‘slope’ parameters for these curves is found to fall within a reasonable confidence interval, the change is said to proceed at the same (‘constant’) rate in all contexts. Variation among the $k$ or ‘intercept’ parameters, on the other hand, is allowed and is where the contextual effects, independent of the underlying grammatical change, are thought to manifest themselves. This is the procedure used in a number of studies that have sought to establish CREs in various processes of change across a number of languages (e.g. Kroch, 1989; Santorini, 1993; Pintzuk, 1995; Kallel, 2005; Pintzuk & Taylor, 2006; Kallel, 2007; Fruehwald et al., 2009; Postma, 2010; Durham et al., 2012; Wallage, 2013; Gardiner, 2015). Henceforth, we shall refer to it as the standard operationalization.$^1$

I.1.2 The non-linking problem

Initially, the logistic function (I.1) was adopted because of its practicability and its success in other disciplines such as population genetics, not because it followed from any established first principles:

[G]iven the mathematical simplicity and widespread use of the logistic, its use in the study of language change seems justified, even though, unlike in the popu-

---

$^1$There exist a number of methods to implement this procedure, such as nonlinear regression on bare frequencies, linear regression on logit-transformed data, and multivariate regression. Which method is chosen is a technical matter; conceptually, all of these implementations share the basic theoretical assumption that the reflexes of one underlying change are described by a family of logistics agreeing in their $s$ parameters but possibly differing in their $k$ parameters.
lation genetic case, no mechanism of change has yet been proposed from which the logistic form can be deduced. (Kroch, 1989, 204)

The logistic has since been derived from mathematical models of language acquisition independently by Niyogi and Berwick (1997) and Yang (2000); Ingason et al. (2013) provide a particularly clear illustration of how syntactic acquisition in successive generations can give rise to logistic change at population level. What has never been explicated in detail, however, is why different contextual reflexes of a single underlying change should be governed by logistics agreeing in their $s$ parameters but freely varying in their $k$ parameters: even though this operationalization has proved useful in gathering empirical support for the Constant Rate Hypothesis, it is not a model of the CRE itself. In short, while the standard operationalization may adequately describe historical data, it fails to explain it, suggesting no mechanism for how contextual reflexes spring from underlying changes. The fact that under the standard formulation the independent contextual reflexes are not linked to each other, or to anything else, in this stronger sense we call the non-linking problem, and there are a number of reasons to believe that the problem is serious enough to warrant that the standard operationalization of CREs should be rejected.

Firstly, note that fitting a number of independent logistics to a number of contexts in some data leaves variation among the $k$ parameters entirely unexplained, even if we assume that the logistics agree in their $s$ parameters as required by the standard operationalization. In principle, it is possible for this variation in $k$ to be arbitrarily large, and it is therefore in principle possible to ‘connect’ two clearly unrelated changes — possibly separated by millennia on the time axis — as long as they happen to share the same growth rate. In principle, then, it is possible to be led to the absurd conclusion that a single change runs to completion in one context before it even takes off in another (Figure I.2).

Secondly, there are reasons to think that not all instances of logistics agreeing in their $s$ parameters are in fact CREs in the sense that a single underlying grammatical change is being modulated in the usage of a speaker or group of speakers by constant contextual (functional, discourse-related, etc.) effects. The relevant evidence comes from studies in which the ‘contextual effects’ are not within-speaker but between-speaker effects or even outright contingencies. Wallenberg (2016) shows that relative clause extraposition is a gradually declining option across the histories of both English and Icelandic, and that the $s$ parameters of the two curves do not differ significantly. Similarly, Willis (2017), in his study of the spread of the innovative
second-person pronoun *chdi* in the recent history of Welsh, finds that in different regions of Wales the change is more or less advanced (i.e. different intercepts) but that the slopes of the changes are not significantly different. Corley (2014) tests for a CRE in the usage of negative concord between female and male speakers of Early Modern English, using data from Nevalainen and Raumolin-Brunberg (2003), and again finds no significant difference in slopes. What these case studies show is that the parameters of two different changes may be similar for reasons other than being reflexes of a single abstract grammatical pattern, and thus that identity of slope parameters is not a sufficient condition for the assumption of a single underlying change. Wallenberg (2016, e244), for instance, notes explicitly that these are different populations, and suggests that the similarity of slopes may indicate that “the same forces are underlying the change” in both the English and the Icelandic populations — but however these forces are to be understood, we cannot be dealing with a CRE in the traditional sense, as all these authors recognize.⁴

These two problems are in most cases only technical in the sense that a researcher will usually have independent reasons for ruling out such fantastical hypotheses: in particular, the inference that two apparently separate changes are reflexes of the same underlying phenomenon is usually motivated by a particular structural analysis which is arrived at on independent grounds. However, the theoretical importance of these problems is great: they demonstrate that the standard independent logistics formulation of the CRE can serve at most as a proxy to CREs, not as a model of them. If the CRE is a phenomenon — and the empirical support gathered for it over the last three decades suggests it is — this means we have so far failed to model one of the more well-established facts about language diachrony. Consequently, we have only a very approximate understanding of the dynamics of language change in the presence of contextual factors, and a number of questions remain wide open: if underlying changes are to be thought of as competition between two or more parametric options or grammars, and if CREs are thought to appear because of some sort of performance effects operating over that process of competition, how, exactly, do the two processes interact? What role does the magnitude of the performance effects play in the overall change? Could contexts that favour the innovatory variant be so favouring as to accelerate the change, and if so, can this accelerating effect be quantified and measured? Similarly, could disfavouring contexts slow the change down? Could they even block change in certain cases? These questions can only be answered with the help of

---

⁴Paolillo (2011) raises a problem that may be related. The standard way of testing for the statistical significance of a putative CRE is to perform a chi-square test of independence on the *s* values of the regressions for the different contexts (Kroch, 1989; Santorini, 1993; Pintzuk, 1995). If the result is not statistically significant, then it is concluded that there is support for a CRE. However, it is not sound to treat a non-significant value as evidence for the null hypothesis, since it was assumed to begin with. We acknowledge this problem and have no solution to it in the present paper, except insofar as our method of modelling CREs does not rely on null-hypothesis significance testing at all.
mathematical models of change that accommodate mechanisms for both grammatical competition and contextual effects, and also define, without equivocation, the possible interactions between these two mechanisms over time.

The non-linking problem has, of course, not gone unnoticed in the literature. As Roberts (2007) puts it:

One might wonder why [the CRE] should hold. It is unlikely to be a fact about the grammars themselves. Instead, it is plausible that it may be a fact either about speech communities or about the ways in which individuals choose among grammars available to them. As such, it may be attributable to sociolinguistic factors or to the dynamics of populations, or both factors acting in tandem. (Roberts, 2007, 313)

Our aim in this paper is to propose a solution to the non-linking problem, and our concrete proposal is that the CRE occurs because of context-specific production biases which serve either to promote or to hinder an underlying change in progress. That is to say, we will argue that the CRE is indeed a fact about the ways in which individuals choose among grammars available to them, and propose a rigorous mathematical model of this kind of speaker behaviour. The result is a first step towards a mechanistic model of the CRE that not only describes the diachronic phenomenon but explains it by deriving it from independently plausible first principles of language acquisition and use.

1.1.3 Plan

The paper is structured as follows. In Section I.2, we augment Yang’s (2000) mathematical model of grammar competition with production biases to account for variability across contexts. This results in a dynamical system in which the evolution of the underlying change — a parameter switch for us — and the evolution of the usage frequencies in a number of linguistic contexts feed into each other iteratively. We then derive analytical expressions for the time evolution of this system and show in Section I.3 that in most cases it can be approximated by a constrained set of equations based on one logistic. In Section I.4, this approximation is used to derive a theorem concerning the possible temporal separation between two reflexes of one underlying change: we show analytically that under our proposed model the time separation between contexts always has a finite upper bound which is inversely proportional to the rate of the underlying change. This solves the problem of unconstrained variation in the \( k \) parameters: under our model, it is no longer possible for the curves of different contexts to be radically distant from each other in time. In the remainder of Section I.4, we proceed to test the model empirically from two complementary angles: (1)
by investigating whether time separations observed in a number of previously established CREs agree with the predictions of our time separation theorem, and (2) by testing whether our model is able to distinguish actual CREs from pseudo-CREs, that is, surface changes that proceed at similar rates accidentally but that are clearly not reflexes of one and the same underlying change.

A side product of this investigation is an extension of some of the analytical results in Yang (2000). In Section I.3, we uncover all possible outcomes of the dynamical interplay between grammatical competition and production biases. A full bifurcation analysis of the two-grammar case shows that production biases can both induce change in settings where Yang’s (2000) model outlaws change, and block change in settings where Yang’s (2000) model predicts change. On the assumption that a model which incorporates the possibility of production biases is more realistic than one that does not, then, the assumption that language change is driven (solely) by distributional differences in the proportion of sentences parsed by different competing grammars is shown to be too simple. A theorem resulting from our analysis of the extended model shows that, when production biases are in operation, such differences are neither necessary nor sufficient for change, though they continue to play an important role in any given change process in a way that can be quantified exactly. In Section I.5, we offer a brief account of the nature of production biases; Section I.6 concludes.

I.2 Grammar competition and production biases

I.2.1 Learning competing grammars

Empirical work on language variation and change has demonstrated the limitations of the traditional view of parameter setting as a once-and-for-all process which leaves the learner with a unique grammar at the point of maturation: speakers have, at least during periods of change, access to more than one grammar (Kroch, 1989, 1994, 2000; Santorini, 1992; Pintzuk, 2003). As pointed out by Santorini (1992, 619), this intra-individual co-existence of multiple grammatical systems is “an ability for which the phenomena of multilingualism, diglossia and intrasentential code-switching provide independent and incontrovertible evidence”; see also the discussion in Roberts (2007, 319–331). This notion has been formalized by Yang (2000, 2002b) in his mathematical model of competition-driven change, on which our model of the CRE is based. We therefore begin by reviewing the operating principles behind this model, focussing on the presentation in Yang (2000).

This model construes language change as a learning process in a homogeneous, well-mixing population with non-overlapping generations. At each iteration, we can therefore think of the population as a single individual who sets parameters based on the linguistic output of the previous generation, abstracting away entirely from
the social and geographical structure of that population. In the competing grammars framework, each biologically possible grammar of human language \( G_i \) is associated with a weight which gives the probability of an individual using that grammar. The framework allows any number of those grammars to compete; however, during well-studied and relatively well-understood periods of language change, it usually seems to be the case that two grammars are in competition. Since, additionally, this renders the mathematics of the model particularly tractable, we focus on the two-grammar case in all that follows.

Let \( G_1 \) and \( G_2 \) be these two grammars, and denote their weights with \( p_t \) and \( q_t \), respectively, indexed for generational time \( t \).

The basic insight behind Yang’s (2000) model is that each grammar has its time-independent (parsing) advantage, which is simply the proportion of sentences the other grammar cannot parse (out of all sentences generated, in abstracto, by either grammar). There are then fundamentally three kinds of sentence: sentences of type \( L_1 \), which \( G_1 \) but not \( G_2 \) parses; sentences of type \( L_2 \), which \( G_2 \) but not \( G_1 \) parses; and sentences of type \( L_X \), which both grammars parse (Figure I.3). The language learner receives primary linguistic data (PLD) — the linguistic output of the generation at time step \( t \) — and his task is to arrive at weights \( p_{t+1} \) and \( q_{t+1} \) for the two competing grammars in his own generation. Letting \( \alpha \) denote the advantage of \( G_1 \) and \( \beta \) that of \( G_2 \), then (assuming he samples his environment uniformly) the learner is confronted with a number of sentences drawn

\[ p_t + q_t = 1. \]

\( \text{Figure I.3} \)
Venn diagram of the sentences generated by \( G_1 \) (the set \( L_1 \cup L_X \)) and by \( G_2 \) (the set \( L_2 \cup L_X \)). Here, \( L_2 \) represents a greater proportion of sentences than \( L_1 \), and so the advantage of \( G_2 \) is greater than that of \( G_1 \).
from the following distribution:

\[
\begin{pmatrix}
L_1 & L_2 & L_X \\
G_1 & \alpha p_t & 0 & (1-\alpha)p_t \\
G_2 & 0 & \beta q_t & (1-\beta)q_t
\end{pmatrix}
\]

(I.2)

Based on this input, the learner is assumed to set parameters in accordance with linear reward–penalty learning, an off-the-shelf learning algorithm from mathematical psychology (Bush & Mosteller, 1951, 1958; Narendra & Thathachar, 1989). Of crucial importance here are the two quantities

\[c_t = \beta q_t\] and \[d_t = \alpha p_t = \alpha(1-q_t)\], known as the penalty probabilities of the two grammars: \(c_t\) is the probability of the learner encountering a sentence which \(G_1\) cannot parse and \(d_t\) the probability of a sentence which \(G_2\) cannot parse. It can be shown (Narendra & Thathachar, 1989, 162–163) that, if the learner’s training sample is large enough, eventually he ends up with a weight \(q_t+1\) which is well approximated by

\[q_{t+1} = \frac{c_t}{c_t + d_t}\] (I.3)

Assuming \(c_t \neq 0\) and \(d_t \neq 0\) without loss of generality, this equation may be reduced to the more useful form

\[q_{t+1} = \left(1 + \frac{d_t}{c_t}\right)^{-1} = \left(1 + \frac{1-q_t}{q_t}\right)^{-1}, \] (I.4)

where we write \(\rho = \alpha/\beta\) for the ratio of the parsing advantages. Equation (I.4), then, relates the grammar weights of the \((t+1)\)th generation to those of the \(t\)th generation, thereby defining the inter-generational or diachronic dynamics of a sequence of (reliable) linear reward–penalty learners.5

It follows that \(\alpha < \beta\), or \(\rho < 1\), is a sufficient condition for grammar \(G_2\) to overtake grammar \(G_1\):

5For two competing grammars, the linear reward–penalty learning algorithm assumes the following form for learning rate \(0 < \gamma < 1\) (see Yang, 2000 for more details). Assuming that the learner’s initial guess for the weight of grammar \(G_2\) is \(Q_0 = 0.5\) (no a priori bias), then, for input sentence \(s = 1, \ldots, N\), the learner picks \(G_2\) with probability \(Q_{t-1}\) (and \(G_1\) with probability \(1 - Q_{t-1}\), attempts to parse the sentence, and sets \(Q_s = Q_{t-1} + \gamma\) if \(G_2\) parses \(s\), and \(Q_s = (1-\gamma)Q_{t-1}\) if \(G_2\) does not parse \(s\). Thus, \(Q\) is increased with successful parsing events and decreased with unsuccessful parsing events. Finally, we set \(q_{t+1} = Q_N\). Under the simplifying assumption that \(N \to \infty\), the learner does not have to contend with a finite dataset or a critical period. It is of course false, but like much work in learnability and modelling we adopt it here in order to derive analytical approximations such as (I.4) which would otherwise be difficult, if not impossible, to derive. This approximation holds in the following sense: \(q_{t+1}\) converges to a normal distribution with mean \(q_{t+1} = c_t/(c_t + d_t)\) and a variance which tends to 0 as \(\gamma \to 0\) and \(N\gamma \to \infty\) (Narendra & Thathachar, 1989, 162–163). Assuming a finite learning sample would introduce a stochastic component (noise) to the system, and exploring the consequences of this falls beyond the scope of the present paper.
Theorem I.1 (The Fundamental Theorem of Language Change; Yang, 2000, 239).

Assume reliable learners, so that (I.4) holds. Then \( q_t \to 1 \) as \( t \to \infty \) if \( \alpha < \beta \), and \( q_t \to 0 \) as \( t \to \infty \) if \( \alpha > \beta \).

In other words, the grammar with the greater parsing advantage will necessarily win out in the long term. The difference equation (I.4) may in fact be solved for \( t \) to yield

\[
q_t = \left(1 + \rho \left(\frac{1 - q_0}{q_0}\right)\right)^{-1},
\]

where \( q_0 \) is the weight of \( G_2 \) at the point of actuation of the change (Appendix I.7.1, Corollary I.4): hence as soon as the value of \( q_0 \) is known, the entire change trajectory can be predicted. Furthermore, it is not difficult to show that this solution is equivalent to

\[
q_t = \left(1 + e^{-s(t-k)}\right)^{-1}
\]

with \( s = -\log(\rho) \) and \( k = -\log(\rho)^{-1} \log(q_0^{-1} - 1) \). Thus, assuming that learners receive representative samples of their linguistic environments, a diachronic sequence of such learners exhibits logistic evolution. In particular, the slope of the trajectory is directly dependent on the advantage ratio \( \rho \) such that the smaller \( \rho \) (the more advantageous \( G_2 \) is), the faster the change from \( G_1 \) to \( G_2 \), and vice versa.

This is the gist of the competing grammars model of language change; for more details, see Kroch (1994), Yang (2002b), Pintzuk (2003) and especially Heycock and Wallenberg (2013), who apply the model to a concrete case study involving the loss of verb movement in Scandinavian.

I.2.2 Competing grammars and contextual biases

To account for contextual effects and the CRE, we now assume the existence of \( K \) linguistic contexts \( 1, \ldots, K \), with each sentence generated by \( G_1 \) or \( G_2 \) belonging to one and only one of these contexts. Each context \( i \) is equipped with a context weight \( \lambda_i \) that gives the proportion of sentences that fall in that context (out of all sentences generated by either \( G_1 \) or \( G_2 \)); clearly, since we are dealing with proportions, we require \( \lambda_1 + \cdots + \lambda_K = 1 \) (Figure I.4). In addition to these weights, each context is associated with a fixed (constant over time) production bias \( b_i \) which can be positive, negative or zero. In the first case, the context favours \( G_2 \); in the second, it favours \( G_1 \); and in the third case, the context is neutral with respect to the two grammars.

---

6Formally, this means that the contexts constitute a partition of the set \( L_1 \cup L_X \cup L_2 \) in the usual set-theoretic sense: the contexts are pairwise disjoint subsets of \( L_1 \cup L_X \cup L_2 \) and their union equals the whole of \( L_1 \cup L_X \cup L_2 \).

7Associating positive production biases with a favour for \( G_2 \) over \( G_1 \) (rather than \( G_1 \) over \( G_2 \)) is but a convention and does not affect the dynamics of our model: reverting the biases would merely swap
Now consider a language learner acquiring his grammar weights based on the output of generation \( t \) of speakers. With Yang (2000), we assume that the \( t \)th generation has internalized grammar weights \( p_t \) and \( q_t \). Where our treatment diverges is the effect these weights have on the language acquisition process of the \((t + 1)\)th generation. Rather than assuming that \( p_t \) and \( q_t \) feed directly into the acquisition process in the following generation, we assume that speakers of the \( t \)th generation may promote or demote the two weights \( p_t \) and \( q_t \) in different linguistic contexts in different ways, subject to the context-specific production biases \( b_i \). It is then on the basis of this usage, modulated by the contextual biases \( b_i \) and the context weights \( \lambda_i \), that the next generation of learners must infer their grammar weights.

Letting \( q_t^{(i)} \) denote the probability with which a speaker of the \( t \)th generation uses grammar \( G_2 \) in context \( i \), and similarly for \( p_t^{(i)} \) and \( G_1 \), a general form of this biasing is

\[
\begin{align*}
    p_t^{(i)} &= p_t + F(b_i, p_t) \\
    q_t^{(i)} &= q_t + G(b_i, q_t)
\end{align*}
\]  

(1.7)

where \( F \) and \( G \) are some (yet undetermined) functions which modulate the effect of the bias \( b_i \) on production. These functions must satisfy two requirements:

\[
\begin{align*}
    p_t^{(i)} + q_t^{(i)} &= 1 \quad \text{(as \( G_1 \) and \( G_2 \) are the only grammars)} \\
    0 &\leq p_t^{(i)}, q_t^{(i)} \leq 1 \quad \text{(as the two quantities are probabilities)}
\end{align*}
\]  

(1.8)

and the following is a theorem.

**Theorem I.2.**

Functions \( F = F(b_i, p_t) \) and \( G = G(b_i, q_t) \) satisfy the conditions (1.7) and (1.8) if,

the labels of the two grammars.
and only if, they satisfy

\[
\begin{align*}
F &= -G \\
|F|, |G| &\leq \min\{p_t, q_t\}
\end{align*}
\]  

(I.9)

Proof. Appendix I.7.2.

Theorem I.2 thus implies that the functions \( F \) and \( G \) are necessarily the additive inverse of each other, and that their absolute value is necessarily bounded from above by the minimum of \( p_t \) and \( q_t \). Technically an infinite number of functions satisfy this pair of conditions, so we need to ask what these functions actually are (Figure I.5). The simplest, most parsimonious choice is to consider the product \( p_t q_t \), which is guaranteed to be bounded from above by both \( p_t \) and \( q_t \) whenever \( 0 \leq p_t, q_t \leq 1 \). In other words, we suggest setting

\[
F = -b_ip_tq_t \quad \text{and} \quad G = b_ip_tq_t
\]

(I.10)

with \(-1 \leq b_i \leq 1\). The contextual usage probabilities in (I.7) then assume the definite forms

\[
\begin{align*}
p_t^{(i)} &= p_t - b_i p_t q_t = p_t - b_i p_t (1 - p_t) \\
q_t^{(i)} &= q_t + b_i p_t q_t = q_t + b_i q_t (1 - q_t)
\end{align*}
\]  

(I.11)

where the contextual biases \( b_i \) range from \(-1 \) (maximally \( G_1 \)-favouring) through \( 0 \) (neutral) to \( 1 \) (maximally \( G_2 \)-favouring).

This choice for the functions \( F \) and \( G \) has a number of intuitively satisfying features. For example, (I.10) implies that if either \( p_t = 1 \) or \( q_t = 1 \), then \( F = G = 0 \) (since in the first case \( q_t = 0 \) and in the second case \( p_t = 0 \) and consequently \( p_t q_t = 0 \)) and no biasing will apply. Empirically, this means that if a grammar has been acquired categorically, no contextual biases will be able to skew usage in the direction of the other grammar. This is intuitively right: if a grammatical option has been acquired categorically, then by definition the competing option does not exist for the speaker and no grammar-external biasing ought to be able to apply. This behaviour of our biasing mechanism in the limits \( q_t \rightarrow 1 \) and \( q_t \rightarrow 0 \) is just one manifestation of a more general feature of the model: that while the biases \( b_i \) themselves are constant and do not change over time, the magnitude of the effect of these biases on usage does depend on the state of the underlying change: the effect is the strongest midway through the change (from Figure I.5, we see that the effect is the strongest when \( q_t = 0.5 \)) and tails off to zero in the limits \( q_t \rightarrow 1 \) (completion) and \( q_t \rightarrow 0 \) (actuation). As we will see in Section I.4, this is what the empirical data also show.\(^8\)

With (I.11) in place, it is possible to work out the diachronic, inter-generational

\(^8\)We further elaborate on the empirical grounding of our biasing mechanism in Section I.5.
The biasing functions $F$ and $G$ have to satisfy the requirement $|F|, |G| \leq \min\{p_t, q_t\} = \min\{q_t, 1 - q_t\}$ (see Appendix I.7.2 for a proof) and thus land in the shaded region of this plot. The parabolic curve shown here gives the most parsimonious such upper bound, the product $p_t q_t = q_t (1 - q_t)$.

dynamics of our model (assuming, again, that learners receive large input samples). What generation $t$ outputs in this extended model is not the distribution given in (I.2), but a combination of grammar advantages ($\alpha$ and $\beta$), grammar weights ($p_t$ and $q_t$), context weights ($\lambda_i$) and context biases ($b_i$). The penalty probability for grammar $G_1$ now becomes

$$c_t = \beta \sum_{i=1}^{K} \lambda_i q_t^{(i)} = \beta \sum_{i=1}^{K} \lambda_i (q_t + b_i p_i q_t) = \beta \left( q_t + \sum_{i=1}^{K} \lambda_i b_i p_i q_t \right) = \beta (q_t + B p_t q_t), \quad (I.12)$$

where the index $i$ runs through the contexts $i = 1, \ldots, K$ and where we write $B = \sum_{i=1}^{K} \lambda_i b_i$ for convenience.\footnote{Note that $-1 \leq B \leq 1$, since $0 \leq \lambda_i \leq 1$ and $-1 \leq b_i \leq 1$.} The quantity $B$, which may be regarded as the net bias operating on the language acquisition process weighted by the context proportions $\lambda_i$, turns out to be a decisive quantity in our model: from (I.12), we immediately see that if $B = 0$, the penalty $c_t$ reduces to the Yangian penalty $c_t = \beta q_t$. Our model, then, generalizes Yang’s (2000) model and reduces to the latter in the special case that the contextual biases are ‘in balance’ — if either all the biases are zero or if $G_2$-favouring (positive) biases cancel out the effect of $G_1$-favouring (negative) biases.
Entirely symmetrically, the penalty for grammar $G_2$ reads

$$d_t = \alpha \sum_{i=1}^{K} \lambda_i p_t^{(i)} = \alpha (p_t - Bp_t q_t). \quad \text{(I.13)}$$

Assuming reliable learners, we may now use these penalty probabilities to write down the inter-generational difference equation that relates $q_{t+1}$ to $q_t$ for the extended model: equation (I.4) becomes

$$q_{t+1} = \left(1 + \frac{d_t}{c_t}\right)^{-1} = \left(1 + \frac{\alpha (p_t - Bp_t q_t)}{\beta (q_t + Bp_t q_t)}\right)^{-1}. \quad \text{(I.14)}$$

Recalling that $p_t = 1 - q_t$, this may be written as

$$q_{t+1} = \left(1 + \Lambda_t \rho \frac{1 - q_t}{q_t}\right)^{-1}, \quad \text{(I.15)}$$

where $\rho = \alpha / \beta$ as before and

$$\Lambda_t = \frac{1 - Bq_t}{1 + B(1 - q_t)}. \quad \text{(I.16)}$$

### I.2.3 The Constant Rate Effect

To summarize, we propose to augment Yang’s (2000) model of grammar competition with a set of production biases $b_i$ which modulate the grammar weights $p_t$ and $q_t$ in actual linguistic production. This modulation is implemented by a mechanism which, we have shown, has to operate within certain analytical bounds. Within those bounds, we have suggested that the most parsimonious mechanism be adopted, corresponding to our particular choice of the bias-modulating functions $F$ and $G$, as explained above. The diachronic behaviour of this extended model is characterized by equations (I.11) and (I.15): the difference equation (I.15) gives the evolution of the underlying grammar weight $q_t$, whilst equation (I.11) supplies the context-specific value of this probability, modulated by the contextual production biases. The flowchart in Figure I.6 illustrates the inter-generational dynamics that result from this mechanism, comparing our extension of Yang’s (2000) model to the original.

Before moving on to an empirical evaluation of our proposed model, we first ask whether it produces, in broad qualitative terms, the right kind of behaviour. To this end, Figure I.7 shows the behaviour of our model in two different situations involving three arbitrarily chosen contexts: in a situation in which the contextual biases are in balance and cancel each other out ($B = 0$; Figure I.7a), and in a situation in which the net effect of biases in favour of the conventional variant $G_1$ conspire against the propagation of the innovative variant $G_2$ ($B < 0$; Figure I.7b). Impressionistically, our
Inter-generational change in Yang’s (2000) model (top) and our model (bottom). After parameter setting, the learner ends up with a weight $q_t$ for grammar $G_2$ (and $p_t$ for grammar $G_1$). In our model, this weight is then attenuated in production by the context-specific production biases $b_i$ so that the actual probability of using $G_2$ in the $i$th context is $q_t^{(i)} = q_t + b_i p_t q_t$ (see text for details). This biased probability, together with the advantage ratio $\rho = \alpha / \beta$ and the context weight $\lambda_i$, then determines the PLD for the following generation.

The model produces a CRE in both cases: the probability of use of $G_2$ increases roughly at the same rate in each context, with a characteristic temporal shift between the propagation curves of the individual contexts. This suggests that our model is able to replicate the central intuition of Kroch (1989) that different reflexes of one underlying change ought to proceed at similar rates, and that the output of our model can, in principle, approximate the empirical situations that have been suggested as CREs in the literature.

In these two cases, the evolution of the underlying probability $q_t$ is different, however, because of the different biasing that applies in each case. In Figure I.7a the contexts are ‘in balance’ ($B = 0$), which by the preceding analysis implies that the evolution of $q_t$ itself is logistic. In Figure I.7b, on the other hand, $G_1$-favouring biases outweigh $G_2$-favouring biases ($B < 0$), hindering the propagation of $G_2$. This is reflected in the fact that the evolution of the underlying $q_t$ is slowed down. Even though $G_2$ still overtakes $G_1$ in the limit, the trajectory of $q_t$ is no longer strictly logistic (careful examination shows that it is not symmetric about the midpoint $q_t = 0.5$, but rather exhibits slower change for $q_t < 0.5$ and faster change for $q_t > 0.5$). This motivates us to consider extreme model parameter regimes, particularly the subspace where $B$ is negative, in more detail.

In equation (I.15), the factor $\Lambda_t$ depends on $q_t$ whenever $B \neq 0$. This complicates
The behaviour of our model with two different sets of production biases, for advantage ratio $\rho = 0.5$ and initial value $q_0 = 0.01$ (1% usage of $G_2$ at the point of actuation): the evolution of both the underlying probability $q_t (\bullet)$ as well as that of the contextual usage probabilities $q_t^{(1)} (\circ)$, $q_t^{(2)} (\times)$ and $q_t^{(3)} (\square)$ is shown up to $q_t = 1 - q_0 = 0.99$. In each case, the context weights are set at $\lambda_1 = 0.2$, $\lambda_2 = 0.4$ and $\lambda_3 = 0.4$. (a) Here the biases are $b_1 = 1$, $b_2 = -1$ and $b_3 = 0.5$. With these choices, $B = \lambda_1 b_1 + \lambda_2 b_2 + \lambda_3 b_3 = 0$, and consequently the positively-biased contexts (\circ and \square) cancel out the effect of the negatively-biased context (\times), resulting in logistic evolution of the underlying probability $q_t (\bullet)$. (b) Here the biases are $b_1 = 1$, $b_2 = -1$ and $b_3 = -0.5$. Now $B = \lambda_1 b_1 + \lambda_2 b_2 + \lambda_3 b_3 = -0.4 < 0$. The two negatively-biased contexts (\square and \times) outweigh the one positively-biased context (\circ), and as a consequence, the change from $G_1$ to $G_2$ takes much longer than in (a). The trajectory of $q_t$ is also not strictly logistic in this case, as is evident from the fact that it is not symmetric about the midpoint $q_t = 0.5$: passage from $q_0 = 0.01$ to $q_t = 0.5$ takes longer than passage from $q_t = 0.5$ to the final value $q_t = 0.99$. 

**Figure I.7**
the analysis of the extended model significantly: while the Yangian equation (I.4) can be solved for \( t \) to yield the logistic function, we are not aware of a closed-form solution to the more complex nonlinear difference equation (I.15) except in the singular case \( B = 0 \), where the equation reduces to (I.4). This has the undesirable practical consequence that there is no trivial way of fitting our model to data — lacking a closed-form curve for the underlying probability \( q_t \) from which to derive curves for the contextual reflexes \( q_t^{(i)} \), there simply are no closed-form contextual curves which to fit. The best one can do is to iterate the model for various choices of model parameter values and initial conditions and compare the resulting trajectories against empirical data, an approach which soon becomes computationally prohibitive as the number of logically possible model parameter combinations grows as a superlinear function of the number of model parameters. To tackle this problem, we will in the next section conduct a full analysis of the behaviour of our model in the limit \( t \to \infty \) and show that, under most empirically meaningful combinations of model parameter values, the underlying trajectory \( q_t \) is well approximated by a logistic curve. Thus, even though we cannot write down the solution of \( q_t \) for arbitrary times \( t \), and even though we know that for some parameter values (such as when \( B < 0 \)) the evolution of \( q_t \) is not logistic, we can use logistic functions to approximate the true value of \( q_t \). This will form the basis of our curve-fitting procedure in Section I.4. A reader who is willing to skip the technicalities of the logistic approximation may advance straight to Section I.4.

I.3 Dynamics of the extended model

I.3.1 Advantage versus bias

As we have noted above, the Fundamental Theorem of Yang’s (2000) model is that a more advantageous grammar will necessarily overtake a less advantageous one: if \( \rho < 1 \) (\( \alpha < \beta \)) and learners are reliable, then \( q_t \to 1 \) as \( t \to \infty \), and thus grammar \( G_2 \) overtakes \( G_1 \) (Theorem I.1). A nontrivial consequence of extending the model with production biases is that this theorem no longer holds: a difference in the proportion of input parsed by the two competing grammars is neither sufficient nor necessary for language change. While this is a minor observation from the point of view of the CRE, which is the main focus of the present paper, the failure of the Fundamental Theorem under suitable combinations of grammar advantages and production biases is an interesting finding from the vantage point of the theory of language change in general, and we will therefore pursue it briefly in this section. The bifurcation scenario here outlined will also play a role in the logistic approximation that we develop in the following subsection for model evaluation purposes.

The production biases \( b_i \) can be positive, negative or zero. In the first case, the
context in question favours $G_2$ over $G_1$; in the second case, $G_1$ is favoured; and in the third case, the context is neutral. The scalar product $B = \sum_{i=1}^{K} \lambda_i b_i$ of context weights and production biases turns out to play a critical role in determining how, and if, change from $G_1$ to $G_2$ happens. If there are negatively biased ($G_2$-disfavouring) contexts, and if their share of all sentences in the language learner’s PLD is large enough, change from $G_1$ to $G_2$ can be blocked even if the advantage of $G_2$ is greater than the advantage of $G_1$. On the other hand, if there are sufficiently strong positively biased ($G_2$-favouring) contexts, $G_2$ may overtake $G_1$ even if the latter’s advantage exceeds that of the former. A critical value $B_c$ of the net bias $B$ in fact exists such that change from $G_1$ to $G_2$ is guaranteed whenever $B > B_c$ but is blocked whenever $B \leq B_c$:

**Theorem I.3 (The Extended Fundamental Theorem of Language Change).**

Assume reliable learners, so that (I.15) holds. Let $q_0$ be the weight of grammar $G_2$ at the point of actuation, let $B = \sum_{i=1}^{K} \lambda_i b_i$, and let

$$B_c = \frac{\rho - 1}{1 + q_0(\rho - 1)}.$$  \hspace{1cm} (I.17)

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, $B > B_c$.

**Proof.** Appendix I.7.3.

In dynamical-systems terminology, the production bias mechanism induces a bifurcation in the parameter space of the extended model: small tweaks made to either the biases ($b_i$) or to the proportion of input falling in each context ($\lambda_i$) can alter the trajectory of language change entirely by determining which of the two grammars will win out (Figure I.8). An immediate consequence of Theorem I.3 is that if no context is negatively biased, $G_2$ will overtake $G_1$ whenever $\rho < 1$:

**Corollary I.1.**

If $\rho < 1$ and $b_i \geq 0$ for all contexts $i$, $q_t \to 1$ as $t \to \infty$.

**Proof.** Since $B_c < 0$ for any choice of $q_0$, if $\rho < 1$. 

108
Even though production biases, then, can induce or block change in parameter regimes where such behaviour is impossible in Yang’s (2000) original model, there are limits to how much of an effect the biases can have over grammar advantages. Briefly put, if \( G_2 \) is much more advantageous than \( G_1 \) \((0 < \rho \ll 1)\), then no amount of negative bias can block change, and, on the other hand, if \( G_1 \) is much more advantageous than \( G_2 \) \((\rho \gg 1)\), no amount of positive bias can make \( G_2 \) overtake \( G_1 \). How much is much depends on the boundary condition \( q_0 \):

**Corollary I.2.**

If \( \rho < q_0/(1 + q_0) \), then \( q_t \to 1 \) as \( t \to \infty \), regardless of the value of \( B \). If \( \rho > (2 - q_0)/(1 - q_0) \), then \( q_t \to 0 \) as \( t \to \infty \), regardless of the value of \( B \).

*Proof.* Clearly \(-1 \leq B \leq 1\) since \(-1 \leq b_i \leq 1\) and \(0 \leq \lambda_i \leq 1\). If \( \rho < q_0/(1 + q_0) \), then \( B_c < -1 \). If \( \rho > (2 - q_0)/(1 - q_0) \), then \( B_c > 1 \).

Figure I.8 illustrates.

**I.3.2 Logistic approximation**

The above results show that the outcome of grammar competition in the presence of context-specific production biases is determined by a complicated interaction between these biases \((b_i)\), the proportion of input that falls in each context \((\lambda_i)\) and the ratio of the parsing advantages of the two competing grammars \((\rho)\). This is because in our model the language acquisition mechanism and the production bias mechanism constitute a feedback loop across iterated applications over multiple generations of language learners, the production biases modulating the acquisition of the grammatical weights \(p_t\) and \(q_t\). As we noted in Section I.2.3 (Figure I.7), this feature of the model also implies that when the production biases are particularly strong, they will cause the evolution of the underlying grammar weights to be non-logistic. The feedback loop gives rise to a nonlinear difference equation for which we have no solution in the general case, and the following problem immediately arises: how can the predictions of our model be tested against empirical data if there is no closed-form curve which to fit?

Even though the evolution of \( q_t \) is, strictly speaking, logistic only when \( B = 0 \), eyeballing trajectories such as the one in Figure I.7b suggests that these trajectories are still S-shaped and perhaps well approximated by logistics. To explore this possibility, we performed a sweep across the model parameter space, generating trajectories of \( q_t \) from the initial condition \( q_0 = 0.01 \) (1% usage of \( G_2 \) at the point of actuation) in the regime \( B > B_c \) (i.e. in the parameter regime where \( G_2 \) is guaranteed to oust \( G_1 \) by Theorem I.3), until \( q_t \) had reached the value \( q_t = 1 - q_0 = 0.99 \). We then proceeded
The outcome of language change for different combinations of advantage ratio $\rho$ and net bias $B$, for two different initial weights for $G_2$: $q_0 = 0.001$ (left) and $q_0 = 0.9$ (right). The thick curve corresponds to the subset of this parameter space where $B = B_c$, the critical bifurcation value. If $B > B_c$, grammar $G_2$ overtakes; if $B < B_c$, grammar $G_1$ prevails; and if $B = B_c$, the system falls in an equilibrium where $q_t = q_0$ for all $t$ (Theorem I.3). Yang’s (2000) model corresponds to the dashed line running at $B = 0$ (no contextual effects, or contexts wholly in balance) and thus predicts that $G_2$ wins for any $0 < \rho < 1$ and loses for any $\rho > 1$. Since $B$ has both a lower and an upper limit ($-1 \leq B \leq 1$), the advantage ratio $\rho$ has critical values, dependent on the boundary condition $q_0$, such that if $\rho$ lands beyond one of these values, no amount of out-of-balance bias can overthrow the advantage-induced dynamical outcome (Corollary I.2). In the figure on the left, for instance, any ratio $\rho$ greater than about 2 guarantees $G_2$ to fail. In the figure on the right, any ratio $\rho$ smaller than about 0.5 guarantees $G_2$ to succeed, no matter what the combination and magnitude of contextual biases.
to fit a logistic curve to each of these trajectories in order to investigate how well the trajectory may be approximated by a logistic. Figure I.9a gives the errors of these fits, showing that the trajectories are closely approximated by logistics whenever $\rho$ is not too large and $B$ is not too close to the critical bifurcation threshold $B_c$.

Figures I.9b–c supply the best slope ($s$) and intercept ($k$) coefficients found by these regressions. We find that $s$ is a decreasing function of $\rho$ and an increasing function of $B$: the more advantageous $G_2$ is, and the more $G_2$ is favoured by the production biases, the steeper the underlying change, as one would expect. The intercept coefficient $k$, in turn, is an increasing function of $\rho$ and a decreasing function of $B$: the less advantage $G_2$ has and the more the production biases tend to disfavour $G_2$, the more the curve of the underlying change is shifted towards positive time.

