THE UNIVERSITY *of* EDINBURGH

Stochastic Evolutionary Modelling of Language Change and Applications to Historical Data

Juan Guerrero Montero

Doctor of Philosophy
The University of Edinburgh
September 2024

Abstract

Language is a complex adaptive system whose properties emerge from a network of communicative interactions between speakers, modulated by their cognitive and social biases. In this conceptualisation, language change has a crucial role as the evolutionary process driving the adaptation of the language system. With the increasing availability of massive digital corpora datasets, stochastic evolutionary models of language change have the potential of providing empirical and quantitative explanations to the nature and origin of human language. In this thesis, I build on and develop such models in order to improve their applicability to historical data and their explanatory potential.

First, I address limitations in the application to historical language data of the Wright-Fisher model from population genetic. By improving on the Beta-with-Spikes approximation to the Wright-Fisher transition probability, I am able to more accurately discern drift from selection in time series of language use. I apply this method to the detection of a phonological bias in the evolution of the past tense of English verbs. I further introduce a methodology able to detect abrupt changes in social dynamics in time series of language use by modelling them as discontinuities in the selective forces acting on the data. I benchmark this method by applying it to the detection of well-documented historical spelling reforms in Spanish.

Secondly, I introduce an Iterated Bayesian Learning model for the evolution of grammatical structure through cultural transmission. Unlike previous evolutionary models of language change, this new model accounts for the co-evolution of interrelated linguistic functions. By accounting for nontrivial communicative effects, the model reproduces features of natural languages absent from simpler models of cultural transmission, like communicative effects in the arising stationary distribution of grammars, and broken detailed balance generating directionality in change. I explore directionality by applying entropy

production, a concept from non-equilibrium Statistical Mechanics, to a variety of models of language change.

Finally, I show that this model of the cultural transmission of grammatical structure is equivalent to a set of co-evolving and interdependent Wright-Fisher processes, facilitating its application to empirical data. Its application to the evolution of relativisers and to the emergence of the periphrastic use of *do* in English highlights the potential of evolutionary models as tools for the empirical and quantitative testing of hypotheses in historical linguistics.

Acknowledgements

I would like to thank the following people, who have all made it possible for me to be where I am today:

First, my principal supervisor Prof Richard Blythe. He has not only greatly influenced my academic thinking, but also been a role model on how to be a good mentor.

Also my second supervisor, Prof Kenny Smith, as well as my collaborators Dr Andres Karjus, Prof Rob Truswell and Dr Dan Lassiter, who have been greatly influential on my work.

The wonderful and brilliant people at the Centre for Language Evolution and the Physics department in Edinburgh, who are now my academic families.

My friends, who are my greatest support in Edinburgh.

And last, but definitely not least, thank you to my parents and sister, without whose inconditional support I would have never been able to be here today.

Widening the lay appreciation of this thesis, I am including Spanish translations of the Acknowledgements and Lay summary sections, titled *Agradecimientos* and *Resumen divulgativo*.

Agradecimientos

Querría agradecer a las siguientes personas, sin cuyo apoyo nunca habría acabado este doctorado:

Primero, a mi supervisor el Prof. Richard Blythe. No sólo ha influenciado mi pensamiento académico, sino que también ha sido un ejemplo de cómo ser un buen mentor.

También a mi segundo supervisor, el Prof. Kenny Smith, y mis colaboradores Dr. Andres Karjus, Prof. Rob Truswell y Dr. Dan Lassiter, que han influenciado mi trabajo profundamente.

A las maravillosas personas del Centre for Language Evolution y el departamento de Física en Edimburgo, que ahora son mis familias académicas.

A mis amigos, que son mi mayor apoyo en Edimburgo.

Y por último, pero no menos importante, gracias a mis padres y mi hermana, sin cuyo apoyo incondicional nunca habría llegado donde estoy hoy.

Lay summary

All human languages change over time. Children use language in ways that differ slightly from the way their parents speak, and even more so from the way their grandparents speak. This often includes their accents, the words they use, or even their grammars. These small inter-generational differences can accumulate over time and give rise to the big changes that led from the unintelligible Old English of *Beowulf* to Modern English, or from Latin into all Romance languages.

At a fundamental level, language change reflects the way human societies and the human mind work. Society may dictate, for example, that a specific pronunciation is more "correct", that a word is offensive, or that a new expression is trendy. The human mind, on the other hand, seems to prefer words, grammars, and pronunciations that are able to express the most information in the simplest ways, which makes us continuously shorten words and streamline grammatical structures. All of these factors affect language change, and in turn mean that understanding language change equals understanding how our societies and minds shape the most basic aspects of the human experience.

Mathematical models are one way that language change can be studied. As it turns out, evolutionary models – those used in biology to understand how organisms evolve over time – can be used for this. This is because languages and biological organisms have a lot more in common than we may initially think: words, vowels and consonants, and grammatical structures are to languages what genes are to biological organisms. Where biologists can use data obtained in the laboratory, we can use corpora – big collections of historical texts and books that give us information on how languages were used in the past.

In this thesis, I apply methods from Statistical Physics to improve on existing evolutionary models and developed new ones with the goal of better capturing properties of language change and better extracting information from corpora. This includes being able to detect changes in social trends by looking at the way language was used, and modelling the interconnection of the evolution of different grammatical structures. With this, I hope to provide linguists with high-quality mathematical and statistical tools to better formulate and test hypotheses about language change.

Resumen divulgativo

Todos los idiomas cambian a lo largo del tiempo. Los niños usan el lenguaje de forma ligeramente diferente de sus padres, y aún más diferente de sus abuelos. A menudo, esto incluye sus acentos, las palabras que usan, e incluso sus gramáticas. Estas pequeñas diferencias intergeneracionales se acumulan a lo largo del tiempo y generan los grandes cambios que llevaron del inglés antiguo de Beowulf al inglés moderno, o del latín a todas las lenguas romances.

A un nivel fundamental, el cambio de las lenguas refleja el funcionamiento the las sociedades y la mente humana. La sociedad puede dictar, por ejemplo, que una pronunciación específica es más "correcta", que una palabra es ofensiva, o que una nueva expresión está de moda. La mente, por otro lado, prefiere palabras, gramáticas y pronunciaciones capaces de expresar la mayor cantidad de información de la manera más sencilla posible, lo cual hace que acortemos palabras y optimicemos estructuras gramaticales. Todos estos factores afectan la evolución de los idiomas, y abren la puerta a entender cómo nuestras sociedades y nuestras mentes moldean los aspectos más básicos de la experiencia humana.

Una manera en la que la evolución de los idiomas puede ser estudiada es a través de modelos matemáticos, en particular modelos evolutivos (los que se usan en biología para entender cómo los organismos evolucionan a lo largo del tiempo). Esto se debe a que los idiomas y los organismos biológicos tienen mucho más en común de lo que uno podría esperar: las palabras, vocales y consonantes, y estructuras gramaticales son a los idiomas lo que los genes son a los organismos. Mientras que los biólogos usan datos obtenidos en el laboratorio, nosotros usamos corpus (colecciones de textos y libros históricos que nos dan información sobre el pasado de los idiomas).

En esta tesis, aplico técnicas de la física estadística para desarrollar y mejorar modelos evolutivos, con el objetivo de mejorar cómo capturan la evolución de los idiomas y cómo extraen información de los corpus. Esto incluye ser capaz de detectar cambios en las tendencias sociales, y modelar la manera en la que la evolución de diferentes estructuras gramaticales está interconectada. Con esto, espero proveer a los linguistas con herramientas matemáticas y estadísticas de alta calidad con las que puedan formular y poner a prueba hipótesis sobre la evolución de los idiomas.

Contents

List of Figures

List of Tables

Chapter 1

Background

Human languages are continuously undergoing change. Language change affects all levels of language structure, from phonology (sounds) to syntax (the rules that are used to combine words into sentences). The accumulation of change is responsible for the large scale alterations that lead to languages becoming unrecognisable over time or diversifying into daughter languages (e.g. the evolution of Old English into Modern English or of Latin into the modern Romance languages). Furthermore, language change can be grounded in theories on the nature and origin of language structure, and contribute to our understanding of it [1–3].

Stochastic modelling and statistical physics can provide us with crucial insight into the dynamics of language change. Firstly, its parallels to evolutionary processes in biology have allowed linguists to borrow well-established models and methods from that field [4, 5]. Secondly, unlike other cultural systems, language data is ubiquitous and easily quantifiable; this is especially true with the advent of digital historical corpora. This availability of data makes empirically applicable models invaluable towards our understanding of language change.

To this end, this thesis aims at developing stochastic models of language change and addressing prominent issues in their empirical application. I do so by introducing methodological advancements in the application of currently existing models, with the goal of making them more reliable and broadly applicable to linguistic data, and able to detect the evolutionary forces driving change. Furthermore, I develop new models that capture features of change absent in current models, like the interconnected evolution of distinct but interdependent

structures, as well as directionality and irreversibility.

In this chapter, I first present some of the fundamental questions and theoretical frameworks in linguistics. Within a usage-based framework, language can be thought of as a complex adaptive system, in which language change is an emergent adaptive phenomenon that can be conceptualised as a generalised evolutionary process. This evolutionary characterisation lays a solid conceptual foundation for the stochastic modelling of language change. I present and discuss two prominent models, iterated Bayesian learning and the Wright-Fisher model of variant competition. I finally discuss the challenges in applying these models to historical data, which will be addressed in the rest of this thesis.

1.1 Language is a complex adaptive system

In this section, I will be introducing some of the most fundamental questions in linguistics, two of the main theoretical paradigms developed in the last century to address them, and how one these theoretical paradigms, usage-based theories, may allow for the characterisation of language as a complex adaptive system. This will lay the foundation for the mathematical characterisation of language change and highlight its connection to relevant questions in the field.

Human language presents a complex structure that is absent or limited in other animal communication systems [6–8]. This structure encompasses features like the existence of a finite set of phonemic building blocks, syntactic categories like nouns and verbs, and compositionality; i.e. the ability to create sentences with complex meanings by combining the meanings of their component words together with syntactic rules [9, 10].

How exactly this structure and its governing rules manifest themselves may vary dramatically from language to language. However, there are constraints to this variation. Some of these constraints, known as universals, can be inferred from studying the distribution of properties in modern languages [11–13]. For example, declarative transitive sentences (i.e. those containing both a subject and a direct object, such as *"The examiners are reading the thesis"*) have dominant word orders where the subject precedes the object in around 96% of the world's languages [14]. From this, it can be inferred that there may be mechanisms shaping human language which favour word orders where the subject precedes

the object. Statistical tendencies are not only a feature of the synchronic level (the properties of languages at one point in time) but also of the diachronic level (languages' histories). The process of grammaticalisation, whereby content words like nouns and verbs lose lexical meaning over time and become function words like adpositions (e.g. prepositions such as *between* and postpositions such as *ago*) or affixes (e.g. suffixes such as ⟨-ing⟩ and prefixes such as ⟨un-⟩), is ubiquitous and believed to be driven by the same underlying cognitive processes in all languages [15, 16].

Explaining the origin of structure in human language, its distributional properties, and how these relate to other aspects of the human experience is one of the deepest questions in modern linguistics. Noam Chomsky addressed this question in his seminal 1965 work *Aspects of the Theory of Syntax* [17], which birthed a family of theories on the origin and nature of human language now known as nativism. Nativist theories posit that language is a genetically encoded entity whose fundamental structural properties, often referred to as Universal Grammar, arise from strong domain-specific cognitive constraints (i.e. facets of cognition that deal specifically with language). That is, the properties of human language are pre-determined by highly specific design features of our neurological wiring. Nativism argues that such a language organ is necessary to explain children's ability to acquire language effectively and quickly in spite of their limited exposure to it. This is often called the poverty of stimulus argument [18]. In nativism, language use in communication is often – but not always [19] – considered to be an evolutionary perk of a language capacity first developed for symbolic thought, and not a factor in the emergence and evolution of its structure in the first place [20].

Nativist theories were dominant for much of the 20th century after their introduction, but advances in linguistics and cognitive science in recent decades have questioned the extent to which their assumptions are necessary to explain language structure. Empirical evidence suggests that social interaction and domain-general cognitive mechanisms such as statistical learning are more crucial to language acquisition in children than predicted in early nativist accounts [21–23]. The poverty of stimulus argument has been questioned, suggesting that the language data children are exposed to may be sufficient to acquire grammar without necessitating a strong domain-specific device [24]. The development of machine learning has highlighted how neural networks can generate language-like systematicity without the need for a language-specific module [25, 26].

These findings have motivated the development of usage-based theories, which stand in opposition to nativist theories by defending a deep relation between the properties of language on one hand, and language use and the communicative needs of speakers on the other. In usage-based theories, individual language knowledge is acquired in social interaction and modulated through domain-general cognitive abilities such as categorisation [27], sequential processing [28], memory [29], and statistical learning [30]. Linguistic knowledge in the speaker's mind consists of a network of associations between linguistic signs and their meanings that is highly sensitive to exposure to others' language use in communication [31, 32]. Language structure then is argued to have emerged from its use as a social communicative tool, streamlined by communicative pressures and sociocognitive abilities such as theory of mind and collective intentionality [33–35]. There is, thus, no need to posit a strong domain-specific language cognitive constraint to explain the origin of language structure.

This emphasis on communication and interaction facilitates the characterisation of language as a complex adaptive system. Since the late 20th century, complex adaptive systems have been a central topic in complexity theory [36, 37], in part due to their applicability in a plethora of interdisciplinary contexts, ranging from ecology [38] to economics [39] and sociocultural dynamics [40, 41]. In their most abstract characterisation, complex adaptive systems are composed of populations of interactive entities, also known as agents. They are complex in that the global behaviour of the system is emergent from the network of agent interactions in ways that cannot be predicted by understanding only the individual behaviour of the agents. They are adaptive in that agents modulate their behaviour by reacting to changes in their local environment in ways that can bring forth changes to the global properties of the system, without the need of a centralised governing entity [42, 43]. Complex adaptive systems (CAS) are thus a powerful explanatory tool that allows us to understand how individual dynamics beget collective properties, and how these collective properties evolve and perpetuate themselves.