I.4 Evaluation

I.4.1 The Time Separation Theorem

Under the logistic approximation from Section I.3, the usage of grammar $G_2$ in the contexts $i = 1, \ldots, K$ is described by a set of $K$ equations

$$
\begin{align*}
q_t^{(1)} &= \tilde{q}_t + b_1 \tilde{q}_t (1 - \tilde{q}_t) \\
&\vdots\\
q_t^{(K)} &= \tilde{q}_t + b_K \tilde{q}_t (1 - \tilde{q}_t)
\end{align*}
$$

(I.18)

where $\tilde{q}_t$ is a logistic function approximating the true underlying probability $q_t$. What historical language corpora give us are usage frequencies in various contexts, and we therefore wish to fit curves of the form (I.18) to such data. The fact that under the logistic approximation all such curves are tied to $\tilde{q}_t$, which itself has a closed-form solution, now facilitates this empirical evaluation: even though the individual context curves $q_t^{(i)}$ themselves are not logistic (unless $b_i = 0$), they are easily derived from one that is. In what follows, we will take a look at a number of case studies, fitting context curves with the help of a nonlinear least squares optimization algorithm.

Estimating the goodness of fit of these regressions is one important goal of this exercise. However, our main aim is to solve the non-linking problem identified in Section I.1.2. Specifically, we wish to demonstrate that our model does not allow arbitrarily large time separations between contexts, operationalized as the difference between the points in time at which different context curves reach their tipping point, or the point in time at which the context frequency of the overtaking grammar equals 0.5 when (I.18) is generalized for real-valued $t$. The logistic approximation gives us a straightforward proof of this.
Figure I.9

(a) Error of fit (sum of squared residuals; nonlinear least squares regression) of a logistic function to trajectories of the underlying probability $q_t$ generated by our model for various combinations of advantage ratio $\rho$ and net bias $B$, for initial condition $q_0 = 0.01$ (1% usage of $G_2$ at the point of actuation). The dashed vertical lines give the critical value $B_c$ of the bifurcation parameter for each selection of $\rho$. We find that trajectories of $q_t$ are closely approximated by logistics except in the immediate vicinity of the bifurcation threshold $B_c$ at which change from $G_1$ to $G_2$ is blocked. (b–c) Best-fitting slope ($s$) and intercept ($k$) values found by these regressions.
Theorem I.4 (The Time Separation Theorem).

For any two contextual reflexes of an underlying change from \( G_1 \) to \( G_2 \) approximated by a logistic \( \tilde{q}_t \) with slope \( s \), the maximal time separation at tipping points is

\[
\Delta(s) = \frac{2}{|s|} \log \left( \frac{1}{\sqrt{2} - 1} \right) \approx 1.76 \frac{1}{|s|}.
\]  

(I.19)

Proof. Appendix I.7.4.

It is to be noted that \( \Delta(s) \) is inversely proportional to \( s \) — the slower the rate of change, the more time separation is allowed between any two contexts and vice versa.

To fit the system (I.18) to a set of data points, we first define reasonable ranges of variation for the \( s \) and \( k \) parameters of the logistic \( \tilde{q}_t \) that we wish to probe. We then loop through the values contained in these ranges, finding the best fitting bias parameters \( b_i \) for each pair \((s, k)\) using a nonlinear least squares optimization algorithm such as the Gauss–Newton procedure (Bates & Watts, 1988), bearing in mind the bounds \(-1 \leq b_i \leq 1\). Finally, out of all these regressions, we pick the combination of \( s, k \) and \( b_i \) that provides the best fit to the data in question. The whole procedure is detailed in pseudocode in Appendix I.7.5.\(^{10}\)

We now proceed to an evaluation of the model by comparing its predictions against three historical changes for which a CRE has been reported in the literature: the emergence of periphrastic \textit{do} in the history of English (Kroch, 1989), the earliest stages of the English Jespersen Cycle (Wallage, 2013), and, to take a phonological example to illustrate the generality of the procedure, the loss of final fortition in Early New High German (Fruehwald et al., 2009). In Sections I.4.2–I.4.4 we first briefly summarize the linguistics of each change, reproduce the relevant empirical data, and visualize the fit of our model to the data when the regression is conducted using the procedure outlined above. In Section I.4.5, we take a more quantitative angle and report the numerical errors of these fits, comparing them to the errors that an application of the standard procedure based on individual logistics (cf. Section I.1.1) would produce. Finally, in Section I.4.6, we take a look at a pseudo-CRE — a case where the standard independent logistics operationalization reports a CRE but where this conclusion is patently absurd from other considerations (cf. Section I.1.2) — in order to investigate whether or not our model, too, is prone to report false positives in such cases.

\[1.4.2\] Periphrastic \textit{do} in English

The first case study we will consider is perhaps the best known instance of a CRE: Kroch’s (1989) interpretation of Ellegård’s (1953) data on the rise of periphrastic \textit{do}
in Early Modern English. The variable in question is whether a form of do is used in certain contexts, as in I.20, or not, as in I.21 (examples from Kroch, 1989, 216).

(I.20) Where doth the grene knyght holde hym?
   ‘Where does the Green Knight hold him?’

(I.21) How great and greuous tribulations suffered the Holy Appostyls . . . ?
   ‘How great and grievous tribulations did the Holy Apostles suffer?’

In modern standard English, a form of do is required in a number of contexts, including all interrogatives as well as negative declaratives. What Ellegård’s (1953) data show is that, on the surface, the use of do appears to ‘take off’ in the different contexts at different rates: for instance, between around 1500 and 1650, negative questions exhibit a much higher proportion of do than affirmative wh-object questions or negative declaratives, though the latter contexts eventually catch up (Table I.1). Kroch (1989) conducted a regression on these contexts and showed that logistic curves fitted to them individually did not differ much in their slope (s) parameters, although they did differ in terms of the intercept (k) parameter. This particular example, while perhaps the most celebrated instance of a CRE, is in fact not the most straightforward instance found in the literature, primarily due to a ‘dip’ in the later portion of the change in some contexts, which makes the progression of do non-monotonic; see Warner (2005) and Ecay (2015) for detailed discussion, concluding that other factors (and possibly another grammar) are at play. Kroch (1989) also identifies the dip and consequently focusses on the first seven data points of Ellegård’s (1953) data only; we follow his practice here.

When the algorithm from Appendix I.7.5 is used to fit a model of the form (I.18) to these data, the picture in Figure I.10 emerges. On visual inspection, the fit is a good one. Crucially, our model is constrained to allow only so much time separation between any two contexts of one change (Theorem I.4), illustrated in Figure I.10 as the horizontal bar extending both ways from the tipping point of the curve of the underlying grammar probability. We find that fitting the model to Ellegård’s (1953) data drives two of the context curves (negative questions and affirmative object questions) to the very extremes of the range licensed by the model — in other words, for these contexts, the production biases need to be maximal in order for the model to fit the data. Crucially, however, the data are described well by the regression curves so obtained, an observation which we back up quantitatively in Section I.4.5.

I.4.3 English Jespersen Cycle

Our second case study, also from the history of English, involves the replacement of preverbal ne/ni by postverbal not during the Middle English period. This change
Table I.1
Proportion of *do* in five different contexts: negative declaratives, negative questions, affirmative transitive questions, affirmative intransitive questions, affirmative *wh*-object questions. From Kroch (1989, 224, Table 3).

<table>
<thead>
<tr>
<th>Period</th>
<th>neg. dec.</th>
<th>neg. q.</th>
<th>aff. tr. q.</th>
<th>aff. intr. q.</th>
<th>aff. obj. q.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1400–1425</td>
<td>0.000</td>
<td>0.117</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1425–1475</td>
<td>0.012</td>
<td>0.080</td>
<td>0.107</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1475–1500</td>
<td>0.048</td>
<td>0.111</td>
<td>0.135</td>
<td>0.000</td>
<td>0.020</td>
</tr>
<tr>
<td>1500–1525</td>
<td>0.078</td>
<td>0.590</td>
<td>0.242</td>
<td>0.211</td>
<td>0.113</td>
</tr>
<tr>
<td>1525–1535</td>
<td>0.137</td>
<td>0.607</td>
<td>0.692</td>
<td>0.197</td>
<td>0.095</td>
</tr>
<tr>
<td>1535–1550</td>
<td>0.279</td>
<td>0.750</td>
<td>0.615</td>
<td>0.319</td>
<td>0.110</td>
</tr>
<tr>
<td>1550–1575</td>
<td>0.380</td>
<td>0.854</td>
<td>0.737</td>
<td>0.423</td>
<td>0.360</td>
</tr>
<tr>
<td>1575–1600</td>
<td>0.238</td>
<td>0.648</td>
<td>0.792</td>
<td>0.444</td>
<td>0.383</td>
</tr>
<tr>
<td>1600–1625</td>
<td>0.367</td>
<td>0.937</td>
<td>0.773</td>
<td>0.619</td>
<td>0.298</td>
</tr>
<tr>
<td>1625–1650</td>
<td>0.317</td>
<td>0.842</td>
<td>0.909</td>
<td>0.757</td>
<td>0.530</td>
</tr>
<tr>
<td>1650–1700</td>
<td>0.460</td>
<td>0.923</td>
<td>0.947</td>
<td>0.702</td>
<td>0.549</td>
</tr>
</tbody>
</table>

Figure I.10
Fit of our model to the data on English periphrastic *do* (curves: model; points: data from Table I.1, first seven periods). On visual inspection, the fit to each context is a good one. Theorem I.4 implies a maximal time separation, illustrated here as the horizontal bar extending both ways from the tipping point of the theoretical curve for the underlying grammatical change (no production biases). The best-fitting parameters found by the regression are $s = 0.031$, $k = 1547.677$, with bias sizes $b_i$ as follows: $-0.885$ for negative declaratives, 1.000 for negative questions, 0.656 for affirmative transitive questions, $-0.647$ for affirmative intransitive questions, and $-1.000$ for affirmative *wh*-object questions. With $s = 0.031$, the maximal time separation licensed by the model is roughly 56.5 years.
involves an intermediate stage in which both *ne* and *not* co-occur. The three stages are illustrated in I.22–I.24 (examples from Wallage, 2008, 644).

(I.22) *we ne moten halden Moses e lichamliche*
   *we neg need observe Moses’ law bodily*
   ‘we need not observe Moses’ law in body’

(I.23) *ac of hem ne speke ic noht*
   *but of them neg spoke I not*
   ‘but I did not speak of them’

(I.24) *I know nat the cause*
   *I know not the cause*
   ‘I do not know the cause’

This replacement of negators, a cross-linguistically common diachronic pathway, is referred to as Jespersen’s Cycle; see Wallage (2008) and Ingham (2013) for detailed discussion of the English development. For our purposes, the change that is important is the replacement of Stage 1 of Jespersen’s Cycle — negation by *ne* alone, as in I.22 — with Stage 2, bipartite negation, as exemplified by I.23. Wallage (2013) shows that Stage 2 is favoured with discourse-old propositions during the middle of the change, but that a CRE obtains (Table I.2). Again, on purely visual inspection, our model fits the data well, and the variation observed between discourse-old and discourse-new propositions falls, roughly, within the time bounds prescribed by the Time Separation Theorem (Figure I.11).

### I.4.4 Loss of final fortition in Early New High German

CREs are not found only with syntactic variables. Fruehwald et al. (2009) reanalyse data from Glaser (1985) on the loss of final fortition in (Bavarian) Early New High German, which is observable in orthographic variation of the period, e.g. *tak* vs. *tag* ‘day (acc. sg.)’, *rat* vs. *rad* ‘counsel (acc. sg.)’. They argue that the orthographic variation clearly represents a phonological change in progress rather than shifting

---

**Table I.2**

Proportion of Stage 1 negation (preverbal *ne*) in the English Jespersen Cycle in discourse-old and discourse-new contexts. From Wallage (2013, 12, Table 1).

<table>
<thead>
<tr>
<th>Period</th>
<th>discourse-old</th>
<th>discourse-new</th>
</tr>
</thead>
<tbody>
<tr>
<td>1150–1250</td>
<td>38/245 (0.155)</td>
<td>335/393 (0.852)</td>
</tr>
<tr>
<td>1250–1350</td>
<td>9/338 (0.027)</td>
<td>135/346 (0.390)</td>
</tr>
<tr>
<td>1350–1420</td>
<td>0/244 (0.000)</td>
<td>2/294 (0.007)</td>
</tr>
</tbody>
</table>
Figure I.11
Fit of our model to the data on the first two stages of the English Jespersen Cycle (curves: model; points: data from Table I.2). On visual inspection, the fit to each context is a good one, though the poor time resolution of the data is a problem. Theorem I.4 implies a maximal time separation, illustrated here as the horizontal bar extending both ways from the tipping point of the theoretical curve for the underlying grammatical change (no production biases). The best-fitting parameters found by the regression are \( s = -0.016 \), \( k = 1203.788 \), with bias sizes \( b_i \) as follows: \(-1.000\) for discourse-old propositions and \(1.000\) for discourse-new propositions. (Note that in a case like this where the slope \( s \) is negative, a negative context bias means a preference for the overtaking grammar, whereas a positive bias indicates preference for the receding one.) With \( s = -0.016 \), the maximal time separation licensed by the model is roughly 110 years.
Table I.3
Proportion of fortition of the three plosives /b/, /d/ and /g/ in Early New High German. From Fruehwald et al. (2009, 4, Table 1).

<table>
<thead>
<tr>
<th>Year</th>
<th>/b/</th>
<th>/d/</th>
<th>/g/</th>
</tr>
</thead>
<tbody>
<tr>
<td>1276</td>
<td>18/18 (1.00)</td>
<td>29/29 (1.00)</td>
<td>54/73 (0.74)</td>
</tr>
<tr>
<td>1373</td>
<td>10/18 (0.56)</td>
<td>24/29 (0.83)</td>
<td>17/76 (0.22)</td>
</tr>
<tr>
<td>1483</td>
<td>2/18 (0.11)</td>
<td>2/24 (0.08)</td>
<td>0/78 (0.00)</td>
</tr>
<tr>
<td>1523</td>
<td>2/16 (0.12)</td>
<td>3/9 (0.33)</td>
<td>0/73 (0.00)</td>
</tr>
</tbody>
</table>

scribal tradition, and that fortition is the result of a single phonological rule whose loss is visible to different degrees in different contexts during the period of the change: /d/ exhibits fortition the most and /g/ the least, with /b/ showing an intermediate pattern (Table I.3). Our model describes the data well, with the observed CRE again falling within the time bounds implied by the model (Figure I.12).

This example illustrates that even though our model is based on the variational learner in Yang (2000), which is essentially a hypothesis about parameter setting in syntax, the logistic approximation (I.18) which underlies the curve-fitting procedure can legitimately be used to model CREs in any domain as long as the assumption of an underlying logistic change is justifiable. As Fruehwald et al. (2009, 9) point out, “[t]he discovery of the Constant Rate Effect in phonological change is perfectly expected under normal generative theories of phonology when the mechanism of change is grammar or rule competition” — the learning algorithm a language learner uses in this case may (or may not) be different from the one assumed in Yang’s (2000) model, but this notwithstanding, as long as some sort of underlying representation similar to the Yangian weights \( p_t \) and \( q_t \) can be assumed to exist, our production bias mechanism may be applied.

I.4.5 Comparison with standard procedure

Sections I.4.2–I.4.4 have adduced evidence to the effect that the model introduced in this paper can account for the CRE: the model gives fair fits to historical data even though it is constrained by the chronological bounds set by the Time Separation Theorem (Theorem I.4). To make this argument more quantitatively, in this section we compare these three fits to the standard operationalization of the CRE: that is, logistic curves agreeing in the \( s \) (slope) parameter but differing in their \( k \) (intercept) parameters. For each of the three case studies considered in Sections I.4.2–I.4.4, then, we carry out two fits: one for our model, and another one for a model consisting of a set of logistic curves (one per context) where the \( s \) parameter is not allowed to vary between contexts but where such variation is allowed for the \( k \) parameter.
Fit of our model to the data on loss of final fortition in Early New High German (curves: model; points: data from Table I.3). On visual inspection, the fit to each context is a good one. Theorem I.4 implies a maximal time separation, illustrated here as the horizontal bar extending both ways from the tipping point of the theoretical curve for the underlying grammatical change (no production biases). The best-fitting parameters found by the regression are $s = -0.019$, $k = 1374.747$, with bias sizes $b_i$ as follows: 0.353 for /b/, 1.000 for /d/, and $-1.000$ for /g/. With $s = -0.019$, the maximal time separation licensed by the model is roughly 93 years.
We quantify the goodness of fit of these regressions in the usual way, by the normalized sum of squared residuals: in other words, for each context of a given change, for each time period, we calculate the displacement between the empirically attested frequency and the value predicted by the model, square this displacement, sum over all time periods and over all contexts, and divide by the number of data points. Thus, the better the fit, the closer the sum of squared residuals is to zero. Some deviance from zero is always to be expected because of the noisy nature of historical language data. However, this measure is still able to capture the difference between models which are good fits, but subject to noise, from models which are simply bad fits to the data in question.

Figure I.13 shows a comparison of the goodness of fit of the two models for each of the three case studies, operationalized using the sum of squared residuals. The crucial finding is that our model, which places more constraints on the shape and placement of the regression curves, fares no worse than the latter in two out of three cases: in other words, a more constrained, theoretically motivated model which generates empirical predictions (in the form of the Time Separation Theorem) performs no worse than a less constrained, theoretically unmotivated model. The exception to this are the data on the English Jespersen Cycle, where the less constrained standard formulation reports a very low error. This appears to be a result of the very small number of data points — just three time periods and two contexts — for this particular case study. Low data resolution necessarily gives a disproportionate advantage to the less constrained model over any model that incorporates more assumptions.

I.4.6 A pseudo-CRE

Above, we have shown that the proposed model can account for the CRE, in the sense that it gives good fits to three historical changes — fits which are, in two of these cases, no worse than fits conducted using the standard operationalization of CREs. It remains to be shown that introducing this more constrained model can actually solve some of the underspecification issues the standard operationalization suffers from. As discussed in Section I.1.2, the method of ‘same slopes, different intercepts’ is susceptible to false positives: the fact that a number of logistics agree in their slope parameters is insufficient evidence that a single underlying change is at hand (see Corley, 2014, Wallenberg, 2016, and Willis, 2017 for examples where the ‘contexts’ cannot possibly be assumed to be evidence of underlying grammatical unity).

Here, we construct a pseudo-CRE by combining the two changes investigated in Sections I.4.3 and I.4.4: the early stages of the English Jespersen Cycle and loss of fortition in Early New High German. As it turns out, these two changes happen to propagate at very similar paces by accident (Figure I.14). The standard operationalization of CREs is, then, expected to report a CRE between changes to English sentential
Error (sum of squared residuals normalized by number of data points) of the fit of our model (grey) and the standard procedure (black) for the three changes examined in Sections I.4.2–I.4.4: periphrastic do in Early Modern English, Jespersen Cycle in Middle English, and loss of final fortition in Early New High German. Generally speaking, the more constrained model defined in this paper does not fare worse than the less constrained, theoretically unmotivated standard operationalization. We suspect that the exceptionally good fit of the standard operationalization for the Jespersen Cycle is accounted for by sparsity of data (6 data points only), which means that any model that is little constrained will be favoured disproportionately.

The pseudo-CRE thus sheds light on the manner in which our model constrains variation in the time dimension — a constraint that is not built into the model as a premise but that follows from first principles in the form of the Time Separation Theorem. Ultimately, the amount of time separation allowed between any two contextual reflexes of a single underlying change depends on $s$, the slope of the underlying logistic $\tilde{q}_t$. To obtain a more intuitive interpretation of the relationship between contextual time separations and the rate of the underlying change, it is useful to convert the slope parameter into a quantity that measures the time the change needs to go from actuation to completion, using the time-to-completion calculations proposed by Ingason.
Figure I.14
An ‘Anglo-Bavarian pseudo-CRE’ that attempts to combine Jespersen’s Cycle in Middle English with loss of final fortition in Early New High German: data from Tables I.2 and I.3. The different contexts exhibit similar rates of change across the two historical changes by accident: the slope of the underlying change is \(-0.016\) for the English Jespersen Cycle and \(-0.019\) for Early New High German fortition (see captions to Figures I.11 and I.12). This means that the standard ‘same slope, different intercepts’ procedure for detecting CREs in historical data is liable to produce a false positive in this case. Our model, which implies an upper bound on the time separation possible between any two contexts of one underlying change (Theorem I.4), can help to diagnose a ‘change’ such as this as a pseudo-change.
et al. (2013, 96–97). Namely, it can be shown that for slope $s$,

$$T_{\tilde{q}_0} (s) = \frac{2}{|s|} \log \left( \frac{1 - \tilde{q}_0}{\tilde{q}_0} \right) \quad (I.25)$$

gives the time it takes for a change to proceed from initial frequency $\tilde{q}_0$ to final frequency $1 - \tilde{q}_0$, for any (small) $\tilde{q}_0$ with $0 < \tilde{q}_0 \leq 0.5$. Choosing $\tilde{q}_0 = 0.01$, a reasonable choice corresponding to 1% usage of the new variant at the point of actuation, this ‘inverse slope’ then gives a time-to-completion of

$$T_{0.01} (s) = \frac{2}{|s|} \log \left( \frac{0.99}{0.01} \right) = \frac{2}{|s|} \log(99) \approx 9.2 \frac{1}{|s|} \quad (I.26)$$

time units (e.g. years) for any slope $s$. Theorem I.4, on the other hand, implies a maximal time separation of

$$\Delta(s) = \frac{2}{|s|} \log \left( \frac{1}{\sqrt{2} - 1} \right) \approx 1.8 \frac{1}{|s|} \quad (I.27)$$

units between any two contexts of a change proceeding at rate $s$. Since $1.8/9.2 \approx 0.2$, this means that the maximal time separation between any two contexts of a single underlying change is roughly a fifth of the time it takes for the change to go to completion in any context individually.

This fact can be used as a heuristic to evaluate purported CREs. For the Anglo-
Bavarian pseudo-CRE, for instance, the time-to-completion for $\tilde{q}_0 = 0.01$ is

$$T_{0.01}(-0.0175) = \frac{2}{0.0175} \log(99) \approx 525 \text{ years}$$

years when calculated for slope $s = -0.0175$, which is the arithmetic mean of the slopes found by our regressions for the two changes previously (see captions to Figures I.11 and I.12). This implies that the time separation between any two contexts should be no more than $0.2 \cdot 525 = 105$ years. On visual inspection, however, the empirical time separation between the discourse-old and /d/ ‘contexts’ must be at least 300 years (Figure I.14). This, essentially, is why the model is able to diagnose the pseudo-CRE.

**I.5 Discussion**

In this paper, we have augmented Yang’s (2000, 2002b) variational learner with production biases that vary by context — the first time, to our knowledge, that this has been done.\textsuperscript{11} We have used this model to make precise the important intuition of Kroch (1989) that variation between contexts in the increasing use of a new variant may, under certain circumstances, be due to the interaction of a single underlying change with fixed contextual biases. Two important issues remain to be discussed: the diachronic implications of the interaction between language acquisition and production biases, and the nature of the production biases themselves.

**I.5.1 Which grammar wins?**

Yang’s (2000) Fundamental Theorem of Language Change, given earlier as Theorem I.1, can be paraphrased as follows: when grammars compete, the one with the greater parsing advantage will win. In Section I.3 we have shown that this result does not hold in our model. Instead, bias and advantage together determine which grammar will triumph: the precise way in which this works is given in our Extended Fundamental Theorem (Theorem I.3).

In one sense this result is unsurprising: one of the great virtues of Yang’s (2000, 2002b) model of the learner and of diachronic change is its simplicity, and our model introduces additional complexity. It is therefore not a particular surprise that our more complex model does not yield the same intuitive generalization. On the other hand, it is not a necessary consequence of this complication that the Fundamental Theorem fails to hold. As Figure I.8 shows, if we add to our model the stipulation that contextual weights must always be precisely in balance ($B = 0$), then Yang’s

\footnote{B. Clark, Goldrick and Konopka (2008) present a filtered version of Yang’s model to account for typological skews. In this model universal biases play a role, but crucially the biases apply to grammars as a whole rather than to particular contexts of use.}
Fundamental Theorem *does* hold. Such a stipulation would be wholly unmotivated, as far as we are aware, and represents a more complex model than ours.

Moreover, we have not, of course, shown that the Fundamental Theorem of Language Change is false — merely that it is false under the assumptions we make. Whether or not it is false, empirically speaking, depends on how well our model, and Yang’s (2000, 2002b) model, correspond to reality: specifically, whether a model that incorporates the effect of contextual biases as ours does is more realistic than one that does not, and more realistic than one that constrains the net bias. We think that is right, but it is likely that a full consideration of the facts of real-life acquisition and change will require a model that is substantially more complex than any that has been proposed thus far. One feature of Yang’s model, with or without our extension, is that it is completely impossible for a grammar $G_2$ to overtake and defeat another grammar $G_1$ if the weak generative capacity of $G_2$ is a proper subset of that of $G_1$; yet a preference for exactly this kind of subset is often invoked in the context of acquisition in the form of the Subset Principle (Berwick, 1985; Manzini & Wexler, 1987), and, in the domain of phonology at least, retreat to the subset is a frequently-attested diachronic pathway, since unconditioned mergers are well-attested and have precisely the effect of reducing the number of forms generated by the grammar (see e.g. Labov, 1994, 551). Future work will need to address these questions of realism, as well as pursuing further analytical consequences of simpler (and thus more tractable) models like this one.

### I.5.2 The nature of production biases

Up to now we have remained mute with respect to the ontology of production biases, beyond stating that they are biases that affect production. In principle, such biases could assume a number of forms. In a word order change such as OV to VO, for instance, one possibility for interpreting the fixed biases we have proposed is as a reflection of performance pressures in the sense of Hawkins (1994, 2004). Hawkins’s (2004, 38) principle of Minimize Domains states that “[t]he human processor prefers to minimise the connected sequences of linguistic forms and their conventionally associated syntactic and semantic properties in which relations of combination and/or dependency are processed”. This general principle is made concrete using a metric of Early Immediate Constituents (EIC), which serves to favour syntactic structures with a uniform directionality of branching. Importantly, EIC does not penalize right-branching (e.g. VO) or left-branching (e.g. OV) grammars directly, instead disfavouring individual structures with a disharmonic directionality of branching, for instance when a head-final VP is embedded under a head-initial TP. This is equivalent to a context-specific production bias in our sense. Hawkins conceptualizes Minimize Domains and EIC as principles of parsing rather than of production, but notes that there is evidence
that EIC might be involved in production too (Hawkins, 2004, 106), and states that “if EIC can be systematically generalized from a model of comprehension to a model of production [...] then so much the better” (Hawkins, 1994, 427). Hawkins’s principles have also been reformulated as principles of derivational/computational complexity by Mobbs (2008) and Walkden (2009).

In phonological change, meanwhile, the biases can be interpreted as well established articulatory phonetic effects. Final fortition, for example, is known to be more likely to apply to velar consonants than to coronal consonants and more likely to apply to coronal consonants than labial consonants (Ohala, 1983), and this order of preference seems to be observed diachronically as final fortition emerged in the history of Frisian (Tiersma, 1985). These two phonological and syntactic examples are intended to give a flavour of how contextual production biases can be interpreted, not to exhaust the range of possibilities. For other changes, other biases might be necessary: for instance, in Wallage’s (2013) data, discourse-old propositions favour Stage 2 of the Jespersen Cycle during the change, and the biases here could plausibly reflect Gricean maxims of cooperative communication.

The above-mentioned biases — constraints on syntactic processing, articulatory pressures, pragmatic principles — are plausibly innate in the sense that they are shared by all speakers across all languages and are not subject to change. This is why in our model definition we maintained that the biases $b_i$ be diachronically constant. Note that the logic here is not just that constant biases imply Constant Rate Effects — we are actually defending the stronger claim that Constant Rate Effects occur if, and only if, diachronically constant biases impinge on an underlying change. Time-dependent biases, or random biases, would result in change processes in which the trajectories of different linguistic contexts are not parallel to each other — something we might refer to as an ‘Inconstant Rate Effect’. Having said that, it is not inconceivable that some non-innate biases are constant on sufficiently long timescales so that they may give rise to Constant Rate Effects: this will be the case when the underlying change itself is fast enough to be carried to completion within the timeframe in which the biases stay fixed. This could be true of certain sociolinguistic biases, and here the biasing mechanism of our model is in agreement with sociolinguistic work (e.g. Labov, 2001, ch. 9) which has found some types of sociolinguistic bias modulation to be strongest midway through the change, just as in our model (cf. Figure I.5).

12We might expect to find the reverse effect in the data from Fruehwald et al. (2009) discussed above, but, curiously, we do not: /g/ is the most favouring context for the loss of devoicing.

13We are grateful to Charles Yang for bringing this point to our attention.
I.6 Conclusion

Building on earlier work that derives logistic evolution as a population-level property of language change (Niyogi & Berwick, 1997; Yang, 2000, 2002b), we have provided a mechanism for the Constant Rate Effect proposed by Kroch (1989). We have done this by enriching Yang’s (2000, 2002b) model of acquisition and change with a contextual bias mechanism that links different context curves to a single underlying change. The work also provides a method of testing for CREs that is demonstrably superior to the traditional method of ‘same slope, different intercepts’, since it is a consequence of the model that there is a fixed upper bound on the time separation of contextual curves. We have shown that this enables us to distinguish certain types of pseudo-CREs from instances in which a single underlying grammatical change is actually plausible. We have also shown that advantage, in the Yangian sense, is not the only factor at play in determining the ultimate outcome of a situation of grammatical competition, if the basic assumptions of our model hold true.

The upshot of all this is that it is now possible to test whether divergent usage frequencies in corpora across different contexts during the course of a change in fact mask a deeper underlying grammatical homogeneity, and to do so in a more restricted and principled way than has been possible to date. Crucially, the method we propose is not only methodologically superior to the standard operationalization of CRE testing: our model in fact derives the possibility of CREs, and sets tight bounds on the kind of empirically observed situation that can be said to constitute a CRE.

Acknowledgements

We thank Ricardo Bermúdez-Otero, Laurel MacKenzie, David Willis, Charles Yang and the reviewers of this paper for comments that led to important improvements, as well as David Denison, Aaron Ecay, Tobias Galla, Anthony Kroch, Gertjan Postma, Alexandra Simonenko and Richard Zimmermann for valuable discussions on the subject. Early versions of the paper were delivered at the 22nd International Conference on Historical Linguistics (Naples, 27–31 July 2015), the workshop on Dealing with Bad Data in Linguistic Theory (Amsterdam, 17–18 March 2016), the 39th Generative Linguistics in the Old World conference (Göttingen, 5–8 April 2016), New Ways of Analyzing Syntactic Variation 2 (Ghent, 19–20 May 2016), and the 18th Diachronic Generative Syntax conference (Ghent, 29 June–1 July 2016); we thank our audiences, conference organizing committees and reviewers. HK gratefully acknowledges financial support from Emil Aaltonen Foundation and The Ella and Georg Ehrnrooth Foundation.
I.7 Appendix: Derivations

I.7.1 Evolution of $q_t$

**Theorem I.5.**

Let $\Lambda_0 = 1$ and

$$q_{t+1} = \left(1 + \Lambda_t \rho \frac{1 - q_t}{q_t}\right)^{-1}. \quad \text{(I.29)}$$

Then, for all $t$,

$$q_t = \frac{q_0}{q_0 + \mathcal{L}_t \rho (1 - q_0)}, \quad \text{(I.30)}$$

where $\mathcal{L}_t = \prod_{\tau=0}^{t-1} \Lambda_\tau$ for $t \geq 1$ and $\mathcal{L}_0 = 1$.

**Proof.** Induction. For $t = 0$,

$$q_0 = \frac{q_0}{q_0 + \mathcal{L}_0 \rho (1 - q_0)} = \frac{q_0}{q_0 + 1 - q_0} = q_0. \quad \text{(I.31)}$$

Now assume that the claim holds for $t$. Then

$$q_{t+1} = \left(1 + \Lambda_t \rho \frac{1 - q_t}{q_t}\right)^{-1}$$

$$= \frac{q_t}{q_t + \Lambda_t \rho (1 - q_t)}$$

$$= \frac{\frac{q_0}{q_0 + \mathcal{L}_t \rho (1 - q_0)}}{\frac{q_0}{q_0 + \mathcal{L}_t \rho (1 - q_0)} + \Lambda_t \rho \left(1 - \frac{q_0}{q_0 + \mathcal{L}_t \rho (1 - q_0)}\right)} \quad \text{(I.32)}$$

$$= \frac{q_0 + \Lambda_t \rho (q_0 + \mathcal{L}_t \rho (1 - q_0) - q_0)}{q_0}$$

$$= \frac{q_0 + \mathcal{L}_t \rho \rho (1 - q_0)}{q_0}$$

$$= \frac{q_0 + \mathcal{L}_{t+1} \rho (1 - q_0)}{q_0}$$

as desired.

**Corollary I.3.**

Let $B = 0$. Then the evolution of $q_t$ is logistic.

**Proof.** Let $B = 0$. Then from equation (I.16) $\Lambda_t = 1$ for all $t$, so that

$$q_{t+1} = \left(1 + \rho \frac{1 - q_t}{q_t}\right)^{-1}. \quad \text{(I.33)}$$

Theorem I.5 now implies

$$q_t = \frac{q_0}{q_0 + \rho (1 - q_0)}. \quad \text{(I.34)}$$
Now assume \( q_t \neq 0 \) and divide both the numerator and the denominator by \( q_0 \):

\[
q_t = \left( 1 + \rho^t \frac{1 - q_0}{q_0} \right)^{-1} \\
= \left( 1 + \exp \left( \log \left( \frac{1 - q_0}{q_0} \right) \right) \right)^{-1} \\
= \left( 1 + \exp \left( t \log(\rho) + \log \left( \frac{1 - q_0}{q_0} \right) \right) \right)^{-1} \\
= \left( 1 + \exp \left( \log(\rho) \left( t + \frac{1}{\log(\rho)} \log \left( \frac{1 - q_0}{q_0} \right) \right) \right) \right)^{-1}.
\]

(I.35)

Hence, \( q_t \) is logistic with \( s = -\log(\rho) \) and \( k = -\frac{1}{\log(\rho)} \log \left( \frac{1-q_0}{q_0} \right) \).

**Corollary I.4.**

In Yang’s (2000) model, the weight \( q_t \) evolves as

\[
q_t = \left( 1 + \rho^t \frac{1 - q_0}{q_0} \right)^{-1},
\]

(I.36)

which is a logistic function of \( t \).

**Proof.** In this model, \( q_t \) obeys (I.29) with \( \Lambda_t = 1 \) for all \( t \). From Theorem I.5,

\[
q_t = \frac{q_0}{q_0 + \rho^t (1 - q_0)}.
\]

(I.37)

Assuming, without loss of generality, that \( q_0 \neq 0 \), we derive

\[
q_t = \left( 1 + \rho^t \frac{1 - q_0}{q_0} \right)^{-1}.
\]

(I.38)

This is a logistic function by Corollary I.3.

### I.7.2 The bias-modulating functions \( F \) and \( G \)

**Theorem I.6.**

Let

\[
\begin{align*}
p_t^{(i)} &= p_t + F(b_i, p_t) \\
q_t^{(i)} &= q_t + G(b_i, q_t)
\end{align*}
\]

(I.39)

Then

\[
\begin{align*}
p_t^{(i)} + q_t^{(i)} &= 1 \\
0 &\leq p_t^{(i)}, q_t^{(i)} \leq 1
\end{align*}
\]

(I.40)
if and only if
\[
\begin{cases}
F = -G \\
|F|, |G| \leq \min\{p_t, q_t\}
\end{cases}
\] (I.41)

Proof. Writing \( F = F(b_i, p_t), \ G = G(b_i, q_t), \ p = p_t \) and \( q = q_t \), the first requirement in (I.40) implies that
\[ F = -G, \] (I.42)

since
\[ p + F + q + G = 1 \quad \text{only if} \quad p + F + 1 - p + G = 1 \quad \text{only if} \quad F + G = 0. \] (I.43)

The second requirement in (I.40), on the other hand, implies \( F \leq q \) and \( -F \leq p \), since
\[ p + F \leq 1 \quad \text{only if} \quad F \leq 1 - p = q \] (I.44)

and
\[ 0 \leq p + F \quad \text{only if} \quad -F \leq p; \] (I.45)

hence
\[ |F| \leq \min\{p, q\}, \] (I.46)

and an exactly symmetric argument shows that
\[ |G| \leq \min\{p, q\}. \] (I.47)

Thus
\[
\begin{cases}
F = -G \\
|F|, |G| \leq \min\{p, q\}
\end{cases}
\] (I.48)

On the other hand, if (I.48) holds, then
\[ p^{(i)}_t + q^{(i)}_t = p + F + q + G = p + F + q - F = p + q = 1, \] (I.49)

and
\[ |F| \leq \min\{p, q\} \quad \text{only if} \quad -F \leq p \quad \text{only if} \quad 0 \leq p + F = p^{(i)}_t. \] (I.50)

Similarly,
\[ |F| \leq \min\{p, q\} \quad \text{only if} \quad F \leq q \quad \text{only if} \quad F \leq 1 - p \quad \text{only if} \quad p^{(i)}_t \leq 1, \] (I.51)

and an analogous argument shows that \( 0 \leq q^{(i)}_t \leq 1. \)

I.7.3 Proof of the Extended Fundamental Theorem

Let
\[ B_c(\rho, q_t) = \frac{\rho - 1}{1 + q_t(\rho - 1)}. \] (I.52)
To prove Theorem I.3, we make use of the following auxiliary result:

**Theorem I.7.**

For all $t$:

1. $q_{t+1} > q_t$ if $B > B_c(\rho, q_0)$;
2. $q_{t+1} = q_t$ if $B = B_c(\rho, q_0)$;
3. $q_{t+1} < q_t$ if $B < B_c(\rho, q_0)$.

**Proof.** From (I.14) and (I.16),

$$q_{t+1} = \frac{1}{1 + \Lambda_t \rho \frac{1 - q_t}{q_t}}$$  \hspace{1cm} (I.53)

with

$$\Lambda_t = \frac{1 - q_t B}{1 + (1 - q_t) B}.$$  \hspace{1cm} (I.54)

To prove the first claim, we examine the difference $q_{t+1} - q_t$. Now $q_{t+1} - q_t > 0$ if and only if

$$\frac{1 - q_t - \Lambda_t \rho (1 - q_t)}{1 + \Lambda_t \rho \frac{1 - q_t}{q_t}} > 0 \quad \text{iff} \quad 1 - q_t - \Lambda_t \rho (1 - q_t) > 0 \quad \text{iff} \quad \Lambda_t \rho (1 - q_t) < 1 - q_t \quad \text{iff} \quad \Lambda_t < \frac{1}{\rho} \quad \text{iff} \quad \frac{1 - q_t B}{1 + (1 - q_t) B} < \frac{1}{\rho}.$$  \hspace{1cm} (I.55)

Now, $-1 \leq B \leq 1$, so always $1 - q_t B > 0$ and $1 + (1 - q_t) B \geq q_t > 0$. Hence

$$\frac{1 - q_t B}{1 + (1 - q_t) B} < \frac{1}{\rho} \quad \text{iff} \quad \rho (1 - q_t B) < 1 + (1 - q_t) B \quad \text{iff} \quad \rho - \rho q_t B < 1 + (1 - q_t) B \quad \text{iff} \quad B(1 - q_t + \rho q_t) > \rho - 1 \quad \text{iff} \quad B(1 + q_t(\rho - 1)) > \rho - 1 \quad \text{iff} \quad B > \frac{\rho - 1}{1 + q_t(\rho - 1)}.$$  \hspace{1cm} (I.56)

as desired. Claims 2 and 3 are handled similarly.

**Corollary I.5.**

For all $t$:

1. $q_{t+1} > q_t$ if $B > B_c(\rho, q_0)$;
2. $q_{t+1} = q_t$ if $B = B_c(\rho, q_0)$;

3. $q_{t+1} < q_t$ if $B < B_c(\rho, q_0)$.

Proof. First, we note that $B_c(\rho, q_1)$ is a decreasing function of $q_1$: if $q < Q$, then $B_c(\rho, q) \geq B_c(\rho, Q)$.

Now let $B > B_c(\rho, q_0)$. Then by Theorem I.7 $q_1 > q_0$. Since $B_c$ is decreasing, $B_c(\rho, q_1) \leq B_c(\rho, q_0) < B$. Hence $q_2 > q_1$ by Theorem I.7. By full induction, $q_{t+1} > q_t$ for all $t$.

Let $B = B_c(\rho, q_0)$. Then by Theorem I.7 $q_1 = q_0$. Then $B_c(\rho, q_1) = B_c(\rho, q_0) = B$, so that $q_2 = q_1$ by Theorem I.7. By full induction, $q_{t+1} = q_t$ for all $t$.

Finally, let $B < B_c(\rho, q_0)$. Then by Theorem I.7 $q_1 < q_0$. Since $B_c$ is decreasing, $B_c(\rho, q_1) \geq B_c(\rho, q_0) > B$, so that $q_2 < q_1$ by Theorem I.7. By full induction, $q_{t+1} < q_t$ for all $t$.

Theorem I.3 now follows from Corollary I.5: since $q_t$ is bounded between 0 and 1 and $q_0$ may be chosen arbitrarily close to 1 or 0, $q_t \to 1$ if $B > B_c(\rho, q_0)$ and $q_t \to 0$ if $B < B_c(\rho, q_0)$ in the limit $t \to \infty$.

1.7.4 Proof of the Time Separation Theorem

We shall now prove Theorem I.4, reproduced here as Theorem I.8:

**Theorem I.8.**

For any two contextual reflexes of an underlying change from $G_1$ to $G_2$ approximated by a logistic $\tilde{q}_t$ with slope $s$, the maximal time separation at tipping points is

$$\Delta(s) = \frac{2}{|s|} \log \left( \frac{1}{\sqrt{2} - 1} \right).$$

(I.57)

Proof. Assume two contexts 1 and 2. The maximal separation will of course be attained with maximal biases $b_1 = 1$ and $b_2 = -1$ (or vice versa). Assume this. Then $q_t^{(1)} = \tilde{q}_t - \tilde{q}_t(1 - \tilde{q}_t) = \tilde{q}_t^2$, and so $q_t^{(2)} = 1/2$ when $\tilde{q}_t = 1/\sqrt{2}$. On the other hand, $\tilde{q}_t$ is given by the logistic

$$\tilde{q}_t = \frac{1}{1 + e^{-st}},$$

(I.58)

where we assume $k = 0$ because translation along the time axis obviously makes no difference to the argument here. A little algebra now shows that $\tilde{q}_t = 1/\sqrt{2}$ if and only if $t = -\frac{1}{s} \log(\sqrt{2} - 1)$. Thus, $q_t^{(2)}$ attains its tipping point at time $t^*_2 = -\frac{1}{s} \log(\sqrt{2} - 1)$. Since the logistic $\tilde{q}_t$ itself attains its tipping point at $t = 0$ and since $b_1 = -b_2$, symmetry implies that $q_t^{(1)}$ attains its tipping point at $t^*_1 = 0 - t^*_2 = \frac{1}{s} \log(\sqrt{2} - 1)$. Hence

$$\Delta(s) = |t^*_1 - t^*_2| = |2t^*_1| = \frac{2}{|s|} |\log(\sqrt{2} - 1)| = -\frac{2}{|s|} \log(\sqrt{2} - 1) = \frac{2}{|s|} \log \left( \frac{1}{\sqrt{2} - 1} \right)$$

(I.59)

as wished.
I.7.5 Curve-fitting algorithm for the extended model

This appendix provides the curve-fitting algorithm used for model evaluation in Section I.4, in pseudocode. We assume the existence of three subroutines: REGRESS, PARAM and ERROR. The first of these can be any optimization algorithm that performs nonlinear regression on the $i$th context for given $s$ and $k$, fitting a curve of the form (I.18) subject to the lower and upper bounds $-1 \leq b_i \leq 1$. The second routine is assumed to return the bias size $b_i$ found by this regression, and the third to give the error (in terms of the normalized sum of squared residuals) of the fit.

1: procedure Fit-CRE
2:    $K \leftarrow$ number of contexts
3:    $S \leftarrow$ range of $s$ values
4:    $\mathcal{K} \leftarrow$ range of $k$ values
5:    $s^* \leftarrow$ current best-fitting $s$
6:    $k^* \leftarrow$ current best-fitting $k$
7:    $b^* \leftarrow$ vector of length $K$ to hold best-fitting bias sizes
8:    $E^* \leftarrow$ error of fit, initialized to a large value
9:    for $s \in S$ do
10:       for $k \in \mathcal{K}$ do
11:          $E \leftarrow 0$ (current error of fit)
12:          $b = (b_1, \ldots, b_K) \leftarrow$ vector of length $K$ to hold bias sizes
13:             for $i \in \{1, \ldots, K\}$ do
14:                $F \leftarrow$ REGRESS($s, k, i$)
15:                $b_i \leftarrow$ PARAM($F$)
16:                $E \leftarrow E + ERROR(F)$
17:             end for
18:          if $E < E^*$ then
19:             $s^* \leftarrow s$
20:            $k^* \leftarrow k$
21:            $b^* \leftarrow b$
22:            $E^* \leftarrow E$
23:         end if
24:     end for
25: end for
26: return $s^*, k^*, b^*, E^*$
27: end procedure
II
Stable variation in multidimensional competition

Henri Kauhanen

Abstract

The Fundamental Theorem of Language Change (Yang, 2000) implies the impossibility of stable variation in the Variational Learning framework, but only in the special case where two, and not more, grammatical variants compete. Introducing the notion of an advantage matrix, I generalize Variational Learning to situations where the learner receives input generated by more than two grammars, and show that diachronically stable variation is an intrinsic feature of several types of such multiple-grammar systems. This invites experimentalists to take the possibility of stable variation seriously and identifies one possible place where to look for it: situations of complex language contact.

II.1 Variation, learning and diachronic stability

Since its introduction in a series of publications by Yang in the early noughties (Yang, 1999, 2000, 2002a, 2002b, 2004), the Variational Learner has stirred much interest among those working in the field of language variation and change: given its inher-

---

*This paper is to appear in A. Breitbarth, M. Bouzouita, M. Farasyn & L. Danckaert (eds.), The determinants of diachronic stability (Amsterdam: John Benjamins). This version is a post-print reflecting changes made in the peer reviewing process, but not the final publisher’s version. When citing, please use the page numbers given in the publisher’s version. © John Benjamins Publishing Company. The publisher should be contacted for permission to reuse or reprint the material in any form.
ently probabilistic nature, the Variational Learning paradigm successfully formalizes many aspects of the competing grammars framework (Kroch, 1994), in which the simultaneous existence of a number of grammatical options in the mind of a speaker is taken for granted. As far as change is concerned, however, this intra-speaker existence of multiple grammars has been considered diachronically unstable, in the sense that over iterated generational learning interactions, grammar competition leads, ultimately, to a stable state of dominance by some single grammar. This mathematical fact, formulated as the Fundamental Theorem of Language Change by Yang (2000), dovetails with the theoretico-empirical claim that all morphosyntactic variation between two forms competing for a single function results, over time, in either the extinction of one form, or a functional specialization of the two forms by which the competition is escaped (Kroch, 1994; Wallenberg, 2016) — in either case, diachronically stable variation between two values of a single variable is thought to be impossible because of a general cognitively motivated blocking effect that militates against stable doublets (Aronoff, 1976).

In this paper, I wish to draw attention to the fact that Variational Learning only predicts this outcome in the case where two, and not more than two, variants compete in a speaker population. An analysis of both the classical Variational Learner and its parametrically constrained variation, the Naive Parameter Learner, reveals that in the general case — when more than two grammars compete — the situation is strikingly different. The Fundamental Theorem gives way to more complicated, even non-monotonic trajectories of change; to bifurcations; and, in many cases, to truly stable variation in which the competing variants do not (or need not) specialize functionally. Since language learners need to set the values of multiple parameters and hence make a choice in a high-dimensional space of possible grammars, these results question whether the Variational Learner can, in fact, explain the (purported) non-occurrence of stable variation. On the other hand, the results invite experimentalists to consider the possibility that when more than two variants come to compete, stable variation may in fact be predicted by general human learning mechanisms (assuming, ex hypothesi, that the reinforcement learning algorithm at the heart of the Variational Learner carries psychological realism).

To begin, it is incumbent on us to make the relevant notions of variation and stability as precise as possible. Any system capable of change is a dynamical system whose behaviour may be modelled using a set of difference equations — if the time variable is taken as discrete — or a set of differential equations — if time is considered continuous. The choice of one or the other description is largely arbitrary; in this paper, I will stick to discrete time, but all the results are valid for a continuous-time description as well (by letting the inter-generational time step tend to zero and examining the resulting differential equations). I then define a language system to
be a probability distribution \( p = (p_1, \ldots, p_n) \) over a finite set of possible grammars \( G_1, \ldots, G_n \), together with a set of difference equations

\[
p'_i = f_i(p) \quad (i = 1, \ldots, n)
\]

which define the system’s dynamics. Here, \( p'_i \) is the successor of \( p_i \); in other words, \( p'_i \) is the value of the \( i \)th variable at time \( t+1 \) given that the state of the entire system at time \( t \) was \( p = (p_1, \ldots, p_n) \). The functions \( f_i \) are, in the general case, real-valued functions; they assume some concrete form as soon as concrete assumptions are made about learning, linguistic interaction, the existence of a critical period, and so on. The probabilities \( p_i \) themselves, \( 0 \leq p_i \leq 1 \), describe the probability of use of the different competing grammars, in the usual sense: in a sequence of \( k \) utterances, roughly \( p_i k \) utterances will be produced by grammar \( G_i \) if \( k \) is large. These probabilities may be taken to describe either a single individual or an entire community of speakers: clearly, both individual and community-level probabilities may change over time, but the corresponding functions \( f_i \) in (II.1) may be rather different in the two cases. In what follows, I will always take \( p_i \) to refer to community-level probabilities and will denote probabilities at the level of individuals with corresponding Greek letters, \( \pi_i \).

Taking the \( p_i \) as community-level probabilities, then, let us proceed to define the notions of variation and stability on the level of speech communities. Intuitively, variation exists if at least two grammars are used with non-zero probability. It then makes sense to define a state of variation as a probability state \( p = (p_1, \ldots, p_n) \) which satisfies \( p_i < 1 \) for all \( i \), for it is precisely under this condition that no single grammar gets to claim all of the available probability mass. Defining the concomitant notion of diachronic stability is a bit trickier, and I shall begin by presenting a physical analogue.

Consider a non-ideal pendulum (Figure II.1A). By non-ideal, I mean to imply that we are not excluding frictional forces by way of idealization. Such a pendulum is also known as a damped pendulum, and the defining characteristic of its dynamics is the existence of a rest point directly below the point of attachment: if the pendulum is ever found in this position, it will not move, barring application of an external force. Moreover, if the pendulum is set in motion from some other initial state, it will ultimately come to a halt at this rest point after a period of diminishing oscillation. Such a rest point is said to be asymptotically stable. More precisely, a rest point \( x \) in the state space of a dynamical system is asymptotically stable if a neighbourhood of states around \( x \) exists such that all trajectories from this neighbourhood converge to

\[x' = x \quad \text{or equivalently} \quad x' - x = 0,\]

that is, as a zero-change state. In the vast literature on dynamical systems, rest points are also known as rest states, fixed points, equilibria, and steady states. The last term, sometimes encountered in discussions of language change, is somewhat unfortunate because of the semantic similarity of the pre-theoretical terms ‘steady’ and ‘stable’ — as we will see presently, not all steady states are stable, in the technical sense.

---

1 In the corresponding mathematical description, a rest point is identified as a state \( x \) which satisfies \( x' = x \) or equivalently \( x' - x = 0 \), that is, as a zero-change state. In the vast literature on dynamical systems, rest points are also known as rest states, fixed points, equilibria, and steady states. The last term, sometimes encountered in discussions of language change, is somewhat unfortunate because of the semantic similarity of the pre-theoretical terms ‘steady’ and ‘stable’ — as we will see presently, not all steady states are stable, in the technical sense.
Now consider the inverted pendulum of Figure II.1B. This pendulum, too, has a rest point, now directly above the point of attachment. Theoretically, if it were possible to balance the pendulum with infinite precision at this rest point, it would not move, since the horizontal component of the sum of the forces acting on the pendulum is zero at this point (we assume the pendulum is fixed to a stiff rod). Even a slight disturbance to the inverted pendulum will, however, nudge it away from the rest point. Such a rest point is unstable, since all trajectories from any local neighbourhood around the rest point take the system state away from the rest point.

Finally, consider the “goo pendulum” of Figure II.1C. Here the pendulum is submerged in a hypothetical goo of infinite viscosity which supports the pendulum but allows its movement when a suitable external force is applied (for a physically realistic approximation, we may think of a low-mass pendulum, such as a needle, submerged in a high-viscosity fluid such as honey). This pendulum will not move from any initial condition. Every possible position of the pendulum is a rest point, and they are all neither asymptotically stable nor unstable. The characteristic behaviour of these non-asymptotically stable states is that, given a perturbation, the system will move to a different, close-by point, but is not “actively” repelled by the rest point nor attracted back to it.

These notions translate directly into our framework of language systems and may now be used to explicate the idea of stable variation. I define a state of stable variation to be a probability vector \( p = (p_1, \ldots, p_n) \) satisfying the following three conditions simultaneously:

1. \( p \) is a state of variation \( (p_i < 1 \text{ for all } i) \)
2. \( p \) is a rest point (\( p'_i - p_i = 0 \) for all \( i \))

3. \( p \) is asymptotically stable

I do not include non-asymptotically stable rest points in this definition since, as per the above discussion, they are not resilient to perturbations. Crucially, given that real-life systems always contain a source of noise, which we may think of as a perturbation to the state of a deterministic system such as (II.1), such states do not count as truly stable.

II.2 Two grammars

With these notions in hand we may proceed to a formal study of variation and stability in the Variational Learning framework, beginning with a summary restatement of the already familiar two-grammar case.

In Yang (2000), language change is reduced to language acquisition by assuming that language learners employ a specific learning strategy, the linear reward–penalty (henceforth, LRP) learning algorithm originating in Bush and Mosteller’s (1955) early work on reinforcement learning and most usefully synthesized by Narendra and Thathachar (1989). This allows one to close the population-dynamical equations (II.1). Specifically, assume the learner needs to make a decision between two grammars \( G_1 \) and \( G_2 \) which are used in the community with probabilities \( p_1 \) and \( p_2 \). Writing \( \pi_1 \) and \( \pi_2 \) for the learner’s hypothesis (i.e. \( \pi_i \) is the probability with which the learner himself employs \( G_i \)), the LRP algorithm assumes the following form:

**Algorithm II.1 (LRP, \( n = 2 \); Narendra & Thathachar, 1989, 110–111).**

1. Let \( \pi_1 = \pi_2 = 1/2 \) initially.

2. Present an input token (sentence) \( x \) to the learner. This is generated by \( G_1 \) with probability \( p_1 \) and by \( G_2 \) with probability \( p_2 \).

3. Learner picks grammar \( G_i \) with probability \( \pi_i \).

4. Suppose the learner picked \( G_1 \).

   a. If \( G_1 \) parses \( x \), the learner increases \( \pi_1 \) by a small amount and decreases \( \pi_2 \) by a small amount. Concretely, \( \pi_1 \) is replaced with \( \pi_1 + \gamma(1 - \pi_1) \), where \( \gamma \) is a small positive number (the learning rate), whilst \( \pi_2 \) is replaced with \( (1 - \gamma)\pi_2 \).

   b. Conversely, if \( G_1 \) does not parse \( x \), the learner decreases \( \pi_1 \) and increases \( \pi_2 \). Concretely, \( \pi_1 \) is replaced with \( (1 - \gamma)\pi_1 \), whilst \( \pi_2 \) is replaced with \( \pi_2 + \gamma(1 - \pi_2) \).

5. (If the learner picked \( G_2 \) instead, execute the previous step with labels 1 and 2 interchanged.)
6. Steps 2–5 are repeated for \( T \) input tokens.

Thus, during learning, the probabilities \( \pi_i \) change in response to the two grammars’ success in parsing input generated from the community-level distribution \( p = (p_1, p_2) \). For simplicity, the latter is assumed to stay constant for the duration of learning; in learning-theoretic terminology, the learner’s environment is a stationary random environment (Narendra & Thathachar, 1989).

If either \( p_1 = 1 \) or \( p_2 = 1 \), then one of the grammars succeeds in parsing any possible input token the learner may encounter. It then follows that in such a case of a homogeneous community, the learner’s hypothesis tends to the population state with growing \( T \) and the unique target grammar is learnable according to a probabilistic variant of Gold’s (1967) learnability criterion (cf. Niyogi, 2002, 354). If the population state \( p = (p_1, p_2) \) is mixed, i.e. a state of variation, the learner exhibits more interesting behaviour.

Let \( \hat{\pi}_i \) denote the value of \( \pi_i \) at the end of learning (at \( T \) learning steps), and assume that \( T \) is large and that the learning rate \( \gamma \) is small. Such a learner shall be called reliable,\(^2\) and it can be shown (Narendra & Thathachar, 1989, 111–112) that, for a reliable learner,

\[
\hat{\pi}_1 \approx \frac{c_2}{c_1 + c_2} \quad \text{and} \quad \hat{\pi}_2 \approx \frac{c_1}{c_1 + c_2},
\]

where \( c_i \) is the penalty probability of grammar \( G_i \):

\[
c_i = \text{Prob}(x : G_i \text{ does not parse } x).
\]

The penalty probabilities are easily determined: we may write \( c_1 = a_2 p_2 \) and \( c_2 = a_1 p_1 \), where \( a_2 \) is the probability of a sentence parsed by \( G_2 \) but not by \( G_1 \), and vice versa for \( a_1 \). Following Yang (2000), I will call \( a_1 \) the advantage of \( G_1 \) and \( a_2 \) the advantage of \( G_2 \) (Figure II.2).

If learners are now arranged in a sequence of non-overlapping generations, the output of generation \( t \) feeding as input to the learning process of generation \( t + 1 \), we have the population-level difference equations

\[
p_1' = \frac{a_1 p_1}{a_1 p_1 + a_2 p_2} \quad \text{and} \quad p_2' = \frac{a_2 p_2}{a_1 p_1 + a_2 p_2}.
\]

(Bearing in mind that \( p_1 + p_2 = 1 \), it suffices to work with the single equation

\[
p_1' = \frac{a_1 p_1}{a_1 p_1 + a_2 (1 - p_1)}.
\]

\(^2\)All results in this paper pertain to systems of reliable learners. The stochastic effects of unreliable learning – short critical periods or large (“high-temperature”) learning rates – are underinvestigated in the literature but must be set aside here.
The classical two-grammar setting (after Yang, 2000, 238, Figure 2). This Venn diagram illustrates all sentences parsed by either grammar; \(a_1\) is the probability of a sentence uniquely parsed by \(G_1\) and \(a_2\) the probability of a sentence uniquely parsed by its competitor \(G_2\).

The inter-generational increment in \(p_1\) is given by \(p'_1 - p_1\), which by simple algebra is found to equal

\[
p'_1 - p_1 = \frac{(a_1 - a_2)(1 - p_1)p_1}{a_1p_1 + a_2(1 - p_1)}.
\]

(II.6)

Figuring out the rest points of this system is now an easy task: from (II.6) it is readily seen that \(p'_1 - p_1 = 0\) if and only if (1) \(p_1 = 0\), (2) \(p_1 = 1\) or (3) \(a_1 = a_2\). Assume first that \(a_1 > a_2\). Then the sign of \(p'_1 - p_1\) is always strictly positive, which means that \(p_1\) always grows, no matter what the state \(p = (p_1, p_2)\). Hence, the state \((1, 0)\) is asymptotically stable and the state \((0, 1)\) unstable. With this ordering of the two advantage parameters, \(G_1\) will drive \(G_2\) out in diachrony, no matter what the initial state of the system. For \(a_1 < a_2\), the reverse state of affairs obtains: \((1, 0)\) is unstable and \((0, 1)\) stable. Now \(G_2\) is the winner. Finally, if \(a_1 = a_2\), then the rate of change of \(p_1\) (and, by necessity, of \(p_2\)) is zero in every possible state \(p = (p_1, p_2)\). The state space is filled with an infinity of non-asymptotically stable rest points, and the system resembles the goo pendulum of Figure II.1C. With this reasoning, we have proved the following two results:

Theorem II.1 (Fundamental Theorem of Language Change; Yang, 2000, 239).

Suppose learners are reliable. Then, in a two-grammar system, \(G_1\) wins in diachrony if \(a_1 > a_2\), and \(G_2\) wins if \(a_1 < a_2\).

Theorem II.2.

No two-grammar system of reliable learners admits stable variation.
Time evolution of a two-grammar system with $a_1 = 0.2$ and $a_2 = 0.1$, from initial state $(p_1, p_2) = (0.01, 0.99)$, for both theoretically perfectly reliable learners (circles, equation II.5) and for large-sample learners (crosses, from computer simulation; only one realization of the stochastic process shown).

Figure II.3 illustrates a typical trajectory in a system of two grammars with unequal advantages. The grammar with the greater advantage ousts its competitor both in the case of theoretically perfectly reliable learners (equation II.5) and in the case of learners who receive a finite but large sample of primary linguistic data.

II.3 Advantage matrices and the cyclical balance criterion

It is not immediately obvious how, or whether, these results generalize to situations where learners are exposed to input from more than two grammars. In fact, extending the model definition itself to such more general cases turns out to be nontrivial. The main difficulty lies in expressing the penalty probabilities $c_i$, which with an increasing number of competing variants assume an increasingly complicated form. This is because in the general case of $n$ competing grammars one has to consider the relative (pairwise) advantages between any two distinct grammars, the number of these advantage relations being $n(n-1) = n^2 - n$ and hence growing superlinearly with $n$.

In the three-grammar case ($n = 3$), the situation is as depicted in Figure II.4. Each grammar potentially generates sentences which are only parsed by that grammar itself. However, the possibility now arises that two of the three grammars jointly
generate something not parsed by the third grammar. Using the symbolism of Figure II.4, we find that the penalty probability for grammar $G_1$ in this more general three-grammar situation may be expressed as

$$c_1 = \alpha_{\{2\}} p_2 + \alpha_{\{3\}} p_3 + \alpha_{\{2,3\}} (p_2 + p_3) = (\alpha_{\{2\}} + \alpha_{\{2,3\}}) p_2 + (\alpha_{\{3\}} + \alpha_{\{2,3\}}) p_3. \tag{II.7}$$

If we now write $a_{12} = \alpha_{\{2\}} + \alpha_{\{2,3\}}$ and $a_{13} = \alpha_{\{3\}} + \alpha_{\{2,3\}}$, we see that $a_{12}$ gives the relative advantage of $G_2$ over $G_1$ and $a_{13}$ the relative advantage of $G_3$ over $G_1$. Proceeding analogously to derive the penalty probabilities $c_2$ and $c_3$, one finds

$$\begin{cases} 
    c_1 = a_{12} p_2 + a_{13} p_3 \\
    c_2 = a_{21} p_1 + a_{23} p_3 \\
    c_3 = a_{31} p_1 + a_{32} p_2 
\end{cases} \tag{II.8}$$

where each $a_{ij}$ thus gives the probability of a sentence which is parsed by $G_j$ but not by $G_i$. It is these relative advantages $a_{ij}$ that determine the system’s dynamics, and consequently it will be useful to collect them in a matrix,

$$A = [a_{ij}] = \begin{bmatrix} 
    0 & a_{12} & a_{13} \\
    a_{21} & 0 & a_{23} \\
    a_{31} & a_{32} & 0 
\end{bmatrix} \tag{II.9}$$

where the diagonal is zero since obviously $a_{ii} = 0$ for any $i$. In what follows, I will
refer to such a matrix as an *advantage matrix*. It is possible, with greater technical
difficulty, to generalize this procedure for arbitrary $n$, and many of the results to
follow carry over to the general case. Here, I restrict my attention to three grammars
in the interest of readability.

Not every square matrix of real numbers is a valid advantage matrix. As already
mentioned, the diagonal is necessarily zero, since no grammar both parses and does
not parse one and the same sentence. Furthermore, from Figure II.4, we note that
the $\alpha$ quantities must all sum to unity, since the event represented by their union
is “a sentence is produced which some grammar parses”. In three dimensions, this
corresponds to the requirement

$$a_{(1)} + a_{(2)} + a_{(3)} + a_{(1,2)} + a_{(1,3)} + a_{(2,3)} + a_{(1,2,3)} = 1. \quad (\text{II.10})$$

Rearranging the terms on the left hand side, we obtain

$$a_{21} + a_{32} + a_{13} + a_{(1,2,3)} = 1. \quad (\text{II.11})$$

On the other hand, arranging the $\alpha$ terms differently, we have

$$a_{31} + a_{23} + a_{12} + a_{(1,2,3)} = 1. \quad (\text{II.12})$$

From (II.11) and (II.12),

$$a_{21} + a_{32} + a_{13} = a_{31} + a_{23} + a_{12} \quad (\text{II.13})$$

or

$$(a_{21} - a_{12}) + (a_{32} - a_{23}) + (a_{13} - a_{31}) = 0. \quad (\text{II.14})$$

Writing $\delta_{ij} = a_{ji} - a_{ij}$, we have

$$\delta_{12} + \delta_{23} + \delta_{31} = 0 \quad (\text{II.15})$$

which I will refer to as the *cyclical balance criterion*. The advantage matrix of any
3-grammar system, then, has to satisfy this criterion.

Within the remit of the cyclical balance criterion, many qualitatively different
kinds of advantage matrix are possible. In particular, it is possible for some of the
advantage quantities $a_{ij}$ to equal zero — this will be the case if inclusion (subset–
superset) relations exist among the competing grammars, in the sense that one gram-
mar parses everything that another does. In what follows, I will however usually
assume that $a_{ij} > 0$ for all $i$ and $j$ with $i \neq j$, and will say that an advantage matrix
satisfying this condition is *proper*. Assuming advantage matrices to be proper thus
delimits the class of formal systems studied to some extent; the benefit of making this assumption is that it makes available a useful learning-theoretic approximation which is not available in the improper case, as we will shortly see. Without this approximation, the improper cases need to be studied separately, on a case-by-case basis.

II.4 Dynamics: general results

With the penalty probabilities (II.8) in hand, we may now proceed to study the dynamics of the three-grammar case. The general form of the LRP algorithm reads as follows:

**Algorithm II.2 (LRP; Narendra & Thathachar, 1989, 116–117).**

1. Let \( \pi_i = 1/n \) initially.
2. Present an input token (sentence) \( x \) to the learner. This is generated by \( G_i \) with probability \( p_i \).
3. Learner picks grammar \( G_i \) with probability \( \pi_i \).
4. Suppose learner picked \( G_k \).
   a. If \( G_k \) parses \( x \), learner replaces \( \pi_k \) with \( \pi_k + \gamma(1 - \pi_k) \), with learning rate \( \gamma \), and \( \pi_j \) with \( (1 - \gamma)\pi_j, j \neq k \).
   b. If \( G_k \) does not parse \( x \), learner replaces \( \pi_k \) with \( (1 - \gamma)\pi_k \) and \( \pi_j \) with \( \frac{\gamma}{n - 1} + (1 - \gamma)\pi_j, j \neq k \).
5. Steps 2–4 are repeated for \( T \) input tokens.

Assuming reliable learners (large \( T \), small \( \gamma \)), Narendra and Thathachar (1989, 117) show that the following approximation holds for the learner’s hypothesis at the end of the learning cycle:

\[
\hat{\pi}_i \approx \frac{\prod_{j \neq i} c_j}{\sum_j \prod_{k \neq j} c_j}.
\]  
(II.16)

Assuming non-overlapping generations of such learners thus yields the diachronic difference equation

\[
p'_i = \frac{\prod_{j \neq i} c_j}{\sum_j \prod_{k \neq j} c_j}.
\]  
(II.17)

3If all the penalty probabilities are strictly positive, \( c_i > 0 \) for all \( i \), then this slightly unwieldy formula reduces to the more aesthetic \( \hat{\pi}_i \approx c_i^{-1} / \sum_j c_j^{-1} \) upon division of both the numerator and the denominator by \( \prod_i c_i \). Narendra and Thathachar (1989) limit their discussion to this case.
In particular, in three dimensions one has

\[
\begin{align*}
p'_1 &= \frac{c_2 c_3}{c_2 c_3 + c_1 c_3 + c_1 c_2} \\
p'_2 &= \frac{c_2 c_3 + c_1 c_3 + c_1 c_2}{c_1 c_3} \\
p'_3 &= \frac{c_1 c_2}{c_2 c_3 + c_1 c_3 + c_1 c_2}
\end{align*}
\]  

(II.18)

where, it bears stressing, each penalty \( c_i \) is itself a function of the system state \( p = (p_1, p_2, p_3) \), leading to a nonlinear equation. For this to be well-defined, mathematically speaking, we need to check that the denominators never equal zero. This is guaranteed for all proper advantage matrices:

**Theorem II.3.**

*For a proper advantage matrix, \( c_i = 0 \) if and only if \( p_i = 1 \).*

*Proof.* Since \( A \) is proper, \( c_i = \sum_{j \neq i} a_{ij} p_j = 0 \) if and only if \( p_j = 0 \) for all \( j \neq i \). But since \( p \) is a probability distribution, the latter occurs if and only if \( p_i = 1 \).

**Corollary II.1.**

*Given a proper advantage matrix, it is never possible for two penalty probabilities \( c_i \) and \( c_j, i \neq j \), to equal zero at the same time. Consequently, the denominators in (II.18) are never zero.*

The learning-theoretic approximation (II.16) therefore leads to a well-defined intergenerational (diachronic) dynamical system whenever advantages are proper (as pointed out in the preceding discussion, the improper cases would need to be studied separately, a task which I set aside in the present paper).

As the \( p_i \) are probabilities, the system (II.18) is defined on the 3-dimensional simplex

\[ S_3 = \{ p = (p_1, p_2, p_3) : 0 \leq p_1, p_2, p_3 \leq 1 \text{ and } p_1 + p_2 + p_3 = 1 \}. \]  

(II.19)

This set may be partitioned into the *interior*

\[ \text{int} \ S_3 = \{ p \in S_3 : 0 < p_1, p_2, p_3 < 1 \} \]  

(II.20)

and the *boundary*

\[ \text{bd} \ S_3 = \{ p \in S_3 : p_i = 0 \text{ for some } i \}. \]  

(II.21)

Of special interest are the three points \( v_1 = (1, 0, 0), v_2 = (0, 1, 0) \) and \( v_3 = (0, 0, 1) \), corresponding to a state of dominance by one of the three grammars; these points are the *vertices* of the simplex. In what follows, I will illustrate the behaviour of
The state \( p = (p_1, p_2, p_3) \) of a 3-grammar system is defined on the 3-dimensional simplex \( S_3 \), which is best illustrated using a barycentric ternary plot.Shown here are the three vertices \( v_1 = (1, 0, 0) \), \( v_2 = (0, 1, 0) \) and \( v_3 = (0, 0, 1) \) as well as the barycentre \( (1/3, 1/3, 1/3) \).

In the general case, the system (II.18) is too complicated to be solved analytically. In other words we do not have, for an arbitrary advantage matrix \( A \), a closed-form equation that would tell us the exact time evolution of the system from any given initial state. We can, however, arrive at an understanding of the system’s dynamics by finding its rest points and studying their stability. A first result is that each of the three vertices \( v_i \) is a rest point and that no further rest points exist on the boundary \( \text{bd} S_3 \), whenever \( A \) is proper:

**Theorem II.4.**

*The points \( v_1 = (1, 0, 0) \), \( v_2 = (0, 1, 0) \) and \( v_3 = (0, 0, 1) \) are rest points of (II.18) for any proper advantage matrix. No other point in \( \text{bd} S_3 \) is a rest point.*

**Proof.** Using Theorem II.3, inspection of (II.18) immediately shows that \( v'_i - v_i = 0 \), i.e. that each vertex \( v_i \) is a rest point.

Now suppose that \( p = (p_1, p_2, 0) \) is a rest point. Then \( p'_2 - p_3 = 0 \), which by (II.18) implies that \( c_1 c_2 = 0 \), which implies that either \( c_1 = 0 \) or \( c_2 = 0 \). From Theorem II.3, \( p_1 = 1 \) in the first case and \( p_2 = 1 \) in the second. Due to the symmetry of (II.18), the same argument holds for states of the form \( (p_1, 0, p_3) \) and \( (0, p_2, p_3) \). Thus, if \( p \in \text{bd} S_3 \) is a rest point, it is necessarily a vertex.
If an interior rest point exists, it satisfies a stability condition:

**Theorem II.5.**

Let \( p = (p_1, p_2, p_3) \in \text{int} S_3 \). Then \( p \) is a rest point if and only if \( c_1 p_1 = c_2 p_2 = c_3 p_3 \).

**Proof.** Since \( p \in \text{int} S_3 \), Theorem II.3 implies that \( c_i > 0 \) for all \( i \). Division by the \( c_i \) is then possible, and (II.17) reduces, with algebra, to

\[
p_i' = \frac{c_i^{-1}}{\sum_j c_j^{-1}} = \frac{1}{c_i \sum_j c_j^{-1}}.
\]

Now \( p_i' - p_i = 0 \) if and only if

\[
c_i p_i = \frac{1}{\sum_j c_j^{-1}}.
\]

This holds for all \( i \) and the right hand side is independent of \( i \). Hence, the previous is equivalent to \( c_1 p_1 = c_2 p_2 = c_3 p_3 \).

Apart from these simple observations, it is difficult to obtain further results concerning the behaviour of (II.18) in the general case. I will next turn to a consideration of a number of special cases which are considerably easier to analyse, in increasing order of complexity, so as to arrive at a general picture of the diachronic behaviour of multiple-grammar systems based on LRP learning.

### II.5 Babelian systems

Arguably the simplest case occurs when all of the pairwise advantages \( a_{ij} \) are equal — in this case, no single grammar has a net benefit over the rest. Formally, I will say that a system is *Babelian* if its advantage matrix satisfies the following: there is an \( a > 0 \) such that \( a_{ij} = a \) for all \( i, j \) with \( i \neq j \). In three dimensions, this amounts to matrices of the form

\[
A = \begin{bmatrix}
0 & a & a \\
 a & 0 & a \\
 a & a & 0
\end{bmatrix}.
\] (II.22)

Notice that such matrices satisfy the cyclical balance criterion (II.15) and are thus valid advantage matrices.

Any Babelian 3-grammar system turns out to have one interior rest point, namely the maximum entropy state \( (1/3, 1/3, 1/3) \):

**Theorem II.6.**

For any Babelian 3-grammar system, the state \( (1/3, 1/3, 1/3) \) is the only interior rest point.
Proof. That \( (1/3, 1/3, 1/3) \) is a rest point would be easy to establish using Theorem II.5. To prove the stronger result that it is the only interior rest point of a Babelian system, let us look at the difference equation (II.18) directly. In the interior \( \text{int} S_3 \), one has (cf. proof of Theorem II.5)

\[
p'_i - p_i = \frac{c_i^{-1}}{\sum_j c_j^{-1}} - p_i = \frac{(\sum_k a_{ik} p_k)^{-1}}{\sum_j (\sum_k a_{jk} p_k)^{-1}} - p_i = \frac{(\sum_k a p_k)^{-1}}{\sum_j (\sum_k a p_k)^{-1}} - p_i = \frac{a^{-1}(\sum_k p_k)^{-1}}{a^{-1} \sum_j (\sum_k p_k)^{-1}} - p_i
\]

for a Babelian system. But \( \sum_k p_k = 1 \), so the above is equivalent to

\[
p'_i - p_i = \frac{a^{-1}}{3a^{-1}} - p_i = \frac{1}{3} - p_i
\]

in three dimensions. Hence \( p'_i - p_i = 0 \) if and only if \( p_i = 1/3 \), and consequently \( (1/3, 1/3, 1/3) \) is the only interior rest point.

Thus any Babelian three-grammar system has four rest points: the three vertices, corresponding to total dominance by one of the three grammars, and the maximum entropy state in which each grammar has equal representation. It remains to figure out the stability of these rest points. In general, stability analysis hinges on studying how the state of the dynamical system under consideration changes in the immediate vicinity of the rest point in question — whether nearby points in the system’s state space are attracted to the rest point or repelled by it (cf. our discussion of the three pendula in Section II.1). Mathematically, we need to study the partial derivatives of the system’s evolution equations when evaluated at the rest point. For a three-dimensional system, the \textit{Jacobian matrix} is defined as the matrix of partial derivatives

\[
J = \begin{bmatrix}
\frac{\partial f_1}{\partial p_1} & \frac{\partial f_1}{\partial p_2} & \frac{\partial f_1}{\partial p_3} \\
\frac{\partial f_2}{\partial p_1} & \frac{\partial f_2}{\partial p_2} & \frac{\partial f_2}{\partial p_3} \\
\frac{\partial f_3}{\partial p_1} & \frac{\partial f_3}{\partial p_2} & \frac{\partial f_3}{\partial p_3}
\end{bmatrix}
\]  

(II.23)

where the functions \( f_i \) are as in (II.1). When the partial derivatives \( \frac{\partial f_i}{\partial p_j} \) are evaluated at a rest point \( p = (p_1, p_2, p_3) \), the Jacobian reduces to a matrix of real numbers; denote this by \( J(p) \). It can then be shown that, for a discrete-time system, (1) if the modulus of each eigenvalue of \( J(p) \) is strictly less than 1, the rest point \( p \) is asymptotically stable, and (2) if the modulus of at least one eigenvalue is strictly greater than
1, p is unstable (Drazin, 1992, 70–71). While this method is foolproof in the sense that it is purely a matter of mechanical calculation, computing the eigenvalues is in most cases extremely tedious and is best left to a computer. In what follows, I shall consequently only report the end results of these computations, suppressing the gritty details.

Applying the Jacobian method on (II.18) gives us our main result on the stability of Babelian systems.

**Theorem II.7.**

*In a three-dimensional Babelian system, the interior rest point \((1/3, 1/3, 1/3)\) is asymptotically stable. The vertex rest points \((1, 0, 0)\), \((0, 1, 0)\) and \((0, 0, 1)\) are all unstable.*

*Proof.* The Jacobian has eigenvalues 0 and 2 at each of the three vertices, and eigenvalues 0 and 1/2 at the interior rest point \((1/3, 1/3, 1/3)\).

Thus, as expected, the natural tendency in a Babelian system is away from dominance and towards the maximally mixed state \((1/3, 1/3, 1/3)\) in which each grammar is used with probability 1/3 (Figures II.6–II.7). This shows that three-grammar Babelian systems have “built-in” stable variation, in stark contrast to the two-grammar case (Section II.2).

**II.6 Symmetric systems**

The above analysis illustrates the procedure of sketching the qualitative behaviour of a dynamical system by way of analysing the system’s rest points and their stability, when the equations governing the system’s evolution cannot be solved. It also shows that true stable variation is a feature of at least some formal systems based on LRP learning. Babelian systems, of course, are far too trivial to be of any serious linguistic interest, and it remains to show that stable variation may occur in other, more realistic multiple-grammar settings.