With this in mind, characterising language as a CAS at the sociocultural level can provide invaluable insight into the ways in which language learning and use at the individual scale affect the properties of language at the cultural scale [1, 44]. In the CAS paradigm, individual language users are the agents in a complex social network. In it, their individual language learning and use is driven by local interactions with other speakers, and reactive to changes in their environment. Linguistic structure and conventionalised language at the level of the speech

community are then emergent phenomena, driven by cultural transmission and modulated by individuals' communicative needs, sociocognitive abilities, and learning biases. CAS computational models have provided explanatory accounts of the emergence of community-wide linguistic conventions in vowel repertoires [45], vocabularies [46], and syntactic features [47]. Similar approaches have also showed that competing pressures for maximally simple and expressive languages are sufficient for the emergence of compositionality without the need for innate constraints [48]. In all, CAS approaches have shown that communication and cultural transmission have a much greater role in the development of grammar and linguistic structure than initially argued by nativist theories [49–51]

The CAS paradigm can also be extended to the characterisation of individual linguistic knowledge. Usage-based theories, particularly in the framework of Construction Grammar [52–54], argue that the internal representation of language in an individual's mind, also known as their grammar, is made up of a complex network of constructions. Constructions are associations between linguistic expressions (morphemes, lexical items, collocations, syntactic structures, etc) and their corresponding functions (meanings or communicative roles). The network of constructions is adaptively reshaped during communication. Grammar could thus be modelled as a CAS with individual constructions fulfilling the role of the agents in the network. Interaction with the environment occurs through the production and reception of instances of language use, known as *utterances*, which may trigger adaptation and growth of the grammatical network. The reception of a novel signal-meaning pair can cause a new construction to appear in the speaker's grammar. If this signal-meaning pair is used frequently, its associated construction can become entrenched. This can lead to complex interactions in the network, such as the weakening or displacement of other constructions fulfilling similar communicative roles. Language acquisition and further changes during a speaker's lifetime then result from the adaptive behaviour of their grammatical network, modulated by frequency tracking of utterances in their environment [55–57]. Figure 1.1 illustrates the inference process giving rise to constructions from exposure to utterances, and provides an example of a network of closely associated constructions.

Complex adaptive systems provide linguists with tools to generate explanatory accounts of language structure in the usage-based paradigm. Language is an example of a nested CAS [58]: the emergent phenomena in the lower, cognitive layer (grammar, language learning and production) are the agents and local

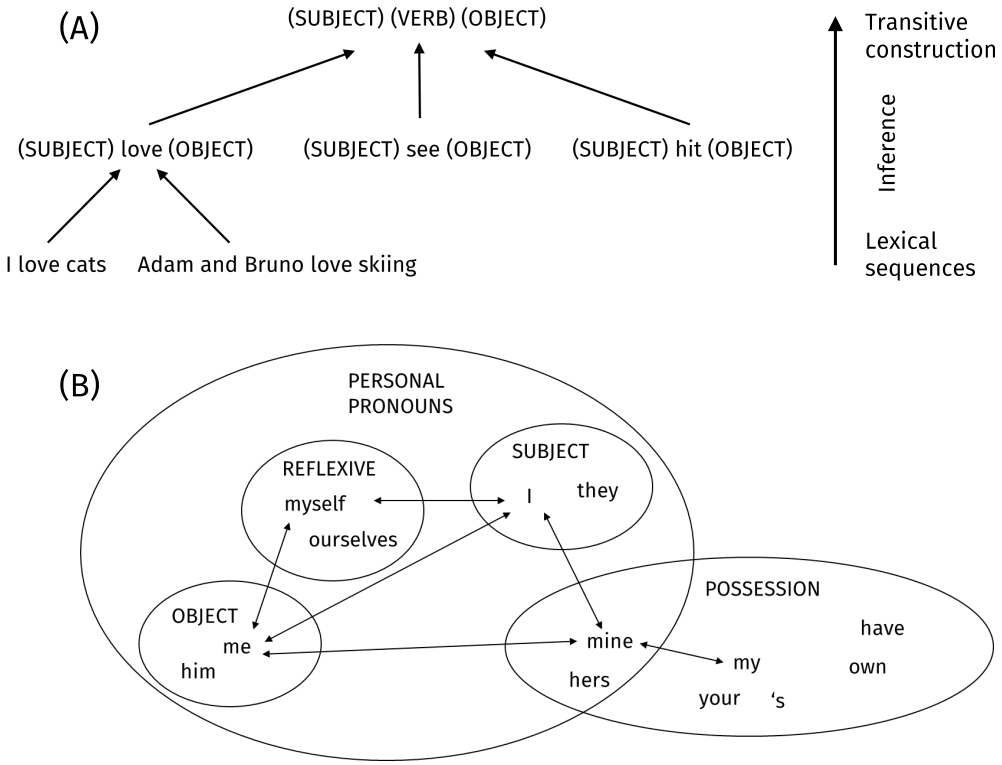

Figure 1.1 *(A) Schematic representation of the inference process of a construction, the formation of transitive sentences in English, through exposure to lexical sequences exemplifying said construction. Adapted from [57]. (B) Example of a network of associations of closely related constructions. Constructions in this example can associate based on grammatical person (arrows), syntactic categories (personal pronouns and within them subject pronouns, reflexive pronouns, object pronouns), or broader syntactic and semantic relations (words expressing possession).*

interactions that make up the basic units of the upper, social layer. Language change through cultural transmission is the key mechanism connecting these local interactions to the emergence of global language structure and distributional universals. Understanding language change thus equals understanding the adaptation and self-regulation of the language system, and provides crucial insight into the social, communicative and cognitive mechanisms causing the emergence of its global properties. Mathematical techniques from statistical physics and nonlinear dynamical systems are key in making this understanding quantitative and empirically testable. As a first step towards this, in the following section, I will delve into language change and present theories that model it qualitatively as a generalised evolutionary process.

1.2 Language change is an evolutionary process

In the past millennium, the English language has lost rich verbal and nominal inflectional systems, adopted thousands of French words after the Norman invasion, and overhauled its vowel system in what is now known as the Great Vowel Shift [59]. These and other diachronic processes, rendering the Late Old English of 1000CE utterly incomprehensible to an average speaker of Modern English, result from the accumulation of a multitude of small-scale processes of innovation in form and meaning, and subsequent competition between new and old forms [60, 61]. These affect all levels of language structure: a phoneme may start being pronounced in a different way; a word may gain a new meaning; a new word order may start being used.

Similar accumulations of diachronic processes are behind all large-scale processes of historical language change, including the diversification of a language into dialects and of those into mutually unintelligible languages. The parallels between these large-scale processes in language and in the evolution and diversification of species were already noticed by Darwin and his contemporaries [62, 63]. However, explorations of the similarities between the smaller underlying scales, i.e. genetic drift in population genetics and competition between linguistic forms in language change, are only a few decades old. A particularly rigorous and thorough characterisation of these parallelisms was introduced in William Croft's 2000 book *Explaining Language Change: An Evolutionary Approach* [64]. Following in the steps of Dawkins [65] and Hull [66], who applied evolutionary concepts to culture and scientific progress, respectively, Croft maps elements of language change to those of a generalised evolutionary theory. These include replicators and replication, alternative replicators and altered replication, interactors, and selection. A schematic representation of these concepts in the context of the evolution of a microorganism are shown in Figure 1.2. These, together with their corresponding instantiations in language change, will be discussed below.

Replicators are the basic units of structure that are perpetuated in time through a replication process. In population genetics, genes are the uncontroversial replicating unit (if epigenetic factors are not considered), and the replication process takes place during cell division or sexual reproduction. For example, when an *E. Coli* bacterium undergoes mitosis, it creates two copies of each of its genes. In Croft's work, he names the basic replicating unit in language change *lingueme*. Coined after Dawkins' memes, linguemes are units of linguistic function whose

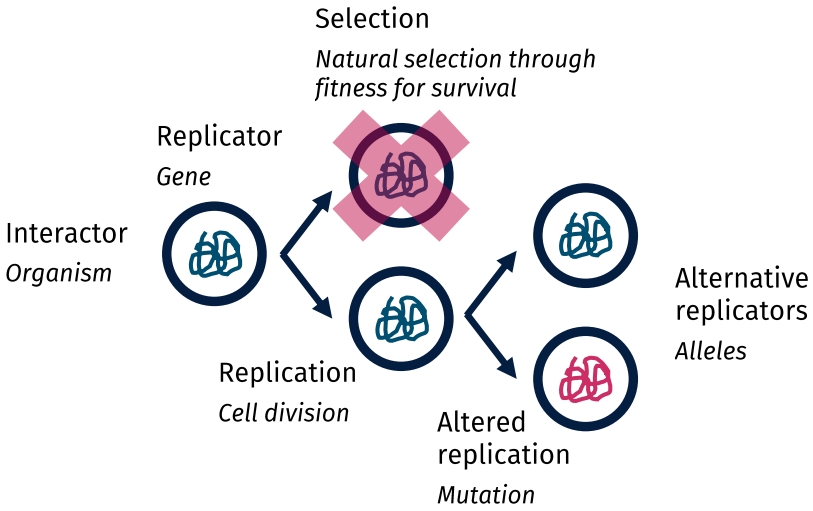

Figure 1.2 *Elements in a generalised evolutionary theory, together with their instantiations in the evolutionary dynamics of a microorganism.*

value can range from the phonological to the morphosyntactic to the semantic. Highlighting the importance of communication in usage-based theories, Croft argues that linguemes are encoded in utterances. A sentence like *"I missed the bus"* encodes, on top of its overall meaning, a plethora of linguistic features such as its word order (Subject Object Verb), the meanings of individual lexical items like *miss* and *bus*, the formation of the past tense of a verb (by adding a suffix ⟨-ed⟩, pronounced /-t/), the conveyance of shared information with the listener (through the definite article *the*), and each of the phonemes that make up the utterance. Each one of these features is a lingueme. A language is then just the pool of all utterances produced by a speech community, equivalent to the gene pool of a population. In this conceptualisation, the replication process happens during communication, whereby the listener incorporates the linguistic content of the linguemes contained in an utterance into their grammar. The listener will then go on to use their updated grammar to produce further lingueme-encoding utterances.

Alternative replicators are each of the possible forms that a replicator can take. New alternative replicators are generated through a process of altered replication. In population genetics, these are the alleles of a gene, with new alleles generated through mutation during cell division. In language, alternative replicators are the variants of a specific lingueme, each one of its possible realisations in the speech community. In the previous sentence, *"I missed the bus"*, the lingueme encoding the phoneme represented by the ⟨u⟩ grapheme has a multitude of realisations

depending on the dialect of the speaker. A typical Southern English speaker will produce it as an unrounded /ʌ/, while a typical Northern speaker will produce it as a rounded /ʊ/ [67]. These are thus two possible variants of the lingueme representing that vowel.

Altered replication takes the form of innovation, which can follow a wide variety of mechanisms ranging from the phonological to the semantic levels (see e.g. [68] for a comprehensive review). Whatever its origin, innovation in language has to do with the introduction of new words or expressions, or the extension of pre-existing ones to new contexts. New forms of lexical items can be generated through phonological alterations like metathesis (the interchange of sounds within a word), or cluster reduction (the simplification of clusters of consonants). Metathesis is responsible for the evolution of French *fromage* "cheese" from Vulgar Latin *formaticus* "formed in a mould". Cluster reduction systematically simplified the /kn/ cluster into /n/ in English words like *knight* and *know*. New variants for a given lingueme can be borrowed through language contact, like the Spanish *izquierda* "left", borrowed from Basque *ezkerra*, which has now replaced the archaic *siniestra*, of Latin origin. Morphological innovation often takes the form of analogy, where an inflectional pattern is extended to words it initially did not apply to. For example, the past tense of *help*, *helped*, arose initially in analogy to the regular paradigm, eventually replacing the original irregular form *holp*. Gradual semantic change through reanalysis is responsible for many common forms of grammatical innovation, including grammaticalisation [16]. A commonly attested process of grammaticalisation involves the innovation of adpositions from words originally referring to body parts. In Daasanach, an Ethiopian Cushitic language, examples of this include *sugu* "behind", originally meaning "back", or *géere* "inside", originally meaning "belly" [69]. Social factors, like the desire to be noticed, amusing, or charming, can also lead to innovation [70].

When alternatives of the same replicator coexist, the evolution of the relative frequency of each of them can be modelled as a competition process, where those that replicate more effectively increase in relative frequency over time. Interactors are the agents whose interaction with their environment causes replication to take place at different rates for different alternative replicators, thus affecting their evolutionary success. The evolutionary forces causing this differential replication are called selection. In biology, interactors are organisms, whose survival and success in reproducing determines whether their genes are passed down to the next generation. Selection occurs when alleles alter their organism's fitness to

Element	Biological evolution	Language change
Interactor	Organism	Speaker and grammar
Replicator	Gene	Lingueme
Alternative replicator	Allele	Variant
Replication	Reproduction	Communication
Altered replication	Mutation	Innovation
Selection	Fitness for survival	Social and cognitive biases

Table 1.1 *Elements of an evolutionary system in generalised evolutionary theory, together with their instantiations in biological evolution and language change. Table adapted from [64].*

survive and multiply in its ecological environment. Organisms with beneficial alleles will reproduce more effectively, increasing the relative proportion of the beneficial allele in the population over generations. In language, the interactors are the speakers in the speech community, together with their grammars which they acquire from language data in their communicative environment and use to produce utterances. While the speakers' survival does not (generally) depend on their language use, their ability to modulate their language use in different social contexts – together with their cognitive biases and articulatory limitations – generates effective selective pressures favouring the use of specific variants over others. For example, shorter or simpler variants tend to be preferred as they minimise production effort [71]. Societal pressures like prestige, taboo, and identity signaling may favour or disfavour specific variants in certain contexts [60, 72, 73].

The evolutionary process of language change can thus be summarised as follows: speakers in a speech community use their grammars to produce utterances encoding linguemes. For each lingueme, different variants may coexist in the speech community. New variants may arise through a variety of innovation processes. Different variants of the same lingueme compete against each other for usage in the speech community. Selective forces of social, cognitive, or linguistic origin may bias the competition process in favour or certain variants by virtue of making speakers differentially likely to use and acquire them, but stochastic effects always play a role in this process. As language is culturally transmitted through communication, the competition process may culminate with one of them taking over as the solely used variant of a lingueme. The elements of language change as an evolutionary process according to Croft's theory can be found in table 1.1. These elements will form the basic features of the stochastic models of language change to be introduced in the following section.

Others have put forward different mappings of linguistic concepts to those in generalised evolutionary theory. In particular, some theories of language change argue that linguistic replicators take the form of cognitive units in the speaker's grammar, like Ritt's competence constituents [74] and Zehentner's constructions [75] (which explicitly link Construction Grammar and evolutionary theories). While they are similar to Croft's linguemes in that they encode basic linguistic features, cognitive replicators shift the focus away from utterances, as they are part of the speaker's internal language. This makes this characterisation less useful for empirical approaches to language change, as a person's grammar is not measurable, only their language use is. However, grammar-focused and utterance-focused approaches are not incompatible, but rather complementary usage-based characterisations of the same process.