A straightforward way of generalizing from Babelian systems is to allow some of the grammars to have unequal advantages but to maintain a symmetry condition: \(a_{ij} = a_{ji}\) for all \(i, j\). In three dimensions, such symmetric systems are thus described by advantage matrices of the form

\[
A = \begin{bmatrix}
0 & a_{12} & a_{13} \\
0 & 0 & a_{23} \\
a_{12} & a_{23} & 0
\end{bmatrix}
\]

\[= \begin{bmatrix}
0 & a & b \\
a & 0 & c \\
b & c & 0
\end{bmatrix}
\]

where I write \(a = a_{12}\), \(b = a_{13}\) and \(c = a_{23}\) for convenience. Again, it is clear that
Figure II.6
Phase space plot of Babelian 3-grammar systems. The three unstable vertex rest points are shown as open circles and the stable interior rest point as a filled circle, as is customary; the line segments give the magnitude and direction of change at various points in the state space. The series of asterisks illustrates one diachronic (inter-generational) trajectory from the initial state \( p = (0.1, 0.9, 0.0) \); see Figure II.7 for a conventional representation of this trajectory in the time dimension.

Figure II.7
The trajectory from Figure II.6 shown in the time dimension.
these matrices satisfy the cyclical balance criterion (II.15) and thus are well-defined.

Setting $p_i' - p_i = 0$ in (II.18) and solving for $p_i$ (in a manner analogous to that in the proof of Theorem II.6 above) reveals that in a symmetric three-grammar system, a rest point exists at

$$p = \left( \frac{c}{a + b + c}, \frac{b}{a + b + c}, \frac{a}{a + b + c} \right). \tag{II.25}$$

Continuing to assume proper advantage matrices, in other words that $a, b, c > 0$, it follows that this rest point is always contained in the interior $\text{int} \, S_3$. It is also the only solution of $p_i' - p_i = 0$ in the interior and hence the only interior rest point of a symmetric system. Furthermore, stability analysis finds that the Jacobian, when evaluated at this rest point, has eigenvalues $0 < 1$ and $1/2 < 1$; hence, the interior rest point is always asymptotically stable. For each of the vertex rest points $v_1$, $v_2$ and $v_3$, the eigenvalues are $0 < 1$ and $2 > 1$. Thus:

**Theorem II.8.**

*Any proper, symmetric three-grammar system* (II.24) *has exactly one interior rest point at*

$$p = \left( \frac{c}{a + b + c}, \frac{b}{a + b + c}, \frac{a}{a + b + c} \right).$$

*This interior rest point is asymptotically stable, while the vertex rest points are all unstable.*

Crucially, the result holds for any values of $a, b, c > 0$. We then conclude:

**Corollary II.2.**

*Any proper, symmetric system of three grammars tends to a state of stable variation.*

Figure II.8 illustrates for a particular choice of the parameters $a$, $b$ and $c$.

**II.7 Quasi-Babelian systems**

Another way of generalizing from the Babelian special case is to explore a more comprehensive class of systems in which some one grammar has either a larger or a smaller advantage than any of its competitors, the latter sharing the same amount of advantage amongst themselves. Formally, I will call a system quasi-Babelian if constants $a, b > 0$ exist such that (1) for some unique $i$, $a_{ji} = b$ for all $j \neq i$, and (2) $a_{kj} = a$ for all $j \neq i$, for all $k \neq j$. By a relabelling of grammars, we may always take $G_1$
Phase space plot of a symmetric 3-grammar system with $a = 0.05$, $b = 0.01$ and $c = 0.02$. Each trajectory not starting at a vertex point tends towards the stable interior rest point at $(c/D, b/D, a/D)$ with $D = a + b + c$.

Figure II.8

To correspond to the grammar having the unique advantage $b$, and I will refer to this as the canonical quasi-Babelian case. In three dimensions, a canonical quasi-Babelian advantage matrix, then, is of the form

$$A = \begin{pmatrix} 0 & a & a \\ b & 0 & a \\ b & a & 0 \end{pmatrix}.$$  \hspace{1cm} (II.26)

Again, it can be checked that the cyclical balance criterion (II.15) is satisfied.

The advantage matrix now has just two independent parameters, $a$ and $b$, and consequently algebraic manipulation of the equations (II.18) becomes easy. Setting $p'_i - p_i = 0$ and solving for $p_i$ reveals that with a canonical quasi-Babelian advantage matrix, (II.18) has either three or four rest points in the simplex $S_3$. In addition to the vertices $v_1$, $v_2$ and $v_3$, a fourth solution exists in the interior at the point

$$p^* = \left( \frac{1}{5 - 2\rho}, \frac{2 - \rho}{5 - 2\rho}, \frac{2 - \rho}{5 - 2\rho} \right)$$ \hspace{1cm} (II.27)

whenever $0 < \rho < 2$, where $\rho = b/a$ gives the ratio of the two advantage parameters. At $\rho = 2$, this solution coalesces with the vertex $v_1$.

This rest point $p^*$ entails a sort of behaviour which is entirely unattested in Babelian and symmetric systems: a bifurcation. For small values of the ratio $\rho = b/a$—that is, for values of $b$ which are small in comparison to $a$—the interior rest point
p* exists. As ρ is increased, this rest point moves towards the vertex v₁ and coincides with the latter at the critical value ρ = ρ_c = 2 of the bifurcation parameter ρ. For ratios ρ ≥ 2, the system consequently only has the three vertex rest points. The following theorem establishes the stability of these rest points in response to the bifurcation; Figures II.9–II.10 illustrate.

**Theorem II.9.**

Assume a canonical quasi-Babelian 3-grammar system with advantage ratio ρ = b/a. Then

1. the vertex rest points v₂ = (0, 1, 0) and v₃ = (0, 0, 1) are always unstable;
2. the vertex rest point v₁ = (1, 0, 0) is asymptotically stable if ρ ≥ 2 and unstable if 0 < ρ < 2;
3. the interior fixed point p* = \( \left( \frac{1}{5-2ρ}, \frac{2-ρ}{5-2ρ}, \frac{2-ρ}{5-2ρ} \right) \) is asymptotically stable whenever it exists, i.e. when 0 < ρ < 2.

**Proof.** For the two vertices v₂ and v₃, the eigenvalues of the Jacobian are 0 and 1 + ρ > 1. Hence, these points are unstable.

At the vertex v₁, the Jacobian has eigenvalues 0 and 2a/b. Hence, this rest point is asymptotically stable if 2a/b < 1, i.e. if b/a = ρ > 2, and unstable if b/a = ρ < 2.

At the interior rest point p*, the Jacobian has eigenvalues 0 < 1, 1/2ρ and 1 − 1/2ρ < 1. Thus, the interior rest point is asymptotically stable whenever 1/2ρ < 1, i.e. when ρ < 2.

**II.8 Naive learning**

Above, I have explored a generalization of the 2-grammar Variational Learner. This generalization has shown that stable variation is an intrinsic feature of many multiple-grammar systems based on LRP learning. The specific systems studied and their interrelationships are summarized in Figure II.11; future work will need to explore systems that lie outside these classes of systems.

Crucially, the preceding analysis relies on the straightforward generalization of LRP learning for n options given in Algorithm II.2. From a psycholinguistic point of view, this way of treating the learner implies, for better or worse, that the learner must keep track of n independent probabilities. Considering that even a few dozens of (binary) grammatical parameters result in an astronomical search space for the learner, the straightforward extension of the LRP algorithm may be argued to be unrealistic on psychological grounds.\(^4\)

\(^4\)The issue is in fact convoluted: on the one hand, the number of grammatical parameters is not known with any certainty (for one recent estimate, see Longobardi & Guardiano, 2009, 1687, who suggest 63 parameters in the DP domain and note that in general "UG parameters number at least in
Phase space plots of the canonical quasi-Babelian 3-grammar system for various advantage ratios $\rho = b/a$; $\rho = 1$ corresponds to the strictly Babelian special case. At $\rho = 2$ a bifurcation occurs in which the interior rest point joins the vertex $v_1$, reversing the latter’s stability.
**Figure II.10**
Orbit diagram of quasi-Babelian 3-grammar systems, illustrating the stable limiting state of the system when started from any non-vertex state. The solid curve gives the value of $p_1$ at the stable rest point, while the dashed curve gives the value of $p_2 = p_3$.

**Figure II.11**
Set relations among the 3-grammar systems studied in this paper, in the universe of all admissible systems (all 3 × 3 advantage matrices satisfying the cyclical balance criterion): all Babelian systems are both symmetric and quasi-Babelian, and all symmetric and quasi-Babelian systems are proper.
An alternative, explored to some extent in Yang (2002b), is to have the learner operate in a parametrically constrained space. That is to say, instead of operating on \(n\) grammar probabilities \(\pi_1, \ldots, \pi_n\), suppose the learner operates on \(N\) parameter probabilities \(\xi_1, \ldots, \xi_N\), where \(n = 2^N\) and \(\xi_i\) gives the probability of the \(i\)th binary parameter being set on. To recover the grammar probabilities, it suffices to multiply the relevant parameter probabilities:

\[
P(G_{\sigma(1)\sigma(2)\ldots\sigma(N)}) = \prod_{i=1}^{N} \xi_{\sigma(i)}^{\sigma(i)}(1 - \xi_{\sigma(i)})^{1-\sigma(i)}
\]  

(II.28)

is the probability of the grammar \(G_{\sigma(1)\sigma(2)\ldots\sigma(N)}\) being selected, with \(\sigma(i) = 1\) if the \(i\)th parameter is to be set on and \(\sigma(i) = 0\) if the \(i\)th parameter is to be set off for this particular grammar.

Since what gets rewarded or punished is the selection of entire grammars and not the selection of individual parameter values, the learner now faces the problem of not knowing which parameter setting(s) to blame in case of parsing failure (Yang, 2002b). One way of attempting to overcome this problem is the following naive learning algorithm.

**Algorithm II.3 (Naive Parameter Learner (NPL); Yang, 2002b).**

1. Set \(\xi_i = 0.5\) for all \(i\) initially.
2. Pick grammar by setting \(i\)th parameter on with probability \(\xi_i\).
3. Receive input sentence \(x\).
4. If grammar parses \(x\):
   a. If \(i\)th parameter was on, increase the value of \(\xi_i\) by replacing \(\xi_i\) with \(\xi_i + \gamma(1 - \xi_i)\), where \(\gamma\) is a learning rate.
   b. Else decrease the value of \(\xi_i\) by replacing it with \((1 - \gamma)\xi_i\).
5. If grammar does not parse \(x\):
   a. If \(i\)th parameter was on, decrease the value of \(\xi_i\) by replacing it with \((1 - \gamma)\xi_i\).
   b. Else increase the value of \(\xi_i\) by replacing it with \(\xi_i + \gamma(1 - \xi_i)\).
6. Repeat steps 2–5 for \(T\) input tokens.

the hundreds”), and on the other hand, the human brain is capable of storing astronomical quantities of information (Bartol Jr et al., 2015). I set the issue aside here — for present purposes, what matters is that stable variation is attested both in the straightforward \(n\)-grammar generalization of LRP learning and in the parametrically constrained Naive Learner, as we will presently see.
Having learners operate in a parametrically constrained space and employing a learning algorithm such as NPL complicates the study of the diachronic behaviour of such a system, since analogues of the learning-theoretic limiting approximations (II.2) and (II.16) are not available. It is, however, possible to study special cases with the help of computer simulations. In what follows, I will explore one such simple special case and show that stable variation is, again, a feature of at least some systems based on Naive Parameter Learning in a parametric space.

For this, suppose for simplicity that Universal Grammar (UG) provides just two elements, a “noun” N and a “determiner” D, and two parameters:

1. whether determiner can be null (on setting) or has to be overt (off setting)
2. whether grammar is head-final (on setting) or head-initial (off setting)

The four grammars then parse, and fail to parse, strings as follows:

<table>
<thead>
<tr>
<th>.parsers fails to parse</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_{11}$</td>
</tr>
<tr>
<td>$G_{10}$</td>
</tr>
<tr>
<td>$G_{01}$</td>
</tr>
<tr>
<td>$G_{00}$</td>
</tr>
</tbody>
</table>

Assuming true optionality, i.e. that grammars $G_{11}$ and $G_{10}$ generate the two types of sentence with probability 0.5, it is easy to work out the probability of each possible input string the learner may encounter:

$$P(N) = 0.5P(G_{11}) + 0.5P(G_{10})$$
$$= 0.5x_1x_2 + 0.5x_1(1 - x_2)$$
$$= 0.5x_1$$

$$P(DN) = 0.5P(G_{11}) + P(G_{01})$$
$$= 0.5x_1x_2 + (1 - x_1)x_2$$
$$= x_2(1 - 0.5x_1) \tag{II.29}$$

$$P(ND) = 0.5P(G_{10}) + P(G_{00})$$
$$= 0.5x_1(1 - x_2) + (1 - x_1)(1 - x_2)$$
$$= (1 - x_2)(1 - 0.5x_1)$$

Here, $x_1$ and $x_2$ are the population-level parameter probabilities (corresponding to $p_i$ in the LRP formulation). The penalty probabilities of the four grammars are then
A two-parameter Naive Parameter Learner at the vertex $x = (1, 1)$ ($G_{11}$ is the unique target grammar); values of $\xi_1$ and $\xi_2$ from one computer simulation. Both of the learner’s parameter probabilities $\xi_1$ and $\xi_2$ tend to 1.

Substituting $x_1 = x_2 = 1$ in (II.30) yields

$$\begin{align*}
c(G_{11}) &= (1 - x_2)(1 - 0.5x_1) \\
c(G_{10}) &= x_2(1 - 0.5x_1) \\
c(G_{01}) &= 0.5x_1 + (1 - x_2)(1 - 0.5x_1) \\
c(G_{00}) &= 0.5x_1 + x_2(1 - 0.5x_1)
\end{align*}$$

which shows that if $G_{11}$ is the unique target grammar, then the NPL algorithm will eventually arrive at the right parameter probabilities $\xi_1 = 1$ and $\xi_2 = 1$, as long as the learner has enough time to tweak the probabilities (Figure II.12). Performing the requisite substitutions shows that the same holds for the remaining three grammars $G_{10}$, $G_{01}$ and $G_{00}$, as well.

The four vertices, at which one of the four grammars has total use, are thus found
to be rest points for the above toy system. What about their stability? To explore this question, we need to set the learner in a mixed environment (at a state in the interior int \( S_4 \) of the four-simplex of grammar probabilities). Figure II.13 shows the behaviour of the learner in the mixed environment \((x_1, x_2) = (0.99, 0.99)\), corresponding to \( P(G_{11}) = 0.99^2 = 0.9801 \) use of the grammar \( G_{11} \). Convergence to the vertex no longer occurs, and the diachronic implications of this become manifest when we set up a sequence of such learners, the output of one generation again feeding as input to the following generation: when started from a mixed state, the system fails to converge to the vertex rest point at which \( G_{11} \) has dominance, and instead appears to be attracted to an interior rest point, that is to say, towards a state of stable variation (Figure II.14).

II.9 Conclusions and conjectures

In this paper, I have shown that diachronically stable variation arises in many kinds of settings of grammar competition, as long as more than two grammars are represented in the learner’s environment. In addition to a systematic study of the \( n \)-grammar LRP learning algorithm in Sections II.2–II.7, the preliminary exploration of a toy parametric UG in Section II.8 points to the conclusion that stable variation occurs in the parametrically constrained Naive Parameter Learner as well.

The results of this paper invite experimentalists to look for evidence of stable
variation in a specific kind of situation — complex language contact. Indeed, given Yang’s (2000) Fundamental Theorem, more than two grammars must be present in the learner’s environment for stable variation to occur, if language acquisition operates along the lines of linear reward–penalty learning. This is a necessary but not a sufficient condition — above we have seen, for example, that quasi-Babelian systems exhibit a phase transition between a phase in which stable variation occurs and one in which it does not occur (the most advantageous grammar instead claiming, eventually, all probability mass). Yet there is a kind of fatalism to these results: all symmetric systems, for instance, always tend to an attractor which is a state of stable variation by Theorem II.8. It thus bears stressing that whenever stable variation occurs in these models, it is not due to extraneous factors such as social evaluations or population dynamics; stable variation follows from the nature of the LRP learning algorithm itself.

It may be instructive to consider this point in a little more detail. Thus consider step 4.b of Algorithm II.2, corresponding to parsing failure. Here the algorithm tells us that whenever the grammar chosen by the learner, $G_k$, fails to parse a sentence, the learner updates the $k$th probability to become $\pi_k = (1 - \gamma)\pi_k$. Thus the probability $\pi_k$ is diminished, and for all the grammar probabilities to keep summing to unity, it follows that some of the remaining probabilities need to be increased. From 4.b, we

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure.png}
\caption{Diachrony for a sequence of Naive Parameter Learners from the initial state $x = (0.99, 0.99)$. Convergence to the vertex $x = (1, 1)$ does not occur, suggesting that this vertex is an unstable rest point.}
\end{figure}
find that the learner actually updates every other probability $\pi_j$, $j \neq k$, to become

$$\frac{\gamma}{n - 1} + (1 - \gamma)\pi_j.$$  

(II.32)

It is not difficult to check that these choices imply $\sum_i \pi_i = 1$, as desired. The consequences of choosing the update (II.32) over other possible choices, however, are non-trivial. Note that this manner of performing the update means that every grammar (apart from $G_k$, which failed) gets boosted by the same amount. This, then, means that the probability vector $\pi = (\pi_1, \ldots, \pi_n)$ that describes the learner’s grammar probabilities is shifted towards the centre $(1/n, \ldots, 1/n)$ of the simplex at every occasion of parsing failure. When this mechanism is iterated over a diachronic sequence of learners, the effect gets amplified and, as we have seen, in some cases leads to diachronically stable variation. This observation also explains why the two-grammar version of the same algorithm behaves so differently: in this case, whenever one of the grammars fails to parse an input sentence, there is just one other grammar whose probability to boost. Consequently the probability vector describing the learner’s state drifts towards dominance by this other grammar rather than towards a mixed state.

I would like to conclude by putting forward the following two conjectures, each supported by the special cases studied above but whose proofs have so far been elusive in the general case: (1) that any $n$-grammar system with a proper advantage matrix has either $n$ rest points (the vertices) or $n + 1$ rest points (the vertices plus one rest point in the interior of the simplex); and (2) that in any proper system, if the interior rest point exists, it is necessarily asymptotically stable. If these results were to carry over to the NPL algorithm, too, the consequence would be clear: diachronic systems of learners operating on linear reward–penalty learning or variants thereof in multiple-grammar environments display a good deal of stable variation. Whether this is acceptable, or whether instead the above results call for a re-evaluation of the assumptions that underlie probabilistic language acquisition algorithms, needs to be answered by empirical work into the occurrence of stable variation in real-life language communities.

**Acknowledgements**

Apart from the DiGS 18 pre-conference workshop on diachronic stability, portions of this work were presented at the 2016 Annual Meeting of the Linguistics Association of Great Britain (York, September 2016). I wish to thank both audiences, as well as Ricardo Bermúdez-Otero, George Walkden and an anonymous reviewer for feedback. All remaining errors and absurdities are, naturally, mine. The research here reported was made possible by generous financial support from Emil Aaltonen Foundation and
The Ella and Georg Ehrnrooth Foundation.
Neutral change

Henri Kauhanen

Abstract

Language change is neutral if the probability of a language learner adopting any given linguistic variant only depends on the frequency of that variant in the learner’s environment. Ruling out non-neutral motivations of change, be they sociolinguistic, computational, articulatory or functional, a theory of neutral change insists that at least some instances of language change are essentially due to random drift, demographic noise and the social dynamics of finite populations; consequently, it has remained little investigated in the historical and sociolinguistics literature, which has generally been on the lookout for more substantial causes of change. Indeed, recent computational studies have argued that a neutral mechanism cannot give rise to ‘well-behaved’ time series of change which would align with historical data, for instance to generate S-curves. In this paper, I point out a methodological shortcoming of those
studies and introduce a mathematical model of neutral change which represents the language community as a dynamic, evolving network of speakers. With computer simulations and a quantitative operationalization of what it means for change to be well-behaved, I show that this model exhibits well-behaved neutral change provided that the language community is suitably clusterized. Thus, neutral change is not only possible but is in fact a characteristic emergent property of a class of social networks. From a theoretical point of view, this finding implies that neutral theories of change deserve more (serious) consideration than they have traditionally received in diachronic and variationist linguistics. Methodologically, it urges that if change is to be successfully modelled, some of the traditional idealizing assumptions employed in much mathematical modelling must be done away with.

**Keywords:** language change; neutrality; mathematical modelling; prestige; S-curves

III.1 Introduction

An outstanding problem in diachronic linguistics concerns the extent to which language change is and can be neutral. Once variation arises, are the competing variants created equal, or is change instead motivated by functional, computational, social or other considerations which favour certain variants over others? Expressing an agnostic take on this question, Lass (1997) observes that

> it’s perfectly possible that both variation and change itself (as a result) are neutral: even selection does not necessarily have to select that which is better “adapted”. In any case, there are even in biology modes of (apparent) selection that are not (in the Darwinian sense) genuinely “selective” or “adaptive” […]. All of these possibilities, given the much better understood nature of variation and change in organisms, need to be considered before any claim for “function” can be made for either variation or change. (Lass, 1997, 354, my emphasis.)

The thrust of this programmatic message, mainly directed at functional explanation but not limited in its scope to explanations of functionalist persuasions, is that the possibility of neutral change has remained, and continues to remain, underinvestigated.

This paper examines that possibility by means of a simple mathematical model of variant competition in a finite population of speakers. Guided by the intuition that language diachrony is typically well-behaved (in a sense to be made precise later), I propose a quantitative metric of well-behavedness and, with the help of computer simulations, investigate how the neutrality hypothesis fares in its light. The upshot
of this investigation is that well-behaved neutral change is, indeed, found to be possible if the social network underlying the language community has a suitable topology and dynamics: briefly, if the language community is strongly clusterized, so that it can be partitioned into more central and more peripheral speakers, neutral change is observed. Moreover, it is found that well-behaved neutral change is a consistent, characteristic emergent feature of such social networks: the effect is not a statistical anomaly, but flows naturally and robustly from the way in which the language community is structured and the way in which that structure evolves over time. On the other hand, in a classical well-mixed (unstructured) population we find that change can rarely be neutral and well-behaved.

The model here studied differs from most previous mathematical models of language change (and, more generally, cultural evolution) in two respects. Firstly, the model does away with the classical idealizing assumption of representing language communities as well-mixed populations, often infinite, and looks instead at finite social networks with non-uniform degree distributions, that is to say networks in which different people have different patterns of connectivity. Secondly, the model takes into account the fact that human social networks are never static but are constantly being rewired by the removal and addition of individuals: friendship and even family ties are not fixed, people move from one social network to another, and deaths and births occur. The fact that neutral change cannot happen in a classical unstructured population but can happen in populations with suitable topologies and rewiring dynamics points to the need to consider the particularities of socialization in language communities at a level of detail which mathematical and computational models of language change have not attempted so far.

The relevance of the possibility of neutral change to diachronic and variationist linguistics is as follows: unless assumptions of non-neutral motivations of change can be supported for independent reasons, a neutral theory of change remains a viable explanatory strategy. More specifically, the observation that well-behaved neutral change is a characteristic feature of certain kinds of social networks suggests that in some cases of change, neutral selection may be at play in addition to, or instead of, non-neutral selection. Moreover, the parsimonious nature of neutral theory holds promise in clearing up certain puzzles which have traditionally received rather ad hoc solutions in the research literature: by removing the notion of (variant) prestige, a neutral mechanism can provide a fresh, bias-free sociolinguistic take on change, as I will argue in Section III.6.
III.2 Neutrality

The possibility that language change might be neutral has, traditionally, received little attention in historical and variationist linguistics. Aside from occasional remarks such as Lass’s (1997) cited above, and Postal’s (1968, 283–285) suggestion that language change is random, non-motivated ‘fashion change’, the neutrality hypothesis has received serious consideration mainly from Trudgill (2008) who, in questioning the role of identity in new-dialect formation, suggests that dialect contact and dialect mixing work in ‘automatic’, non-biased ways. Although Trudgill’s position is avowedly anti-identity, and thus rejects one form of (social) bias, it is not entirely clear whether his account might not admit some other form of bias, however. In fact, whether argued for or against, the neutrality hypothesis is rarely defined in precise, unequivocal terms in the literature. In this section, therefore, I will explicate the hypothesis by putting forward a definition of neutrality and contrasting neutral change with change governed by non-neutral factors.

Throughout this paper, I will focus on a situation in which a fixed number of linguistic variants are in competition in a specific linguistic domain. To keep the discussion maximally general, I will not make further assumptions about the nature or composition of these variants. Depending on the application, a variant could be a complete parametric specification of Universal Grammar, a single value of one particular parameter, an allophone of a phoneme, and so on. What matters is that there is a number of variants, each of which could be adopted, in principle, by any speaker. The neutrality hypothesis can then be stated, in intuitive terms, as follows.

(III.1) **Neutrality**

The probability with which a speaker acquires a certain linguistic variant out of a number of competing variants equals the relative frequency of that variant in the speaker’s neighbourhood, *modulo* a small probability of innovating, uniformly at random, another variant from among all (biologically, cognitively) possible variants.

‘Neighbourhood’ here means the speaker’s linguistic neighbourhood, a term to be explicated in more detail shortly. The content of the neutrality hypothesis, then, is that variant selection in individual speakers is controlled by the frequencies of the competing linguistic variants: apart from a small amount of random noise that accounts for variant innovation, no considerations other than the frequency distribution of variants could affect which variant an individual acquires or adopts. To borrow terminology from biology, under an assumption like (III.1) change (evolution) is *frequency-driven*, and the competing variants do not have differential *fitnesses* which could bias the process of adoption, favouring one variant over another. In contrast to fitness-driven (e.g. Darwinian) selection, in neutral change the adoption of a given
variant is not adaptive in any sense. The innovation events, when they do occur, are likewise neutral: the innovatory variant is chosen from among all possible variants uniformly at random, so there is no bias towards innovating any particular variant.\footnote{Although one should beware of drawing facile cross-disciplinary analogies, it is worthwhile to point out that neutral mechanisms of change have been proposed in evolutionary biology. In biological evolution, a variant (a genotype or a phenotype, or some part of one) is said to be selectively neutral or simply neutral if having that variant confers neither a selective advantage nor a selective disadvantage. Depending on one’s take on the level of selection debate (Reeve & Keller, 1999), this implies that a neutral variant will neither increase nor decrease the fitness of its bearer, of the bearer’s species, or of that variant itself. This mechanism of neutral evolution (Alonso, Etienne & McKane, 2006) is to be contrasted with Darwinian natural selection, which operates on complicated fitness landscapes that confer selective pressures on the competing replicators or vehicles. Although (non-neutral) natural selection remains the de facto mechanism for explaining evolution on various levels of biological organization, neutral theories have been proposed and defended for molecular evolution (Kimura, 1994) as well as in ecology for competition within a trophic level (Hubbell, 2001).}

It will be instructive at this point to briefly consider models and modes of explanation in diachronic linguistics which are either explicitly or implicitly non-neutral, so as to bring the contrast into sharper relief. The typological-functional explanations Lass (1997) alludes to in the quotation cited in Section III.1 form an obvious but important instance: there, it is assumed that a linguistic variant can perform better or worse in any given role; linguistic forms perform communicative, cognitive and other functions, and some do this better than others (e.g. Anttila, 1989). Under a functionalist hypothesis, processes of change will then be guided by people’s intuitions (conscious or subconscious) concerning the performance of different variants in serving these various functions. Change is not neutral, since adopting some variants is deemed, in one sense or another, better than adopting other variants, to the extent even that possible language states are classified into ‘consistent’ and ‘transitional’, or ‘preferred’ and ‘dispreferred’ ones (e.g. Hawkins, 1990; Vennemann, 1993; also see the critical discussion in Lightfoot, 1999, 85–87 and passim). In the most extreme versions of this framework, linguistic systems are viewed as teleological (Itkonen, 1981, 1982), and ‘language change is language improvement’ (Vennemann, 1993, 322).

Another way of flouting the neutrality hypothesis (III.1) is by way of considerations of economy of computation or of production. Starting with Lightfoot’s (1979) Transparency Principle, the diachronic generative syntax literature has generally favoured explanatory frameworks of this kind, where innate principles or third-factor processing constraints are taken to bias the acquisition of syntax. To take a more recent example, Roberts and Roussou (2003) assume a Merge-over-Move principle to account for parametric reanalysis and grammaticalization under the right kind of trigger experience. In much the same vein, theories and models of sound change which appeal to articulatory (e.g. aerodynamic, inertial) constraints as a motivation or cause of change are non-neutral, in that speakers are assumed to be biased to produce certain (e.g. centralized, lenited) phonetic variants over other, possible ones. Such con-
straints are said to give rise to variation in the auditory input available to the listener, eventually causing change in the speaker–listener loop (Ohala, 1989; Pierrehumbert, 2001).

A third mechanism for non-neutral change is constituted by different kinds of social biases. In a typical prestige-based explanation, for example, speakers are said to accommodate towards variants they consider prestigious or associate with a particular social group (Labov, 1972). Again, the acquisition or adoption of a variant under such circumstances will be non-neutral, as it is not driven simply by the frequency distribution of variants in the speakers’ environments. In a prestige-based explanation, the explanatory onus is on speakers’ estimations of the social ‘desirability’ of particular variants, either overt or covert, and even sociolinguists who are careful to consider other components of processes of actuation and propagation, such as variation in social network structure, have usually assumed (often a priori) that at least a small amount of prestige is necessary for innovatory forms to propagate through a language community. Thus

in view of the very general finding of sociolinguistic research that the prestige values attached to language are often quite covert and difficult to tap directly, we may suggest that a successful innovation needs to be evaluated positively, either overtly or covertly. This is of course a necessary but not a sufficient condition for its ultimate adoption (J. Milroy & Milroy, 1985, 368, my emphasis.)

These non-neutral mechanisms have been implemented, in varying degrees of detail, in computational models of language change. For instance, Ke et al. (2008) find that an innovatory variant must in their model be biased over the prevailing conventional variant, sometimes twenty-fold, in order to secure successful propagation. Similarly, in a model of sociolinguistic factors in change, Fagyal et al. (2010) find that speakers must be biased to adopt variants from speakers who are both well-connected and prestigious in order for the model to generate propagation curves that have the broad outline of an S-curve; the S-curve being taken as a basic desideratum which a model of change should be able to replicate. Finally, in what is perhaps the most extensive computational study so far of the effect of various kinds of biases in variant adoption and propagation, Blythe and Croft (2012) find that in their extension of the utterance selection model of language change (Baxter et al., 2006, 2009) a neutral, non-biased mechanism is unable to generate realistic time series of change such as S-curves.

The framework adopted in the latter study deserves more detailed comment, as it provides a useful sociolinguistic taxonomy of selection mechanisms in language

---

2It perhaps needs to be stressed in this connection that the point of contest between neutral and non-neutral theory is not whether things such as computational or articulatory constraints exist, but whether they are operative or causative in language change on a population level.
change more generally. Leaning on Croft’s (2000) evolutionary theory of language change, itself based on Hull’s (1988) general theory of selection processes, Blythe and Croft (2012, 272–277) classify replication mechanisms into four categories: (i) neutral evolution, which is random, frequency-driven drift; (ii) neutral interactor selection, in which speaker–speaker interaction frequencies play a role; (iii) weighted interactor selection, in which interactions between different speakers are weighted differently; and (iv) replicator selection, in which the competing linguistic variants themselves are weighted differently. The neutrality hypothesis (III.1), as here defined, corresponds to (i) and (ii) with the proviso that interaction frequencies play a role at the point of acquisition, not (necessarily) across the lifespan of speakers as in the usage-based model of Blythe and Croft (2012). The difference between the neutral mechanisms (i)–(ii) and the non-neutral ones (iii)–(iv) is that in the former case no social evaluations take place, whereas in the latter case either speakers or the competing linguistic variants themselves receive potentially differential evaluations, which fact is taken to be the motor of change. In fact, Blythe and Croft (2012) find that, in their model, S-curves are reliably obtained only for replicator selection, in other words for selection where the social evaluations mark linguistic variants directly, in the Labovian sense.

Although the aforementioned formal models have elucidated important aspects of variation and change in natural language, they have their limitations. Most importantly from the point of view of our present concerns, each of the three models mentioned above (Ke et al., 2008; Fagyal et al., 2010; Blythe & Croft, 2012) represents language communities as static networks of speakers. That is to say, even though these models are rich enough to represent differences in social network topology, or differences and asymmetries in the probabilities with which different speakers interact, they lack a mechanism for evolving that topology, and consequently fail to model the social dynamics of a language community: in the aforementioned models, there is no way for individual speakers to be removed from or added to the network, or for their connection sets or interaction probabilities to evolve within a single simulation run. Importantly, the models then fail to address the question of whether and how that social dynamics might affect the linguistic variant dynamics operating on the social network. The assumption of a static network clearly does not hold of human societies — and the longer the time spans of any particular changes we may be interested in explaining, the worse this approximation becomes. Moreover, recent research in fields such as mathematical epidemiology and evolutionary game theory has demonstrated that the qualitative features of a dynamical system operating on a network may be significantly altered when that underlying network is endowed with a dynamics of its own (Gross, Dommar D’Lima & Blasius, 2006; Traulsen, Santos & Pacheco, 2009). It is then reasonable to ask whether previous computational studies of language change may not have arrived at the wrong conclusions concerning the
neutrality hypothesis by making the wrong kinds of idealizing assumptions.

### III.3 Model

In this section, I will define the present model in intuitive terms, using as little mathematical notation as possible. A technical, mathematical definition will be found in Appendix III.8.1.

As outlined in Section III.2, I will be focussing on a situation in which some number $C$ of variants are in competition; often, the focus is on cases where $C$ is small, but in general this number could be arbitrarily large. The variants are assumed to be distributed across a language community in the following sense: each one of $N$ speakers will entertain exactly one of the $C$ variants at any time. The speakers themselves are distributed on a social network, and the connections a speaker has in this network will affect their process of variant acquisition or adoption. For simplicity, I assume that the network is symmetric, that connections are binary, and that the network is not multiplex. In other words, if speaker $i$ is connected to speaker $j$, then speaker $j$ will also be connected to speaker $i$; each pair of speakers is either connected or not connected (there is no notion of ‘weight’ of connection); and only one connection is allowed between any two speakers.

The set of speakers to whom a given speaker is connected I shall call the *neighbourhood* of that speaker; using basic graph-theoretical terminology, the cardinality of this set is called the speaker’s *degree* (in other words, the degree of a speaker is simply the number of speakers that speaker is connected to). When new speakers acquire their variant, their neighbourhoods are all important. In line with (III.1), I assume that the probability of acquiring variant $r$ ($r = 1, \ldots, C$) equals the relative frequency of that variant in the speaker’s neighbourhood, *modulo* random noise which is taken to model processes of innovation. This random noise is inserted into the model as an *innovation parameter* $\mu$ which ranges from 0 to 1 and gives the probability that the speaker picks a variant from among the $C$ possible variants uniformly at random. It is clear that this probability has to be rather low for language communities to display a degree of coherence in what variants they use — and this expectation is borne out by the simulations reported in Section III.5, below. The parameter is, however, an essential part of the model, for without it, variation could not arise in the first place.\(^3\)

To model social dynamics, the network of speakers is mixed by a graph-rewiring process over time, as follows: at each iteration step, one of the speakers is selected for removal uniformly at random and is replaced by a new speaker, whose social connec-

\(^3\)To see this, suppose that each individual in the community happens to use the same variant, so that the relative frequency of this variant in the community equals 1. If $\mu = 0$, then, in line with (III.1), any new speaker inserted into the network will acquire the said variant with probability 1, and change is impossible.
tions are set according to a socialization algorithm. The new speaker then acquires their variant as outlined above. The socialization algorithm is modelled on the intuition that human social networks normally contain both more and less connected individuals (cf. Barabási & Albert, 1999), and operates as follows. Let \( \sigma \) be a real number with \( 0 \leq \sigma \leq 1 \) and \( K \) an integer with \( 1 \leq K \leq N - 1 \). The new speaker is then given exactly \( K \) connections according to the following procedure: for each connection, the speakers in the network are first rank-ordered into a queue in terms of decreasing degree in such a way that the order of speakers having the same degree is random, but speakers with higher degree occur earlier in the queue than speakers with lower degree. Then, with probability \( \sigma \), a connection is made to the first speaker in this queue, and with the remaining probability mass \( 1 - \sigma \) a connection is made to a speaker chosen uniformly at random from the queue. Once the connection is established, this speaker is removed from the queue and the procedure iterated until the new speaker has received \( K \) connections. (Note that this does not imply that each speaker will, at any point of time, have exactly \( K \) connections: speakers’ degrees will change during their lifetimes thanks to the graph-rewiring process, as other speakers are removed from the network and replaced by new ones.) Different values of \( \sigma \), a preferentiality parameter, then give rise to networks with different amounts of clusterization around a central component, and different combinations of \( K \) and \( \sigma \) can be used to model different kinds of population structures: for high \( K \) and low \( \sigma \) (\( K \approx N - 1 \) and \( \sigma \approx 0 \)), the population is well-mixing, whereas for small \( K \) and high \( \sigma \), for instance, the network has a star-like appearance, with a clear partitioning into central and peripheral individuals (Figure III.1).

The model assumes invariant and categorical speakers — speakers who fix onto one of the competing variants at the point of acquisition and never change thereafter — this assumption being made in the interest of computational and mathematical tractability. Although some linguistic features are known to remain variable throughout a speaker’s lifetime (e.g. Harrington, 2006; Sankoff & Blondeau, 2007), there is, equally, evidence that for other features late-life change is unlikely or outright impossible. For instance, a number of studies have shown the existence of “hard features” — features which only young children manage to acquire and for which plasticity is lost as the speaker matures (these include, for example, phonological features with lexically irregular conditioning; see Kerswill, 1996 for a review of a number of relevant studies). Moreover, there is evidence that early categoricity predicts late-life stability: in a longitudinal panel study of a dozen phonetic variables undergoing change in a rural Finnish-speaking community, Nahkola and Saanilahti (2004) found significant late-life change only in speakers who had acquired features as variable ones. For categorical or near-categorical features, late-life change appears to be unlikely. These findings suggest that categorical, acquisition-driven change is one way
Figure III.1
Different values of the preferentiality parameter $\sigma$, combined with varying values of $K$, lead to networks with different amounts of clusterization. Note that the networks are not static but are rewired over time by the removal and addition of speakers; as a consequence, individual speakers may at times become disconnected from the rest of the network. For the networks in this figure, $N = 50$. 
in which languages change, and that the assumption of invariant speakers is therefore not unduly unrealistic — but future modelling work should, of course, investigate the consequences of relaxing the assumption.

III.4 Well-behaving

Evaluating the neutrality hypothesis (III.1) requires us to compare the output of the neutral model defined in Section III.3 against some sort of standard. More specifically, our interest is in two questions: (i) does the neutral mechanism give change in the first place? and (ii) if so, do the trajectories of change look anything like real life change trajectories? In this section, I introduce a way of operationalizing these two questions by way of a notion of the ‘well-behavedness’ of change. Following a preliminary, intuitive characterization, I will show how this notion can be formalized in mathematical, quantitative terms, so that the presence or absence of well-behaved change can be detected in simulation data generated by mathematical models of language change. The following section then proceeds to evaluate the current model vis-à-vis this quantitative operationalization in order to investigate the viability of neutral change.