1.3 Stochastic models of language change

The modelling of language change through cultural transmission as an evolutionary process of innovation and competition between variants has created a solid theoretical foundation for its mathematical characterisation, which I will be discussing in the next section. Models of iterated learning – those that focus on the inter-generational chain of transmission of the language system – are better suited to describe changes in the grammar of speakers and how distributional properties emerge from the cultural transmission process. Models of competition focus on changes in the frequency of use of variants and the evolutionary forces that drive historical change. Chiefly, these approaches are not only qualitatively complementary, but have also been proven mathematically equivalent. These models will form the basis of the stochastic characterisation of change and empirical methodologies that will be developed in the rest of this thesis.

1.3.1 Models of grammar change: Iterated Bayesian learning

Cultural transmission is the key mechanism driving language change [50]. In it, new speakers acquire grammar through exposure to the utterances of other speakers. Mathematical models allow for the controlled testing of how different features of human cognition, language use and society affect the outcome of this

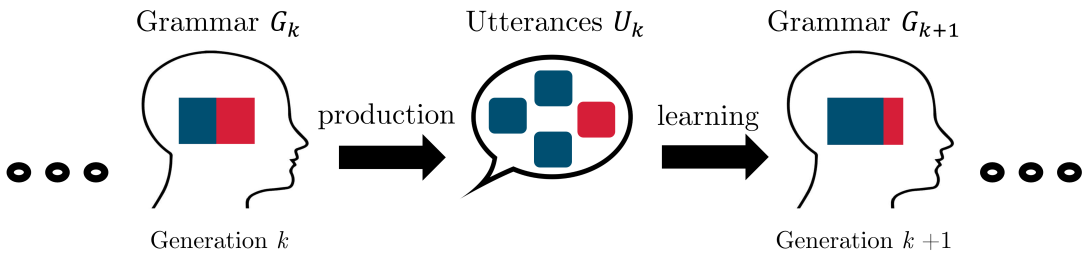

Figure 1.3 *Schematic representation of the simplest implementation of the iterated learning model for a language containing two forms, represented by the colours blue and red. A grammar consists of frequencies with which each form is expected by the speaker to appear in the language. These frequencies are represented here by colour bars, with the length of each colour representing its expected relative frequency. Utterances are chosen between the two forms by sampling from the grammar. In a given generation k, a speaker uses their grammar G_k to produce utterances U_k, which are then used by the speaker in generation $k+1$ to infer the grammar G_{k+1} through a learning process. This process is iterated indefinitely.*

process, leading to interesting insight into the origins of language properties. For more than two decades now, iterated learning has been one of the most prolific mathematical and computational models of the cultural transmission of language [48, 76, 77]. This paradigm has been diversified into a variety of models which capture different aspects of cultural transmission with varying degrees of complexity [78–81]. In its most basic implementation, the temporal complexity of the cultural process is discretised as non-overlapping generations. In each generation, the speech community is further stripped of its social complexity, leaving just a model speaker representing the average speaker in the community. Each generation has one speaker that exists only in that generation. The speaker has a grammar that details the way in which a set of linguistic forms can be used to produce a set of meanings. Here, I will be focusing on the case where grammars simply detail the frequency with which speakers expect a set of forms to be used. Inter-generational transmission occurs when the speaker in generation k produces a set of utterances in which they choose amongst the available forms based on their grammar. These are then used by the speaker of generation $k+1$ to infer the grammar of the language that the utterances were produced with. Cultural transmission is thus reduced to a linear concatenation of learning and production processes taking place at each generation. A schematic representation of this model for the simplest case of a language with two forms appears in figure 1.3.

The production process commonly involves an unbiased sampling from the

speaker's grammar, where a form is produced as frequently as the grammar of the speaker assumes it is in the language. However, it can also involve a number of effects including production biases that favour a form over others beyond what the grammar initially implies, and innovation rates that generate novel forms.

The learning process hinges on the choice of a learning strategy for the model speaker. The most common choice is Bayesian learning, due to empirical evidence that it accurately models human probabilistic learning [82, 83] and to its mathematical features stemming from its connection to Bayes theorem [84]. When this learning strategy is used, the iterated learning algorithm is commonly referred to as iterated Bayesian learning (IBL), first introduced by Griffiths and Kalish [85]. In Bayesian learning, the speaker infers the probability of a grammar G from the utterances of the previous generation U using Bayes theorem:

$$\mathrm{P}\left(G|U\right) = \frac{\mathrm{P}\left(U|G\right)\Pi_0\left(G\right)}{\sum_G \mathrm{P}\left(U|G\right)\Pi_0\left(G\right)}. \tag{1.1}$$

Here, $\mathrm{P}\left(G|U\right)$ is the posterior probability, representing the probability with which the speaker believes that grammar G generated the given data. In many models, the speaker chooses their own grammar by sampling from this posterior probability. $\mathrm{P}\left(U|G\right)$ represents the likelihood of the utterances U under the assumption that grammar G is the one that generated them. Finally, $\Pi_0\left(G\right)$ is the prior probability of G, representing the learning biases of the speaker. A uniform prior would assign the same probability to all grammars. Other priors can be chosen to represent a speaker's preference for simpler forms, or regular, predictable mappings between forms and functions, for example. This is the crucial component of the IBL paradigm, as it allows for the exploration of the effects of different learning biases on the distributional properties and evolutionary dynamics of grammars arising from the cultural transmission process.

With this, the probability that a speaker will infer grammar G' after being exposed to data produced by the previous generation using grammar G defines a Markov process with the following transition probability:

$$\mathrm{P}\left(G'|G\right) = \sum_U \mathrm{P}\left(G'|U\right)\mathrm{P}\left(U|G\right), \tag{1.2}$$

where the transition was expanded using the law of total probability in terms of the possible utterances of the speaker in the previous generation, and the

production probability was assumed to equal the likelihood.

It can be shown that in the simplest case, this transition probability and the learning prior distribution of the speakers satisfy the detailed balance condition:

$$\begin{aligned}
\mathrm{P}\left(G'|G\right)\Pi_0\left(G\right) =& \sum_U \mathrm{P}\left(G'|U\right)\mathrm{P}\left(U|G\right)\Pi_0\left(G\right) \\
=& \frac{\sum_U \mathrm{P}\left(U|G'\right)\Pi_0\left(G'\right)\mathrm{P}\left(U|G\right)\Pi_0\left(G\right)}{\sum_{G''}\mathrm{P}\left(U|G''\right)\mathrm{P}(G'')} \\
=& \sum_U \mathrm{P}\left(G|U\right)\mathrm{P}\left(U|G'\right)\Pi_0\left(G'\right) \\
=& \mathrm{P}\left(G|G'\right)\Pi_0\left(G'\right). \qquad (1.3)
\end{aligned}$$

This condition implies that, assuming that the distribution of languages follows the prior, any trajectory of change would be reversible: the probability of starting with a grammar G' and arriving at a grammar G would be the same as starting with G and arriving at G'. This has strong implications for the statistics of the process. First, it means that the prior probability is a stationary distribution of the Markov chain. Assuming aperiodicity (i.e. that trajectories in the chain do not show cyclic behaviour) and irreducibility (i.e. that every grammar G' can be reached from every other grammar G after a finite number of generations), which generally hold in IBL, this also implies that the process will eventually converge to this distribution [86], meaning that the grammars obtained after an arbitrarily long time from an ensemble of IBL chains with the same prior will follow the prior distribution as if they had been sampled from it, independently of the starting conditions of each of the chains. This is a questionable feature, as it predicts that the distributional features of language are uniquely determined by our learning biases – and not by social and communicative pressures, as usage-based theories would predict.

Several studies have addressed the issue of convergence to the prior by proving that it no longer holds under a variety of more complex realisations of the IBL paradigm. These include modifications to the way that speakers choose a grammar from their posterior probability [78], more complex population structures where speakers learn from more than one independent source [79], grammars that model the prevalence of meanings in the world – and not just the way they are associated with signs [80] – and pragmatic communication where speakers make active efforts to avoid ambiguity [81].

The second statistical implication of equation 1.3 is that the Markov process is reversible, meaning that any inter-generational trajectory in grammar space in the stationary state is equally as likely in the forward and backward directions. This would imply that language change is not directional, which is patently not the case. Grammaticalisation, for example, occurs overwhelmingly in the direction of loss of lexical meaning, and hardly ever in the direction of loss of grammatical meaning [15]. Notably, previous modelling efforts have not directly addressed reversibility with the same vehemence as convergence to the prior, in spite of their connection.

1.3.2 Models of variant competition: Wright-Fisher

As previously discussed, language change can be construed as a competition process between lingueme variants whose outcome is determined by evolutionary forces like selection and variation, together with random stochastic drift. The Wright-Fisher model from population genetics is the fundamental stochastic model of competition dynamics [87–89], and thus an invaluable tool in quantifying the effects of these evolutionary forces on language change that will be central to the rest of this thesis. In it, a population of N individuals is assumed to evolve in discrete generations, where the traits of the population in generation $t+1$, are sampled from those in the population in generation t. Assuming K mutually exclusive traits with counts x_k in the population, the state of the population in generation t is determined by the set $X_t = \{X_{t,k} : k = 1, ..., K\}$ with $0 \leq X_{t,k} \leq N$ and $\sum_k X_{t,k} = N$. The transition probability between traits X_t in generation t and traits X_{t+1} in generation $t+1$ follows a multinomial distribution given by

$$\mathrm{P}\left(X_{t+1}|X_t\right) = \frac{N!}{\prod_k X_{t+1,k}!} \prod_k g_k\left(X_t\right)^{X_{t+1,k}}, \tag{1.4}$$

where g_k represents the fitness function of trait k, the frequency with which trait k is expected to be found in generation $t+1$. Fitness functions satisfy $\sum_k g_k(X) = 1$ for all X.

The fitness function thus parametrises evolutionary forces affecting the competition process. An unbiased system will be represented by fitness functions determined only by the frequency of their associated trait in the previous generation, whereby our expectation to find a trait in generation $t+1$ equals

its incidence in generation t:

$$g_k\left(X_t\right) = x_{t,k}, \tag{1.5}$$

where the normalised frequency $x_{t,k} = X_{t,k}/N$ was introduced. Note that the normalised frequencies satisfy $0 \leq x_{t,k} \leq 1$, $\sum_k x_{t,k} = 1$

Variation, the ability for traits to spontaneously transform, will be reflected in the introduction of variation parameters ϵ_{kj} representing the rate with which trait k will become trait j during inter-generational transmission:

$$g_k\left(X_t\right) = x_{t,k}\left(1 - \sum_j \epsilon_{kj}\right) + \sum_j \epsilon_{jk} x_{t,j}. \tag{1.6}$$

Note that when variation is not trait-dependent ($\epsilon_{kj} = \epsilon/K$), this reduces to

$$g_k\left(X_t\right) = x_{t,k}\left(1 - \epsilon\right) + \frac{\epsilon}{K}. \tag{1.7}$$

Finally, a system with selection, the systematic biasing of the competition process in favour of specific traits, will be best represented using selection parameters s_k quantifying the strength of the bias in favour of trait k. A positive s_k will represent a bias favouring trait k, while a negative s_k will represent a bias against k:

$$g_k\left(X_t\right) = \frac{x_{t,k}\,\mathrm{e}^{s_k}}{x_{t,k}\,\mathrm{e}^{s_k} + 1 - x_{t,k}}. \tag{1.8}$$

Note that all nontrivial fitness functions reduce to the unbiased one, eq. 1.5, when their evolutionary parameters are set to 0. A system evolving under this fitness condition will still experience fluctuations in the relative frequency of the traits over time. These unbiased fluctuations are termed stochastic drift. While drift may refer to directional change in linguistics following Sapir (1921) [9], I use it here in the cultural evolutionary sense, denoting unbiased stochastic change. This effect is always present, independently of the presence of other biasing evolutionary forces. The population size parameter N is linked to the strength of drift; low N will magnify its effects and give rise to greater deviations from the expected trajectory, while high N will result with higher likelihood of dynamics with lower standard deviation from the expected value of the transition.

1.3.3 Equivalence between cultural transmission and variant competition

This parametrisation of evolutionary forces facilitates the quantitative characterisation of diachronic phenomena of competition, at the price of moving away from internal grammatical characterisations typical of the more linguistically justified models of cultural transmission. This does not mean that models of competition cannot be grounded in cultural transmission. This is what Reali and Griffiths set out to do in their 2010 paper *Words as Alleles*, where they proved the equivalence of the simplest instantiation of the IBL model of cultural trasmission presented in Section 1.3.1 and the Wright-Fisher model [4]. I will now proceed to summarise their mathematical argument here.

Following the IBL paradigm, assume that a speaker's grammar consists of an estimation of the relative frequency θ_k with which they expect variant k of the lingueme to be used in the language, for $k = 1, ..., K$, where K is the total number of different lingueme variants in the language. A grammar G is then determined by the choice of the set of relative frequencies $\Theta = \{\theta_k : k = 1, ..., K\}$, with $0 \leq \theta_k \leq 1$ and $\sum_k \theta_k = 1$. Further assume a symmetric Dirichlet prior distribution

$$\mathrm{P}(\Theta; \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \theta_k^{\alpha/K-1} \tag{1.9}$$

where α is a parameter determining the shape of the prior. $0 < \alpha < K$ is the typical choice, representing a bias for regularity, where speakers prefer languages where one variant is dominant.

In this characterisation, a set U containing L utterances of the lingueme is determined by the total production u_k of each of the variants, $U = \{u_k : k = 1, ..., K\}$ with $0 \leq u_k \leq L$ and $\sum_k u_k = L$. Given grammatical frequencies Θ, and assuming unbiased production by the speaker, the production probability and likelihood of a set of utterances is then given by a multinomial distribution

$$\mathrm{P}(U|\Theta) = \frac{L!}{\prod_k u_k!} \prod_{k=1}^{K} \theta_k^{u_k} \,. \tag{1.10}$$

When a speaker in the following generation acquires the grammar (i.e. chooses Θ) after being exposed to utterances U, the posterior probability over Θ will be

given by another Dirichlet distribution:

$$\mathrm{P}\left(\Theta|U\right) = \frac{\Gamma(\alpha + L)}{\prod_k \Gamma(\alpha/K + u_k)} \prod_{k=1}^{K} \theta_k^{u_k + \alpha/K - 1}, \tag{1.11}$$

where eq. 1.1 was used together with eqs. 1.9 and 1.10. Note that this is another Dirichlet distribution with modified parameters, due to it being the conjugate prior of multinomial likelihoods [90].

The exact choice of Θ that a speaker makes is empirically unavailable to us. However, for a speech community, Reali and Griffiths assume that their global language use can be approximated as having been produced by the average $\hat{\Theta}$ of the posterior in eq. 1.11. For each θ_k, this is given by

$$\hat{\theta}_k\left(U\right) = \frac{u_k + \alpha/K}{\alpha + L} \tag{1.12}$$

. When considered together with the production probability eq. 1.10, the utterance transition probability can be reasonably approximated as

$$\mathrm{P}\left(U_{t+1}|U_t\right) = \frac{L!}{\prod_k u_{t+1,k}!} \prod_{k=1}^{K} \hat{\theta}_k\left(U_t\right)^{u_{t+1,k}} \ . \tag{1.13}$$

By identifying $U \equiv X$, $L \equiv N$ and $\alpha \equiv N\epsilon/\left(1-\epsilon\right)$, this is equivalent to a Wright-Fisher process (cf. eq. 1.4) evolving under the effects of uniform variation using the fitness function in eq. 1.7.

The Wright-Fisher model for the competition of linguistic variants is thus derivable from simple models of cultural transmission. This is not just a quirk of the IBL paradigm. The Fokker-Planck equation resulting from the diffusion limit of the Wright-Fisher model has been derived from the utterance selection model, a mathematical realisation of William Croft's evolutionary ideas incorporating a more complex network of speakers [5]. This makes the Wright-Fisher model a natural choice for the study of competition processes in historical language data, particularly in the form of time series of variant frequencies extracted from corpus data.

1.4 Applying competition models to historical data

Collections of historical texts and other forms of language use, also known as corpora, are the fundamental source of data in historical linguistics. This field's inference tools, including the comparative method, the formulation of sound change laws, and language contact analysis, have been vital in informing our understanding of the mechanisms behind language change in historical timescales [68, 91]. Statistical and computational models have the potential to complement these detailed qualitative insights with ways of making their predictions quantifiably and reproducibly testable.

In the context of competition between linguistic variants, the Wright-Fisher model is able to achieve this by allowing for the formulation of hypotheses in terms of the explicitly parametrised evolutionary forces in its fitness functions (eqs. 1.5–1.8). These hypotheses can then be tested by finding the parameter values that maximise the likelihood that the model generates a relevant time series of variant frequencies. For example, social biases favouring one of the variants in a data set may be modelled as selection (eq. 1.8); the strength of these biases may then be measured by finding the optimal value of s in this fitness function, and the validity of this hypothesis may be quantified as a p-value where this model is compared against a simpler null hypothesis.

While promising, the empirical application of the Wright-Fisher model comes with some issues. Some of its problems stem from the interpretation of its parameters. The drift parameter N has a straightforward interpretation as the population size in analyses of allele frequency data in population genetics. Its interpretation in the context of language change is not nearly as clear. Work with agent-based models has found correlations between N and the size of the speech community [5, 92]. However, heterogeneous social network structures can result in N correlating only weakly with the size of the human population [92–94]. The drift parameter N is also related to the total usage of all variants of a linguistic variable in a given generation, and how long it is retained in memory [4, 5]. Intuitively, speakers will be more consistent in their usage of specific variants the more they encounter them, and the longer they recall these encounters. This increased individual consistency will be reflected as smaller fluctuations in time series data. The Wright-Fisher model does not allow these different contributions to N to be distinguished. This is further compounded by sampling error and evolutionary stochastic drift producing similar fluctuations in the dynamics of the data, leading

to time series built using a bigger corpus generally being assigned a higher N.

Another issue stems from the limitations of the model for computational analyses. The Wright-Fisher transition probability (eq. 1.4) can become computationally prohibitive when there are inter-generational gaps in the frequency data. Several approximations have been developed to address this issue in mathematical biology, but they usually become inaccurate close to extreme values of allele frequencies [95]. These issues are even more acute in language change, where extreme variant frequencies and insufficient data leading to inter-generational gaps are especially common. In Chapter 2, I present an approximation of the Wright-Fisher transition probability, based on Tataru et al.'s Beta-with-Spikes approximation [96], that alleviates these issues. I then proceed to benchmark this methodology by applying it to the analysis of time series of language data previously analysed using other approximations of the Wright-Fisher model. I argue that this methodology outperforms previous techniques for the detection of selection in time series of frequency data. I further analyse a novel dataset of English verbs, being able to detect a phonological bias in the competition between their regular and irregular past tense forms. The methodology and results presented in this chapter were published in [97] and [98].

Further issues of current competition models stem from their assumptions about the stability of evolutionary forces. In population genetics, given a stable ecological environment (easily achievable in the laboratory), selection can be assumed to affect each allele of interest in the same way throughout the evolution of the population. In speech communities, the social attitudes towards specific linguistic variants can vary over time [72], meaning that time-independent selection parameters may not be able to adequately capture the evolutionary dynamics of language change. In Chapter 3, I introduce a technique for the detection of drastic changes in evolutionary parameters in time series of language and cultural data. I benchmark this technique by detecting well-documented historical spelling reforms in Spanish. Just like Chapter 2, this chapter is based on results published in [97] and [98].

The Wright-Fisher model further assumes linguemes whose variants compete against each other in isolation from the influence of other linguemes. However, usage-based theories like Construction Grammar assume that linguemes, in the form of constructions, exist in an interconnected network in human cognition, with changes in one construction triggering changes in other constructions through local interactions in the speaker's grammar [56]. These interactions have

consequences at the scale of language change, with effects like morphological analogy only being explainable if interactions between linguemes are taken into account. In Chapter 4, I introduce a new model of iterated Bayesian learning where these interactions between multiple linguemes and variants are taken into account. By introducing an extra *understanding* phase to the inter-generational transmission, the model is able to generate stationary distributions that incorporate not only learning biases, but also communicative effects. Detailed balance is also broken, meaning that the model is able to produce some directional behaviour. This behaviour is explored quantitatively through the application of entropy production, a tool from non-equilibrium statistical physics. In Chapter 5 I further show that, unlike other IBL models that are able to go beyond convergence to the prior, this model remains equivalent to a modified Wright-Fisher model. This enables its application to time series of historical change involving multiple co-evolving linguemes and variants. I illustrate this through applications to two historical data sets, proving the potential of this type of model in the testing of hypotheses in historical linguistics.

Chapter 2

Reliable detection and quantification of selective forces in historical language data

As established in the previous chapter, understanding language change as a process of competition between linguistic variants allows for its modelling in terms of evolutionary forces such as selection, variation, and drift. Hypotheses about the nature of diachronic phenomena can then be distilled into explicitly parametrised effects in evolutionary stochastic models, and tested by restricting the values of said parameters using relevant linguistic data sets. Among all parametrisable evolutionary forces, selection that favours specific variants is particularly interesting, as it has been shown to be the most likely mechanism behind the typical S-shaped curve of language change in time series of historical data [99], and can be interpreted in terms of linguistic [60] and social [72] factors. However, current empirical approaches suffer from issues that limit their applicability to the detection and quantification of selection in historical data sets. This chapter aims at addressing several of these issues by building on current methodologies, and demonstrating their applicability through applications to relevant processes of language change.

The long history of stochastic models of competition in population genetics allows us to inherit a wealth of methods and insight from that field. A first step towards the detection and quantification of selection requires eliminating the possibility that the observed change may be explained by a neutral model [100], that is, one

that appeals to genetic drift operating in the absence of selection. To this end, a variety of classical statistical tests were developed (see e.g. [101] for a review) to detect departure from predictions of the neutral theory. Traditionally, these are based on quantities that can be ascertained from a sample of genetic material taken from a population at a single time, such as the number of nucleotide differences.

Time series, i.e. collections of measures of variant frequencies at different points in time, enable greater inferential power and more precise parameter estimation than samples collected at one point in time only. In biological research, these are usually obtained from evolution experiments conducted in the laboratory [102]. Advances in high-throughput sequencing technologies admit the collection of large genetic datasets [103], facilitating the sampling of microbial populations at multiple points in time. In linguistics, time series are typically built by quantifying the relative frequency of usage of variants of interest at different times in relevant historical corpora.

As discussed in Section 1.3.2, the Wright-Fisher model is a natural choice for the mathematical characterisation of competition processes at multiple points in time, able to explicitly parametrise evolutionary forces like drift, selection, and mutation. For any time series of interest, these parameters can be estimated as those that maximise the likelihood that the Wright-Fisher model generates the data. The computation of the likelihood involves the transition probability between each pair of contiguous frequency data points in the time series. The Wright-Fisher transition probability over a single generation was previously introduced for a system with K variants in equation 1.4. In this chapter, I will be focusing on systems with two variants, in which case the transition probability over a single generation reduces to a binomial distribution:

$$\mathrm{P}_{\mathrm{WF}}\left(x_{t+1}|x_t; N, \Theta\right) = \binom{N}{Nx_{t+1}} g\left(x_t; \Theta\right)^{Nx_{t+1}} \left(1 - g\left(x_t; \Theta\right)\right)^{N(1-x_{t+1})}, \qquad (2.1)$$

where x_{t+1} and x_t are the frequencies of the variant in generations $t+1$ and t, respectively; N is the population size, also known as drift parameter; and $g\left(\cdot\,; \Theta\right)$ is the fitness function, which incorporates evolutionary effects through parameters Θ. Computing this transition probability directly may not always be feasible, as it may prove impractical when the population size of the system is large or also being numerically estimated [95]. This is especially true when contiguous measurements of variant frequency are not one generation apart, in

which case equation 2.1 has to be summed over multiple intermediate generations to obtain the appropriate transition probability. This numerical intractability is a crucial problem in the empirical application of the Wright-Fisher model, and a key issue to be addressed in this chapter.

When attempting to detect deviations from a model of neutral drift, some methods overcome these difficulties by forgoing the computation of the likelihood altogether. The Frequency Increment Test (FIT) [104], applied to time series of language data in Newberry et al. (2017) [105], does so by approximating the transition probability between adjacent data points as a Gaussian distribution, which allows for the application of Student's t-test to obtain a p-value for the neutral model. Conversely, Karsdorp et al. (2020) [106] use pre-trained neural networks to classify time series of variant frequency data as evolving under drift or selection. However, in avoiding the computation of the likelihood, these methods sacrifice the ability to estimate the maximum-likelihood parameters quantifying the evolutionary forces of interest, which may be relevant to the characterisation of the underlying process.

Computing the likelihood efficiently, and with it the model parameters, necessitates approximating the Wright-Fisher transition probability over multiple generations. A variety of schemes have been developed to this end [95, 107]. One possibility is to numerically integrate the Fokker-Planck equation corresponding to the diffusion approximation of the model of interest [108, 109]. Another way to proceed is to approximate the transition probability with an appropriate distribution function, such as a Gaussian [110] or a Beta distribution [111], and fix parameters by matching its moments to those obtained from the Wright-Fisher model. Such distribution functions are well-behaved by construction, but may fail to adequately approximate the true transition probability [95].

In this chapter, I approximate the transition probability with a distribution function that is sufficiently rich to capture the properties of the underlying Wright-Fisher model, but has a small number of parameters that can be estimated efficiently. Specifically, I adopt the Beta-with-Spikes (BwS) distribution introduced in Tataru et al. (2015) [96] as the functional form. The Beta distribution has been found to better describe changes in allele frequencies than a Gaussian distribution [95], primarily because the requirement that variant frequencies lie between 0 and 1 means that frequency differences are necessarily non-Gaussian close to these boundary points. Replacing the Gaussian with a Beta distribution rectifies this problem, but fails to account adequately for the

accumulation of probability at the boundaries associated with the probabilities of loss (i.e. extinction) and fixation (i.e. becoming the sole existing variant of a trait) of a variant. It is precisely this shortcoming that augmenting the Beta distribution with delta functions (spikes) at the boundary points addresses [96].

By estimating loss and fixation probabilities and moments of the Wright-Fisher transition probability, the parameters in the BwS distribution can then be chosen to match: this is then found to improve the approximation to the exact transition probability relative to the Gaussian or standard Beta approximations [95]. In Tataru et al. (2017) [107], the moment-matching procedure is based on recursion relations for the mean and variance of the Wright-Fisher transition probability that are not exact: they are based on a Taylor series expansion of the fitness function around the mean allele frequency in the previous generation. This approximation breaks down when selection is large [110], and errors accumulate over multiple generations, sometimes to the point of exiting the parameter regime for which the BwS distribution is well-defined. In this chapter, I introduce a self-contained approach to estimating the BwS parameters that avoids these problems. The basic idea is to start with a BwS distribution at the beginning of one generation, and to determine the parameter values that best approximate the distribution that results after one generation of evolution within the Wright-Fisher model. This leads to a *self-contained* recursion, in the sense that it maps the BwS parameters directly from one generation to the next via averages with respect to the BwS distribution, rather than via the moments of the target Wright-Fisher distribution.

This estimation scheme is introduced in Section 2.1, together with the methods needed for its application in the detection and quantification of evolutionary forces in historical language data. In Section 2.2, this methodology is applied to data sets of competition between past-tense forms of English verbs. First, in Section 2.2.1, it is bench-marked through an application to a data set from the Corpus of Historical American English (COHA, [112]), where its performance is compared to other prominent methods as applied in previous studies. Finally, Section 2.2.2 is dedicated to its application to a novel data set from the 2019 English Google Books corpus [113]. Our results suggest the presence of a phonological bias in the competition process between irregular and regular forms, highlighting the potential of this and similar methods in the statistical testing of linguistic hypotheses. This chapter is based on work published in Guerrero Montero et al. [97] and Guerrero Montero and Blythe (2023) [98].

2.1 Methods

2.1.1 The Wright-Fisher model for two variants under selection

This chapter is concerned with the analysis of frequency time series, that is, sequences of measurements $X = \{(x_t, t)\} = \{(x_1, t_1), (x_2, t_2), ..., (x_m, t_m)\}$ where $0 \leq x_i \leq 1$ is the fraction of instances of use of a lingueme (e.g., the past tense of a specific verb) in which a variant of interest (e.g., the regular form) was used during a time window centred on time t_i, measured in discrete generations. As previously discussed, I am most interested in detecting and quantifying selection. Under a Wright-Fisher paradigm, the transition probability over a single generation between two data points in the time series reduces to the binomial distribution in equation 2.1, with a fitness function $g(\,\cdot\,; s)$ which incorporates selection through the parameter s. This takes the form:

$$g(x_t; s) = \frac{x_t \mathrm{e}^s}{x_t \mathrm{e}^s + 1 - x_t}\,, \tag{2.2}$$

which has been commonly used in the theoretical characterisation of language change [114, 115]. This leads to a two-variant transition probability over a single generation given by:

$$\mathrm{P_{WF}}(x_{t+1}|x_t; N, s) = \binom{N}{N x_{t+1}} \left(\frac{x_t \mathrm{e}^s}{x_t \mathrm{e}^s + 1 - x_t}\right)^{N x_{t+1}} \left(\frac{1 - x_t}{x_t \mathrm{e}^s + 1 - x_t}\right)^{N(1-x_{t+1})}. \tag{2.3}$$

At $s = 0$, the fitness function reduces to $g(x; 0) = x$, representing a system evolving under pure unbiased drift. The strength of drift is encoded in the population size N. In population genetics, it is well understood that structured populations, where individuals are divided into classes by age, sex, location or other characteristics, can be approximated by a Wright-Fisher model by setting N equal to an appropriate *effective* population size [116]. As reviewed in Chapter 1, the interpretation of N is less obvious in the cultural case; however concrete models of language use have found N to depend on factors like the size and structure of the speech community, the memory lifetime of individual speakers and the number of tokens produced in an utterance [4, 5, 92]. In any case,

N quantifies the effects of drift in transmission: the lower N, the higher the uncertainty in the transmission to the next generation.