Change, in a very general sense, can be said to occur when the distribution of linguistic variants over a language community changes. In most cases of interest to the historical linguist, such changes proceed from a state where one variant is dominant or nearly so in the community (has relative frequency of, or close to, 1) to another such state where another variant has become (near) dominant. Moreover, when diachronic data are consulted, such shifts between two dominance states are, as a general rule, found to be remarkably smooth, or well-behaved. Language change is not a random walk in the frequency space of possible linguistic variants; on the contrary, time series of changes can often be approximated to a good degree using a sigmoid, or S-shaped function (Bailey, 1973; Krock, 1989; Croft, 2000; Blythe & Croft, 2012). Although it remains unknown whether all changes follow an S-curve, and if so, whether the detailed ‘shape of S’ is the same in all changes (Niyogi & Berwick, 1997; Denison, 2003; Ghanbarnejad et al., 2014), propagation curves of linguistic variants tend to be reasonably monotone: they are unlikely to show oscillations by repeatedly inflecting up and down in the time domain, but rather proceed smoothly from one dominance state to another.

With these considerations in mind, I suggest that any model of language change should fulfil the following three criteria, which I here lay, in a programmatic manner, as characteristic properties of language diachrony under normal conditions:

(III.2) Dominance

For most of the time, the language community relaxes into a state in which
one variant is (nearly) dominant, so that most or even all speakers use that variant.

(III.3) *Shifting*

Upon introduction of an innovatory variant, this innovation may nonetheless begin to spread and eventually permeate (most of) the community; thus, the community may shift from a state of dominance by variant $r$ to a state of dominance by variant $r' \neq r$.

(III.4) *Monotonicity*

Such shifts proceed in a monotone manner with the frequency of the invading variant increasing, and the frequency of the receding variant decreasing, along smooth propagation curves.

A language community that fulfils all three criteria I shall call *well-behaved*:

(III.5) *Well-behavedness*

Language change is well-behaved if, and only if, it satisfies dominance, shifting and monotonicity.

To fix these ideas, let us inspect two simulation histories generated by the model defined in Section III.3 qualitatively. Figure III.2 shows a snapshot of an ill-behaved history in a three-variant system violating both dominance and monotonicity; this history was generated by setting $\sigma = 0$ and $\mu = 0.005$, the remaining parameters having the values $N = 100$, $K = 10$ and $C = 3$. Although the language community does display a kind of change, and hence exhibits shifting, this change is not monotone; there is too much zig-zagging movement in the propagation curves of the individual variants for this trajectory to be considered well-behaved. Moreover, the community does not settle on a dominant variant for any extended period of time. The history in Figure III.3, on the other hand, illustrates an entirely different situation, even though it was produced by the very same neutral mechanism. This history, generated with $\sigma = 1$, the other parameter values remaining the same, satisfies all three conditions: dominance, shifting and monotonicity.

Well-behaved neutral change is, then, possible. It remains to show that this is not merely a chance occurrence but a consistent behavioural characteristic of the model for certain ranges of model parameter values. For this, a quantitative analogue of each of the criteria (III.2)–(III.4) is needed, one that can be calculated over a large batch of simulation runs for a number of possible combinations of model parameter values in order to estimate, in a statistically robust manner, to what extent that criterion is satisfied by that combination of model parameters. A formal definition of such quantitative measures is given in Appendix III.8.2; here, I introduce the measures in prose.
**Figure III.2**
Portion of an ill-behaved history that violates dominance and monotonicity in a system of three variants. This trajectory was generated with parameter settings $N = 100$, $K = 10$, $C = 3$, $\mu = 0.005$ and $\sigma = 0$.

**Figure III.3**
Portion of a well-behaved history satisfying dominance, shifting and monotonicity. For this simulation, $N = 100$, $K = 10$, $C = 3$, $\mu = 0.005$ and $\sigma = 1$. 
To estimate to what extent a given simulation run satisfies dominance, I shall use a measure of dominance time $D_\delta$ that ranges from 0 (language community never dominant) to 1 (language community dominant all the time). The parameter $\delta$, a small real number, controls how strictly dominance is to be measured. More precisely, for given $\delta$, I say that a language community is $\delta$-dominant, if some one variant has a relative frequency equal to or greater than $1 - \delta$; the $D_\delta$ measure then gives the proportion of time (in relation to the length of the entire simulation run) the system spends in a state of $\delta$-dominance. Varying $\delta$ allows one to calculate dominance times to varying degrees and, inter alia, to subsume a notion of stable variation under the notion of dominance. For example, setting $\delta = 0.3$, we would call a community dominant if one of the competing variants had a relative frequency of at least 0.7 — allowing the rest of the frequency mass, a number bounded from above by 0.3, to be distributed among the remaining variants in any manner.

To measure shifting, I shall simply determine, for a given simulation run, the number of times the community shifts from a state of $\delta$-dominance by some variant $r$ to a state of $\delta$-dominance by another variant $r' \neq r$, for a predefined dominance level $\delta$. In what follows, I shall denote this shifting measure with $S_\delta$, where $\delta$ gives the desired dominance level.

Finally, to quantify monotonicity, I shall look at the autocorrelation properties of individual simulation histories. The idea is to place a short time window of some $\tau$ time steps at a specific time point $t_0$ of a history (so the time window extends from $t_0$ to $t_0 + \tau$), and then to count how many times the frequency of each competing variant both increases and decreases inside that window. To be more precise, let $m^+_r = m^+_r(t_0, \tau)$ be the number of times the frequency of variant $r$ increases within such a time window, and let $m^-_r = m^-_r(t_0, \tau)$ be the corresponding count of decreases. It is then easy to see that the product $m^+_r m^-_r$ equals 0 if, and only if, the frequency curve of variant $r$ is monotone in the window: the product is zero if and only if at least one of the multiplicands is zero, which is equivalent to monotonicity. For technical reasons explained in Appendix III.8.2, I next take the square root of this product and sum over competing variants, arriving at $\sum_{r=1}^{C} \sqrt{m^+_r m^-_r}$. Finally, an average is taken over different selections of time window start point $t_0$ (in other words, the time window is slid across the entire simulation history) and an inversion and a normalization performed so that the final measure, $M_\tau$, ranges from 0 in the non-monotone case to 1 in the perfectly monotone case, with $\tau$ controlling the resolution at which monotonicity is measured.

Some of the properties of this measure, proved in Appendix III.8.2, are worth mentioning here: it can be shown that $M_\tau = 0$ for any even\(^4\) integer $\tau$ if the frequency

\(^4\)The technical reason for the restriction here, without loss of generality, to even (rather than odd) integers is explained in Appendix III.8.2.
of at least one variant zig-zags persistently, increasing at every other iteration step and decreasing at every other; that $M_\tau = 1/3$ for large $\tau$ if the history is a random walk (so that for any variant and any iteration step, it is equally probable that the frequency of this variant increases, decreases or stays the same at that step); and that $M_\tau = 1$ for any $\tau$ if the frequency curve of each variant is monotone in all windows of size $\tau$. Figure III.4 illustrates a few histories with their corresponding $M_\tau$ scores for different window sizes $\tau$, to give an idea of what amount of smoothness to expect for individual trajectories, given a value of $M_\tau$.

We then have three measures, $D_\delta$, $S_\delta$ and $M_\tau$, to measure dominance, shifting and monotonicity in individual simulation runs. Before proceeding to an application of these measures to the neutral model, it is perhaps in order to clarify the purpose of

---

Figu**re III.4**

Four histories with their corresponding monotonicity scores $M_\tau$ for two different window sizes $\tau = 10, 50$. Note that for a random walk the expected value of $M_\tau$ is $1/3$ (see text), and that $M_\tau$ approaches 1 as the history becomes more and more monotone.
defining the measures in the first place. Importantly, the above operationalization of the well-behavedness of linguistic change should not be taken to imply that no language community is ever ill-behaved. After all, empirically demonstrated cases exist of both stable variation (violation of strict dominance; cf. Wallenberg, 2013), and of zig-zagging or failing changes (violation of monotonicity; Coussé & De Sutter, 2012; Postma, 2010). The purpose of introducing dominance, shifting and monotonicity as conditions of well-behaved change is to fashion a litmus test for the neutral model: insofar as the model satisfies the three conditions, it can be taken seriously as a mathematical model of language change. With the above operationalization, dominance and monotonicity are actually continuous quantities, ranging from 0 to 1, and thus admit of a notion of degree. Requiring the neutral model to satisfy well-behavedness to a large degree is the strictest possible analytical test to which the model can be subjected in this regard, and if real life language communities are found to be less well-behaved than that, then the case for neutral change is correspondingly strengthened.

III.5 Simulations

To investigate to what extent the model defined in Section III.3 satisfies dominance, shifting and monotonicity, or criteria (III.2)–(III.4), a number of computer simulations of the model were run using a range of model parameter settings. For each combination of model parameters investigated, 50 simulations were run to arrive at the averages reported below, and in each simulation, a social network of \( N = 100 \) speakers was assumed. The simulations were run in parallel on a high-throughput computing cluster, with the pseudorandom number generator seeded using environmental noise to ensure statistical independence of simulation runs. Before starting each actual linguistic simulation, the social network algorithm was iterated for \( 100N = 10^4 \) iterations so that the degree distribution of the network settled; each actual linguistic simulation (apart from the simulations reported in Section III.5.5; see below) lasted for \( 5 \times 10^4 \) iterations and started from a state in which one of the competing variants had strict dominance (relative frequency 1, or in other words, \( \delta \)-dominance with \( \delta = 0 \)).

III.5.1 Main result

Figure III.5 gives shifting scores for a system of \( C = 3 \) competing variants, for various values of preferentiality \( \sigma \) and innovation rate \( \mu \), and for two different values of attachment set size \( K \), using a dominance threshold of \( \delta = 0.1 \). The results indicate that each of these model parameters has their effect on shifting ability: keeping \( K \) and \( \mu \) constant, the effect of increasing \( \sigma \) from 0 towards 1 is a monotonic increase in shifting; for \( \mu \), on the other hand, an optimal value exists that supports shifting ability
Shifting $S_{0.1}$ in a system of $C = 3$ competing variants, for various values of preferentiality $\sigma$ and innovation rate $\mu$ and for attachment set sizes $K = 10, 30$, calculated using a dominance threshold of $\delta = 0.1$; averages over 50 simulation runs. Neutral change is supported the best by tightly clusterized communities (small $K$, high $\sigma$).

Increasing $K$, in turn, has the effect of flattening the shifting measure with respect to $\sigma$: as the population becomes more and more well-mixing, preferential connectivity naturally ceases to have an effect, and shifting becomes rarer. In sum, change is the more probable the smaller $K$ is and the larger $\sigma$ is — the more clusterized the community is around a central component (cf. Figure III.1) — provided that innovations ($\mu$) occur at a suitable rate.

Figure III.6 plots dominance times and monotonicity, the former calculated assuming $\delta = 0.1$, the latter computed using a window size of $\tau = 10$; variation in $\tau$ has only a minor effect on the monotonicity measure (not reported; but cf. Figure III.4). The main finding with respect to dominance is that increasing the innovation rate $\mu$ results in a sharp drop in this measure, with the value of $\sigma$ attenuating the effect a bit so that communities with higher preferentiality $\sigma$ remain dominant for larger $\mu$ than communities with lower preferentiality $\sigma$. A similar, but much less drastic drop as a response to variation in $\mu$ is observed for monotonicity.

To gauge what combinations of model parameter values support well-behaved neutral change the best overall, we can consider the product of the three measures, namely $S_{0.1}D_{0.1}M_{10}$. Figure III.7 gives this product, and we find that communities with low $K$, high $\sigma$ and intermediate $\mu$ are the most likely to exhibit well-behaved neutral change.
Figure III.6
Dominance $D_{0.1}$ (bottom surface) and monotonicity $M_{10}$ (top surface) in a system of $C = 3$ competing variants; averages over 50 simulation runs. Both dominance and monotonicity drop as the innovation rate $\mu$ is increased, with large preferentialities $\sigma$ attenuating this effect.

Figure III.7
The combined well-behavedness measure $S_{0.1}D_{0.1}M_{10}$ for a system of $C = 3$ variants; averages over 50 simulation runs. Overall, well-behaved neutral change is supported best by tightly clusterized language communities and by innovation rates $\mu$ which are low but not too low.
III.5.2 Effect of number of variants

In the above simulations, the number of competing variants was fixed at $C = 3$. This is a rather small number, and it is reasonable to ask whether the behaviour of the system would not change if more variants were available to speakers. To investigate this, another batch of simulations was run using identical model parameter settings except that the number of competing variants was now fixed at $C = 30$.

Increasing the number of variants turns out to have a nontrivial effect on the well-behavedness of a neutral system. Figure III.8 gives the difference between the shifting scores received by the new batch of simulations and those received by the simulations of Section III.5.1. Here, we find that for certain combinations of preferentiality $\sigma$ and innovation rate $\mu$, the community with $C = 30$ shifts more than the community with $C = 3$, whereas for other model parameter combinations the reverse is true: increasing the number of competing variants improves shifting for large $\sigma$, but only if $\mu$ has a modest value.

Figure III.9 reports, similarly, the difference in dominance and monotonicity scores received by the two batches of simulation runs. Increasing $C$ leads to slightly lower dominance and monotonicity overall, an effect which is the strongest for an intermediate range of values of $\mu$.

Thus, overall, allowing speakers a larger space of grammatical options can have the effect of increasing the probability of change, but only at the cost of some reduction in how well-behaved that change is in terms of dominance and monotonicity.

III.5.3 Effect of dominance threshold

The dominance threshold $\delta = 0.1$ used above is rather strict: it demands a variant to have a relative frequency of more than 0.9 in order for that variant to be considered dominant. Lowering the dominance threshold is expected to increase both shifting and dominance, and this expectation is confirmed by calculations of $S_\delta$ and $D_\delta$ using a less stringent dominance threshold of $\delta = 0.3$ (Figures III.10 and III.11). A nontrivial finding is that the preferentiality parameter $\sigma$ has a strong effect on dominance for less extreme dominance thresholds: for $\delta = 0.3$ and $K = 10$, for instance, $\sigma = 0$ implies practically no dominance if $\mu$ is on the order of 0.1, while for $\sigma = 1$ dominance times remain in the $> 0.5$ region for such innovation rates. Thus, the model predicts that when change is neutral, stable variation is supported best by language communities which are tightly clusterized.

III.5.4 Effect of rewiring dynamics

We can also ask whether it is just the topology of the social network that licenses well-behaved neutral change for certain ranges of parameter values, or whether the
FIGURE III.8
Difference \((B - A)\) in shifting \(S_{0.1}\) between \((A)\) the 3-variant system of Section III.5.1 (Figure III.5) and \((B)\) another system with \(C = 30\) competing variants \(ceteris paribus\). For large \(\sigma\), the 30-variant community shifts more than the 3-variant system if \(\mu\) has a modest value; for larger \(\mu\), the reverse obtains.

FIGURE III.9
Difference \((B - A)\) in dominance \(D_{0.1}\) and monotonicity \(M_{10}\) between \((A)\) the 3-variant system (Figure III.6) and \((B)\) a system with \(C = 30\) competing variants \(ceteris paribus\). Increasing the number of competing variants leads to slightly lower dominance and monotonicity overall, the effect being the most pronounced for intermediate values of innovation rate \(\mu\).
Figure III.10
Shifting $S_{0.3}$ for a system with model parameter values identical to those of the system of Figure III.5, calculated using a less stringent dominance threshold of $\delta = 0.3$. Adjusting the threshold in this way leads to more shifting events across all of the model parameter space.

Figure III.11
Dominance $D_{0.3}$ (bottom surface) and monotonicity $M_{10}$ (top surface) for a system with model parameter values identical to those of the system of Section III.5.1 (cf. Figure III.6), calculated using a less stringent dominance threshold of $\delta = 0.3$. For this laxer dominance threshold, network preferentiality $\sigma$ has a strong effect on dominance: $\sigma = 0$ implies essentially no dominance if innovations occur at a rate of about $\mu = 0.1$, whereas for more tightly clusterized communities ($\sigma \approx 1$) dominance times remain in the $>0.5$ region for such innovation rates. This means that stable variation — $\delta$-dominance with a lax dominance threshold such as $\delta = 0.3$ — is supported best by language communities which are tightly clusterized, when change is neutral.
social dynamics induced by the removal and addition of speakers plays a role. To 
investigate this, another batch of simulations was run with parameter settings identical 
to those of the first ensemble (Section III.5.1), but with the rewiring dynamics turned 
off. (In other words, the network was first rewired for $10^4$ iteration steps, as above, to 
give it the topology induced by the particular choice of $K$ and $\sigma$ in each case, so that 
the network had the same topology as in the rewired case. However, the rewiring 
dynamics was turned off at this point, so that during the actual linguistic simulation 
no rewirings took place and the network was thus static.) Figure III.12 gives the dif-
ference in the overall well-behavedness score — the product $S_{0.1}D_{0.1}M_{10}$ — between 
these two ensembles. For less clusterized networks (large $K$ or small $\sigma$) the differ-
ence is negligible, as would be expected. For strongly clusterized networks, however, 
an entirely different picture emerges when the rewiring dynamics is removed: the 
community without rewiring displays consistently lower well-behavedness scores.

This finding may appear puzzling at first sight, but is actually connected in a nat-
ural way to one of the central idealizing assumptions of the model, that speakers 
stabilize and do not change after initial acquisition.\footnote{I am much indebted to an anonymous reviewer for raising this point.} With this assumption, a tightly 
clusterized network gives rise to a central hub consisting of speakers who are con-
ected to most other speakers in the network, and whose role in the competition of 
linguistic variants depends on whether the network is rewired or not. With rewiring, 
in a highly clusterized network new speakers always receive many connections to 
these central speakers who, thanks to the critical period assumption, do not them-
selves change after maturation. Central speakers therefore become the vehicle of 
change, conserving their own variant while distributing it to speakers newly joined 
to the network. If network rewiring is suppressed, however, the central speakers of a 
clusterized network effectively sample from the majority of the population and thus 
get a very representative picture of the frequency of variants that exist in the net-
work. The central speakers, rather than advancing a change, serve to hinder changes 
in this setting: as the frequency of an innovation is necessarily low, any innovation 
event is likely to be quelled by speakers in the central hub, for when these speakers 
do update their variant, they are unlikely to adopt the innovatory one.

This observation, then, reveals that interactions between features of within-
speaker dynamics (here, the critical period assumption) and between-speakers dy-
namics (here, the degree of clusterization of the social network) may be important 
ough to affect causation in language change, by adjusting the probability of an 
innovation surviving and propagating through a language community.
Figure III.12

Difference \( (B - A) \) in the overall measure of well-behavedness, \( S_{0.1}D_{0.1}M_{10} \), between (A) the system of Section III.5.1 (Figure III.7) and (B) another one with the rewiring dynamics turned off, model parameter settings remaining the same. When the community is tightly clustered, suppression of rewiring suppresses well-behaved neutral change. (Note that in this figure, in contrast to previous ones, both the \( \sigma \) axis and the \( \mu \) axis have been inverted to better exhibit the dip in the high-\( \sigma \) regime.)

III.5.5 Rate of change

A comparison of Figures III.2 and III.3 suggests, impressionistically, that the speed with which an innovation spreads through a community can depend quite drastically on the structure of the community. To investigate this dependence systematically, a final batch of simulations was run (200 simulations for each combination of model parameters), this time with a number of innovative speakers inserted ‘by hand’ into an otherwise homogeneous community at the start of each simulation. Out of all simulation histories so generated, the ones where change from this initial state to a state of \( \delta \)-dominance with \( \delta = 0 \) by the innovative variant occurred were then selected for further investigation by recording the number of iteration steps it took the community to traverse from the former state to the latter. Figure III.13 gives this \textit{time-to-dominance} for various combinations of \( K \) and \( \sigma \) for a network of size \( N = 100 \), with 10 innovators. We find that the presence of a central, well-connected hub of speakers in the network has the effect of speeding up change; for small \( K \), the decrease in time-to-dominance is as much as tenfold when moving from \( \sigma = 0 \) (no clusterization) to \( \sigma = 1 \) (maximal clusterization).
Figure III.13
A log-lin plot of time-to-dominance for various values of attachment set size $K$ and preferentiality $\sigma$, quantified as the number of iterations it takes for an innovatory variant to permeate the community from an initial state where a number $m_0$ of speakers entertain the innovatory variant. Here, for each pair of $K$ and $\sigma$, network size was fixed at $N = 100$ and number of innovators at $m_0 = 10$, and the latter were picked uniformly at random from among all speakers. $C = 3$ competing variants were assumed throughout with innovation rate $\mu = 0.01$. Time-to-dominance is found to be an exponential function of $\sigma$, so that increasing $\sigma$ leads to a speed-up in change for small $K$. 
III.6 Discussion

The above simulations show that the form a linguistic trajectory assumes — whether well-behaved or not — can depend crucially on the social structure and social dynamics of the language community, if none of the competing linguistic variants are biased over others. The results demonstrate well-behaved neutral change for certain types of preferentially attached societies and show that such change is much less likely in societies lacking preferential connections. Whether real language communities exist with these parameter settings is an empirical matter; the above considerations imply that if such communities exist, well-behaved neutral change is a characteristic property of them.

It is worthwhile to point out explicitly how these results differ from earlier ones, particularly those obtained by Fagyal et al. (2010). While both studies investigate the role of network effects in language change, the model here studied is neutral in the sense that variant acquisition is determined by frequency and does not depend on sociolinguistic considerations. In the model of Fagyal et al. (2010), by contrast, speakers give more weight to speakers who have high degree centrality, so a linguistic variant becomes the fitter the more it is adopted by such central speakers, and their model is thus classified as weighted interactor selection in the Blythe–Croft taxonomy (see p. 169, above). This difference has nontrivial sociolinguistic implications. With a biased model, one assumes that speakers are able to evaluate the centrality or prestige, or both, of each speaker to whom they are connected, and that they in fact pay attention to such evaluations. In a neutral model, the only causative social factor in language change is the way speakers are (happen to be) connected, and one need not (or does not) assume that speakers have access to or make use of prestige evaluations.

An important feature of the framework adopted in this paper is not exhibited by previous mathematical models of language change: it models evolution on and of a network simultaneously. Infinite-population models have considered non-overlapping, well-mixed generations (e.g. Niyogi & Berwick, 1997; Yang, 2000; Komarova, Niyogi & Nowak, 2001; Mitchener, 2006; Niyogi & Berwick, 2009), and in most if not all finite-population models (including those of Ke et al., 2008, Fagyal et al., 2010, and Blythe & Croft, 2012) the social network is not allowed to evolve as the linguistic variants compete on that network. In the present model, the generations of speakers are overlapping and the network is updated in accordance with the socialization algorithm in use, at each iteration. The simulation results demonstrate that this interplay of the social network rewiring dynamics and the linguistic variant dynamics has an effect on the probability of a language community shifting, as well as on the well-behavedness of any such shifts (Section III.5.4); importantly, this refutes previous claims (based on static population modelling) that neutral change cannot be
well-behaved (Fagyal et al., 2010; Blythe & Croft, 2012).

An obvious criticism of the model is that there is, as yet, no independent evidence for the sort of social network structure the model presupposes. Although the role and importance of social network effects in language change have been noted before (L. Milroy, 1980; J. Milroy & Milroy, 1985), we still lack a deep understanding of the basic properties of human social networks, both topological and dynamic. Two immediate goals can be discerned in this regard. Firstly, empirical studies are needed to establish what the connectivity patterns of actual language communities are — how exactly are they clusterized, what are their typical degree distributions, are they possibly multiplex, do inter-speaker links have weights on them or is a binary characterization sufficient, and so on. Secondly, these patterns have to be captured in mathematical models that are considerably more complex than the algorithms currently in use in the complex systems and network science literature (for a review of the state of the art and some suggestions for future directions, see Kivelä et al., 2014).

That said, it is possible to interpret the present model, in what is perhaps a promising and productive way, in the light of earlier proposals concerning social factors in linguistic change. We have seen the preferentiality parameter $\sigma$ to control the clusterization of the social network, and it is possible to take this as an operationalization of the degree to which a language community is closeknit, in the terminology of J. Milroy and Milroy (1985): networks with large (close to 1) $\sigma$ will then correspond to communities which are closeknit. Now, we may well imagine several such communities to be connected along inter-community links, composing thereby a network of networks, so that many links are found within the subcommunities but between the subcommunities a much smaller number of links exist. The intra-community links can then be thought to correspond to the Milroys’ strong ties, the inter-community links corresponding to weak ties. In the present model networks with large $\sigma$ act as both strong conservers and rapid distributors of linguistic variants: for instance, it can be shown that in the limiting case of $\sigma = 1$, the probability of a speaker in the central cluster of the highly clusterized network distributing their variant to at least one other speaker during the former’s lifetime is given by

$$q = 1 - \left(1 - \frac{1}{K} + \mu \left(\frac{1}{K} - \frac{1}{C}\right)\right)^N$$  \hspace{1cm} (III.6)

as long as $\mu < 1/K$. Importantly, this number is bounded from below by $1 - 1/e \approx 0.63$, irrespective of the values of $N$ (network size) and $K$ (attachment set size), and tends to 1 as $K$ tends to 1 and $\mu$ tends to 0. Thus, it is always more probable for variants flowing from the centre of the network to be replicated than not to be replicated, and the probability is the greater the more clusterized the network (Figure III.14). This explains both conservatism and progressivism: on the one hand, if no innovatory
The probability, $q$, of a central speaker distributing their variant to at least one other speaker before the former is removed from the network by the network-rewiring algorithm, for innovation rate $\mu = 0.01$ and number of competing variants $C = 30$ (eqn. III.6). Note that $q \to 1$ as $K \to 1$ and $\mu \to 0$, and that $q > 1 - 1/e \approx 0.63$ for any choice of $K$ and $N$ satisfying $\mu < 1/K$.

variants happen to be introduced into the centre of the strongly clusterized network, the centre acts as a strong suppressor against innovations which occur in intermediately (but not strongly) connected speakers, and on the other hand, if an innovatory variant happens to invade the centre of the network, it is almost certainly distributed to at least one other speaker before the bearer of that innovatory variant is removed from the network by the network-rewiring process.

This analogy between the present model and the Milroys’ framework can be pressed further. J. Milroy and Milroy (1985) draw, following Rogers and Shoemaker (1971), a distinction between the innovators and the early adopters of a change. In the present model, all innovation events occur in speakers whose degree is $K$; these speakers, who correspond to the Milroys’ innovators, do not belong to the central cluster of the social network. Clearly, language change only happens if, following this initial actuation of an innovatory variant, the variant is subsequently propagated through the layers of the social network and becomes, eventually, dominant. In the present model, this happens typically if the social network comes to be so rewired that the innovating speaker is ‘promoted’ to the centre of the clusterized, closeknit community, i.e. if their degree increases due to rewirings of other speakers; this occurs with a finite probability which increases as $\sigma$ is increased. Once in the centre, the probability of this innovating speaker influencing the variant adoption processes of new speakers is significantly increased; these speakers adopting the new variant then correspond to the Milroys’ early adopters, and propagation of the innovatory variant is successful if the number of early adopters is large enough.

Yet the present model does not serve merely as a computational implementation
or (partial) corroboration of the Milroys’ framework; it adds a positive contribution thanks to the neutrality assumption. As I have noted above (p. 168), J. Milroy and Milroy (1985) assume that innovatory variants must have a non-zero prestige value attached to them, if they are to propagate successfully through a language community. This is prima facie puzzling, for it raises the further question of how (and why) language communities should be able to agree on the social valuation of invading variants:

The puzzle is of course how young people living in the closed communities of Ballymacarrett, Clonard and Hammer, whose contact with others outside their areas has been only of a very tenuous kind, have come to reach cross-community consensus on the social value to be assigned to the two variants of the (pull) variable (J. Milroy & Milroy, 1985, 374).

The above simulation results suggest that such cross-community consensus may, in fact, be unnecessary. Prestige need not be attached either to linguistic variants or to individual speakers; in order to have well-behaved neutral change, it suffices to have a non-uniform, but dynamic population structure containing hubs of speakers. Prestige reduces to degree centrality: the influence of individual speakers lies in the number of connections they have in their language community, not in a social evaluation assigned on top of that number of connections.

III.7 Conclusion

In this paper, I have investigated the possibility that language change is, in some cases, neutral and not motivated by functional, social, articulatory or other biases. I have defined a simple model of variant competition in a finite network of speakers in which variant adoption is neutral, and have tested this model against three criteria that together constitute well-behavedness of change, viz. dominance, shifting ability and monotonicity. Results from computer simulations show that if the network of speakers is suitably clusterized, so that it has a central component with some very well connected speakers, well-behaved neutral change is observed in this model. I have proposed a way of interpreting this finding in the framework of J. Milroy and Milroy (1985) and have suggested that a neutral mechanism, such as the one here considered, calls for a re-evaluation of the role of prestige as a causal factor in at least some cases of change. I have stressed the importance of approaching language diachrony from the viewpoint of mathematical models and the need to increase the

---

6Assuming again, as the model does, that speakers are categorical and invariant after a critical period. The results in Section III.5.4 suggest that the interaction of this assumption with the (language-external) social dynamics of the language community is nontrivial; the consequences of relaxing the assumption need to be systematically investigated in future research.
complexity and realism of these models, and hope that results like those reported in this paper can go some way towards justifying this angle of attack. Subsequent work on the neutrality hypothesis should both incorporate more realistic models of social dynamics and relax some of the simplifying assumptions made in this paper, to see if well-behaved neutral change continues to be observed under such modifications.

The discrepancy seen between the results here reported and those obtained by Ke et al. (2008), Fagyal et al. (2010) and Blythe and Croft (2012) is explained by the different assumptions that go into the definition of each of these four models of language change. In the latter three models, the social network structure underlying the linguistic variant dynamics is not allowed to evolve during a simulation run, so that speakers’ neighbourhoods remain fixed. In the present model, speakers are added to and removed from the social network in accordance with the network-rewiring algorithm described in Section III.3 and Appendix III.8.1, and the neighbourhood of a speaker may change during their lifetime if speakers in that neighbourhood are removed, or if new speakers are added thereto. The simulation results show that this interplay between the network-rewiring dynamics and the linguistic variant dynamics, together with the assumption of having speakers who latch onto one or another variant early on and do not change thereafter, is instrumental in supporting well-behaved neutral change in communities which are tightly clustered (Figure III.12).

The model here studied makes a number of predictions which are, in principle, open to investigation and empirical testing. Firstly, the above simulations predict that increasing the number of linguistic variants available to speakers makes neutral change more likely if the innovation rate has a moderate value — but only at the expense of a slight drop in the well-behavedness of change, when quantified using the notions of dominance and monotonicity (Section III.5.2). Secondly, the simulations predict that stable variation should be more likely in clustered communities than in well-mixing ones (Section III.5.3). Finally, change in a clustered community is much faster than in a well-mixing one of corresponding size (Section III.5.5).

The possibility of well-behaved neutral change has implications for diachronic work that seeks to establish non-neutral motivations for language change. While the possibility of neutral change does not imply its probability and does not, per se, undermine non-neutral theory in instances where sound reasons exist for believing in the presence of non-neutral motivations, the results here reported do warn against appealing to non-neutral explanations when such reasons are lacking; ‘these possibilities […] need to be considered before any claim for “function” can be made for either variation or change’ (Lass, 1997, 354). Any particular case of change may in fact be a constellation of neutral and non-neutral factors, and one important goal for research in language diachrony must be to tease apart the relative contributions of these two modes of change.
III.8 Appendix

III.8.1 Formal definition of the model

Consider a language community of \( N \) speakers distributed on an undirected graph \((V, E_t)\), where \( V = \{1, \ldots, N\} \) is the set of speakers (vertices) and \( E_t \) is an irreflexive, symmetric relation giving the speaker adjacencies (edges), indexed for time \( t \). Denote by \( E_t(i) = \{ j \in V : (i, j) \in E_t \} \) the neighbourhood of speaker \( i \) and by \( \text{deg}_t(i) = |E_t(i)| \) the degree of speaker \( i \) at time \( t \). Let \( C = \{1, \ldots, C\} \) be the set of linguistic variants, and for each time \( t \) define a function \( v_t: V \rightarrow C \) which gives the variant of speaker \( i \) at time \( t \). Then define an indicator function

\[
\chi_t(i, r) = \begin{cases} 
1 & \text{if } v_t(i) = r \\
0 & \text{otherwise}
\end{cases}, \tag{III.7}
\]

and let the graph \((V, E_t)\) be rewired in discrete time by the following algorithm.

Algorithm III.1.

Define a stochastic process to shuffle the graph \((V, E_t)\) as follows:

1. Let \( 0 \leq \mu, \sigma \leq 1 \) and \( K \) be a positive integer with \( K \leq N - 1 \).

2. At time \( 0 \), the relation \( E_0 \) is initialized randomly; say, every speaker has a probability of \( 1/2 \) to be connected to any other speaker.

3. Choosing a simulation length \( n \), iterate from \( t = 1 \) to \( t = n \):

   (a) Select a speaker \( i \) from \( V \) uniformly at random.

   (b) Remove all of \( i \)'s connections.

   (c) For each \( d = 0, \ldots, N - 1 \), take each speaker other than \( i \) having a degree of exactly \( d \); put these speakers into a set \( Q_d \); and shuffle \( Q_d \) to make an ordered tuple \( \hat{Q}_d \).

   (d) Define an ordered set \( Q \), the queue, as follows, where \( \circ \) denotes concatenation:

\[
Q = \hat{Q}_{N-1} \circ \hat{Q}_{N-2} \circ \cdots \circ \hat{Q}_0. \tag{III.8}
\]

   (e) Give the speaker \( i \) a connection as follows:

      i. With probability \( \sigma \), connect \( i \) to the first speaker in \( Q \), and delete this speaker from \( Q \)

      ii. With probability \( 1 - \sigma \), connect \( i \) to a speaker selected uniformly at random from \( Q \), and delete this speaker from \( Q \)

   (f) Repeat the previous step until \( i \) has received exactly \( K \) connections.
(g) Set the variant of speaker $i$ as follows: for each possible variant $r$, the probability of setting $\nu_i(t) = r$ is to equal

$$\frac{\mu}{C} + \frac{1 - \mu}{K} \sum_{j \in E_i} \chi_j(t, r).$$

(III.9)

III.8.2 Quantifying well-behavedness

In quantifying well-behavedness of change, our interest is in how the frequencies of the $C$ competing variants unfold in time. For this, let $x_r(t)$ denote the relative frequency of the $r$th variant at time $t$, and let $x(t) = (x_1(t), \ldots, x_C(t))$ be the frequency-state of the system. A sequence of frequency states $x(1), \ldots, x(n)$ I shall call a history or (frequency) trajectory.

**Dominance.** Let $0 \leq \delta \leq 1$. I shall call a frequency-state $x(t) = (x_1(t), \ldots, x_C(t))$ $\delta$-dominant if $x_r(t) \geq 1 - \delta$ for some $r$. Dominance times for a history $x(1), \ldots, x(n)$ are then obtained by the time-averaged measure

$$D_\delta = \frac{1}{n} \sum_{t=1}^{n} \Delta_\delta(t),$$

(III.10)

where

$$\Delta_\delta(t) = \begin{cases} 1 & \text{if } x(t) \text{ is } \delta\text{-dominant} \\ 0 & \text{otherwise} \end{cases}. \quad (III.11)$$

**Shifting.** To measure shifting ability, I shall record, for a given simulation run, the number of shifts from $\delta$-dominance by variant $r$ to $\delta$-dominance by another variant $r' \neq r$, for a predefined dominance threshold $\delta$. More formally, for a history $x(1), \ldots, x(n)$, the shifting measure, $S_\delta$, is defined as the number of time points $t \in \{1, \ldots, n\}$ such that $x_r(t) \geq 1 - \delta$ for some $t$, some $r$, and $x_{r'}(t') \geq 1 - \delta$ for some $t' < t$, some $r' \neq r$.

**Monotonicity.** A sequence $x(1), \ldots, x(n)$ is monotone if $t < t'$ implies either $x(t) \leq x(t')$ or $x(t) \geq x(t')$. A history $x(1), \ldots, x(n)$ will be called monotone if each variant frequency sequence $x_r(1), \ldots, x_r(n)$ is monotone.

Generally, it is possible to estimate the monotonicity of a history by the following measure, for integer $\tau > 0$ and real $\alpha > 0$:

$$W_{\tau, \alpha} = \frac{1}{n - \tau} \sum_{t_0=1}^{n-\tau} \sum_{r=1}^{C} \left( \sum_{t=t_0}^{t_0+\tau-1} s^+_r(t) \right)^\alpha \left( \sum_{t=t_0}^{t_0+\tau-1} s^-_r(t) \right)^\alpha,$$

(III.12)
where
\[ s^+_r(t) = \begin{cases} 1 & \text{if } x_r(t) < x_r(t+1) \\ 0 & \text{if } x_r(t) \geq x_r(t+1) \end{cases}, \]  
(III.13)
and
\[ s^-_r(t) = \begin{cases} 1 & \text{if } x_r(t) > x_r(t+1) \\ 0 & \text{if } x_r(t) \leq x_r(t+1) \end{cases}. \]  
(III.14)

(For an intuitive characterization of this equation in terms of the quantities \( m^+_r(t_0, \tau) \) and \( m^-_r(t_0, \tau) \), see p. 176, above.) This has the following properties under our model:

**Proposition III.1.**

For a simulation operating under Algorithm III.1 (Appendix III.8.1),

(i) \( 0 \leq W_{r,\alpha} \leq \tau^{2\alpha}/2^{2\alpha-1} \) for all \( \tau, \alpha; \)

(ii) \( W_{r,\alpha} = 0 \) for any \( \tau \) and \( \alpha \) if and only if the history is monotone;

(iii) \( W_{r,\alpha} = 0 \) for some (sufficiently small) \( \tau \) and all \( \alpha \) if and only if the history is piecewise monotone;

(iv) \( W_{r,\alpha} = \tau^{2\alpha}/2^{2\alpha-1} \) for \( \tau \) even, and \( W_{r,\alpha} = (\tau^2 - 1)^\alpha/2^{2\alpha-1} \) for \( \tau \) odd, if and only if the history zig-zags persistently;

(v) For large \( \tau \), the expected value of \( W_{r,\alpha} \) is \((2/3)^{2\alpha} \tau^{2\alpha}/C^{2\alpha-1}\) if the history is a random walk.

**Proof.** Let \( m^+(r, t_0, \tau) = m^+_r(t_0, \tau) = \sum_{t=t_0}^{t_0+\tau-1} s^+_r(t) \) and \( m^-(r, t_0, \tau) = m^-_r(t_0, \tau) = \sum_{t=t_0}^{t_0+\tau-1} s^-_r(t) \).

(i) That \( W_{r,\alpha} \geq 0 \) is plain. The maximum is achieved when two variants \( r_1 \) and \( r_2 \) alternate in upward and downward inflections. For \( \tau \) even, this means that \( m^+(r_i, t_0, \tau) = m^-(r_i, t_0, \tau) = \tau/2 \) for \( i = 1, 2 \), for all \( t \), and \( m^+(r_i, t_0, \tau) = m^-(r_i, t_0, \tau) = 0 \) for \( i \neq 1, 2 \), and therefore
\[
W_{r,\alpha} = \frac{1}{n-\tau} \sum_{t=t_0+\tau}^{t_0+n-1} 2^{\alpha} \left( \frac{\tau}{2} \right)^2 = \frac{\tau^{2\alpha}}{2^{2\alpha-1}}. \]  
(III.15)

If \( \tau \) is odd, then
\[
\begin{align*}
  m^+(r_i, t_0, \tau) &= \frac{\tau-1}{2} \\
  m^-(r_i, t_0, \tau) &= \frac{\tau-1}{2} + 1 \\
  m^+(r_j, t_0, \tau) &= \frac{\tau-1}{2} + 1 \\
  m^-(r_j, t_0, \tau) &= \frac{\tau-1}{2}
\end{align*} \]  
(III.16)
either for \( i = 1, j = 2 \) or for \( i = 2, j = 1 \). In either case,
\[
W_{r,\alpha} = \frac{1}{n-\tau} \sum_{t=t_0+\tau}^{t_0+n-1} 2^{\alpha} \left( \frac{\tau-1}{2} \right)^2 \left( \frac{\tau-1}{2} + 1 \right)^2 = \frac{(\tau^2-1)^\alpha}{2^{2\alpha-1}} < \frac{\tau^{2\alpha}}{2^{2\alpha-1}}. \]  
(III.17)
(ii) If a history is monotone, then either 
\[ m^+(r, t_0, \tau) = 0 \] or 
\[ m^-(r, t_0, \tau) = 0 \] or both for each variant \( r \), for each time \( t_0 \). Hence \( W_{r, t_0, \tau} = 0 \). Conversely, if \( W_{r, t_0, \tau} = 0 \), then 
\[ m^+(r, t_0, \tau) = 0 \] or 
\[ m^-(r, t_0, \tau) = 0 \] for each \( r \), for each \( t_0 \), which implies that the history is monotone.