In the literature, one can find relationships between the selection parameter s and the fitness function $g(x, s)$ different to the one specified in equation 2.2, particularly of the form $g(x, s') = \frac{x(s'+1)}{xs'+1}$ [89, 107]. Equation 2.2 has two appealing features over this more common parametrisation. First, it satisfies

$$g(x; s) + g(1 - x; -s) = 1\,, \tag{2.4}$$

which means that if one of two variants in a population has a selection strength s, the other one implicitly has a selection strength $-s$. This choice thus lends a symmetry between positive and negative selection strengths of the same magnitude, which aids in the interpretation of the results. In particular, it also means that the fitness function is well defined for every real-valued s, whereas the traditional parametrisation is restricted to $s' \geq -1$. The second property of interest involves the low-fluctuation limit. In the limit $N \to \infty$, the system behaves deterministically: variant frequency x_{t+1} at time $t + 1$ will equal the fitness function of the variant frequency in the previous generation, $g(x_t; s)$, with probability 1. This property does not depend on the functional form of g. However, the fitness function in equation 2.2 further satisfies the functional relation

$$g\left(g(x; s_1); s_2\right) = g\left(x; s_1 + s_2\right)\,. \tag{2.5}$$

Thus, in the deterministic regime, the evolution over k generations with selection coefficients $s_1, s_2, \ldots, s_k$ is the same as a single generation of evolution with selection coefficient $s_1 + s_2 + \cdots + s_k$. Later, I will explain how this property can be exploited to speed up the inference of evolutionary parameters by aggregating multiple generations into one. Finally, it should be noted that the two different specifications of the fitness function with selection can be mapped onto each other via the relation $s = \ln(1 + s')$.

Note that, while a two-variant model is being used here for convenience and simplicity, its application is not limited to systems presenting only two variants of a lingueme. A K-variant system can always be reduced to a 2-variant one for the purpose of analysis under a Wright-Fisher paradigm, as long as one variant of interest is chosen and the frequencies of usage of the remaining $K - 1$ are considered as a whole. Mathematically, this is equivalent to summing over all 'uninteresting' variants in the K variant model, equation 1.4. For a system with

3 variants with frequency vector $\vec{x} = (x_1, x_2, 1 - x_1 - x_2)$ and fitness vector $\vec{g} = (g_1, g_2, 1 - g_1 - g_2)$, this would amount to the following, assuming the variant of interest is the first one:

$$\begin{aligned}
\sum_{Nx_2=0}^{N(1-x_1)} \mathrm{P}_{\mathrm{WF}}\left(\vec{x} \,|\, \vec{g}\right) = & \sum_{Nx_2=0}^{N(1-x_1)} \binom{N}{N\vec{x}} g_1^{Nx_1} g_2^{Nx_2} \left(1 - g_1 - g_2\right)^{N(1-x_1-x_2)} \\
= & \binom{N}{Nx_1} g_1^{Nx_1} \sum_{Nx_2=0}^{N(1-x_1)} \binom{N(1-x_1)}{Nx_2} g_2^{Nx_2} \left(1 - g_1 - g_2\right)^{N(1-x_1-x_2)} \\
= & \binom{N}{Nx_1} g_1^{Nx_1} \left(g_2 + 1 - g_1 - g_2\right)^{N(1-x_1)} \\
= & \binom{N}{Nx_1} g_1^{Nx_1} \left(1 - g_1\right)^{N(1-x_1)} \\
= & \mathrm{P}_{\mathrm{WF}}\left(x_1 \,|\, g_1\right) . \qquad (2.6)
\end{aligned}$$

This can of course be iterated for any number of variants. Notably, it leaves the fitness function associated with the variant of interest unchanged, meaning that both models are quantitatively equivalent when studying the evolutionary forces affecting this variant.

2.1.2 The Beta-with-Spikes approximation to the Wright-Fisher transition probability

Variant frequency data sampled at different time points may not be one generation apart. In this case, it is necessary to sum equation 2.3 over multiple intermediate generations to obtain the appropriate Wright-Fisher transition probability. This is not viable in practice for large population size N and large numbers of intermediate generations k, as the memory requirements for this procedure scale as $\mathcal{O}(N^2)$, and its computational complexity as $\mathcal{O}(kN^3)$ [95]. These considerations motivate approximating the transition probability with some distribution that has a small number of parameters, and using equation 2.3 to determine how those parameters change over multiple generations.

Tataru et al. (2015) [96] introduced the Beta-with-Spikes (BwS) distribution for this purpose. Starting at generation t with a fixed frequency x_t, the distribution

after k generations is assumed to be well-described by the form

$$\mathrm{P}_{\mathrm{BwS}}(x_{t+k}|x_t; N, s) = P_{0,k}\delta(x_{t+k}) + P_{1,k}\delta(1 - x_{t+k}) + (1 - P_{1,k} - P_{0,k})\,\frac{x_{t+k}^{\alpha_k - 1}(1 - x_{t+k})^{\beta_k - 1}}{\mathrm{B}(\alpha_k, \beta_k)}\,, \quad (2.7)$$

which has four parameters α_k, β_k, $P_{0,k}$ and $P_{1,k}$. The central part of this distribution is a Beta distribution, whose shape is controlled by α_k and β_k and whose normalization is given by the Beta function $\mathrm{B}(\alpha_k, \beta_k)$. A variety of shapes can be accessed by tuning α_k and β_k, including a uniform distribution (corresponding to $\alpha_k = 1$, $\beta_k = 1$), distributions that are strongly peaked around the mean ($\alpha_k \gg 1$, $\beta \gg 1$), and those that have an integrable divergence at the boundaries ($\alpha_k < 1$, $\beta_k < 1$).

This flexibility is however insufficient to capture the accumulation of probability at the boundary points which occurs when a variant undergoes loss or fixation. This possibility is incorporated into the transition probability in equation 2.7 through the two Dirac delta functions at the extremes of the interval (similarly to *inflated* models used in Beta regression [117]). The quantity $P_{0,k}$ represents the probability that the variant of interest has gone extinct after k generations. Conversely, $P_{1,k}$ represents the probability that it has fixated, becoming the only available variant for its lingueme. The Beta distribution contribution then describes the transition probability *conditioned* on fixation not yet having occurred.

The crucial step in applying the BwS approximation to data is to estimate the parameters α_k, β_k, $P_{0,k}$ and $P_{1,k}$. The general approach is to estimate moments of the transition probability and the fixation probabilities in the Wright-Fisher model, and choose the parameters in the BwS distribution so that they match up. To this end, note first that the mean and variance of a Beta-distributed variable y with parameters α and β are given by [118]:

$$\mathbb{E}\left[y\right] = \frac{\alpha}{\alpha + \beta} \quad (2.8)$$

$$\mathrm{Var}\left[y\right] = \frac{\alpha\beta}{(\alpha + \beta)^2\,(\alpha + \beta + 1)} \quad (2.9)$$

Then, by defining α_k and β_k as

$$\alpha_k = \left(\frac{E_k^*(1-E_k^*)}{V_k^*} - 1 \right) E_k^* \tag{2.10}$$

$$\beta_k = \left(\frac{E_k^*(1-E_k^*)}{V_k^*} - 1 \right) (1-E_k^*)\,, \tag{2.11}$$

the mean and variance of the Beta part of the BwS approximation will match the mean E_k^* and variance V_k^* of the Wright-Fisher transition probability after k generations, conditioned on fixation not having occurred. These can be obtained from the mean E_k and variance V_k of the full distribution, as well as the fixation probabilities $P_{0,k}$ and $P_{1,k}$, via

$$E_k^* = \frac{E_k - P_{1,k}}{1 - P_{0,k} - P_{1,k}} \tag{2.12}$$

$$V_k^* = \frac{V_k + E_k^2 - P_{1,k}}{1 - P_{0,k} - P_{1,k}} - (E_k^*)^2\,. \tag{2.13}$$

It remains to estimate E_k, V_k and the loss and fixation probabilities.

2.1.3 Estimations of the Wright-Fisher moments

Previous approaches to estimating E_k, V_k, $P_{0,k}$ and $P_{1,k}$[95, 107] have been based around a truncated Taylor-series expansion of the fitness function. Here I take a different approach to estimating the BwS parameters by introducing a scheme which relies on approximating the shape of the transition probability rather than its fitness function. With this, I aim at reducing the accumulation of error that stems from the truncation of the Taylor expansion. In this section, I proceed to explicitly derive both estimation schemes.

Both schemes start with exact recursions for the mean and variance of the transition probability at generation $t+k+1$. Given the transition probability $\mathrm{P}_{\mathrm{WF}}(x_{t+k}|x_t)$ representing the Wright-Fisher probability distribution of variants after k time steps with known starting frequency x_t, the distribution at the next time step can be found using the Chapman-Kolmogorov equation for the transition probabilities of the process:

$$\mathrm{P}_{\mathrm{WF}}(x_{t+k+1}|x_t) = \sum_{x_{t+k}} \mathrm{P}_{\mathrm{WF}}(x_{t+k+1}|x_{t+k})\mathrm{P}_{\mathrm{WF}}(x_{t+k}|x_t) \tag{2.14}$$

which arises from the law of total probability, using the one-step Wright-Fisher transition probability. Note that for simplicity, I'm letting the subscripts specify the number of time steps between data points. The associated transition probabilities are different for each value of k.

From this, the mean of the frequency at generation $t+k+1$ is given by

$$\begin{aligned}
E_{k+1} &= \sum_{n=0}^{N} \frac{n}{N} \mathrm{P}_{\mathrm{WF}} \left(x_{t+k+1} = \frac{n}{N} | x_t \right) \\
&= \sum_{n=0}^{N} \frac{n}{N} \sum_{x_{t+k}} \mathrm{P}_{\mathrm{WF}} \left(\frac{n}{N} | x_{t+k} \right) \mathrm{P}_{\mathrm{WF}} \left(x_{t+k} | x_t \right) \\
&= \sum_{x_{t+k}} \sum_{n=0}^{N} \frac{n}{N} \mathrm{P}_{\mathrm{WF}} \left(\frac{n}{N} | x_{t+k} \right) \mathrm{P}_{\mathrm{WF}} \left(x_{t+k} | x_t \right) \\
&= \sum_{x_{t+k}} g\left(x_{t+k} \right) \mathrm{P}_{\mathrm{WF}} \left(x_{t+k} | x_t \right) \\
&= \mathbb{E}_{\mathrm{WF}} \left[g(x_{t+k}) | x_t \right] , \qquad (2.15)
\end{aligned}$$

where the fourth equality uses the analytical result for the mean of the binomial distribution, the Wright-Fisher transition probability, thus eliminating the explicit dependence of the final result on this probability function. In the last equality, $\mathbb{E}_{\mathrm{WF}}$ represents the expectation under the k-step Wright-Fisher transition probability.

The variance, similarly, can be found as:

$$\begin{aligned}
V_{k+1} =& \sum_{n=0}^{N} \left(\frac{n}{N} - E_{k+1} \right)^2 \mathrm{P}_{\mathrm{WF}} \left(x_{t+k+1} = \frac{n}{N} | x_t \right) \\
=& \sum_{x_{t+k}} \sum_{n=0}^{N} \left(\frac{n}{N} \right)^2 \mathrm{P}_{\mathrm{WF}} \left(\frac{n}{N} | x_{t+k} \right) \mathrm{P}_{\mathrm{WF}}(x_{t+k} | x_t) - E_{k+1}^2 \\
=& \sum_{x_{t+k}} \left[\left(1 - \frac{1}{N} \right) g(x_{t+k})^2 + \frac{1}{N} g(x_{t+k}) \right] \mathrm{P}_{\mathrm{WF}}(x_{t+k} | x_t) - E_{k+1}^2 \\
=& \left(1 - \frac{1}{N} \right) \left(\mathbb{E}_{\mathrm{WF}}[g(x_{t+k})^2 | x_t] - E_{k+1}^2 \right) + \frac{1}{N} \left(\mathbb{E}_{\mathrm{WF}}[g(x_{t+k}) | x_t] - E_{k+1}^2 \right) \\
=& \left(1 - \frac{1}{N} \right) \mathrm{Var}_{\mathrm{WF}}[g(x_{t+k}) | x_t] \\
&+ \frac{1}{N} \mathbb{E}_{\mathrm{WF}}[g(x_{t+k}) | x_t] \left(1 - \mathbb{E}_{\mathrm{WF}}[g(x_{t+k}) | x_t] \right) , \qquad (2.16)
\end{aligned}$$

where the third equality uses the analytical form of the second moment of the

binomial distribution, and the last equality uses equation 2.15 together with the definition of the variance. In it, $\mathrm{Var}_{\mathrm{WF}}$ represents the variance under the k-step Wright-Fisher transition probability.

The loss probability is given by:

$$\begin{aligned} P_{0,k+1} =& \mathrm{P}_{\mathrm{WF}}\left(x_{t+k+1}=0|x_t\right) \\ =& \sum_{x_{t+k}} \mathrm{P}_{\mathrm{WF}}\left(x_{t+k+1}=0|x_{t+k}\right)\mathrm{P}_{\mathrm{WF}}(x_{t+k}|x_t) \\ =& \sum_{x_{t+k}} \left(1-g(x_{t+k})\right)^N \mathrm{P}_{\mathrm{WF}}(x_{t+k}|x_t) \\ =& \mathbb{E}_{\mathrm{WF}}\left[\left(1-g(x_{t+k})\right)^N|x_t\right], \end{aligned} \tag{2.17}$$

and equivalently for the fixation probability we have:

$$\begin{aligned} P_{1,k+1} =& \mathrm{P}_{\mathrm{WF}}\left(x_{t+k+1}=1|x_t\right) \\ =& \sum_{x_{t+k}} \mathrm{P}_{\mathrm{WF}}\left(x_{t+k+1}=1|x_{t+k}\right)\mathrm{P}_{\mathrm{WF}}(x_{t+k}|x_t) \\ =& \sum_{x_{t+k}} g(x_{t+k})^N \mathrm{P}_{\mathrm{WF}}(x_{t+k}|x_t) \\ =& \mathbb{E}_{\mathrm{WF}}\left[g(x_{t+k})^N|x_t\right], \end{aligned} \tag{2.18}$$

The truncated Taylor scheme

Equations 2.15 to 2.18 do not have closed analytical forms for arbitrary $g(x_{k+t})$, and their direct integration using a k-generation Wright-Fisher transition probability is computationally intractable in general. From equations 2.15–2.18, recursive relations for the moments, loss and fixation probabilities after $k+1$ generations can be obtained by Taylor expanding $g(x_{t+k})$ about E_k up to second order, and dropping all moments of higher order than the variance. This is the previously used approach to estimation [95, 107] that I refer to here as the *truncated Taylor scheme*.

For the mean (from equation 2.15) one gets:

$$\begin{aligned} E_{k+1} \approx & g(E_k) + g'(E_k)\mathbb{E}_{\mathrm{WF}}\left[(x_{t+k} - E_k)\,|x_t\right] \\ & + \frac{g''(E_k)}{2}\mathbb{E}_{\mathrm{WF}}\left[(x_{t+k} - E_k)^2\,|x_t\right] \\ = & g(E_k) + \frac{g''(E_k)}{2}V_k \end{aligned} \tag{2.19}$$

where we have used the identities $\mathbb{E}_{\mathrm{WF}}\left[x_{t+k}|x_t\right] = E_k$ and $\mathbb{E}_{\mathrm{WF}}\left[(x_{t+k} - E_k)^2\,|x_t\right] = V_k$.

For the variance:

$$\begin{aligned} \mathrm{Var}_{\mathrm{WF}}[g(x_{t+k})|x_t] = & \mathbb{E}_{\mathrm{WF}}\left[g(x_{t+k})^2|x_t\right] - E_{k+1}^2 \\ \approx & g(E_k)^2 - E_{k+1}^2 + 2g'(E_k)g(E_k)\mathbb{E}_{\mathrm{WF}}\left[(x_{t+k} - E_k)|x_t\right] \\ & + \left(g'(E_k)^2 + g''(E_k)g(E_k)\right)\mathbb{E}_{\mathrm{WF}}\left[(x_{t+k} - E_k)^2|x_t\right] \\ = & -\left(\frac{g''(E_k)}{2}V_k\right)^2 - g''(E_k)g(E_k)V_k \\ & + \left(g'(E_k)^2 + g''(E_k)g(E_k)\right)V_k \\ = & -\left(\frac{g''(E_k)}{2}V_k\right)^2 + g'(E_k)^2V_k \\ \approx & g'(E_k)^2V_k \end{aligned} \tag{2.20}$$

where, in line with previous work (see supplementary material in Paris et al. (2019) [95]), the last equality assumes V_k^2 to be of the same order as $\mathbb{E}_{\mathrm{WF}}\left[(x_{t+k} - E_k)^4\,|x_t\right]$ and thus negligible. By introducing this result into equation 2.16, we obtain

$$V_{k+1} \approx \frac{1}{N}E_{k+1}\left(1 - E_{k+1}\right) + \left(1 - \frac{1}{N}\right)V_k\, g'(E_k)^2\,. \tag{2.21}$$

For the loss and fixation probabilities, Tataru et al. [107] propose linearizing the fitness function as $g(x) = x$ and approximating the transition probability $\mathrm{P}_{\mathrm{WF}}(x_{t+k}|x_t)$ as a Beta-with-Spikes (eq. 2.7). With that, starting from equation

2.17:

$$\begin{aligned} P_{0,k+1} \approx & \mathbb{E}_{\mathrm{BwS}}\left[(1-x_{t+k})^N \,|x_t\right] \\ = & P_{0,k} + (1 - P_{0,k} - P_{1,k}) \int \frac{x^{\alpha_k+1}(1-x)^{\beta_k+N+1}\mathrm{d}x}{B(\alpha_k,\beta_k)} \\ = & P_{0,k} + (1 - P_{0,k} - P_{1,k}) \frac{B(\alpha_k,\beta_k+N)}{B(\alpha_k,\beta_k)} \end{aligned} \tag{2.22}$$

and similarly for the fixation probability (equation 2.18):

$$P_{1,k+1} \approx P_{1,k} + (1 - P_{0,k} - P_{1,k}) \frac{B(\alpha_k+N,\beta_k)}{B(\alpha_k,\beta_k)}\,. \tag{2.23}$$

The error of these approximations increase with the selection coefficient s, as stronger selection makes the dropped higher-order terms less negligible. Consequently, this recursion is expected only to be accurate for small s. Despite this, these recursion relations benefit from being simple and quick to apply.

The self-contained scheme

The self-contained estimation scheme introduced here is motivated by the expectation that it will keep the accumulation of error under control. The basic idea is to approximate the Chapman-Kolmogorov equation (eq. 2.14) as a BwS distribution obtained after k generations, and generate the intermediate distribution

$$\mathrm{P}_{\mathrm{int}}(x_{t+k+1}|x_t) = \int \mathrm{P}_{\mathrm{WF}}(x_{t+k+1}|x_{t+k})\mathrm{P}_{\mathrm{BwS}}(x_{t+k}|x_t)\mathrm{d}x_{t+k} \tag{2.24}$$

by applying just one step of the Wright-Fisher process (2.3). We then examine the moments and fixation probabilities of this intermediate distribution, rather than those obtained from an exact Wright-Fisher transition probability, and use the values obtained to set the BwS parameters for generation $k+1$. We view this as a *self-contained scheme*, as it maps directly from one set of BwS parameters to the next.

Figure 2.1 compares the Wright-Fisher transition probability with the intermediate and Beta-with-Spikes distributions for the case of high selection ($s = 0.2$) and high drift ($N = 50$). Even in this challenging regime, we find the intermediate and Beta-with-Spikes distributions remain similar to the exact Wright-Fisher

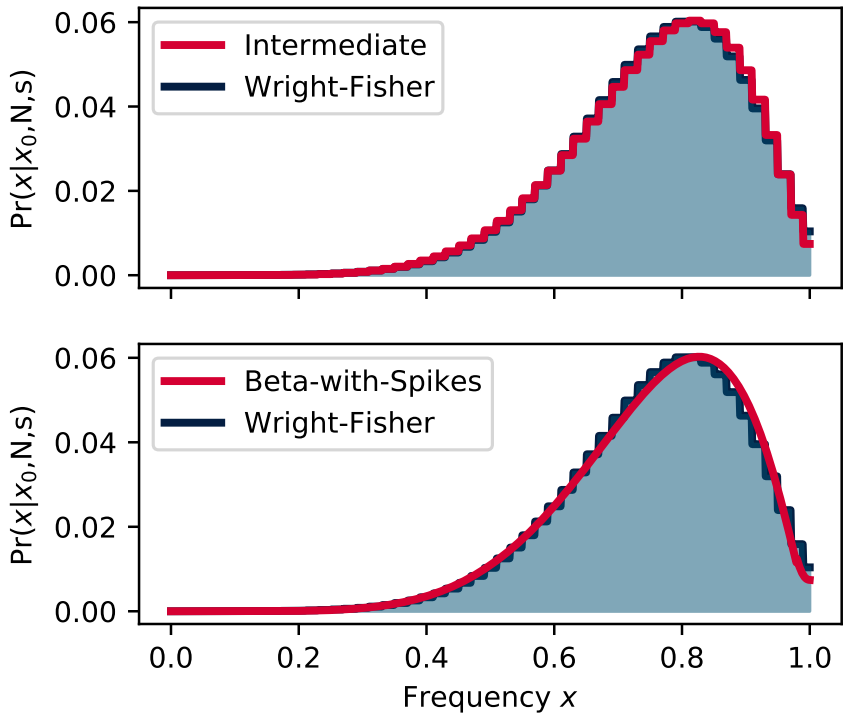

Figure 2.1 *Top panel: Comparison of the intermediate (red) and Wright-Fisher (blue) distributions. Bottom panel: Comparison of the Beta-with-Spikes (red) and Wright-Fisher (blue) distributions. Distributions generated with* $N = 50$, $s = 0.2$, $x_0 = 0.5$ *after* $k = 8$ *generations.*

transition probability after $k = 8$ generations. The continuous Beta-with-Spikes distribution is generated by mapping its mean, variance, loss and fixation probabilities to those of the intermediate distribution.

Computing the desired parameters under the approximation in equation 2.24 amounts to replacing the WF subscript in the final expressions in equations 2.15-2.18 with a BwS subscript. The resulting integrals arising from the expectations in those expressions are computationally tractable without having to resort to approximations of the fitness function. From equation 2.15, the mean of the intermediate distribution with population size N is given by

$$\begin{aligned} E_{k+1} &= \mathbb{E}_{\mathrm{BwS}}\left[g(x_{t+k})|x_t\right] \\ &= P_{1,k} + (1 - P_{0,k} - P_{1,k})\,\frac{\int g(x)x^{\alpha_k - 1}(1-x)^{\beta_k - 1}}{\mathrm{B}(\alpha_k, \beta_k)}\,, \end{aligned} \tag{2.25}$$

where the last equality uses equation 2.7 together with the assumptions $g(0) = 0$ and $g(1) = 1$.

The variance, similarly, can be found as

$$\begin{aligned} V_{k+1} =& \left(1-\frac{1}{N}\right)\left(\mathbb{E}_{\mathrm{BwS}}[g(x_{t+k})^2|x_t]-\mathbb{E}_{\mathrm{BwS}}[g(x_{t+k})|x_t]^2\right) \\ &+\frac{1}{N}\mathbb{E}_{\mathrm{BwS}}[g(x_{t+k})|x_t]\left(1-\mathbb{E}_{\mathrm{BwS}}[g(x_{t+k})|x_t]\right) \\ =& \left(1-\frac{1}{N}\right)\left[P_{1,k}+(1-P_{0,k}-P_{1,k})\,\frac{\int g(x)^2 x^{\alpha_k-1}(1-x)^{\beta_k-1}}{\mathrm{B}(\alpha_k,\beta_k)}\right] \\ &+\frac{1}{N}E_{k+1}-E_{k+1}^2\,, \end{aligned} \tag{2.26}$$

where equations 2.16, 2.7 and 2.25 have been used.

The loss probability (equation 2.17) is given by

$$\begin{aligned} P_{0,k+1} =& \mathbb{E}_{\mathrm{BwS}}\left[(1-g(x_{t+k}))^N\,|x_t\right] \\ =& P_{0,k}+(1-P_{0,k}-P_{1,k})\,\frac{\int (1-g(x))^N\, x^{\alpha_k-1}(1-x)^{\beta_k-1}}{\mathrm{B}(\alpha_k,\beta_k)} \end{aligned} \tag{2.27}$$

and equivalently for the fixation probability (equation 2.18) we have

$$\begin{aligned} P_{1,k+1} =& \mathbb{E}_{\mathrm{P}}\left[g(x_{t+k})^N|x_t\right] \\ =& P_{1,k}+(1-P_{0,k}-P_{1,k})\,\frac{\int g(x)^N x^{\alpha_k-1}(1-x)^{\beta_k-1}}{\mathrm{B}(\alpha_k,\beta_k)} \end{aligned} \tag{2.28}$$

These parameters may be used now to generate the parameters α_{k+1} and β_{k+1} of the Beta-with-Spikes transition probability after $k+1$ generations. This process can then be iterated indefinitely. While error in the estimation does accumulate as k increases, due to the use of the approximate intermediate transition probability in equation 2.24, this is generally of less magnitude than the error accumulated due to the truncation of the Taylor expansion. This is discussed in depth below.

This method does involve more computation than the Taylor-series approach, in that four numerical integrals have to be performed in each iteration, which can increase the computation time by a up to a factor of order 10, compared to an iteration of the Taylor-series approach. However, this effort is manageable in practice, and if necessary can be reduced by appealing to scaling properties identified below.

Comparison of approximation schemes

Comparing the BwS transition probabilities obtained from the two estimation schemes to the exact Wright-Fisher distribution would be difficult. This is chiefly because the Wright-Fisher distribution is defined only at discrete frequency values, while the BwS distribution is well defined over the entire interval of frequencies. Thus, to assess the relative quality of the two estimation schemes, I construct a baseline BwS distribution which is obtained by computing the moments and fixation probabilities within the Wright-Fisher model using numerical methods that are exact to machine precision. Then, we can measure the distance between this baseline and each of the BwS distributions obtained through the estimation schemes set out above. For this purpose, I use the Wasserstein distance [119], bearing in mind that a smaller distance indicates a better approximation to the baseline. The Wasserstein distance between probability distributions f_1 and f_2 defined in the interval $[0, 1]$ is defined in terms of their cumulative distribution functions F_1 and F_2 as:

$$W(f_1, f_2) = \int_0^1 |F_1(x) - F_2(x)| \, \mathrm{d}x \tag{2.29}$$

The results are shown in Figure 2.2. In all cases the population size $N = 100$, and I present a comparison of performance as a function of the number of generations k, the initial frequency x_0 and for three different values of the selection coefficient. The estimation based on the truncated Taylor expansion quickly accumulates error in its estimation of the moments when the selection coefficient is sufficiently large, which may lead to negative values of α and β and an undefined distribution. This happens at generation $k = 23$ for the distribution with intermediate selection and as early as $k = 7$ for the distribution with strong selection. This is reflected in the figure as points where the Wasserstein distance reaches a maximal value of 0.1. The self-contained estimation never leads to an undefined distribution, provided sufficiently accurate integration schemes are used, as α and β as defined in equation 2.10 are always non-negative given that P_0, P_1, E and V are all obtained from the same well-defined distribution.