(iii) Suppose that a history is monotone when viewed through a window of size \( \tau_0 \). Then with the above reasoning 
\[ m^+(r, t_0, \tau) = 0 \] or 
\[ m^-(r, t_0, \tau) = 0 \] in such windows, and consequently we have \( W_{r, t_0, \tau} = 0 \) for the average. Conversely, if \( W_{r, t_0, \tau} = 0 \), the history is piecewise monotone in windows of size at most \( \tau_0 \).

(iv) This was shown in (i).

(v) Consider an arbitrary variant \( r \) at any time \( t \). Then \( x_r(t) \) can inflect upwards in two ways: either \( r \) is selected to change so that \( x_r \) increases while some \( r' \neq r \) decreases, or some \( r' \neq r \) is selected so that \( x_{r'} \) decreases and \( x_r \) increases. Since \( x(t) \) is assumed to be a random walk, the probability for \( x_r(t) \) to increase is then given by

\[
p = \frac{1}{C} \cdot \frac{1}{3} + \frac{C - 1}{C} \cdot \frac{1}{3} = \frac{2}{3C}
\]  
(III.18)

(probability of picking \( r \) times probability of \( x_r \) increasing (rather than decreasing or not changing), plus probability of picking \( r' \) times probability of \( x_{r'} \) decreasing so that \( x_r \) increases). By symmetry, the probability for \( r \) to inflect downward is the same. In a window of length \( \tau \), we would then expect \( pr \) upward and \( pr \) downward inflections for variant \( r \), if \( \tau \) is sufficiently large. This gives us

\[
F := \sum_{r=1}^{C} \left( \left( \frac{2}{3C} \right)^2 \right) = C \left( \frac{2}{3C} \right)^2 \left( \tau^{2a} \right) \frac{2a}{(2a-1)}
\]  
(III.19)

hence

\[
W_{r, t_0, \tau} = \frac{1}{n - \tau} \sum_{t_0=1}^{n-\tau} F = \left( \frac{2}{3} \right)^{2a} \frac{\tau^{2a}}{(2a-1)},
\]  
(III.20)

as wished.

Now let

\[
M_r = 1 - \frac{W_{r,1/2}}{\tau}.
\]  
(III.21)

Then

**Proposition III.2.**

For a simulation operating under Algorithm III.1 (Appendix III.8.1),

(i) \( 0 \leq M_r \leq 1 \) for any \( \tau \);

(ii) \( M_r = 1 \) for any \( \tau \) if and only if the history is monotone;

(iii) \( M_r = 1 \) for some (sufficiently small) \( \tau \) if and only if the history is piecewise monotone.
(iv) \( M_\tau = 0 \) for even \( \tau \) if and only if the history zig-zags persistently;

(v) For large \( \tau \), the expected value of \( M_\tau \) is \( 1/3 \) if the history is a random walk.

Proof. From Proposition III.1 by simple substitution via (III.21).

Thus, the value of \( M_\tau \) will range from 0 (inclusive) to 1 (inclusive) for even window sizes \( \tau \). The closer this value is to 1 the more monotone the history; the closer the value is to 0, the less monotone the history. Having these desirable properties, \( M_\tau \) (restricted, without loss of generality, to even \( \tau \)) will serve as our measure of monotonicity.
Linguistic stability estimation without families

Henri Kauhanen, Deepthi Gopal, Tobias Galla & Ricardo Bermúdez-Otero

Abstract

Different structural features of human language change at different rates and thus exhibit different temporal stabilities. Existing methods of linguistic stability estimation rely on genealogical techniques which require a prior classification of the world’s languages into language families; these methods result in unreliable stability estimates for features which are sensitive to horizontal transfer between families and whenever data are aggregated from families of divergent time depths. To overcome these problems, we describe a method of stability estimation without family classifications, based on mathematical modelling and the analysis of contemporary geospatial distributions of linguistic features. Regressing the estimates produced by our model against those of a genealogical method, we report broad agreement but also important differences. In particular, we show that our approach is not liable to some of the false positives incurred by the genealogical method. Our results suggest that the historical evolution of a linguistic trait leaves a footprint in its global geospatial distribution, and that rates of evolution can be recovered from these distributions by treating language dynamics as a spatially extended stochastic process, without prior genealogical classifications.

Keywords: linguistic typology; stability estimation; complex systems

*A somewhat updated version of this paper has now appeared on arXiv (1801.09637) under the title “Geospatial distributions reflect rates of evolution of features of language”. © The Authors.
IV.1 Stability estimation in linguistics

The languages of the world vary in a number of purely structural features: some place the subject before the verb, others after it; some have nasal vowels, others lack them; some have a future tense, others do not. The set of structural linguistic features gives rise to the state space of human language in which every attested as well as possible language is situated, and which sets the stage for language change, i.e. the cultural evolution of linguistic features. Currently around 7,000 languages are known to be spoken across the world (Simons & Fennig, 2017), but linguistic change is, on the whole, fast: Old English texts from little more than a thousand years ago, for instance, are completely unintelligible to untrained speakers of Modern English.

Some linguistic features seem more prone to change than others: for example, while Modern English has lost most of the noun inflection of Old English, it continues to place the genitive before the noun, as Old English did in the majority of instances (Allen, 2008). Modern work in linguistic typology has sought to explain why such differences in the susceptibility of individual features to change should exist; this work ranges from traditional qualitative typological accounts (Hawkins, 1983; Nichols, 1992) to more recent proposals to quantify linguistic stability with the help of data drawn from large typological databases such as the World Atlas of Language Structures (WALS) (Dryer & Haspelmath, 2013).

The majority of existing linguistic stability estimation techniques rely on a genealogical grouping of the world’s languages into language families, established by standard comparative reconstruction techniques (Maslova, 2004; Parkvall, 2008; Dediu, 2011; Dediu & Cysouw, 2013; Greenhill et al., 2017). Although implementation details vary, all genealogical stability estimation methods work on the basic assumption that stable linguistic features ought to be conserved within language families, while unstable features exhibit within-family variation. To illustrate, the basic order of possessor and possessed (as in *John’s car*) is possessed–possessor in all Romance languages (e.g. Italian *la macchina di Gianni*), inherited from Proto-Romance, whilst the extent of verbal inflection in Romance varies tremendously from the extremely conservative Sardinian to the extremely innovative French, in which verbal inflection has eroded to such an extent that the singular forms of most verbs are pronounced the same. Based on such considerations, the former feature may be classified as stable and the latter as unstable, insofar as this conclusion is supported by data from several languages and several language families.

In general, the genealogical classification procedures used in modern linguistics are highly reliable: the family trees they yield are rarely in dispute, except in respect of the fine structure of otherwise uncontroversial families, or the very distant putative kinship relationships lying beyond the reach of the standard comparative method.
of linguistic reconstruction (Bowern & Evans, 2014). The high reliability of linguistic genealogies, however, does not by itself render them an appropriate tool for estimating the stability of linguistic features. We here outline two limitations shared by all genealogical methods: the problem of time depths, and the problem of horizontal transfer.

Firstly, whilst language families established by comparative means are not generally in dispute, there is no agreement on how to estimate their time depths — no consensus exists on the dating of protolanguages (McMahon & McMahon, 2005; Pereltsvaig & Lewis, 2015; Bowern, 2017), and existing genealogical stability estimation methods do not attempt this. However, since the extent of variation among the surviving members of a family must be proportional to the age of the family (the temporal distance between the root and the leaves of the relevant family tree), the stability estimates produced by a genealogical estimation technique will be highly sensitive to the ages of the families employed. To take an extreme example, if the feature of lip rounding in front vowels (the existence of vowel sounds such as /y/ and /ø/ as in the German words kühl and schön) was estimated from Germanic languages alone, it would be found to be highly stable, since most existing Germanic languages do have these sounds. Zooming out to consider the entire Indo-European family, we however find a number of languages lacking front rounded vowels (such as most existing Romance languages); if applied to the Indo-European family in toto, a genealogical stability estimation method would report less stability for front rounded vowels. The problem is not that genealogical methods are sensitive to time depths, but that no controls have been established to ascertain the reliability of across-family comparisons; the time depths of the WALS genera employed in some methods (Dediu & Cysouw, 2013), in particular, are described by the WALS editors as ‘highly tentative’, ‘based on meagre initial impressions’ and consisting of no more than ‘educated guesses’ (Dryer, 1989; Dryer & Haspelmath, 2013). When stability estimates are aggregated from several such genealogical stocks possibly reflecting largely divergent time depths, the resulting stability estimates may end up highly unreliable.

Secondly, genealogical stability estimation techniques necessarily miss out the effect of horizontal transfer — the borrowing of a feature from one family to another — on a feature’s stability. Yet some linguistic features (e.g. inflectional markers) are more resistant to borrowing than others (Gardani, Arkadiev & Amiridze, 2014), while some (e.g. case systems) are highly vulnerable to simplification in contact situations involving large numbers of second-language learners (Bentz & Winter, 2013). Combining genealogical and areal groupings (Parkvall, 2008) is not a solution, however, as no agreed method exists for delimiting linguistic areas or for estimating the time depths of areal relationships.

All of this motivates the search for a stability measure that reflects the relatedness
of languages without presorting them into predefined groupings and can take into account the existence of horizontal transfer effects. Here, we propose such a technique by modelling language dynamics as a stochastic process defined over a spatial substrate. We show that the resulting model has a stationary distribution which can be found analytically: in particular, this solution allows us to estimate the tendency of individual linguistic features to change based solely on their contemporary geospatial distributions, measured from the WALS atlas. Regressing our stability estimates against those generated by a genealogical method (Dediu, 2011), we report broad agreement but also interesting divergences. Importantly, our method avoids some of the apparent false positives incurred by the genealogical method, which, for example, identifies front rounded vowels as a highly stable feature. When these problematic outliers are removed from the regression, however, the stability estimates predicted by our method correlate highly with those of the genealogical method. We conclude that methods based on pre-established genealogical groupings not only mispredict the stability of certain problematic features, but may also be unnecessary — a model that relies solely on directly observable geospatial information fares no worse.

### IV.2 Stability estimation from geospatial distributions

The simplest model of language dynamics treats the evolution of binary linguistic features as a stochastic process, memoryless and independent from feature to feature (Greenberg, 1978). The dynamics of each single feature are then given by a Markov chain with two parameters, $p_I$ and $p_E$ (Fig. IV.1). The former parameter gives the probability of a language adopting the feature in question; we will call it the feature’s *ingress rate*. The latter parameter, in turn, gives the probability of a language losing the feature; we will refer to it as the feature’s *egress rate*. The stability of a feature may then be equated either with its non-egress, i.e. the tendency of languages to remain in a state in which the feature is attested (Greenberg, 1978), or with its combined non-ingress-and-non-egress, i.e. the tendency of languages not to transit between the two states (Maslova, 2004; Cysouw, 2011).
It is possible to estimate linguistic transition probabilities under certain assumptions, for instance by examining divergence rates of pairs of languages which are related by common descent, and thereby arrive at estimates of stability (Maslova, 2004). This approach is, of course, vulnerable to the two objections outlined above: the problems of relative time depths and horizontal transfer. In order to overcome the problems and develop a stability estimation method which does not rely on prior genealogical groupings, we extend the basic Markov chain model by implementing it on a spatial substrate and by offering an explicit account of interaction dynamics between neighbouring language communities. This model leads to a definition of featural stability in terms of spatial correlations; estimating the latter from an empirical database such as WALS then affords estimates of stability for any feature for which sufficient geospatial data exist.

Specifically, we assume language communities to be distributed on a spatial substrate which, for reasons of mathematical tractability, we take to be a two-dimensional regular lattice. In line with the Markov chain model of Fig. IV.1, we assume that each language community is subject to Markovian ingress–egress dynamics, for each linguistic feature independently of the rest. In addition to this dynamics, operating independently in each cell of the lattice, we assume the existence of a spatially extended interaction process whereby the state of one language community may affect the state of a neighbouring community. While in principle many kinds of interaction dynamics could be studied, we have so far explored what is arguably the simplest such process, the so-called voter model (Liggett, 1997; Castellano et al., 2009). In this model, the feature value of a randomly selected lattice neighbour is simply copied into the target cell; the model, and various of its extensions, have been applied to problems as diverse as ecological conflict (Clifford & Sudbury, 1973), kinetics of catalytic reactions (Krapivsky, 1992), the spread of opinions in social networks (Zschaler, Böhme, Seißinger, Huepe & Gross, 2012) and elections (Fernández-Gracia, Suchecki, Ramasco, San Miguel & Eguíluz, 2014). We initialize the lattice in a random state (for each feature \( F \) and community \( C \), \( F \) is present in \( C \) with probability 0.5), and iterate over the following steps:

1. Pick a random community \( C \) and a random feature \( F \).

2a. With probability \( q \): pick a random lattice neighbour \( C' \) of \( C \), and set the value for feature \( F \) in \( C \) to that in \( C' \).

2b. With probability \( 1 - q \): if \( F \) is absent from \( C \), acquire \( F \) with probability \( p_I \) (ingress); if \( F \) is present in \( C \), lose \( F \) with probability \( p_E \) (egress).

Inevitably, this model idealizes away from many of the complexities of real-world language dynamics. What matters for present purposes is that the model should be
In the model, linguistic features are distributed over a two-dimensional regular lattice with periodic boundary conditions, illustrated here by three *ad hoc* lattices (yellow: feature present; blue: feature absent). In each case, the feature frequency is $\rho = 0.5$, but the isogloss density $\sigma$, measured as the proportion of disagreeing lattice interfaces (red circles), varies with the spatial distribution of the feature and grows as that distribution becomes more and more scattered.

able to capture the effects of both faithful linguistic transmission and linguistic modification or “mutation”, both in endogenous historical developments and in exogenous contact situations, on a global scale and over the long term. Our model does this through the simple expedient of assuming that the lattice spacing is sufficiently fine, so that many neighbouring cells are occupied by communities speaking different varieties of the same language, whilst in other cases neighbouring cells harbour different languages. When transmission occurs between two cells occupied by dialects of the same language, and the donor cell possesses the ancestral value of the feature, then the interaction process (probability $q$) models descent without modification; if, in contrast, the donor cell possesses an innovative value of the feature, then the interaction process models processes of transfer under contact. On the other hand, the ingress–egress dynamics (probabilities $p_I$ and $p_E$) covers processes of mutation, in which transmission fails, causing a feature to flip in the absence of an already existing exemplar. Crucially, this can happen endogenously, as in the case of phonetically motivated sound change (Ohala, 1981), or exogenously, as when contact between two languages with different case systems leads to the emergence of a simplified system, different from both its predecessors (Bentz & Winter, 2013).

It can be shown that our model tends to a stationary distribution at which the statistical properties of the lattice are described by analytically soluble quantities, dependent on the model parameters $p_I$, $p_E$ and $q$ (see SI). For the purposes of stability estimation, we are interested in two quantities in particular (illustrated in Fig. IV.2): the overall frequency of a feature across the lattice (denoted by $\rho$), and the feature’s associated isogloss density, defined as the proportion of pairs of adjacent lattice cells that differ in the feature value (known as the density of reactive interfaces in other litera-
The state of the lattice at the stationary distribution: computer simulations (points) and analytical solution (curves) for various values of $\tau$.

The feature frequency $\rho$ and isogloss density $\sigma$ of a linguistic feature may be measured empirically from a typological database equipped with geolocation information. In line with previous research, we here use the WALS database, focussing on 35 binary features (see Methods, below). For each feature, the frequency $\rho$ is given simply by

$$\rho = \frac{p_I}{p_I + p_E},$$  \hspace{1cm} (IV.1)

while the stationary-state isogloss density is

$$\sigma = 2H(\tau)\rho(1 - \rho),$$  \hspace{1cm} (IV.2)

with

$$\tau = \frac{(1 - q)(p_I + p_E)}{q}$$  \hspace{1cm} (IV.3)

and

$$H(\tau) = \frac{\pi(1 + \tau)}{2K\left(\frac{1}{1+\tau}\right)} - \tau,$$  \hspace{1cm} (IV.4)

where $K(\cdot)$ is the complete elliptic integral of the first kind (see SI for full technical details). Thus, from Eq. (IV.2), the stationary-state isogloss density $\sigma$ is found to be a parabolic function of the feature’s overall frequency $\rho$, the height of this parabola being controlled by the parameter $\tau$ (Fig. IV.3).
Empirical measurements of feature frequency $\rho$ and isogloss density $\sigma$ for 35 WALS features. The crosshairs give 95% bootstrap confidence intervals (see Methods).

the proportion of languages attesting that feature in the feature’s WALS language sample. Isogloss density $\sigma$ is calculated as follows: for each language, we first establish its nearest geographical neighbour in the relevant language sample. The empirical isogloss density is then given by the proportion of nearest-neighbour language pairs differing in their values for the feature in question. Our data are summarized in Fig. IV.4, which supplies $\rho$ and $\sigma$ for each of the 35 features; a lower value of isogloss density $\sigma$ signals a geographically clustered feature, whilst a higher value implies a feature with a scattered geographical distribution. Figs. IV.5A–B illustrate this difference with two features: tone (the encoding of meaning by voice pitch, as in Mandarin Chinese $m\ddot{a}$ ‘mother’ vs. $m\acute{a}$ ‘horse’) and indefinite article (e.g. English $a$ linguist).

For a given feature frequency $\rho$, the isogloss density $\sigma$ is fixed by the value of $H(\tau)$, itself a monotonic increasing function of $\tau$ (Eq. IV.4). The parameter $\tau$, in turn, gives the relative rate of the ingress–egress dynamics over the interaction dynamics (Eq. IV.3). Given that in our model the interaction dynamics serves to transmit features faithfully, while the ingress–egress dynamics embodies processes of mutation, $\tau$ is a temperature: lower values of $\tau$ signify a stable feature, higher values indicating instability. Since each of our 35 empirical features lies on a unique parabola in the space spanned by $\rho$ and $\sigma$ (Fig. IV.4), estimating its temperature is now a simple matter of inverting the function $H(\tau)$. For each feature, we measure $\rho$ and $\sigma$ as described above. From Eq. (IV.2), we then obtain the value of $H(\tau)$ and invert this to recover $\tau$. Although the elliptic integral in Eq. (IV.4) cannot be expressed in terms of elementary functions and $H(\tau)$ thus cannot be inverted analytically, the inversion
FIGURE IV.5
Empirical and simulated geospatial distributions of two linguistic features on the hemisphere from 30° W to 150° E (yellow: feature present, blue: feature absent): (A) tone, empirical, (B) indefinite article, empirical, (C) tone, simulated, (D) indefinite article, simulated. Shown are both individual empirical/simulated data points and a spatial interpolation (inverse distance weighting) on these points. Map projection: Albers equal-area.
Table IV.1
The seven least stable and most stable features in our dataset, as ranked by decreasing temperature ($\tau$).

<table>
<thead>
<tr>
<th>Feature (WALS ID)</th>
<th>Temperature $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>verbal person marking (100A)</td>
<td>$5.06 \times 10^{-1}$</td>
</tr>
<tr>
<td>definite article (37A)</td>
<td>$3.54 \times 10^{-1}$</td>
</tr>
<tr>
<td>question particle (92A)</td>
<td>$3.43 \times 10^{-1}$</td>
</tr>
<tr>
<td>gender in independent personal pronouns (44A)</td>
<td>$2.46 \times 10^{-1}$</td>
</tr>
<tr>
<td>indefinite article (38A)</td>
<td>$2.30 \times 10^{-1}$</td>
</tr>
<tr>
<td>lateral consonants (8A)</td>
<td>$2.15 \times 10^{-1}$</td>
</tr>
<tr>
<td>front rounded vowels (11A)</td>
<td>$1.95 \times 10^{-1}$</td>
</tr>
</tbody>
</table>

... ...

| velar nasal (9A)                              | $1.83 \times 10^{-3}$      |
| tone (13A)                                    | $1.28 \times 10^{-3}$      |
| order of adjective and noun is AdjN (87A)     | $1.21 \times 10^{-3}$      |
| order of subject and verb is SV (82A)         | $4.93 \times 10^{-4}$      |
| order of genitive and noun is GenN (86A)      | $3.34 \times 10^{-5}$      |
| order of numeral and noun is NumN (89A)       | $3.31 \times 10^{-5}$      |
| order of object and verb is OV (83A)          | $1.91 \times 10^{-5}$      |

can be performed numerically. Using this procedure we obtain an estimate of $\tau$ for any feature for which empirical measurements of frequency $\rho$ and isogloss density $\sigma$ exist. Table IV.1 supplies these estimates for the least stable and most stable features in our dataset (for a full listing of $\tau$ estimates for the entire dataset, see Table IV.3 in the SI).

To get an impression of the extent of agreement between our model and empirical feature distributions, we took the $\tau$ estimates predicted by our model for the two features tone and indefinite article mapped in Figs. IV.5A–B, and ran the model on a lattice superposed on the same geographical area using these values of $\tau$. Figs. IV.5C–D show the state of the simulation at the stationary distribution, indicating broad agreement with the empirical distributions (note that we are interested in explaining the overall shape of the geospatial distributions — scattered or clustered — and not their locations, which are contingent upon the locations of the relevant protolanguages).

IV.3 Comparison with a genealogical method

The technique proposed by Dediu (2011) represents the state of the art in genealogy-based stability estimation. Using a Bayesian phylogenetic algorithm, this method produces a posterior distribution of rates of evolution for each linguistic feature within a predefined genealogical grouping. Dediu tests two phylogenetic algorithms and draws data from two sources — WALS and Ethnologue (Simons & Fennig, 2017) —
Regression of our temperature ($\tau$) estimates against Dediu’s PC1. (A) Red line: regression with all 24 data points (Pearson’s $r = 0.53$, $p = 0.008$). Black line: regression with five outliers (red crosses) removed (Pearson’s $r = 0.90$, $p = 1.613 \times 10^{-7}$). (B) Outliers were detected by pruning the dataset recursively for those data points that contribute most to the regression error, quantified as the sum of squared residuals. This identified features 11A, 107A, 8A, 44A and 57A (see text) as outliers.

Regression of our temperature ($\tau$) estimates against Dediu’s PC1. (A) Red line: regression with all 24 data points (Pearson’s $r = 0.53$, $p = 0.008$). Black line: regression with five outliers (red crosses) removed (Pearson’s $r = 0.90$, $p = 1.613 \times 10^{-7}$). (B) Outliers were detected by pruning the dataset recursively for those data points that contribute most to the regression error, quantified as the sum of squared residuals. This identified features 11A, 107A, 8A, 44A and 57A (see text) as outliers.

In Fig. IV.6A, we take the 24 features in the intersection of our list of 35 features and Dediu’s, and regress our $\tau$ estimates against Dediu’s PC1 (red regression line). There is a moderate significant correlation between the estimates predicted by the two methods (Pearson’s $r = 0.53$, $p = 0.008$), but a number of features are clearly outliers of the regression. To detect these outliers objectively, we pruned the regression recursively by removing those data points that contribute the greatest error in terms of sum of squared residuals; Fig. IV.6B gives the reduction of error at each step of this pruning. The reduction profile prompts us to classify as outliers the first five data points, corresponding to the following WALS features: 11A (front rounded vowels), 107A (passive construction), 8A (lateral consonants), 44A (gender in independent personal pronouns) and 57A (possessive affixes). Regressing the pruned dataset (Fig. IV.6A, black regression line), we find a high, significant correlation between our $\tau$ estimates and Dediu’s PC1 (Pearson’s $r = 0.90$, $p = 1.613 \times 10^{-7}$).

We suggest that, rather than representing different views on stability, these outliers are false positives of the genealogical method. We illustrate this with the case of
(the presence or absence of) front rounded vowels (WALS feature 11A), i.e. the vowels /y/ (e.g. Finnish kyy), /ø/ (German schön), /œ/ (French bœuf) and /Œ/ (Danish grøn).

This is one of the most stable features in the genealogical analysis but one of the least stable features in ours (Table IV.1); we argue that evidence from both language change and language acquisition supports our position. On the one hand, front rounded vowels are frequently innovated: historical fronting of the back rounded vowel /u/ to [y] (with or without subsequent phonemicization to /y/) has been documented in a number of languages, including but not limited to Armenian, Attic-Ionic Greek, French, Frisian, Lithuanian, Old Scandinavian, Oscan, Parachi, Umbrian, West Syriac, Yiddish, Zuberoan Basque, and numerous dialects of English (Ohala, 1981; Labov, Yaeger & Steiner, 1972; Dressler, 1974; Lass, 1988; Labov, 1994; Egurtzegi, 2017; Samuels, 2017); additionally, front rounded vowels can arise through the influence of /i/ or /j/ on a neighbouring back rounded vowel (Ohala, 1981; Iverson & Salmons, 2003).

On the other hand, front rounded vowels are difficult to acquire in situations of language contact: there is experimental evidence that second-language learners whose first language lacks these vowels perceive them as more similar to back vowels than front vowels (Strange, Levy & Law, 2009). This perceptual assimilation is, moreover, mirrored in speech production: productions of /y/ by second-language learners are far less advanced in phonetic space than native speakers’ productions, and are indeed often perceived as /u/ by the latter (Rochet, 1995). The fact that front rounded vowels are readily innovated points to a high ingress rate, while frequent acquisition failure by second-language learners in situations of language contact points to a high egress rate; these facts are inconsistent with the high stability predicted by the genealogical method, but consistent with our approach, in which high ingress and high egress make for high temperature (Eq. IV.3) and thus low stability.

IV.4 Discussion

The challenge of linguistic stability estimation arises, essentially, from having to work with a poor signal. Although evolutionary and anthropological evidence suggests that human language in its modern form has existed for at least 100,000 years (Bickerton, 2007; Tallerman, 2012), the historical evolution of languages is (necessarily) poorly documented: such documentation only covers a few thousand years for languages with the best coverage, cannot in principle go beyond the introduction of the first writing systems, and does not exist at all for the majority of the world’s languages. The rest of the historical evolution of language must be reconstructed based on available data. In this paper, we have argued that stability estimation methods relying

---

1Front rounded vowels are the fourth most stable feature (out of 86) in Dediu’s (2011, Table S4) study and the second most stable (out of 62) in Dediu and Cysouw’s (2013, Table 7) metastudy.
on even the most accurate reconstructive methods have their shortcomings: there is no guarantee that the genealogical classifications assumed in such estimation reflect equivalent time depths between different lineages, and the methods do not control for horizontal transfer between languages belonging to different lineages. We have put forward a stability estimation procedure that does without genealogies and infers stability estimates from contemporary geospatial distributions of linguistic features alone. The method is relatively easy to implement: all that is needed are measures of feature frequency and isogloss density for a large enough sample of languages, and inversion of Eq. (IV.4).

We have offered some evidence in support of our method. Turning now to its limitations, we note that our approach currently only applies to binary features. Most genealogical methods do not suffer from this limitation: Dediu’s (2011) procedure, in particular, can be applied to polyvalent as well as binary features. Interestingly, however, Dediu finds a correlation between estimates generated for polyvalent and binary (or binarized) features. This suggests that the resolution at which the values of a linguistic variable are recorded may be a minor issue: after all, any polyvalent classification can be reduced to a hierarchy of binary oppositions. Such a reduction implies that features may not be independent, but feature interactions are known to exist at any rate (Greenberg, 1963; Dryer, 2003), and no method of stability estimation so far proposed can handle these. In principle, it would be possible to implement feature interactions in our lattice model by relaxing the assumption that the ingress–egress process embodies rates of change which are independent from feature to feature (Axelrod, 1997), but it is at the moment unclear whether such a relaxation would still result in a model whose behaviour is soluble so that quantitative stability estimates may be reliably inferred. We leave this as an open problem for future study.

Methods

Data preparation

The WALS Online database (Dryer & Haspelmath, 2013) was downloaded on 18th July 2017 and used as the empirical basis for measures of feature frequency $\rho$ and isogloss density $\sigma$. Since WALS employs a polyvalent coding for most features, a manual pass was made, retaining only those features that are binary or binarizable on uncontroversial linguistic grounds. Features with fewer than 300 languages in their language sample were discarded to ensure statistically robust results. This procedure resulted in 35 binary features (see Table IV.2 in the SI for a listing, together with full information about our binarization scheme). Nearest neighbours of languages were determined by the great-circle distance, calculated from WALS coordinate data using the haversine formula with the assumption of a perfectly spherical earth.
Analysis

To eliminate any possible effect the differing language sample sizes of different WALS features might have on our statistics, a fixed number of 300 languages was considered for each feature, with languages sampled uniformly at random from the feature’s language sample. This procedure was repeated 10,000 times for each feature to generate the bootstrap averages and confidence intervals reported in Fig. IV.4. For comparison with the genealogical method, Table S1 in the SI to Dediu (2011) was consulted and only those features were selected for comparison for which our binarization schemes agreed; the PC1 values for the intersecting features were then gathered from Table S4. Correlation strengths were measured using the Pearson correlation coefficient; significance was tested with a two-tailed $t$-test. The analytical solution of the lattice model (Eqs. IV.1–IV.4) appears in the SI.

Data availability

WALS is freely available at http://wals.info; the binarization scheme used to prepare the data appears in the SI.

Code availability

Computer code for data analysis, stability estimation and numerical simulations may be obtained from the corresponding author.

Author contributions

Designed the study: HK, DG, TG and RB-O. Analysed the data: HK, DG, TG and RB-O. Solved the mathematical model: HK, DG and TG. Wrote the paper: HK, DG, TG and RB-O. Wrote visualization routines: HK and DG. Wrote the data analysis and simulation code: HK.

Author information

The authors declare no conflict of interest and no competing financial interests. Correspondence and requests for materials should be addressed to henri.kauhanen@unikonstanz.de.

Acknowledgements

We thank George Walkden for discussions. HK acknowledges financial support from Emil Aaltonen Foundation and The Ella and Georg Ehrnrooth Foundation.
IV.5 Supplementary information

IV.5.1 Features consulted

Table IV.2 provides a listing of the 35 WALS (Dryer & Haspelmath, 2013) features consulted in this study, together with our scheme for feature binarization. Each WALS feature is a variable of either nominal or ordinal level, whose possible values are recorded in the WALS database using integer labels. The meanings of these labels are explained at length in Section IV.5.4, below; Table IV.2 indicates which values of each variable were folded into the ‘feature absent’ category and which values to the ‘feature present’ category in our binarization.

IV.5.2 Temperature estimates

Table IV.3 gives the temperature ($\tau$) estimates found by our method for the 35 features.

IV.5.3 Analytical solution of lattice model

We will treat the model as a spin system on a two-dimensional regular square lattice with $N = L \times L$ sites and periodic boundary conditions (for comparable approaches to the voter model without an ingress–egress dynamics, see Krapivsky, 1992; Krapivsky et al., 2010). Our model is conceptually similar to a voter model with noise, which has been treated with similar methods in de Oliveira (2003). We write $s(x) \in \{-1, 1\}$ for the spin at lattice site $x = (x_1, x_2)$; $S(x) = \langle s(x) \rangle$ for the average spin of $x$ (over realizations of the stochastic process); and $m = \sum_x S(x)/N$ for the mean magnetization over the entire lattice. The feature frequency $\rho$, or fraction of up-spins in the system, is related to $m$ by the identity $\rho = (m + 1)/2$. We further write $S(x, y) = \langle s(x)s(y) \rangle$ for the pair correlation of $s(x)$ and $s(y)$. In summations, $x'$ is understood to index the set of von Neumann neighbours of site $x$, i.e. the set

$$\{(x_1 - 1, x_2), (x_1 + 1, x_2), (x_1, x_2 - 1), (x_1, x_2 + 1)\}. \quad (IV.5)$$

Spin flip probability. Central to our analytical derivations is the spin flip probability, i.e. the probability with which the spin at site $x$ changes its state from $-1$ to $+1$ or vice versa, if it is selected for potential update. In our model this is of the form

$$w(x) = (1 - q)A(x) + qB(x), \quad (IV.6)$$

where $A(x)$ is the contribution of the ingress–egress process and $B(x)$ the contribution of the spatial (voter) process. These are

$$A(x) = \frac{1 - s(x)}{2}p_t + \frac{1 + s(x)}{2}p_E = \frac{1}{2} [p_t + p_E + (p_E - p_t)s(x)] \quad (IV.7)$$
The 35 binary features considered in this study. The ‘WALS’ column gives the WALS feature mined; values indicated in the ‘abs.’ column were folded into our ‘binary feature absent’ value, whilst values indicated in the ‘pres.’ column were folded into our ‘binary feature present’ value (see Section IV.5.4, below). The final column gives the size of the WALS language sample (number of languages) for each feature.

<table>
<thead>
<tr>
<th>description</th>
<th>WALS</th>
<th>abs.</th>
<th>pres.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. adpositions</td>
<td>48A</td>
<td>1</td>
<td>2–4</td>
<td>378</td>
</tr>
<tr>
<td>2. definite article</td>
<td>37A</td>
<td>4–5</td>
<td>1–3</td>
<td>620</td>
</tr>
<tr>
<td>3. <em>hand</em> and <em>arm</em> identical</td>
<td>129A</td>
<td>2</td>
<td>1</td>
<td>617</td>
</tr>
<tr>
<td>4. <em>hand</em> and <em>finger(s)</em> identical</td>
<td>130A</td>
<td>2</td>
<td>1</td>
<td>593</td>
</tr>
<tr>
<td>5. front rounded vowels</td>
<td>11A</td>
<td>1</td>
<td>2–4</td>
<td>562</td>
</tr>
<tr>
<td>6. gender in independent personal pronouns</td>
<td>44A</td>
<td>6</td>
<td>1–5</td>
<td>378</td>
</tr>
<tr>
<td>7. glottalized consonants</td>
<td>7A</td>
<td>1</td>
<td>2–8</td>
<td>567</td>
</tr>
<tr>
<td>8. grammatical evidentials</td>
<td>77A</td>
<td>1</td>
<td>2–3</td>
<td>418</td>
</tr>
<tr>
<td>9. indefinite article</td>
<td>38A</td>
<td>4–5</td>
<td>1–3</td>
<td>534</td>
</tr>
<tr>
<td>10. inflectional morphology</td>
<td>26A</td>
<td>1</td>
<td>2–6</td>
<td>969</td>
</tr>
<tr>
<td>11. inflectional optative</td>
<td>73A</td>
<td>2</td>
<td>1</td>
<td>319</td>
</tr>
<tr>
<td>12. lateral consonants</td>
<td>8A</td>
<td>1</td>
<td>2–5</td>
<td>567</td>
</tr>
<tr>
<td>13. morphological second-person imperative</td>
<td>70A</td>
<td>5</td>
<td>1–4</td>
<td>547</td>
</tr>
<tr>
<td>14. order of adjective and noun is AdjN</td>
<td>87A</td>
<td>2</td>
<td>1</td>
<td>1366</td>
</tr>
<tr>
<td>15. order of degree word and adjective is DegAdj</td>
<td>91A</td>
<td>2</td>
<td>1</td>
<td>481</td>
</tr>
<tr>
<td>16. order of genitive and noun is GenN</td>
<td>86A</td>
<td>2</td>
<td>1</td>
<td>1249</td>
</tr>
<tr>
<td>17. order of numeral and noun is NumN</td>
<td>89A</td>
<td>2</td>
<td>1</td>
<td>1153</td>
</tr>
<tr>
<td>18. order of object and verb is OV</td>
<td>83A</td>
<td>2</td>
<td>1</td>
<td>1519</td>
</tr>
<tr>
<td>19. order of subject and verb is SV</td>
<td>82A</td>
<td>2</td>
<td>1</td>
<td>1497</td>
</tr>
<tr>
<td>20. ordinal numerals</td>
<td>53A</td>
<td>1</td>
<td>2–8</td>
<td>321</td>
</tr>
<tr>
<td>21. passive construction</td>
<td>107A</td>
<td>2</td>
<td>1</td>
<td>373</td>
</tr>
<tr>
<td>22. plural</td>
<td>33A</td>
<td>9</td>
<td>1–8</td>
<td>1066</td>
</tr>
<tr>
<td>23. possessive affixes</td>
<td>57A</td>
<td>4</td>
<td>1–3</td>
<td>902</td>
</tr>
<tr>
<td>24. postverbal negative morpheme</td>
<td>143F</td>
<td>4</td>
<td>1–3</td>
<td>1324</td>
</tr>
<tr>
<td>25. preverbal negative morpheme</td>
<td>143E</td>
<td>4</td>
<td>1–3</td>
<td>1324</td>
</tr>
<tr>
<td>26. productive reduplication</td>
<td>27A</td>
<td>3</td>
<td>1–2</td>
<td>368</td>
</tr>
<tr>
<td>27. question particle</td>
<td>92A</td>
<td>6</td>
<td>1–5</td>
<td>884</td>
</tr>
<tr>
<td>28. shared encoding of nominal and locational predication</td>
<td>119A</td>
<td>1</td>
<td>2</td>
<td>386</td>
</tr>
<tr>
<td>29. tense-aspect inflection</td>
<td>69A</td>
<td>5</td>
<td>1–4</td>
<td>1131</td>
</tr>
<tr>
<td>30. tone</td>
<td>13A</td>
<td>1</td>
<td>2–3</td>
<td>527</td>
</tr>
<tr>
<td>31. uvular consonants</td>
<td>6A</td>
<td>1</td>
<td>2–4</td>
<td>567</td>
</tr>
<tr>
<td>32. velar nasal</td>
<td>9A</td>
<td>3</td>
<td>1–2</td>
<td>469</td>
</tr>
<tr>
<td>33. verbal person marking</td>
<td>100A</td>
<td>1</td>
<td>2–6</td>
<td>380</td>
</tr>
<tr>
<td>34. voicing contrast</td>
<td>4A</td>
<td>1</td>
<td>2–4</td>
<td>567</td>
</tr>
<tr>
<td>35. zero copula for predicate nominals</td>
<td>120A</td>
<td>1</td>
<td>2</td>
<td>386</td>
</tr>
</tbody>
</table>
Table IV.3
Feature frequency $\rho$, isogloss density $\sigma$ and estimated temperature $\tau$ for the 35 features considered in this study (to eight significant decimals), ordered by increasing $\tau$ (from stable to unstable).