Figure 2.3 provides a closer look at the accuracy in the estimation of the mean E, variance V, loss probability P_0 and fixation probability P_1 for both the self-contained and truncated Taylor estimation schemes, with $N = 100$ and $s = 0.1$. For each one of the four parameters, the absolute difference between their estimated and exact values is plotted as a function of the initial probability

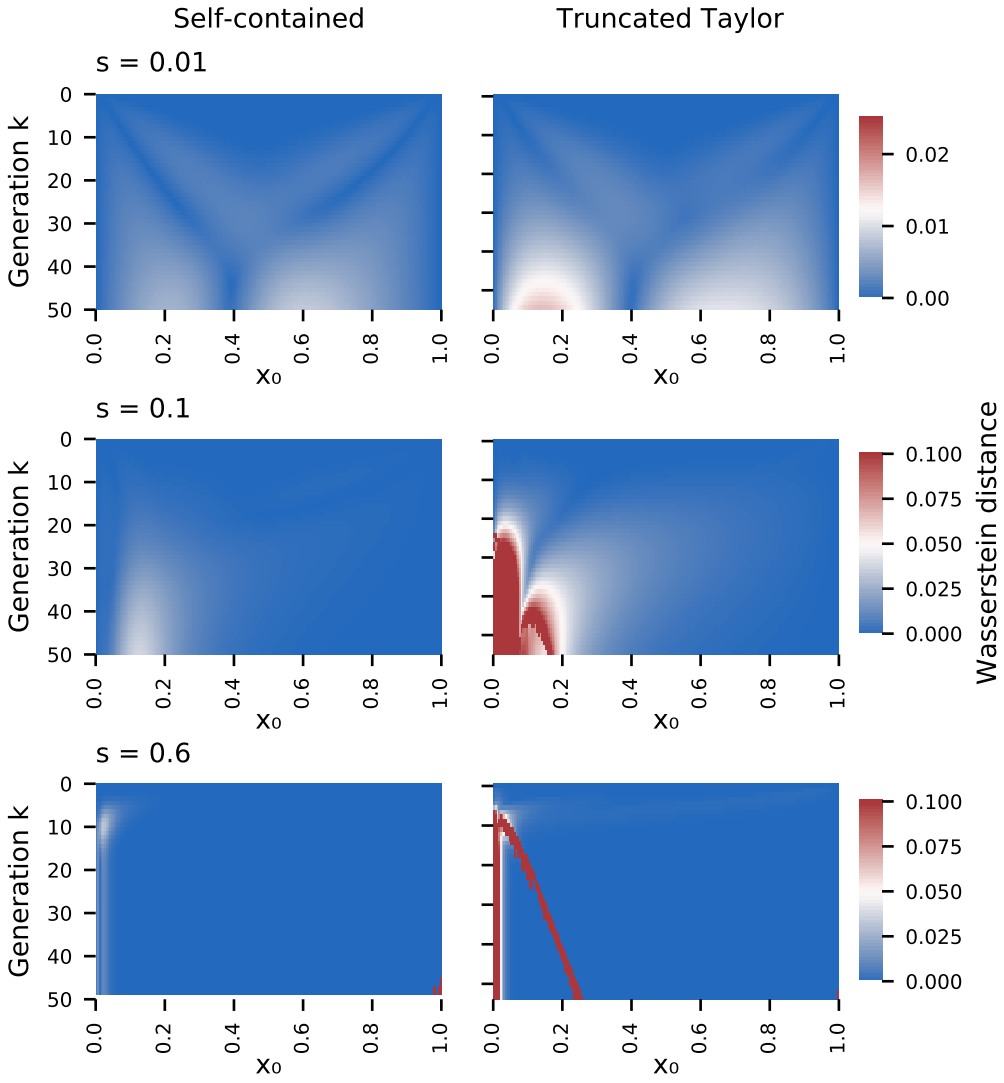

Figure 2.2 *Wasserstein distance between the Beta-with-Spikes distribution with numerically exact moments and with approximated moments for two approximation schemes and three values of the selection strength s, as a function of the initial frequency x_0 and the generation k. Left: results for the self-contained approximation. Right: results for the approximation based on the truncated Taylor expansion. Top figures: weak selection ($s = 0.01$). Middle figures: intermediate selection ($s = 0.1$). Low figures: strong selection ($s = 0.6$). Sudden increase of the Wasserstein distance to high values (capped at 0.1 for readability) in the approximation based on the truncated Taylor expansion for intermediate and strong selection is due to accumulation of error that leads to an undefined distribution. Results for $N = 100$.*

x_0 and generation k. Again, the plots demonstrate the robustness of the self-contained estimation, keeping an absolute error below 0.1 for all data points, whereas the truncated Taylor expansion surpasses this value for all parameters before generation $k = 25$.

2.1.4 Maximum-likelihood estimation and model comparison

As established in the introduction, the social, linguistic and cognitive forces driving language change are very diverse. Still, their measurable effects can be broadly characterised as belonging to one of two types. Systematic biases drive the evolutionary process in a specific direction, and can be modelled as selective forces, parametrised through s in the Wright-Fisher model. Frequency effects and stochasticity in transmission produce random, unbiased drift parametrised through N in Wright-Fisher, whose effects are always present, albeit not always sufficient to explain the behaviour of the data. With a robust computational approximation to the Wright-Fisher model in place, we can apply maximum-likelihood estimation methods to the detection and quantification of these evolutionary forces in historical data.

Maximum likelihood estimation is a conceptually simple yet powerful technique for estimating parameter values in a model. Given a time series of variant frequency data $X = \{(x_t, t)\} = \{(x_1, t_1), (x_2, t_2), ..., (x_m, t_m)\}$, the likelihood of a Wright-Fisher model with parameters N and s will be given by

$$\mathcal{L}\left(N, s|X\right) = \prod_{i=1}^{m-1} \mathrm{P}_{\mathrm{WF}}\left(x_{i+1}|x_i; N, s\right) , \tag{2.30}$$

where the Markov property of the Wright-Fisher model allows us to factorise the likelihood in terms of the individual transition probabilities between consecutive data points. These transition probabilities will be, in practice, approximated using a self-contained BwS scheme as previously laid out.

The likelihood then allows us to obtain empirical estimations $\hat{N}$ and $\hat{s}$ of the evolutionary parameters as those that maximise the likelihood $\mathcal{L}\left(N, s|X\right)$ of the model given the data,

$$(\hat{N}, \hat{s}) = \arg\max L\left(X|N, s\right) . \tag{2.31}$$

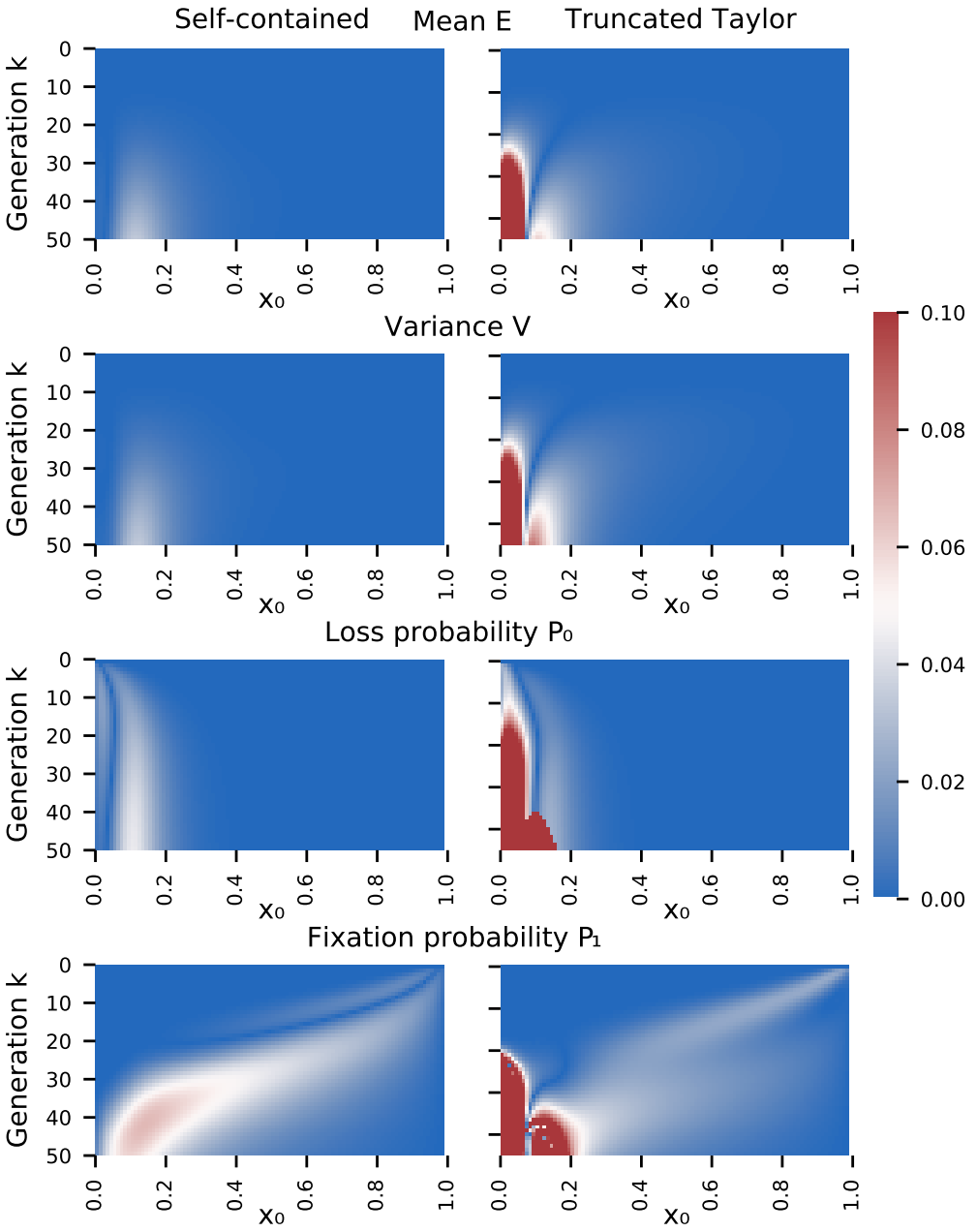

Figure 2.3 *Absolute value of the difference between exact and approximated values of the mean, variance, loss probability and fixation probability, as a function of the initial frequency x_0 and the generation k. Left sub-panels: self-contained approximation. Right sub-panels: truncated Taylor expansion approximation. Results for $N = 100$, $s = 0.1$.*

In practice, I find that the likelihood $\mathcal{L}(N, s|X)$ has a single maximum, which can be located by successively optimising on N at fixed s and vice versa following a coordinate descent algorithm [120].

Measuring selection is not sufficient to *ascertain* it, however. First, one must establish that a null model of pure stochastic drift (i.e. $s = 0$) is not sufficient to explain the behaviour of the data. Maximum-likelihood methods can also be applied in this context of model comparison and model selection, in what is typically known as the *likelihood-ratio test*. Given a data set X and its maximum-likelihood parameters $(\hat{N}, \hat{s})$ for the test hypothesis of selection, as well as a maximum-likelihood parameter $\hat{N}_0$ under the null hypothesis of pure drift, the test statistic of the two models is defined in terms of the ratio of their maximal likelihoods as:

$$\lambda = 2 \ln \left(\frac{\mathcal{L}\left(\hat{N}, \hat{s}|X\right)}{\mathcal{L}\left(\hat{N}_0, 0|X\right)} \right) . \tag{2.32}$$

The model with the greater range of parameters will always provide a fit at least as good as the null hypothesis, meaning that its likelihood will always be equal or greater than that of the null hypothesis and thus $\lambda \geq 0$. λ can then be used to obtain a p-value reflecting how likely one would be to observe that difference in likelihood by chance alone, assuming that the null hypothesis indeed generated the observations. A p-value threshold is then used to decide whether the null hypothesis of pure drift should be confidently rejected, with the conventional criterion for rejection being $p < 0.05$.

Typically, the computation of the p-value relies on the assumption that the test statistic λ can be approximated as being χ^2-distributed if the null hypothesis happens to be true. The p-value can then be computed as the probability to sample a value higher than λ from this distribution. This assumption is based on Wilks' theorem [121, 122], which states that the distribution of λ's converges to a χ^2-distribution as the number of observations in the data set goes to infinity. Its applicability to finite data sets, thus, can produce inaccurate results.

To avoid this issue and maximise the accuracy of my analyses, I use the empirical likelihood-ratio test (ELRT), as introduced in Feder et al. (2014) [104]. I thus generate 1,000 artificial time series spanning the same time period as the empirical data X with parameter values $s = 0$ and $N = \hat{N}_0$. For each of these, I find the parameters that maximise the likelihoods of both the test and null models, and compute their test statistics λ using equation 2.32. The p-value for the null

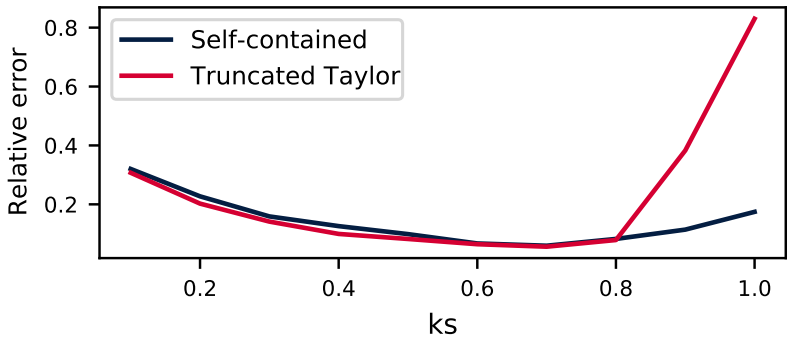

Figure 2.4 *Relative error in the estimation of the selection parameter s in artificially generated time series, using a BwS-based maximum-likelihood inference with both the self-contained and the tructated Taylor approximations of the moments, as a function of ks, the product of the true selection parameter s and the number of generations k between data points.*

hypothesis of drift can then be estimated as the fraction of artificially-generated time series whose λ is higher than that of the historical time series of interest.

Figure 2.4 compares the performance of both the self-contained and the truncated Taylor approximations of the moments in the estimation of the selection parameter s from artificially generated time series, using maximum-likelihood inference. While both approximations perform similarly at low ks (the product of the true selection parameter and the number of generations between data points) the numerical stability of the self-contained scheme keeps the error in the estimation decidedly lower starting at $ks = 0.8$.

2.1.5 The choice of Wright-Fisher generation time

The Wright-Fisher model discretises time into generations of a given length. When dealing with allele frequency data in microbial populations, the generation time is usually estimated as the time it takes for a population to double in size under conditions of exponential growth [123]. In linguistic and cultural contexts, however, the generation time is a more elusive concept: an individual may change their lingueme variant usage throughout their lifetime in ways that are often difficult to measure. Thus, defining a 'typical' timescale for competition dynamics in language may be close to impossible.

When dealing with frequency data sampled at times t_1, t_2,..., t_m, there are then

two possibilities to consider. The first one is that the underlying generation time of the process is shorter than the time between consecutive samples, meaning that there is more than one generation between t_i and t_{i+1}. The second one is that generations are long enough to encompass more than one data point in the time series.