<table>
<thead>
<tr>
<th>feature</th>
<th>$\rho$</th>
<th>$\sigma$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. order of object and verb is OV</td>
<td>0.50352113</td>
<td>0.12087912</td>
<td>0.00001910</td>
</tr>
<tr>
<td>2. order of numeral and noun is NumN</td>
<td>0.44097222</td>
<td>0.12454212</td>
<td>0.00003306</td>
</tr>
<tr>
<td>3. order of genitive and noun is GenN</td>
<td>0.59420290</td>
<td>0.12204724</td>
<td>0.00003339</td>
</tr>
<tr>
<td>4. order of subject and verb is SV</td>
<td>0.86071429</td>
<td>0.07722008</td>
<td>0.00049325</td>
</tr>
<tr>
<td>5. order of adjective and noun is AdjN</td>
<td>0.29651115</td>
<td>0.14843750</td>
<td>0.00120923</td>
</tr>
<tr>
<td>6. tone</td>
<td>0.41666667</td>
<td>0.17333333</td>
<td>0.00128340</td>
</tr>
<tr>
<td>7. velar nasal</td>
<td>0.50000000</td>
<td>0.18333333</td>
<td>0.00183200</td>
</tr>
<tr>
<td>8. order of degree word and adjective is DegAdj</td>
<td>0.54135338</td>
<td>0.21212121</td>
<td>0.00569870</td>
</tr>
<tr>
<td>9. uvular consonants</td>
<td>0.17000000</td>
<td>0.13000000</td>
<td>0.00979939</td>
</tr>
<tr>
<td>10. ordinal numerals</td>
<td>0.89666667</td>
<td>0.08666667</td>
<td>0.01181972</td>
</tr>
<tr>
<td>11. shared encoding of nominal and locational predication</td>
<td>0.30333333</td>
<td>0.21000000</td>
<td>0.01732702</td>
</tr>
<tr>
<td>12. glottalized consonants</td>
<td>0.27666667</td>
<td>0.21000000</td>
<td>0.02605203</td>
</tr>
<tr>
<td>13. inflectional optative</td>
<td>0.15000000</td>
<td>0.14333333</td>
<td>0.04120610</td>
</tr>
<tr>
<td>14. inflectional morphology</td>
<td>0.85333333</td>
<td>0.14000000</td>
<td>0.04215630</td>
</tr>
<tr>
<td>15. possessive affixes</td>
<td>0.71333333</td>
<td>0.23666667</td>
<td>0.04784846</td>
</tr>
<tr>
<td>16. passive construction</td>
<td>0.43333333</td>
<td>0.29333333</td>
<td>0.05949459</td>
</tr>
<tr>
<td>17. tense-aspect inflection</td>
<td>0.86666667</td>
<td>0.13666667</td>
<td>0.05994843</td>
</tr>
<tr>
<td>18. <em>hand and arm</em> identical</td>
<td>0.37000000</td>
<td>0.28000000</td>
<td>0.06211254</td>
</tr>
<tr>
<td>19. productive reduplication</td>
<td>0.85000000</td>
<td>0.15333333</td>
<td>0.06274509</td>
</tr>
<tr>
<td>20. voicing contrast</td>
<td>0.68000000</td>
<td>0.26333333</td>
<td>0.06799913</td>
</tr>
<tr>
<td>21. adpositions</td>
<td>0.83333333</td>
<td>0.17000000</td>
<td>0.06908503</td>
</tr>
<tr>
<td>22. postverbal negative morpheme</td>
<td>0.46333333</td>
<td>0.30333333</td>
<td>0.06961203</td>
</tr>
<tr>
<td>23. <em>hand and finger(s)</em> identical</td>
<td>0.12000000</td>
<td>0.13000000</td>
<td>0.07032095</td>
</tr>
<tr>
<td>24. morphological second-person imperative</td>
<td>0.77666667</td>
<td>0.21666667</td>
<td>0.08655531</td>
</tr>
<tr>
<td>25. preverbal negative morpheme</td>
<td>0.70666667</td>
<td>0.27333333</td>
<td>0.11292895</td>
</tr>
<tr>
<td>26. plural</td>
<td>0.91000000</td>
<td>0.11000000</td>
<td>0.12817720</td>
</tr>
<tr>
<td>27. zero copula for predicate nominals</td>
<td>0.45333333</td>
<td>0.33333333</td>
<td>0.13213325</td>
</tr>
<tr>
<td>28. grammatical evidentials</td>
<td>0.56666667</td>
<td>0.33333333</td>
<td>0.14401769</td>
</tr>
<tr>
<td>29. front rounded vowels</td>
<td>0.06666667</td>
<td>0.08666667</td>
<td>0.19468340</td>
</tr>
<tr>
<td>30. lateral consonants</td>
<td>0.83333333</td>
<td>0.20000000</td>
<td>0.21489842</td>
</tr>
<tr>
<td>31. indefinite article</td>
<td>0.44666667</td>
<td>0.35666667</td>
<td>0.22952822</td>
</tr>
<tr>
<td>32. gender in independent personal pronouns</td>
<td>0.33000000</td>
<td>0.32333333</td>
<td>0.24577576</td>
</tr>
<tr>
<td>33. question particle</td>
<td>0.59666667</td>
<td>0.36666667</td>
<td>0.34336306</td>
</tr>
<tr>
<td>34. definite article</td>
<td>0.60666667</td>
<td>0.36333333</td>
<td>0.35396058</td>
</tr>
<tr>
<td>35. verbal person marking</td>
<td>0.78000000</td>
<td>0.27666667</td>
<td>0.50590667</td>
</tr>
</tbody>
</table>
and

$$B(x) = \frac{1}{4} \sum_{x'} \frac{1 - s(x)s'(x')}{2} = \frac{1}{2} \left[ 1 - \frac{1}{4} \sum_{x'} s(x)s'(x') \right], \quad (IV.8)$$

where it is important to remember that the summation over $x'$ runs over the four nearest neighbours of $x$. Hence we have

$$w(x) = \frac{1}{2} - q \left[ p_l + p_E - (p_I - p_l) s(x) \right] + \frac{q}{2} \left[ 1 - \frac{1}{4} \sum_{x'} s(x)s'(x') \right]. \quad (IV.9)$$

**Stationary-state feature frequency $\rho$.** The spin at $x$ changes by the amount $-2s(x)$ with probability $\frac{1}{N}w(x)$, the prefactor $1/N$ representing the probability of $x$ being picked for update. Consequently, the mean spin $S(x) = \langle s(x) \rangle$ evolves as

$$S(x, t + \Delta t) - S(x, t) = \frac{1}{N}(-2w(x)s(x)), \quad (IV.10)$$

where $\Delta t$ is the time step associated with each attempted spin flip. Bearing in mind that $s(x)s(x) = 1$ and plugging Eq. (IV.9) in, this implies

$$\frac{S(x, t + \Delta t) - S(x, t)}{1/N} = (1 - q)(p_l - p_E - (p_I + p_E)S(x)) + q \left[ -S(x) + \frac{1}{4} \sum_{x'} S(x') \right]. \quad (IV.11)$$

Taking the sum over all sites $x$, one has

$$\frac{m(t + \Delta t) - m(t)}{1/N} = (1 - q)(p_l - p_E) - (1 - q)(p_I + p_E) \frac{1}{N} \sum_x S(x) -
- \frac{q}{N} \sum_x S(x) + \frac{q}{4N} \sum_x \sum_{x'} S(x'). \quad (IV.12)$$

Now $\sum_x \sum_{x'} S(x') = 4 \sum_x S(x)$, since the LHS is the sum of the four von Neumann neighbours of all lattice sites, so that each site, having four neighbours, gets counted four times. Using $m = \sum_x S(x)/N$, we then have

$$\frac{m(t + \Delta t) - m(t)}{1/N} = (1 - q)(p_l - p_E - (p_I + p_E)m). \quad (IV.13)$$

With the standard choice $\Delta t = 1/N$, and taking the limit $N \to \infty$ (i.e. the continuous-time limit $\Delta t \to 0$), we thus find

$$\frac{dm}{dt} = (1 - q)(p_l - p_E - (p_I + p_E)m). \quad (IV.14)$$
Hence the mean magnetization in the stationary state ($dm/dt = 0$) is

$$m = \frac{p_I - p_E}{p_I + p_E}. \quad (IV.15)$$

From this, using $\rho = (m + 1)/2$, we find

$$\rho = \frac{p_I}{p_I + p_E} \quad (IV.16)$$

for the fraction of up-spins in the stationary state.

**Pair correlation function.** To compute the pair correlation $S(x, y) = \langle s(x)s(y) \rangle$, we note that $s(x)s(y)$ changes by the amount $-2s(x)(y)$ if either $x$ or $y$ flips spin. Assuming $x \neq y$, and working directly in the continuous-time limit, we have

$$\frac{dS(x, y)}{dt} = \langle -2 [w(x) + w(y)] s(x)s(y) \rangle. \quad (IV.17)$$

After some algebra we find

$$\frac{dS(x, y)}{dt} = (1 - q) [(p_I - p_E)[S(x) + S(y)] - 2(p_I + p_E)S(x, y)] +$$

$$+ q \left[ -2S(x, y) + \frac{1}{4} \sum x' S(x', y) + \frac{1}{4} \sum y' S(x, y') \right], \quad (IV.18)$$

where the summation over $y'$ is over the four nearest neighbours of $y$.

We now assume translation invariance and write $C(r) = C(x - y) = S(x, y)$. Then, for $r \neq 0$,

$$\frac{dC(r)}{dt} = (1 - q) [(p_I - p_E)[S(x) + S(y)] - 2(p_I + p_E)C(r)] +$$

$$+ q \left[ -2C(r) + \frac{1}{4} \sum x' C(x' - y) + \frac{1}{4} \sum y' C(x - y') \right]. \quad (IV.19)$$

Due to translation invariance the two summations on the RHS coincide, and we have (always restricting $r \neq 0$)

$$\frac{dC(r)}{dt} = (1 - q) [(p_I - p_E)[S(x) + S(y)] - 2(p_I + p_E)C(r)] + 2q\Delta C(r), \quad (IV.20)$$

where $\Delta$ is the lattice Laplace operator

$$\Delta f(x) = -f(x) + \frac{1}{4} \sum x' f(x'). \quad (IV.21)$$

215
We note that we always have the boundary condition \(C(0, t) = 1\) for all \(t\) (self-correlation is 1 at all times, as \(s(x)^2 = 1\)).

**Stationary-state isogloss density \(\sigma\).** If \(C(e_1, t)\) were known, where \(e_1\) is the unit vector \((1, 0)\), the isogloss density could be obtained via the identity

\[
\sigma(t) = \frac{1 - C(e_1, t)}{2}.
\]

(IV.22)

Thus, knowing the limiting \((t \to \infty)\) value of \(C(e_1)\) would imply the stationary-state isogloss density \(\sigma\).

At the stationary state, \(S(x) + S(y) = 2m\) and \(\frac{dC(r)}{dt} = 0\). Assuming \(r \neq 0\), Eq. (IV.20) then implies

\[
0 = (1 - q) \left[ (p_I - p_E)m - (p_I + p_E)C(r) \right] + 2q \Delta C(r),
\]

(IV.23)

in other words,

\[
0 = (1 - q) \left[ (p_I - p_E)m - (p_I + p_E)C(r) \right] + q \Delta C(r)
\]

\[
= (1 - q)(p_I + p_E) \left[ \frac{m^2 - C(r)}{p_I + p_E} \right] + q \Delta C(r)
\]

(IV.24)

In the special case \(q = 0\) (i.e., no spatial interaction between neighbouring sites), this implies \(C(r) = m^2\) for all \(r \neq 0\). Thus for any two sites \(x \neq y\), \(\langle s(x)s(y) \rangle = m^2\), indicating that spins at different sites are fully uncorrelated. This is what one would expect, as all sites operate independently when \(q = 0\).

In the special case \(q = 1\) (no ingress–egress dynamics within the sites), on the other hand, we obtain \(\Delta C(r) = 0\) for \(r \neq 0\). We also have \(C(0) = 0\). This implies \(C(r) = 1\) everywhere, so either all spins are up or all are down. This is the only possible stationary state when the only dynamics is through nearest-neighbour interactions.

Now suppose \(0 < q < 1\). Dividing Eq. (IV.24) by \(q\) we obtain

\[
0 = \frac{1 - q}{q} \left( p_I + p_E \right) \left[ m^2 - C(r) \right] + \Delta C(r)
\]

\[
= \left( \frac{q - 1}{q} \right) \left( p_I + p_E \right) \left[ C(r) - m^2 \right] + \Delta C(r)
\]

(IV.25)

\[
= \alpha C(r) - \alpha m^2 + \Delta C(r),
\]

where we write

\[
\alpha = \frac{(q - 1)(p_I + p_E)}{q}.
\]

(IV.26)
Now let
\[ c(r) = C(r) - m^2. \] (IV.27)

We note that \( \Delta C(r) = \Delta c(r) \). To solve Eq. (IV.25), it then suffices to solve (for \( r \neq 0 \))
\[ \Delta c(r) + \alpha c(r) = 0 \] (IV.28)
subject to the condition
\[ c(0) = 1 - m^2. \] (IV.29)

We now first focus on the equation
\[ \Delta G(r) + \alpha G(r) = -\delta_{r,0}, \] (IV.30)
for all \( r \) (including \( r = 0 \)), and where \( \delta_{xy} \) is the Kronecker delta.

Let \( G_\alpha(r) \) be a solution of Eq. (IV.30). Then
\[ c(r) = (1 - m^2) \frac{G_\alpha(r)}{G_\alpha(0)} \] (IV.31)
is a solution of Eqs. (IV.28) and (IV.29). This can be seen as follows: first, from Eq. (IV.31),
\[ c(0) = (1 - m^2) \frac{G_\alpha(0)}{G_\alpha(0)} = 1 - m^2, \] (IV.32)
so the condition in Eq. (IV.29) is met. Second, we need to show that Eqs. (IV.30) and (IV.31) imply \( \Delta c(r) + \alpha c(r) = 0 \) for \( r \neq 0 \). For \( r \neq 0 \) we have
\[ \Delta G(r) + \alpha G(r) = 0 \] (IV.33)
from Eq. (IV.30). The quantity \( c(r) \) in Eq. (IV.31) is proportional to \( G_\alpha(r) \) with a proportionality constant \( (1 - m^2)/G_\alpha(0) \) which is independent of \( r \). Given that \( G_\alpha(r) \) fulfills Eq. (IV.33) for \( r \neq 0 \) it is then clear that \( c(r) \) fulfills \( \Delta c(r) + \alpha c(r) = 0 \) for \( r \neq 0 \), i.e. Eq. (IV.28).

So we are left with the task of finding a solution of
\[ \Delta G(r) + \alpha G(r) = -\delta_{r,0}. \] (IV.34)

Let us write \( G(r) \) in Fourier representation:
\[ G(r) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} dk_1 dk_2 e^{ik \cdot r} \hat{G}(k), \] (IV.35)
where \( k = (k_1, k_2) \) and \( \hat{G}(k) \) is the Fourier transform of \( G(r) \), i.e.

\[
\hat{G}(k) = \sum_r e^{-i k \cdot r} G(r) = \sum_x \sum_y e^{-i x k_1 - i y k_2} G(x, y). \tag{IV.36}
\]

Applying this to both sides of Eq. (IV.34) leads to

\[
\sum_{x,y} e^{-i x k_1 - i y k_2} [\Delta G(r) + \alpha G(r)] = -1. \tag{IV.37}
\]

Next, notice

\[
\sum_{x,y} e^{-i x k_1 - i y k_2} \Delta G(x, y) = \frac{1}{4} \sum_{x,y} e^{-i x k_1 - i y k_2} [G(x + 1, y) + G(x - 1, y) + G(x, y + 1) + G(x, y - 1) - 4G(x, y)]
= \frac{1}{4} \sum_{x,y} \left[ e^{-i(x-1)k_1 - i y k_2} + e^{-i(x+1)k_1 - i y k_2} + e^{-i x k_1 - i(y-1)k_2} + e^{-i x k_1 - i(y+1)k_2} - 4 e^{-i x k_1 - i y k_2} \right] G(x, y)
= \frac{1}{4} \sum_{x,y} \left[ e^{i k_1} + e^{-i k_1} + e^{i k_2} + e^{-i k_2} - 4 \right] e^{-i x k_1 - i y k_2} G(x, y)
= \frac{1}{2} [\cos k_1 + \cos k_2 - 2] \sum_{x,y} e^{-i x k_1 - i y k_2} G(x, y)
= \frac{1}{2} [\cos k_1 + \cos k_2 - 2] \hat{G}(k).
\tag{IV.38}
\]

Using this in Eq. (IV.37) gives

\[
\left[ \frac{1}{2} (\cos k_1 + \cos k_2 - 2) + \alpha \right] \hat{G}(k) = -1, \tag{IV.39}
\]

in other words

\[
\hat{G}(k) = \frac{1}{1 - \alpha - \frac{1}{2} [\cos k_1 + \cos k_2]]. \tag{IV.40}
\]

Using Eq. (IV.35) we then have

\[
G_\alpha(r) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{i k \cdot r} dk_1 dk_2 \frac{1}{1 - \alpha - \frac{1}{2} (\cos k_1 + \cos k_2)}. \tag{IV.41}
\]

Hence

\[
G_\alpha(0) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{dk_1 dk_2}{1 - \alpha - \frac{1}{2} (\cos k_1 + \cos k_2)}. \tag{IV.42}
\]

The integrand is symmetric with respect to \( k_1 \leftrightarrow -k_1 \) and \( k_2 \leftrightarrow -k_2 \) respectively, due
to the identity $\cos(k) = \cos(-k)$. Hence

$$G_\alpha(0) = \frac{1}{\pi^2} \int_0^\pi \int_0^\pi \frac{dk_1 dk_2}{1 - \alpha - \frac{1}{2}(\cos k_1 + \cos k_2)} \int_0^\pi \int_0^\pi \frac{dk_1 dk_2}{1 - \frac{1}{2(1 - \alpha)}(\cos k_1 + \cos k_2)}. \quad \text{(IV.43)}$$

This expression is related to the complete elliptic integral of the first kind, see for example Morita (1971). We use the following notation:

$$K(z) = \frac{1}{2\pi} \int_0^\pi du \int_0^\pi dv \frac{1}{1 - \frac{z}{2}(\cos u + \cos v)}. \quad \text{(IV.44)}$$

Hence we have

$$G_\alpha(0) = \frac{2}{\pi(1 - \alpha)} K\left( \frac{1}{1 - \alpha} \right) = \frac{2K_\alpha}{\pi(1 - \alpha)}, \quad \text{(IV.45)}$$

where we abbreviate

$$K_\alpha = K\left( \frac{1}{1 - \alpha} \right) \quad \text{(IV.46)}$$

for convenience.

Next, we write the Laplacian $\Delta G_\alpha(0)$ in full:

$$\Delta G_\alpha(0) = -G_\alpha(0) + \frac{1}{4} [G_\alpha(1, 0) + G_\alpha(-1, 0) + G_\alpha(0, 1) + G_\alpha(0, -1)]. \quad \text{(IV.47)}$$

Assuming isotropy, each term in the square brackets is equal to $G_\alpha(e_1) = G_\alpha(1, 0)$. Hence Eq. (IV.34), evaluated at $r = 0$, takes the form

$$G_\alpha(e_1) + (\alpha - 1)G_\alpha(0) = -1, \quad \text{(IV.48)}$$

in other words

$$G_\alpha(0) - G_\alpha(e_1) = 1 + \alpha G_\alpha(0). \quad \text{(IV.49)}$$

From this we find

$$1 - \frac{G_\alpha(e_1)}{G_\alpha(0)} = \left[ \alpha + \frac{1}{G_\alpha(0)} \right]. \quad \text{(IV.50)}$$

Now using Eqs. (IV.27) and (IV.31),

$$C(r) = m^2 + c(r) = m^2 + (1 - m^2) \frac{G_\alpha(r)}{G_\alpha(0)}. \quad \text{(IV.51)}$$
The stationary-state isogloss density \( \sigma \) is then

\[
\sigma = \frac{1}{2} \left[ 1 - C(e_1) \right]
= \frac{1}{2} \left[ 1 - m^2 - (1 - m^2) \frac{G_\alpha(e_1)}{G_\alpha(0)} \right]
= \frac{1}{2} \left( 1 - m^2 \right) \left[ 1 - \frac{G_\alpha(e_1)}{G_\alpha(0)} \right]
= \frac{1}{2} \left( 1 - m^2 \right) \left[ \alpha + \frac{1}{G_\alpha(0)} \right]
= \frac{1}{2} \left( 1 - m^2 \right) \left[ \alpha + \frac{\pi(1 - \alpha)}{2K_\alpha} \right],
\]

where we have used Eq. (IV.45). On the other hand,

\[
1 - m^2 = 1 - \frac{(p_I - p_E)^2}{(p_I + p_E)^2} = 4 \left( \frac{p_I}{p_I + p_E} \right) \left( \frac{p_E}{p_I + p_E} \right) = 4 \rho (1 - \rho),
\]

so that

\[
\sigma = 2 \rho (1 - \rho) \left[ \alpha + \frac{\pi(1 - \alpha)}{2K_\alpha} \right].
\]

Recalling that

\[
\alpha = -\tau = -\frac{(1 - q)(p_I + p_E)}{q},
\]

see Eq. (IV.26), and defining \( \tau = -\alpha \), i.e. \( K_\alpha = K_{\tau} \), yields our final equation for the stationary-state isogloss density:

\[
\sigma = 2H(\tau) \rho (1 - \rho),
\]

where

\[
H(\tau) = \frac{\pi(1 + \tau)}{2K \left( \frac{1}{1 + \tau} \right)} - \tau.
\]

The expression in Eq. (IV.56) is a downward-opening parabola in \( \rho \) whose height is fixed by \( H(\tau) \), where

\[
\tau = \frac{(1 - q)(p_I + p_E)}{q}.
\]

**Sanity check: the limits** \( \tau \to 0 \) and \( \tau \to \infty \). The complete elliptic integral \( K(z) \) has the following known properties (Fernow, 2016):

\[
\lim_{z \to 0} K(z) = \frac{\pi}{2} \quad \text{and} \quad \lim_{z \to 1} K(z) = \infty.
\]

Now consider the limit \( \tau \to 0 \). This limit is relevant when either \( q \to 1 \), or both \( p_E \to 0 \) and \( p_I \to 0 \), see Eq. (IV.58). These are situations in which the spatial (voter)
process dominates the ingress–egress dynamics. In this limit $1/(1 + \tau) \to 1$, so that $K_\tau = K(1/(1 + \tau)) \to \infty$. Noting that Eq. (IV.57) can be written in the form

$$H(\tau) = \tau \left[ \frac{\pi}{2K_\tau} - 1 \right] + \frac{\pi}{2K_\tau}.$$  \hspace{1cm} (IV.60)

we then find that $H(\tau) \to 0$. Thus, in the limit where the spatial (voter) process completely overtakes the ingress–egress process, the stationary-state isogloss density is zero, indicating that all sites agree in their spin.

Next consider $\tau \to \infty$. This limit occurs when $p_I + p_E > 0$ and $q \to 0$, i.e. the ingress–egress process dominates. Then $1/(1 + \tau) \to 0$, so that $K_\tau \to \pi/2$. From Eq. (IV.60), we then find that $H(\tau) \to 1$ in this case. Thus, in the limit where the ingress–egress process completely overtakes the spatial process, the stationary-state isogloss density is given by the parabola $\sigma = 2\rho(1 - \rho)$, indicating complete independence of the individual spins.

### IV.5.4 WALS feature levels

The following list gives the values of the WALS features mined; the italicized part after each value gives its value in our binarization (present for ‘feature present’, absent for ‘feature absent’ and N/A if the WALS value was excluded from the binarization as irrelevant). Languages attesting irrelevant values were excluded from the corresponding feature language sample for the purposes of calculating our statistics.

1. adpositions

   - WALS feature mined: ‘Person Marking on Adpositions’ (48A)

   - Values:

     1: No adpositions (absent)
     2: No person marking (present)
     3: Pronouns only (present)
     4: Pronouns and nouns (present)

2. definite article

   - WALS feature mined: ‘Definite Articles’ (37A)

   - Values:

     1: Definite word distinct from demonstrative (present)
2: Demonstrative word used as definite article (present)
3: Definite affix (present)
4: No definite, but indefinite article (absent)
5: No definite or indefinite article (absent)

3. hand and arm identical
   - WALS feature mined: ‘Hand and Arm’ (129A)
   - Values:
     1: Identical (present)
     2: Different (absent)

4. hand and finger(s) identical
   - WALS feature mined: ‘Finger and Hand’ (130A)
   - Values:
     1: Identical (present)
     2: Different (absent)

5. front rounded vowels
   - WALS feature mined: ‘Front Rounded Vowels’ (11A)
   - Values:
     1: None (absent)
     2: High and mid (present)
     3: High only (present)
     4: Mid only (present)

6. gender in independent personal pronouns
   - WALS feature mined: ‘Gender Distinctions in Independent Personal Pronouns’ (44A)
   - Values:
     1: In 3rd person + 1st and/or 2nd person (present)
2: 3rd person only, but also non-singular (present)
3: 3rd person singular only (present)
4: 1st or 2nd person but not 3rd (present)
5: 3rd person non-singular only (present)
6: No gender distinctions (absent)

7. glottalized consonants
   - WALS feature mined: ‘Glottalized Consonants’ (7A)
   - Values:
     1: No glottalized consonants (absent)
     2: Ejectives only (present)
     3: Implosives only (present)
     4: Glottalized resonants only (present)
     5: Ejectives and implosives (present)
     6: Ejectives and glottalized resonants (present)
     7: Implosives and glottalized resonants (present)
     8: Ejectives, implosives, and glottalized resonants (present)

8. grammatical evidentials
   - WALS feature mined: ‘Semantic Distinctions of Evidentiality’ (77A)
   - Values:
     1: No grammatical evidentials (absent)
     2: Indirect only (present)
     3: Direct and indirect (present)

9. indefinite article
   - WALS feature mined: ‘Indefinite Articles’ (38A)
   - Values:
     1: Indefinite word distinct from ‘one’ (present)
     2: Indefinite word same as ‘one’ (present)
3: Indefinite affix (*present*)
4: No indefinite, but definite article (*absent*)
5: No definite or indefinite article (*absent*)

10. inflectional morphology
   - WALS feature mined: ‘Prefixing vs. Suffixing in Inflectional Morphology’ (26A)
   - Values:
     1: Little affixation (*absent*)
     2: Strongly suffixing (*present*)
     3: Weakly suffixing (*present*)
     4: Equal prefixing and suffixing (*present*)
     5: Weakly prefixing (*present*)
     6: Strong prefixing (*present*)

11. inflectional optative
   - WALS feature mined: ‘The Optative’ (73A)
   - Values:
     1: Inflectional optative present (*present*)
     2: Inflectional optative absent (*absent*)

12. lateral consonants
   - WALS feature mined: ‘Lateral Consonants’ (8A)
   - Values:
     1: No laterals (*absent*)
     2: /l/, no obstruent laterals (*present*)
     3: Laterals, but no /l/, no obstruent laterals (*present*)
     4: /l/ and lateral obstruent (*present*)
     5: No /l/, but lateral obstruents (*present*)
13. morphological second-person imperative

- WALS feature mined: 'The Morphological Imperative' (70A)
- Values:
  1: Second singular and second plural (present)
  2: Second singular (present)
  3: Second plural (present)
  4: Second person number-neutral (present)
  5: No second-person imperatives (absent)

14. order of adjective and noun is AdjN

- WALS feature mined: ‘Order of Adjective and Noun’ (87A)
- Values:
  1: Adjective-Noun (present)
  2: Noun-Adjective (absent)
  3: No dominant order (N/A)
  4: Only internally-headed relative clauses (N/A)

15. order of degree word and adjective is DegAdj

- WALS feature mined: 'Order of Degree Word and Adjective' (91A)
- Values:
  1: Degree word-Adjective (present)
  2: Adjective-Degree word (absent)
  3: No dominant order (N/A)

16. order of genitive and noun is GenN

- WALS feature mined: 'Order of Genitive and Noun' (86A)
- Values:
  1: Genitive-Noun (present)
  2: Noun-Genitive (absent)
  3: No dominant order (N/A)
17. order of numeral and noun is NumN

- WALS feature mined: 'Order of Numeral and Noun' (89A)

- Values:

  1: Numeral-Noun (present)
  2: Noun-Numeral (absent)
  3: No dominant order (N/A)
  4: Numeral only modifies verb (N/A)

18. order of object and verb is OV

- WALS feature mined: 'Order of Object and Verb' (83A)

- Values:

  1: OV (present)
  2: VO (absent)
  3: No dominant order (N/A)

19. order of subject and verb is SV

- WALS feature mined: 'Order of Subject and Verb' (82A)

- Values:

  1: SV (present)
  2: VS (absent)
  3: No dominant order (N/A)

20. ordinal numerals

- WALS feature mined: 'Ordinal Numerals' (53A)

- Values:

  1: None (absent)
  2: One, two, three (present)
  3: First, two, three (present)
  4: One-th, two-th, three-th (present)
  5: First/one-th, two-th, three-th (present)
6: First, two-th, three-th (present)
7: First, second, three-th (present)
8: Various (present)

21. passive construction
   - WALS feature mined: 'Passive Constructions’ (107A)
   - Values:
     1: Present (present)
     2: Absent (absent)

22. plural
   - WALS feature mined: 'Coding of Nominal Plurality’ (33A)
   - Values:
     1: Plural prefix (present)
     2: Plural suffix (present)
     3: Plural stem change (present)
     4: Plural tone (present)
     5: Plural complete reduplication (present)
     6: Mixed morphological plural (present)
     7: Plural word (present)
     8: Plural clitic (present)
     9: No plural (absent)

23. possessive affixes
   - WALS feature mined: 'Position of Pronominal Possessive Affixes’ (57A)
   - Values:
     1: Possessive prefixes (present)
     2: Possessive suffixes (present)
     3: Prefixes and suffixes (present)
     4: No possessive affixes (absent)
24. postverbal negative morpheme

- WALS feature mined: 'Postverbal Negative Morphemes’ (143F)
- Values:
  1: VNeg (*present*)
  2: [V-Neg] (*present*)
  3: VNeg&[V-Neg] (*present*)
  4: None (*absent*)

25. preverbal negative morpheme

- WALS feature mined: 'Preverbal Negative Morphemes’ (143E)
- Values:
  1: NegV (*present*)
  2: [Neg-V] (*present*)
  3: NegV&[Neg-V] (*present*)
  4: None (*absent*)

26. productive reduplication

- WALS feature mined: 'Reduplication’ (27A)
- Values:
  1: Productive full and partial reduplication (*present*)
  2: Full reduplication only (*present*)
  3: No productive reduplication (*absent*)

27. question particle

- WALS feature mined: 'Position of Polar Question Particles’ (92A)
- Values:
  1: Initial (*present*)
  2: Final (*present*)
  3: Second position (*present*)
  4: Other position (*present*)
  5: In either of two positions (*present*)
  6: No question particle (*absent*)
28. shared encoding of nominal and locational predication

- WALS feature mined: 'Nominal and Locational Predication' (119A)
- Values:
  1: Different (*absent*)
  2: Identical (*present*)

29. tense-aspect inflection

- WALS feature mined: 'Position of Tense-Aspect Affixes' (69A)
- Values:
  1: Tense-aspect prefixes (*present*)
  2: Tense-aspect suffixes (*present*)
  3: Tense-aspect tone (*present*)
  4: Mixed type (*present*)
  5: No tense-aspect inflection (*absent*)

30. tone

- WALS feature mined: 'Tone' (13A)
- Values:
  1: No tones (*absent*)
  2: Simple tone system (*present*)
  3: Complex tone system (*present*)

31. uvular consonants

- WALS feature mined: 'Uvular Consonants' (6A)
- Values:
  1: None (*absent*)
  2: Uvular stops only (*present*)
  3: Uvular continuants only (*present*)
  4: Uvular stops and continuants (*present*)
32. velar nasal

- WALS feature mined: 'The Velar Nasal' (9A)

- Values:

  1: Initial velar nasal (present)
  2: No initial velar nasal (present)
  3: No velar nasal (absent)

33. verbal person marking

- WALS feature mined: 'Alignment of Verbal Person Marking' (100A)

- Values:

  1: Neutral (absent)
  2: Accusative (present)
  3: Ergative (present)
  4: Active (present)
  5: Hierarchical (present)
  6: Split (present)

34. voicing contrast

- WALS feature mined: 'Voicing in Plosives and Fricatives' (4A)

- Values:

  1: No voicing contrast (absent)
  2: In plosives alone (present)
  3: In fricatives alone (present)
  4: In both plosives and fricatives (present)

35. zero copula for predicate nominals

- WALS feature mined: 'Zero Copula for Predicate Nominals' (120A)

- Values:

  1: Impossible (absent)
  2: Possible (present)
Is there a quasigrmam equation?

Taking the multidimensional inter-generational difference equation of language change to the continuous-time limit reveals an unexpected analogy between models of language change and models of biological evolution. This helps to address a polemical discussion of whether or not language change is like biological evolution: differences in ontology aside, the two are similar to the extent that they are governed by similar equations.

7.1 Biological evolution and language change

Chapter 1 began with a simple thermodynamical analogy and a statement of a reductionist philosophy: just as statistical mechanics derives high-level thermodynamical regularities from lower-level particle ensembles, so we may hope a theory of language acquisition and language change to derive linguistic regularities from the behaviour of speakers and language communities. I will return to a revised consideration of this analogy in the following, closing chapter, in which we will find that the picture is in fact considerably more complicated than the simple analogy suggests. In the present chapter, I will however first spend a few pages in exploring another favourite analogy of recent years, that between language change and biological evolution. The prospect of establishing a link between these two domains has attracted both elated optimism (Croft, 2000; Nowak, 2006) and stark criticism (Andersen, 2006), and deserves to be explored for that reason alone. More importantly, however, we will see that whatever the ontological status of the analogy, the mathematics show certain interesting parallelisms between the two domains; with this, I hope to show that diachronic linguistics interfaces meaningfully with evolutionary biology, as long as the most obvious pitfalls are avoided.

A genome is (a possibly very long) sequence of the nucleotides adenine, thymine, cytosine and guanine; it defines an organism’s genotype, or genetic makeup. Abstract-
ing away from environmental factors, the genotype of an organism is what mostly determines the organism’s phenotype – its "shape, behavior, performance and any kind of ecological interaction" (Nowak, 2006, 30). The phenotype, in turn, has a fitness relative to the environment in which it lives, understood simply as the reproductive rate of this phenotype. Cutting corners, we may speak directly of the genotype’s fitness, and say that the genotype operates in a fitness landscape defined by its environment. A population of genotypes evolves by natural selection in the following sense: a process of replication places copies of already existing genomes in the population in proportion to the relative fitnesses of the different genotypes; a process of mutation produces variation within the population by transmogrifying specific points in genomes;¹ and a process of elimination removes existing genomes by some (constant) rate. Change in the structure of the population follows from the combination of all these processes: elimination ensures a struggle for survival, reproduction by differential fitness ensures an optimization to local ecological conditions, and mutation ensures the availability of variation on which selection can operate. The dynamics of this genetic competition may be extremely complicated in real life; progress, however, can be had if certain idealizing assumptions are made.

Specifically, assume that the population is infinite and that the fitness landscape is constant, so that the fitness of a given genotype is independent of what other genotypes are present in the population and at what frequencies. The competition of genomes may then be modelled by a set of \( n \) differential equations,

\[
\frac{dx_i}{dt} = \sum_{j=1}^{n} x_j f_j q_{ji} - x_i \phi,
\]  

one for each \( i = 1, \ldots, n \). Here, \( x_i \) is the relative abundance (frequency) of the \( i \)th genotype, \( f_i \) is its (constant) fitness, and \( q_{ji} \) gives the probability of a mutation from the \( j \)th genotype to the \( i \)th genotype. The quantity \( \phi \) is the average fitness in the entire population, defined as

\[
\phi = \sum_i x_i f_i.
\]  

The role of this normalization factor is to model the process of elimination by preserving population size. The interpretation of equation (7.1), then, is as follows: the increase in the abundance of the \( i \)th genotype is accounted for by mutation from other genotypes \( j \neq i \) \( (q_{ji}) \), weighted by their abundances \( (x_j) \) and fitnesses \( (f_j) \), as well as from replication of the \( i \)th genotype itself, subsumed as the \( x_i f_i q_{ii} \) term in

¹In most eukaryotes, additional phenotypic variation is produced by sexual recombination. As a full discussion of this intricate process would result in a disproportionate distraction, and as it would not change the basic analogy or disanalogy between the two domains of biological evolution and language change, in this section I restrict my attention to the evolution of asexually reproducing entities such as bacteria and viruses.
the summation. The rate of decrease in the abundance of the $i$th genotype, in turn, is a function of the entire population ($\psi$): in the context of limited resources or finite physical space, the population has a carrying capacity which blocks exponential replication and forces a ceiling.

Equation (7.1), known as the *quasispecies equation* (Eigen, McCaskill & Schuster, 1989; Nowak, 2006), is defined on the $n$-dimensional simplex $S_n$ of genotype frequencies, just like the inter-generational difference equation (17) of Paper II (equation (4.11) of Section 4.1, above) is defined on the $n$-dimensional simplex $S_n$ of grammar probabilities. It is then tempting to ask: what similarities exist between the two equations? Can the quasispecies equation — whose behaviour is well understood — be reinterpreted in a linguistically meaningful way? Here we run immediately into a number of problems: even though grammars (I-languages) may be straightforwardly thought of as sequences of Parameters and thereby analogized with genomic sequences, grammars are not copied from generation to generation the way genomes are; it is doubtful whether grammars have any kind of fitnesses to speak of; even if they do, these fitnesses would not normally be expected to be constant (as required by the quasispecies equation), but rather to depend on what other grammars are present in the population.

In an attempt to address the latter problem, Martin Nowak and colleagues in a sequence of publications (Komarova et al., 2001; Mitchener, 2003; Mitchener & Nowak, 2004; Nowak, 2006) have studied the *replicator–mutator equation*, a generalization of the quasispecies equation:

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} x_j f_j(x) q_{ji} - \phi x_i.$$  

(7.3)

This generalization allows for variable, population-dependent fitnesses: $f_j(x)$ is now some function of the population state $x = (x_1, \ldots, x_n)$, where $x_i$ is the frequency of grammar $G_i$. The increase in the frequency of the $i$th grammar is proportional to the abundance of the various grammars in the population ($x_j$), their "fitnesses" ($f_j(x)$) and the probability with which one grammar "mutates" into another one ($q_{ji}$). Mutation is interpreted, perhaps unproblematically, as erratic (not converging to the target) learning, but it is less clear how to think about the fitness of a grammar. Nowak (2006, 272) puts

$$f_i(x) = \sum_{j=1}^{n} h_{ij} + \frac{h_{ji}}{2} x_j,$$  

(7.4)

where $h_{ij}$ is the probability of grammar $G_i$ generating a sentence which is parsed by grammar $G_j$ — an inverse quantity vis-à-vis the advantage parameters $a_{ij}$ of Paper II
and Chapter 4. The intuition here is that grammar $G_i$ receives a payoff upon each successful communication with a mutually intelligible grammar $G_j$ — or rather, the payoff is the greater the more mutually intelligible the two grammars are. The dynamics of (7.3), then, are defined by two matrices, a parsability matrix $H = [h_{ij}]$ and a mutation matrix $Q = [q_{ij}]$.

While the proponents of this approach must be lauded for their innovative thinking and mathematical rigour, I submit that the underlying ontological commitments of the approach are suspect. Briefly, the problem is that while genomes are transmitted by a process of copying, grammars are not transmitted (modulo fitness and copying errors) but acquired or learnt or “grown” (Lightfoot, 1999) by children. Each human child has the entire set of humanly possible grammars as a search space, and it is not clear that the intricacies of language acquisition can be captured by a mutation matrix. What would be needed in order to install the replicator–mutator equation as a fixture of the theory of language change is, in fact, a bridging theory that derived the components of that equation from some restrictive, independently plausible theory of language acquisition.

It is unlikely that such a bridging theory is forthcoming, however, due to the way communicative fitness is interpreted. Nowak (2006) assumes that communicative fitness translates directly into reproductive success, and that these differences in reproductive rates are the motor of language change: “[w]e assume that payoff translates into reproductive success: individuals with a higher payoff produce more offspring” (Nowak, 2006, 272). This is dubious, since human reproductive success rarely depends on the setting of grammatical Parameters. Thus consider a population of speakers of a V2 language. By definition, a non-V2 “mutant” should not be able to invade such a population, since the mutant’s communicative fitness will be strictly lower than the communicative fitness of any V2 speaker. Yet in real life, a person’s reproductive rate depends less on things like the setting of the V2 Parameter than on many other — mostly extralinguistic — considerations. The reason why a certain grammar is favoured in a population cannot be because speakers who have internalized this grammar produce more offspring than do others; rather (as I will argue extensively below) a grammar may have a fitness in the looser sense that a learner is more likely to acquire this grammar than another grammar, given a specific linguistic environment. In translating communicative fitness into reproductive success, the replicator–mutator equation puts the cart — not before, but entirely beside the horse.

Table 7.1 pools together some similarities and differences between biological evolution and language change. In addition to the factors already mentioned, we may

---

2In Nowak’s (2006) notation these $h_{ij}$ are $a_{ij}$: I have changed the notation to avoid confusion with Yangian advantages which, as just pointed out, have to do with the generation of a sentence which the competing grammar fails to parse.
note a difference in the size of the state space: discounting viruses, genome lengths range from on the order of $1.5 \times 10^5$ base pairs in some bacteria (Nakabachi et al., 2006) to $1.5 \times 10^{11}$ base pairs in the flowering perennial *Paris japonica* (Pellicer, Fay & Leitch, 2010). This contrasts starkly with the linguistic domain, in which the number of Parameters — while not known with any certainty at the moment — must be orders of magnitude smaller (Longobardi & Guardiano, 2009, for example, suggest 63 binary Parameters for nominal syntax). A further disanalogy, arising directly from the different modes of transmission, has to do with the largest possible difference between the parent and the offspring: for genomes, this upper bound is small, as mutations are typically limited to a few point mutations in the sequence, whereas in language change the upper bound of the analogous difference between the child and its “cul-

<table>
<thead>
<tr>
<th>Biological evolution</th>
<th>Language change</th>
</tr>
</thead>
<tbody>
<tr>
<td>replicator</td>
<td>genotype</td>
</tr>
<tr>
<td>nature of replicator</td>
<td>discrete information-encoding structure</td>
</tr>
<tr>
<td>vehicle</td>
<td>organism</td>
</tr>
<tr>
<td>mode of transmission</td>
<td>reproduction</td>
</tr>
<tr>
<td>source of variation</td>
<td>random mutations, secondary stochastic factors (gene flow, genetic drift)</td>
</tr>
<tr>
<td>mode of selection</td>
<td>reproductive fitness</td>
</tr>
<tr>
<td>size of variation space</td>
<td>super-astronomical</td>
</tr>
<tr>
<td>upper bound on intergenerational difference</td>
<td>small</td>
</tr>
<tr>
<td>amenable to mathematical treatment?</td>
<td>yes</td>
</tr>
<tr>
<td>standard theory</td>
<td>population genetics, quasispecies theory, evolutionary game theory</td>
</tr>
</tbody>
</table>

Table 7.1
Similarities and differences between (asexual) biological evolution and language change. “Standard theory” is meant in Kuhn’s (1962) sense of normal science; a Kuhnian normal-scientific paradigm of language change arguably does not exist.
tural parents” is not known with any certainty, although it is known that in extreme cases such as in creolization (Bickerton, 1984; Arends, Muysken & Smith, 1995) there may be great discrepancy between the child’s input and output. Thus genomes take tightly constrained walks in extremely large spaces, while grammars sometimes make large leaps in much smaller spaces.

Considering the vast dissimilarities recorded in Table 7.1, the prospects of an evolution-inspired account of language change may appear rather bleak. I will nevertheless argue in the next section that the two fields are fundamentally connected, due in part to the fact that both deal with change in discrete information structures, and in part to the fact that the notion of fitness may, after all, be appropriated for use in diachronic linguistics in a way that rings true to the basic ontological commitments of linguistic theory.

7.2 Quasigrammars

The quasispecies of the quasispecies equation (7.1) should not be conflated with the ordinary notion of biological species. Defined originally in a molecular evolution setting (see Eigen et al., 1989; Biebricher & Eigen, 2006), a quasispecies is a collection of replicating entities which almost (quasi) form a kind (species); it is, essentially, a region in genomic sequence space, a collection of organisms with almost but not entirely identical genomes. Thus a population of viruses may be thought of as a quasispecies: stemming from a single master sequence by way of a series of mutations, each member of this population resembles each other member but also differs from it. In a popular article, Eigen (1993) puts it rather eloquently:

This region in sequence space can be visualized as a cloud with a center of gravity at the sequence from which all the mutations arose. It is a self-sustaining population of sequences that reproduce themselves imperfectly but well enough to retain a collective identity over time. Like a real cloud, it need not be symmetric, and its protrusions can reach far from the center because some mutations are more likely than others or may have higher survival values that allow them to produce more offspring. That cloud is a quasispecies. (Eigen, 1993, 45.)

If we think of grammars as analogous to genome sequences, then it seems that the notion of quasispecies also has a linguistic analogue: the everyday or “sociopolitical” (Hale, 2007, 9) notion of language. That is, languages are quasispecies of I-languages or, as I shall from now on say, languages are quasigrammars.

The point is not new — Lass (1997, 375ff.), to my knowledge, is the first to have observed this similarity. He gives the above quote from Eigen and goes on to explore some of the implications of the analogy, but in a non-mathematical way. The
question I wish to ask is whether the analogy can be put to actual, rather than just metaphorical, use.

In this endeavour, we have roughly two options vis-à-vis the nature of linguistic representation in the individual: either individuals internalize sequences of Parameter values (much like organisms “internalize” sequences of nucleotides), as per the classical tradition, or they internalize probabilities attached to these sequences, as per the Variational Learning framework. It turns out that the consequences of the choice between one or the other kind of theory may not be felt on the population level: for note that a population of Yangian individuals who use a particular grammatical option $X$ per cent of the time and its alternative $100 - X$ per cent of the time is equivalent, *qua population state*, to a population of classical individuals $X$ per cent of whom use the option and $100 - X$ per cent of whom use the alternative.\(^3\) Thus quasigrams, like quasispecies, live on a simplex of probabilities, frequencies or relative abundances (whichever term one wishes to use), whose dimension is given by the total number of humanly possible grammars.

To ask whether the quasispecies analogy can be put to actual use is, in effect, to ask whether the equivalent of equation (7.1) exists for language change. Is there a quasigrammar equation? A reasonable candidate is equation (4.11) (equation (17) of Paper II), reproduced here as equation (7.5):

$$p_i(t + 1) = \frac{\prod_{j \neq i} (A p(t))_j}{\sum_j \prod_{k \neq j} (A p(t))_k}.$$  

(7.5)

This difference equation gives the population-level probability of the $i$th grammar $G_i$ for generation $t + 1$ as a function of the population state $p$ in generation $t$. Recall that $A$ is a (time-independent) advantage matrix $A = [a_{ij}]$ such that $a_{ij}$ is the probability of a sentence parsed by $G_j$ but not by $G_i$, and that $c_j(t) = (A p(t))_j$ is the penalty probability of $G_j$. Taking account of the latter fact, and limiting ourselves to the simplex interior $int S_n$ so that division by the $p_i$ is possible, (7.5) may be expressed in the concise form

$$p_i(t + 1) = \frac{c_i(t)^{-1}}{\sum_j c_j(t)^{-1}}.$$  

(7.6)

by division of both the numerator and the denominator by $\prod_i c_i(t)$. Paper II studied the properties of this difference equation for a number of classes of advantage matrix. To facilitate comparison with the continuous-time quasispecies equation (7.1), I will now rewrite equation (7.6) as a differential equation. Writing $D = \sum_j c_j(t)^{-1}$ for the

---

\(^3\)Strictly speaking this equivalence only holds in the limit of infinite populations. Further discussion of the tension between classical and VL accounts of acquisition follows in Chapter 8.
denominator, the inter-generational difference is
\[ p_i(t + 1) - p_i(t) = \frac{1}{D} \left[ c_i(t)^{-1} - p_i(t) \sum_j c_j(t)^{-1} \right]. \] (7.7)

Assuming a time increment of arbitrary size \( \Delta t \), we have
\[ p_i(t + \Delta t) - p_i(t) = \Delta t \frac{1}{D} \left[ c_i(t)^{-1} - p_i(t) \sum_j c_j(t)^{-1} \right]. \] (7.8)

Taking the continuous-time limit \( \Delta t \to 0 \) and disregarding the prefactor \( 1/D \), whose value is independent of \( i \), we arrive at the differential equation
\[ \frac{dp_i}{dt} = c_i^{-1} - p_i \sum_j c_j^{-1} \] (7.9)
or, what amounts to the same thing,
\[ \frac{dp_i}{dt} = \left[ \sum_j a_{ij}p_j \right]^{-1} - \phi p_i, \] (7.10)
where
\[ \phi = \sum_j c_j^{-1} = \sum_j \frac{1}{\sum_k a_{jk}p_k}. \] (7.11)

Equation (7.10) bears some striking similarities to the quasispecies equation (7.1) (see Figure 7.1). The quasispecies equation gives the rate of change of the relative abundance of the \( i \)th genotype in a population of carriers of genotypes; equation (7.10) gives the rate of change in the probability of grammar \( G_i \) in a population of carriers of grammars. The quasispecies equation has an increase term and a decrease term; so does equation (7.10). In the quasispecies equation, the increase in the abundance of the \( i \)th genotype is a function of current abundances, fitnesses (reproduction rates) and mutation rates between different genotypes. In equation (7.10), the increase in the use of grammar \( G_i \) is identical to the inverse of that grammar’s penalty \( c_i = \sum_j a_{ij}p_j \), which itself depends on the current population composition \( p = (p_1, \ldots, p_n) \) and the advantage matrix \( A \); the smaller this penalty, the larger the increase and the “fitter” grammar \( G_i \) is, and conversely, the larger the penalty, the smaller the increase and the less fit \( G_i \) is (relative to the current environment). In the quasispecies equation, the decrease term is a function of the population state, namely, a function of the average fitness \( \phi \); in equation (7.10), the decrease term is similarly a function of the population state, namely, the prefactor \( \phi \) is the sum of inverse penalties across all grammars — in effect, an average grammatical fitness. Note that, these similarities notwithstanding,
equation (7.10) has neither a replication nor a mutation term — because such processes are not relevant to language change. The increase term in (7.10) accounts solely for the effects of language acquisition, which is analogous to biological replication and mutation in a loose sense only: crucially, this term does not include $p_i$ as a factor (since $a_{ii} = 0$ for any advantage matrix $A = [a_{ij}]$), unlike the quasispecies equation. Thus there is no sense in which individual copies of grammar $G_i$ are replicated as further copies of $G_i$. Rather, each new language learner explores the entire space of possible grammars and lands on a grammatical hypothesis based on the input he or she gets — or in this Yangian framework, based on the current population state $p$ (which is variable) and the advantage matrix $A$ (which is constant). This is exactly in line with the qualitative differences between language change and biological evolution noted in the previous section.

I conclude that equation (7.10) is a candidate quasigrammar equation. Importantly, unlike the replicator–mutator equation (7.3), which does not have independent linguistic motivation, equation (7.10) was derived from a theory of language acquisition — Variational Learning in multiple dimensions (see Paper II and Chapter 4, above). Like the replicator–mutator and quasispecies equations, this equation speaks of the fitnesses of competing replicators (grammars) — yet, crucially, these
fitnesses are not related to reproductive rates but rather to how likely it is for a language learner to hypothesize a given grammar in a given linguistic environment, a linguistically sensible reinterpretation of the notion of fitness. Like the replicator–mutator equation, but unlike the quasispecies equation, equation (7.10) makes these fitnesses frequency-dependent: the fitness landscape is not constant, but depends on what other grammars are present in the population. The consequence is that an exploration of the properties of the quasigrammar equation (7.10), or some extension thereof, will almost surely benefit from consideration of the techniques and results of evolutionary game theory (where fitnesses are likewise dependent on the population composition; see Hofbauer & Sigmund, 1998; Sandholm, 2010).

Will this approach prove fruitful in explaining how and why languages change? I believe so, and in the next chapter will give some immediate directions for future research. Before that some caveats are, however, in order. The quasispecies equation (7.1) does not cover all of quasispecies theory — recall the assumptions of an infinite population and a constant fitness landscape. Relaxing these assumptions, as well as adding more realistic components, adds significant mathematical complexity and may even overthrow some of the results that hold for equation (7.1). (The dynamics of the replicator–mutator equation, which allows non-constant fitness landscapes, are significantly richer than those of the quasispecies equation.) For example, it is now accepted in evolutionary game theory that deviations from the classical, continuous infinite-population framework may have qualitative, not just quantitative consequences, in the sense that the properties of games change when they are played in finite populations (Nowak & May, 1992), on networks (Lieberman, Hauert & Nowak, 2005), or by players who have memory limitations (Galla, 2009), and so on. Similarly, a candidate quasigrammar equation must not be interpreted as encompassing — or pretending to encompass — all there is to say about language change. Various “restrictions to the finite” would still need to be studied separately, as extensions of the basic, idealized, approximate equation, as would many contingent historical events such as linguistic bottlenecks and “speciation” created by population movements and other random extra-linguistic events, and so on. In this respect, language change is no different from biological evolution.

The epistemic role of a quasigrammar equation within the broader fabric of diachronic linguistics must then be no different from the epistemic role of the quasispecies equation in molecular biology: it is a well-behaved, relatively simple approximation — yet one derived from first principles of the domain of inquiry in question — which makes it possible to think clearly about complex dynamic processes occurring in evolving populations, and which acts as a basic explanatory ansatz which may be extended and refined in various ways.
8 Models and explanations

Language acquisition and language use constitute a feedback loop whose properties give rise to patterns of linguistic diachrony. Accounts of language change must, then, pay special attention to the interaction between acquisition and usage effects. This is best done by formulating mathematical models which make the assumed mechanisms explicit and derive as theorems quantitative predictions which may be tested against empirical data. Such models illustrate how covering-law-type explanation can be carried out in diachronic linguistics.

8.1 Acquisition, use and change

Figure 8.1 appropriates the Z-model (Figure 2.1) in a slightly simpler representation, retaining the essential idea: that language change reduces to language acquisition, the population level to the level of the individual. The discussion in the foregoing chapters and, especially, in the four Papers forming the core of this thesis, now invites us to update this simple diagram in a number of ways. Let us take a step back and see what we have learnt and what future directions this research opens up.

8.1.1 Production biases

Paper I introduced production biases as a mechanism for deriving Kroch’s (1989) Constant Rate Effect. These production biases were assumed (1) to apply after primary language acquisition, (2) to be tied to specific linguistic contexts and not apply to grammatical options directly, and (3) not to be learned (i.e., to be either innate or purely physiological). The production bias mechanism proposed in Paper I resulted in a model that derives the CRE — with this model, the CRE is no longer an unexplained statistical occurrence, but predicted from first principles. Specifically, the fact that the model was found to accord with a number of historical data sets both in
terms of overall fit and in terms of the predictions of the Time Separation Theorem (Theorem 3.1) lends empirical support to the hypothesis that production biases are in operation.

Yet there is no logical necessity for production biases either to be tied to particular contexts or to be non-learned. Plausibly, some such biases are neither, and these could, conceivably, explain phenomena such as sociolinguistically motivated S-curves, where a bias is usually needed (Blythe & Croft, 2012) but where that bias normally attaches to competing variants directly (and not to specific linguistic contexts) and clearly has to be learned rather than not. The two dimensions of context-specificity and learned/not learned give rise to a $2 \times 2$ cross-tabulation outlining four types of production biases overall (Table 8.1). What the empirical implications of the remaining two types are must be settled by future research — though it has already been suggested that non-learned biases targeting entire grammatical options

<table>
<thead>
<tr>
<th>context-specific</th>
<th>general</th>
</tr>
</thead>
<tbody>
<tr>
<td>learned</td>
<td>?</td>
</tr>
<tr>
<td>not learned</td>
<td>CREs</td>
</tr>
</tbody>
</table>

Table 8.1
Production biases are cross-categorized by two dimensions: context-specific vs. general and learned vs. not learned. On the interpretation of context-specific learned and general non-learned biases, see text.
Language change as inter-generational change: the inclusion of production biases.

Figure 8.2

May be responsible for certain typological universals (B. Clark et al., 2008; Culbertson, Smolensky & Legendre, 2012), and in the case of learned but context-specific biases we can at least speculate the appearance of something like a quasi-CRE, where different contexts show similar but not entirely parallel behaviour as is seen in true CREs.

It is now possible to update Figure 8.1 with the inclusion of production biases (Figure 8.2). If an amount of analogical licence is allowed, these production biases act as a prism on the beam of the underlying grammatical state, turning that state into actual usage frequencies observed in language production.\(^1\) It is crucially important to note that, diachronically, the production biases constitute a feedback loop with that underlying grammatical state, rather than operating independently and on top of it. This implies that production biases can have diachronic consequences, as they may shift the grammatical state in the following generation; this effect, moreover, can be amplified over iterated generational interactions so that the result is a sort of linguistic butterfly effect whereby an initially innocuous factor comes to change an entire diachronic trajectory. How this works was figured out in Paper I for context-specific non-learned biases in the Yangian framework; working out the dynamics of the other cases remains an important goal for future research.

The specific mechanism for production biases proposed in Paper I requires that

---

\(^1\)Unlike in the optical case, the contextual usage (given by the production biases) is in no way contained in the grammar itself. The analogy is useful at any rate, I believe, as it illustrates how production biases "project" the underlying grammatical state into context-specific usages.
where \( p \) is the probability attached to a grammatical Parameter and \( b_i \) is a contextual bias in context \( C_i \), the *de facto* usage of this grammatical Parameter in context \( C_i \) is given by

\[
p^{(i)} = p + b_i p (1 - p)
\]

(8.1)

where \(-1 \leq b_i \leq 1\). This choice was motivated by an *a priori* bound on possible forms of contextual modulation of the underlying grammatical probability, based on the requirement that \( p^{(i)} \) must, at all times, remain a probability (thus, bounded between 0 and 1 and adding up to unity with any other grammatical probabilities). As discussed in Paper I, equation (8.1) is not the only form of biasing consistent with these requirements, but it is the most parsimonious one, in the sense that it represents the lowest-degree smooth polynomial of \( p \) that satisfies the requirements. This leads me to propose the following conjecture for confirmation or refutation by future research:

**Conjecture 8.1 (The Lazy Biases Conjecture).**

Where \( p \) is a grammatical probability and \( b \) a production bias \((-1 \leq b \leq 1\)), the actual effect of \( b \) on \( p \) is

\[
p + bp(1 - p).
\]

(8.2)

Future research will be needed to test this conjecture and to explore the merits of the production bias hypothesis in general. Here it should be noted that a straightforward psycholinguistic experiment is available, at least in principle. Since the production biases in the framework of Paper I are bounded between \(-1\) and 1, the Lazy Biases Conjecture implies that the production of any individual speaker must be bounded by \( p - p(1 - p) \) and \( p + p(1 - p) \), so that the maximal vertical (synchronic) separation between two linguistic contexts is

\[
|p - p(1 - p) - p - p(1 - p)| = 2p(1 - p).
\]

(8.3)

Although the historical case studies considered in Paper I do not afford us any data at the level of individual speakers, identifying a suitable change in progress would make it possible to test the conjecture at the individual level by way of a simple elicitation experiment, by finding out whether or not each individual speaker’s production is contained within the separation (8.3).

### 8.1.2 Acquisition biases

Paper IV investigated spatial distributions of linguistic features across the globe, proposing that these distributions, as well as feature frequency distributions, are determined in part by constraints on learnability, reflected in the mathematical model as the ingress and egress probabilities of individual features. This suggests that Figure
8.2 should be further extended to cater for *acquisition biases*, biases operating during the process of language acquisition (Figure 8.3). Such biases are relevant to L2 as well as L1 acquisition — as discussed in Section 6.2, the instability and consequent scattered spatial distribution of inflectional systems of various kinds, for example, can be explained by assuming that L2 learners are biased against the acquisition of complex morphology.

### 8.1.3 Networks

Much work on the mathematical modelling of language change has operated in the infinite-population limit (e.g. Niyogi & Berwick, 1997; Yang, 2000). Paper III studied the effects of modelling a finite population instead — specifically, effects of social network clusterization. It was found that strong clustering in a dynamically evolving network of speakers may give rise to neutral change, or at least decrease the amount of bias necessary for an innovatory variant to overtake a conventional one. More generally, this work suggests that connectivity patterns in social networks may affect language dynamics in ways to which infinite-population modelling is entirely oblivious, and that future models of diachronic change must also attempt to model the varied sources of PLD that language acquirers have as members of social networks.
8.1.4 Quasigrammars

Paper II presented a generalization of the Variational Learner (Yang, 2002b) into multidimensional settings in which acquirers have evidence for more than two grammars in the parental population. Working in the limit of infinite populations of reliable learners, an inter-generational difference equation was derived that describes the evolution of the population-level vector of grammar probabilities. This equation was further taken to the continuous-time limit in Chapter 7 and a close connection with the biological quasispecies equation noted. There, echoing Lass (1997, 375ff.) but demonstrating for the first time a mathematical connection between the two formulations, I pointed out that much as biological quasispecies form a population of similar replicating information-encoding entities, languages may be fruitfully thought of as quasigrammars, i.e. as populations of similar information-encoding entities (I-languages) which reproduce (even though in a manner very different from biological
reproduction).

Specifically, for \( n \) grammars \( G_1, \ldots, G_n \), the quasigrammar equation reads

\[
\frac{dp_i}{dt} = \left[ \sum_j a_{ij}p_j \right]^{-1} - \phi p_i
\]

(8.4)

where \( p_i \) is the probability of grammar \( G_i \), \( a_{ij} \) is the relative parsing advantage of grammar \( G_j \) over grammar \( G_i \), and

\[
\phi = \sum_j \frac{1}{\sum_k a_{jk}p_k}
\]

(8.5)

is the average inverse parsing penalty in the population (each summation ranging from 1 to \( n \)). This equation meets two important desiderata: (1) it allows one to relate the population state of a language community to that state at a previous time, thereby obtaining diachronic trajectories, yet (2) the equation itself is derived from a model of first language acquisition at the level of the individual. In this way diachronic trajectories become understandable as systems of interactions between I-languages. In a sense, (8.4) is a linguistic “equation of motion”.

The quasigrammar equation in its present form does not incorporate production or acquisition biases, and future work should explore ways in which this may be done. Incorporating acquisition biases into the equation, for example, amounts to replacing each occurrence of penalty probability in equations (8.4)–(8.5) (i.e., both terms \( a_{ij}p_j \) and \( a_{jk}p_k \)) with a function \( F_j(p, A, \beta) \) or \( F_k(p, A, \beta) \), where \( A = [a_{ij}] \) is the advantage matrix and \( \beta = (\beta_1, \ldots, \beta_n) \) is a vector of acquisition biases somehow convolved with the vector of grammar probabilities \( p = (p_1, \ldots, p_n) \). Both mathematical modelling and psycholinguistic theory are required to establish a definite form for such functions \( F_j \).

8.1.5 Linguistics is a pre-Mendelian science

There is, then, cause for optimism in the mathematical modelling of language change: the past twenty years have witnessed, in fact, a gradual progression from simple proof-of-concept or toy models to more complex and realistic models with empirically testable predictions. But challenges remain. Perhaps the most pressing has to do with language acquisition — and, somewhat disappointingly, the predicament is not one of detail, but one of broad outlines. By this I refer to the tension between classical, non-probabilistic accounts of acquisition as Parameter setting (as exemplified by the classical TLA, for example) on the one hand, and explicitly probabilistic models of acquisition (such as VL, but also stochastic variants of Optimality Theory in phonology; Boersma, 1997) on the other.
The challenge arises directly from the reductionist philosophy which this dissertation has built upon: if change reduces to acquisition, then any tweak made to one’s theory of acquisition may result in a tweak to the resulting theory of change. The tension between classical models of acquisition and the VL framework is particularly acute, as the latter approach allows (in fact, requires) probabilities to constitute linguistic representation, or competence, to use the classical term — something that classical approaches explicitly deny. This implies that while the VL framework can easily incorporate a theory of usage effects, such as the production bias model put forward in Paper I, it simultaneously runs the risk of negating accepted wisdom on the competence–performance divide.

While I would not urge dropping pure research within these two modelling approaches pending further empirical evidence in favour of one or the other approach, it is clear that such evidence will be crucial in shaping the future of language acquisition and change research. Harking back to the biological analogy one last time, we could diagnose the present state of affairs as follows: linguistics is a pre-Mendelian (if post-Darwinian) science — even though it is now generally acknowledged that language change is a dynamical process, or even an evolutionary process (with the right provisos in place; see Chapter 7), our understanding of the all important subprocess of inter-generational transmission remains relatively poor.

8.1.6 Acquisition and use

The acquisition–use divide discussed in Chapter 2 turns out to be a false dichotomy: explaining complex processes of linguistic diachrony demands not only that we take both aspects seriously, but that both aspects be united in comprehensive models that make their interrelationships explicit. Acquisition and use constitute a diachronic feedback loop characteristic of complex systems: the former feeds into the latter, the latter into the former. The equations characterizing such feedback are usually nonlinear — this was also the case in Paper I, in which a full bifurcation analysis of the proposed model showed that production biases may sometimes undo the effects of acquisition. From an epistemological point of view, these nonlinearities are important, for they result in models that exhibit bifurcations — large-scale, qualitative, “behaviour-altering” responses to minute variation in control parameters. This may help to address the actuation problem (Weinreich, Labov & Herzog, 1968) — why a change is observed in situation or community A while situation/community B, even though largely similar to A, resists undergoing the same change process. In nonlinear systems, even tiny perturbations, resulting for instance from a changing bias tied to language use, may reverse the stability of fixed points and throw the system onto a qualitatively different trajectory.
8.1.7 Deterministic and stochastic effects

The difference and differential equations studied in Papers I and II as well as Chapter 7 are deterministic, in the sense that a population-level grammar probability vector $p(t)$ at some time $t$ completely determines the value of that vector at a later time $t' > t$, $p(t')$. This determinism resulted from taking both the population size and the length of the critical period to the infinite limit. The approach assumed in Papers III and IV, by contrast, was explicitly stochastic: in the former, the connectivity patterns of social networks are established in part at random, whilst in the latter, what event occurs (ingress, egress or spatial interaction) at any iteration of the model is established at random.

Since every natural system is finite and therefore susceptible to effects of stochasticity — in other words, since these effects must exist — one may ask what their import is. At present little work exists to address these issues — what is needed is systematic studies of how both acquisition at the individual level and diachrony at the population level are affected by (i) variation in the length of acquisition, (ii) variation in overall population size, (iii) variation in the number of speakers the learner is connected to, and (iv) variation in the topology of the social network of the speech community.

In Variational Learning of one Parameter, for instance, the learning rate and the length of the learning period interact to produce the eventual state attained by the learner. For any fixed length of learning (number of tokens of PLD heard) $T$, very small values of the learning rate $\gamma$ tend to produce a cautious learner whose state remains close to the initial state $\pi = 0.5$; intermediate values of $\gamma$ result in a learner who lands close to the value of $\pi$ expected from the deterministic theory of reliable learners; whilst high values of $\gamma$ result in a “stochastic” learner for whom $\pi$ attains a value close to either 0 or 1, depending on the vagaries of the stream of PLD (Figure 8.5). Thus for any value of $T$, a “sweet spot” of $\gamma$ exists that aligns the predictions of the infinite and finite models; future work is needed to establish whether values of $\gamma$ in this sweet spot are psychologically realistic.

8.2 Envoi: the role of mathematical models in explaining language change

Throughout this thesis, I have advertised the use of mathematical modelling as a viable technique for relating the level of the individual to the level of the population and for explaining processes of language change. I would like to conclude by defending, very briefly, my position that adopting a modelling approach improves the explanatory adequacy of diachronic linguistic theory.

It is not an unreasonable approximation to say that most of scientific explanation...
is well characterized by the classical deductive–nomological (DN) or covering law model of explanation (Hempel & Oppenheim, 1948): the logical deduction of statements describing what is to be explained (the explanandum) from a set of statements describing general laws and specific initial conditions (the explanans).\(^3\) As a straight-
forward extension of this model, a probabilistic explanation derives an explanandum, not with logical necessity, but with some probability, from an explanans at least some of whose laws are probabilistic (Hempel, 1965).

The DN model was originally developed with the natural sciences in mind, and its applicability in fields such as linguistics has been questioned. Most famously, Lass (1980) arrived at the pessimistic conclusion (later moderated in Lass, 1997) that there are no DN explanations in historical linguistics, and that probabilistic versions of these explanations do not, in fact, explain anything. Thus take his example of homorganic nasal+obstruent clusters (Lass, 1980, ch. 2): here, the explanandum concerns both the synchronic distribution of homorganicity as well as any given single change heterorganic > homorganic we may observe. Lass quickly dismisses the possibility of finding a DN explanation for the emergence of homorganic clusters, on the grounds that the relevant generalization (that homorganic clusters are easier to pronounce, i.e. "more natural", than heterorganic ones) does not have the status of a law.

The question then becomes whether a probabilistic explanation of the facts can be found. I quote verbatim (C = condition, L = law, E = explanandum; the double line signifies probabilistic inference as opposed to deduction, "makes likely" rather than "entails"): 

\[
\begin{align*}
C_1 & \text{ a sequence [nb]} \\
L_1 & \text{ [mb] is ‘easier than’ [nb]} \\
L_2 & \text{ it is (overwhelmingly, … etc.) probable that, given a choice, speakers} \\
& \text{ will choose ‘easier’ articulations over ‘harder’ ones} \\
E & \text{ [nb] > [mb]} \\
\end{align*}
\]

(Lass, 1980, 19)

Given this argument schema, neither the overall distribution of homorganic clusters nor any individual occurrence of the change [nb] > [mb] can be explained:

In what sense then can the notion of ‘ease’ be said to explain either a change [nb] > [mb] or the cross-language distribution of homorganic NC [nasal+obstruent – HK] clusters? My answer would be: in no sense. What we have [in the above-quoted argument schema – HK] is merely a semi-formalized paraphrase of an observed distribution, which incorporates a non-empirical (ultimately invulnerable) causal (or perhaps better ‘pseudo-causal’) hypothesis. The formalization produces an illusion of informativeness, but nothing is really ‘explained’: neither the distribution itself (which the probabilistic schema in fact brings to our attention as the real explanandum, a problem we skirt by invoking a ‘probabilistic law’), nor of course the particular instance we were trying to explain in the first place. Causality, for any instance or for the aggregate, is still as opaque as
it was before, and we are left only with what in any case we already know we had: a taxonomic generalization. (Lass, 1980, 21, emphasis in the original.)

That is, the distribution itself is not explained since the explanans is merely a restatement of the explanandum; particular instances of assimilation of the type [nb] > [mb] are not explained since, under probabilistic inference, they cannot be predicted.

Although the latter problem is surely insurmountable in that logic dictates the impossibility of deductive prediction of particular instances under probabilistic laws, there is reason to be less pessimistic about the prediction of distributions or, more generally, of general propensities (on this, see also Bermúdez-Otero, 2017). To illustrate this, the work in Paper IV represents some first steps towards a probabilistic explanation of precisely the kinds of distribution facts Lass (1980) is looking to explain: there, we saw that synchronic distributions (both frequency and spatial) are predicted from first principles of transmission and mutation, captured in the two dynamics of spatial interaction and ingress–egress. The crucial difference to Lass’s argument schema is that here the explanans is not simply a reformulation of the explanandum, but in fact a (reasonably complex) model, avoiding the problem of circularity. Distributions are predicted from combinations of model parameter values, and since the model is mathematical, a number of quantitative predictions (“if I tweak model parameter A in way X, then a change of amount Y is expected in distribution B”) are made available.

Similar remarks can be made about the production bias model of Paper I. This model is the first to show how an independently motivated theory of language acquisition may be wedded with an account of language use so that an empirical phenomenon (the Constant Rate Effect) emerges as a mathematical theorem of the combined model. In fact, the production bias model can be viewed as a classical (deterministic) DN explanation of the CRE in which the model gives the covering laws and the conditions are specified by the number and strength of the production biases, as well as by the parsing advantage ratio (or whichever other general bias is assumed to provide the thrust for the underlying change). This model, too, makes available predictions which are not just qualitative but quantitative: the Time Separation Theorem predicts a specific maximal temporal contextual separation from the initial conditions, and the biasing mechanism predicts propagation curves of specific shapes and offsets from those same initial conditions. The model has empirical content in the sense that these predictions can be evaluated against data.

In this case, our understanding of the explanandum (the CRE) is improved the more we can glimpse into the internal structure and operation of a model that meets the minimal explanatory requirement of predicting the correct empirical facts. None of this is to say that the model is “true”, just that it is an improvement over existing purely statistical treatments of the phenomenon, which describe but cannot claim to explain. To some extent, however, the fruitfulness of a model is measured not just
by its ability to explain existing explananda but also by its ability to suggest new ones, i.e. to present *new problems*. In the case of the production bias model of the CRE, one now has to ask: supposing that the CRE is an actual fact (thus, a genuine explanandum) and supposing that production biases derive that fact (thus, supposing that the model builders have not made mistakes in their derivations, and that the explanans concerning the production biases is a true statement), can anything explain those production biases? Can they be related to an independently plausible theory of psycholinguistics, for example?

While I do not have answers to these questions at the moment, the questions can now be asked; this, of course, is just another instance of the general regressive nature of explanation, of onetime explanantia becoming new explananda. The peculiar effectiveness of mathematical modelling in pushing the boundaries of ignorance lies, ultimately, in the fact that — since it forces us to be explicit and exact — mathematics expedites this process considerably.
References


Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of
Chicago Press.
vania Press.


Appendix: Some code

The following Fortran 90 program was used to run simulations of the lattice model put forward in Paper IV.

```fortran
! littlesq.f90
!
! Henri Kauhanen 2017
!
! Run some simulations of our "little squares" model on a regular lattice with periodic boundary conditions. This code runs such simulations for a given value of q (probability of voter event) and a given value of tau (temperature), for _RESOLUTION_ values of p_i and p_e. Output is in the form of a whitespace separated table.
!
! The program takes three command-line arguments: the first argument is tau, the second q, the third the filename for the outfile. Other parameters need to be set using preprocessor directives, see directly below.

! Some preprocessor directives here. If compiling with gfortran, don't forget to use the -cpp flag so that these get processed. In order of appearance: length of lattice side; multiplier M, to decide number of iterations, which will be N^2*M; length of the measurement phase L; rho and sigma will be calculated for the last L iterations; how many values of p_i (and p_e) to consider.
#define _N_ 100
#define _ITERMULT_ 10000
#define _MPLENGTH_ 100
#define _RESOLUTION_ 20

! Subroutine to return a random von Neumann neighbour of a lattice node subroutine randneigh(sx, sy, sN, snx, sny)
implicit none
integer, intent(in) :: sx, sy, sN
integer, intent(out) :: snx, sny
real :: randnob
!
! get random snx, sny
call random_number(randnob)
if (randnob <= 0.25) then
! north
snx = sx
sny = sy - 1
else if (randnob > 0.25 .and. randnob <= 0.50) then
! east
snx = sx + 1
sny = sy
else if (randnob > 0.50 .and. randnob <= 0.75) then
! south
snx = sx
sny = sy + 1
```
else
  ! west
  snx = sx - 1
  sny = sy
end if

! take care of lattice boundaries
if (snx < 1) then
  snx = sN
else if (snx > sN) then
  snx = 1
end if
if (sny < 1) then
  sny = sN
else if (sny > sN) then
  sny = 1
end if
end subroutine randneigh

! Subroutine to set the RNG seed from environmental noise; essential if
! this is run parallel on Condor
subroutine seed_from_urandom()
  implicit none
  integer :: i(12)
  open(99, file='\dev/urandom', access='stream', form='UNFORMATTED')
  read(99) i
  close(99)
  call random_seed(put=i)
end subroutine seed_from_urandom

! The program itself
program littlesq
  ! Declare variables
  implicit none
  character(len=32) :: comarg1, comarg2, comarg3 ! command-line arguments
  integer :: N, iter, mp_length, measurement_phase, resotop, nact, ups ! stuff
  integer :: i, j ! counters
  integer :: x, y ! lattice coordinates
  integer :: nx, ny ! lattice coordinates for neighbour
  real :: q, rho, sigma, pe, tau, intended_tau
  real, dimension(_RESOLUTION_) :: pin
  integer, dimension(_N_, _N_) :: spins
  real, dimension(_N_, _N_) :: rands

  ! Read command-line argument
  call getarg(1, comarg1)
  call getarg(2, comarg2)
  call getarg(3, comarg3)

  ! Initialize variables
  N = _N_
  nx = 1
  ny = 1
  iter = _ITERMULT_*N*N
  mp_length = _MPLENGTH_
  measurement_phase = iter - mp_length
resotop = _RESOLUTION_
read(comarg1, '*') intended_tau
read(comarg2, '*') q
pin = seq(real(0), (q/(1-q))*intended_tau, _RESOLUTION_)

! Open file connection, write csv header
open(1, file=comarg3)
write(1, '"pi" "pe" "q" "N" "ups" "nact" "rho" "sigma" "tau" "itau"

call seed_from_urandom()

! Loop through pi values, running simulation
do i = 1, resotop
  ! Initialize lattice
call random_number(rands)
do x = 1, N
do y = 1, N
  if (rands(x,y) <= 0.5) then
    spins(x,y) = -1
  else
    spins(x,y) = 1
  end if
end do
end do

! Set pe
pe = (q/(1-q))*'intended_taul - pin(i)

! Run simulation
nact = 0
ups = 0
do j = 1, iter
  ! Pick a random node of the lattice
call random_number(randno)
randno = randno*N
x = int(randno) + 1
call random_number(randno)
randno = randno*N
y = int(randno) + 1

! Which event to execute?
call random_number(randno)
call random_number(randno2)
if (randno <= q) then
  ! voter event: first, pick one von Neumann neighbour of (x,y)
call randneigh(x, y, N, nx, ny)
  ! then copy spin from (nx,ny) into (x,y)
  spins(x,y) = spins(nx,ny)
else
  if (spins(x,y) == 1) then
    if (randno2 <= pe) then
      ! egress event
      spins(x,y) = -1
    end if
  else
    if (randno2 <= pin(i)) then
      ! ingress event
      spins(x,y) = 1
    end if
  end if
end if
end if
end if
end if

! Find the proportion of active surfaces during the latter part of the run (measurement_phase)
if (j > measurement_phase) then
do x = 1,N
do y = 1,N
! East
nx = x - 1
ny = y
if (nx < 1) then
  nx = N
end if
if (spins(x,y) /= spins(nx,ny)) then
  nact = nact + 1
end if
! South
nx = x
ny = y - 1
if (ny < 1) then
  ny = N
end if
if (spins(x,y) /= spins(nx,ny)) then
  nact = nact + 1
end if
end do

! Find density of up spin
do x = 1,N
do y = 1,N
if (spins(x,y) == 1) then
  ups = ups + 1
end if
end do

! Calculate rho, sigma and tau
sigma = real(nact)/real(2*mp_length*N*N)
rho = real(ups)/real(mp_length*N*N)
tau = (((1-q)*(pin(i) + pe))/q

! Writeout
write(1,) pin(i), pe, q, N, ups, nact, rho, sigma, tau, intended_tau
end do

! Close file connection and goodbye
close(1)

! Helper functions
contains

! Returns a sequence of real numbers from 'from' to 'to' of length 'len'
pure function seq(from, to, len) result(arr)
  implicit none
real, intent(in) :: from, to
integer, intent(in) :: len
real, dimension(len) :: arr
integer :: i
arr(1) = from
do i = 2, len
    arr(i) = arr(i-1) + (to - from)/(len - 1)
end do
double precision function seq
end function seq
!

! Returns a logarithmic sequence from 'from' to 'to' of length 'len'
pure function logseq(from, to, len) result(arr)
implicit none
real, intent(in) :: from, to
integer, intent(in) :: len
real, dimension(len) :: arr
integer :: i
arr = seq(log(from), log(to), len)
do i = 1, len
    arr(i) = exp(arr(i))
end do
double precision end function logseq

double precision end program littlesq