Let's first address the former possibility, concerned with the issue of matching a real time interval, Δt, to a number of generations, k. In situations where this is not known, we can appeal to scaling properties of N and s with k to obtain values whose scales are set primarily by Δt and only weakly by k. As mentioned in Section 2.1.1, the fitness function defined by equation 2.2 satisfies in the deterministic limit ($N \to \infty$) an exact scaling relation whereby k generations, each lasting $\frac{\Delta t}{k}$, with selection coefficient s_k are equivalent to a single generation, lasting Δt, with coefficient $s_1 = ks_k$. Meanwhile, in the diffusion limit ($N \gg 1$) and pure drift ($s = 0$), k generations with drift coefficient N_k are equivalent to a single generation with drift coefficient $N_1 = N_k/k$ [see e.g. 89]. In the general case, I propose that for any two choices of number of generations k and $\tilde{k}$ between two data points separated by a time interval Δt, we have the scaling behaviors

$$ks_k = \tilde{k}s_{\tilde{k}} \quad N_k/k = N_{\tilde{k}}/\tilde{k}\,. \tag{2.33}$$

These relations allow us to divide a time interval into k steps for the purpose of performing the analysis, and quote effective population sizes and selection coefficients appropriately for a standardized time interval determined by $\tilde{k}$. For example, data could be presented at intervals of ten years, divided into $k = 2$ steps of five years for the purposes of analysis, and quoted for a standardized interval of one year ($\tilde{k} = 10$) to facilitate comparison of analyses performed for different time series.

The success of this approach depends on equation 2.33 holding with reasonable accuracy for general N and s, beyond the special limits described above. Figure 2.5 confirms this for the case of synthetic data generated by iterating equation 2.3 ten times between updates. In this case, the true number of generations between sample points is $\tilde{k} = 10$, but I choose to analyze with a different number, k. The error in the maximum likelihood estimates of N_k/k and ks_k, relative to the true values, is shown as a function of k in Fig. 2.5. The error is found to be modest (around 10% or less) even when the parameters are far from the values that make the scaling behaviour exact (low N for the scaling behaviour of s, high selection strength Ns for the scaling behaviour of N). What

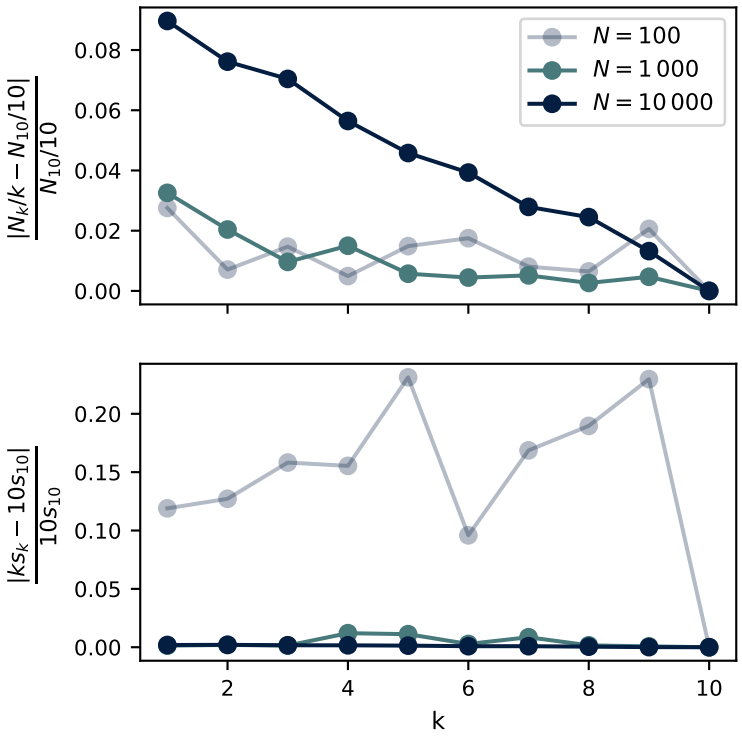

Figure 2.5 *Relative error in the scaling behaviour of N and s as time series with 10 generations between data points are reanalysed as having $k < 10$ generations between data points. Statistics generated as the average over 2000 artificially-generated time series with $s = 0.05$ and $N = 100$, $N = 1000$, $N = 10000$.*

this means in practice is that one can reduce the number of iterations of the self-contained estimation scheme to a small number k, by dividing the time between data points into k generations, whilst retaining reasonable parameter estimates for some chosen standardized generation time.

The other possibility, i.e. that each 'true' generation contains several data points of variant frequency data, proves to be trickier. As shown in figure 2.6, different choices of generation time may affect the features of the time series arising from their corresponding temporal binning strategies of the data. Larger bins will offer greater precision as they contain larger samples of historical data, minimising sampling noise in the estimation of the variant frequencies. Smaller bins will result in a greater number of them, increasing the resolution of the time series at the cost of precision. These effects can result in different estimations of the evolutionary parameters and p-value for the null model of pure drift, beyond what is to be reasonably expected given the scaling laws in equations 2.33.

This issue was empirically explored in the context of language change by Karjus

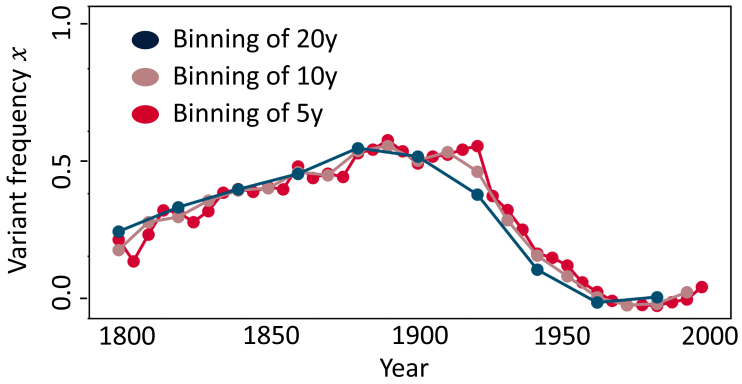

Figure 2.6 *Time series resulting from the same data set of variant usage data under different binning strategies resulting from different choices of generation time. Time series with more fine-grained temporal binning offer greater resolution at the expense of greater sampling noise. Conversely, coarse-grained temporal binning increases precision at the expense of resolution.*

et al. (2020) [124]. The empirical unavailability of the 'true' generation time means that the variability under binning needs to be accounted for in order to properly characterise the behaviour of the data. In Sections 2.2.1 and 2.2.2, I will demonstrate how to do so by applying Principal Component Analysis [125] to visualise the variability and assess the robustness results obtained from a variety of choices of generation time.

2.2 Applications

Having validated the methods of the previous section with synthetic data, I proceed to apply them now to historical language data, focusing on the competition dynamics between past-tense forms of English verbs. I first revisit the set of verb time series from the Corpus of Historical American English (COHA) [112] to benchmark my approach against those of Newberry et al. (2017) [105] and Karsdorp et al. (2020) [106]. These results demonstrate that the BwS method is both more robust than a similar likelihood-based approach [105] and more informative than a neural network trained to perform a binary classification [106].

Then, I go one step further by applying the method to a novel data set of verbs from the 2019 English Google Books corpus [113] and using my methodology to formulate and test a hypothesis on the origin of selective forces, rather than limiting myself to detecting and quantifying such forces. I also introduce and

apply to both data sets a method for assessing the variability of parameter values under different binning strategies, which facilitates judging the robustness of results.

2.2.1 Drift versus selection in past-tense English verbs

In the context of language change, a simple example of competition between two variants is provided by English verbs with an irregular past tense form which in many cases coexists with a regular form. For example, the past tense of *dream* may be produced as the regular *dreamed* or the irregular *dreamt*. This competition has been studied from a variety of quantitative perspectives [105, 106, 124, 126–128]. Of greatest relevance to this work are those studies that aimed to distinguish drift from selection as the mechanism behind changes in the relative frequencies of the regular and irregular forms over time.

Newberry et al. (2017) [105] applied the Frequency Increment Test (FIT) [104] to a set of verbs from the Corpus of Historical American English (COHA) [112]. This is a maximum-likelihood method that relies on a normal approximation to the Wright-Fisher transition probabilities. Like the Beta-with-Spikes maximum-likelihood method in Section 2.1.2, this method yields estimates of the effective population size and selection strength, along with a p-value for the null hypothesis of pure drift. However, there are situations where results are flagged as unreliable due to the frequency increments failing a normality test [105]. Karjus et al. (2020) [124] further noted that the results can also be sensitive to the size of the window over which frequencies are estimated, as discussed in Section 2.1.5.

Karsdorp et al. (2020) [106] avoid these issues by taking the rather different approach of training a neural network on simulated time series generated by the Wright-Fisher transition probabilities (eq. 2.3) for different values of s (but fixed $N = 1,000$). Each time series in the training set is labelled according to whether it was generated purely by drift ($s = 0$) or if selection was operating ($s \neq 0$). Once trained, the network yields a binary classification of empirical time series, according to whether they are more similar to the examples of drift or selection in the training data. I refer to this as the Time Series Classification (TSC) approach. The advantage of TSC is that no approximation to the Wright-Fisher transition probabilities is made. Moreover, one can manipulate the training data so that it displays artifacts of binning or finite sample sizes that are features of real time series, which in turn should improve the reliability of the classification. This

approach does however come with some drawbacks. While the output from the classification algorithm is a value between 0 and 1, it does not have an obvious interpretation as a probability. Karsdorp et al. (2020) [106] used a threshold of 0.5 to label time series as arising from drift or selection. The method further does not provide an estimate of the strength of s, and since N was fixed in the training set, this amounts to an assumption that this single value of N was appropriate for all empirical time series. This could be an issue since Newberry et al. (2017) [105] report a wide range of values of N for this data set (from around 80 to around $22,500$).

In Appendix A.1 I report the maximum-likelihood estimates of N and s, along with the p-value for the drift hypothesis, obtained using the self-contained BwS method for the same set of verbs that were considered by Newberry et al. (2017) [105] and Karjus et al. (2020) [124] using FIT and by Karsdorp et al. (2020) [106] using TSC. I perform the analysis by extracting annual relative frequency data of the irregular forms of each of the verbs from COHA and aggregating it into 10-, 20- and 40-year bins. The reason for this is a trade-off between the more precise frequency estimates that derive from larger bins and the greater temporal resolution obtained from a larger number of bins over the relevant historical period. By employing different binning strategies, we can gain insights into the consequences of this trade-off. Variable-width binning strategies have also been successfully applied in previous studies [105]. In these, the number of tokens per bin is kept roughly constant at an arbitrarily chosen value, at the expense of varying their temporal width. For the purpose of comparing the different methods, I have chosen to look only at fixed-binning strategies, although the self-contained BwS method could be combined with variable-width binning.

In Figure 2.7 I compare the results obtained from the three different methods by ordering the verbs from left to right by decreasing BwS p-value, averaged over the three temporal binnings. Each panel corresponds to a different analysis method, and indicates the p-value for the hypothesis of pure drift for each verb and binning protocol. Higher p-values are more suggestive of the historical changes being due to drift: these are represented with colours ranging from light to dark blue, with darker colours representing higher p-values. Meanwhile, low p-values point towards other forces (such as selection) being present and are represented with different shades of red. Notably, this figure only reports the significance level of selection, and not its direction – it may favour the irregular form at the expense of the regular one (irregularisation) just as much as it may favour the regular

form at the expense of the regular one (regularisation). While I use the standard p-value threshold of 0.05 in the transition between blue (drift) and red (selection) in this representation, I acknowledge that these mechanisms lie in a continuum by making the transition between these extremes smooth.

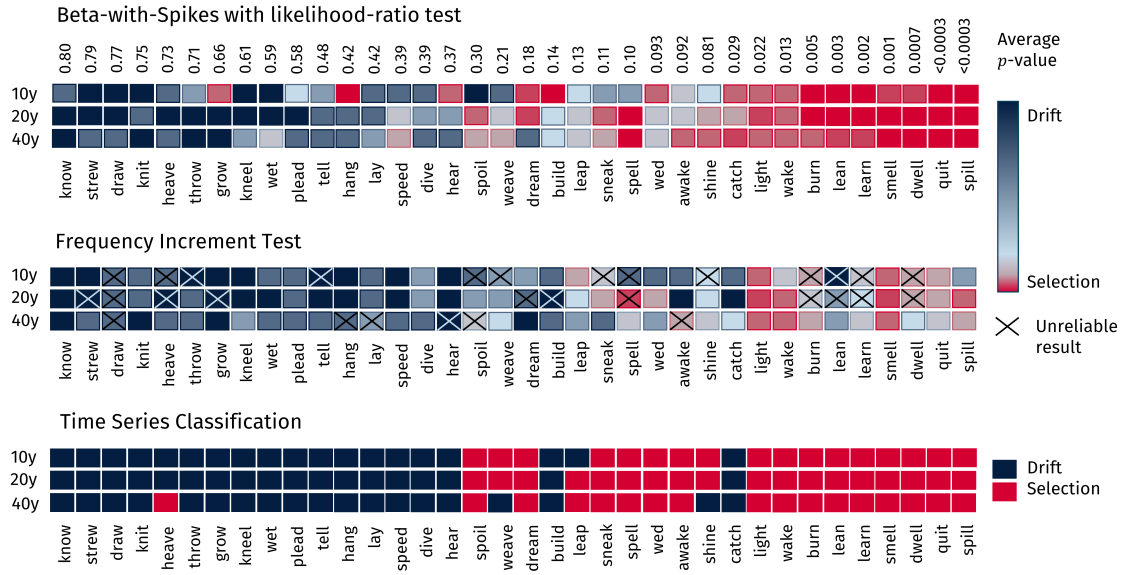

Figure 2.7 *Results for the detection of selective forces in 36 COHA verbs, with three different methods and for three different temporal binnings of 10, 20 and 40 years. Results for both the FIT and BwS likelihood-ratio algorithms produce a p-value for the pure drift hypothesis. Blue shades represent higher p-values (i.e. similar likelihoods of the models with pure drift and with selection), while red shades represent p-values under the traditional* 0.05 *threshold of significance for selection. Time series where the normal approximation that FIT relies on is inaccurate are crossed out. Results for the TSC method from Karsdorp et al. (2020) [106] are classified in a binary way as either drift or selection. The average p-value across the three bins widths obtained through the BwS algorithm is shown along the horizontal axis. The correlation coefficients between the p-values obtained with different methods are 0.63 (Pearson) between FIT and BwS, 0.68 (biserial) between TSC and FIT, and 0.62 (biserial) between BwS and TSC. The BwS method gives results consistent with TSC when FIT is unreliable.*

We can see from Figure 2.7 that the three distinct methods give broadly consistent results, with those verbs towards the left being more compatible with change through pure drift, and those to the right with change from selection. Results obtained through the FIT method are generally consistent with those obtained with the BwS method. However, 30 of the FIT results (27.8% of the total) are flagged as 'unreliable' due to a failed normality test [124]. These reliability issues

are designed out of the BwS method, as it does not require normally-distributed increments. The higher precision of the BwS at high selection strengths leads to higher significance (lower p-values) in its detection of selective forces when compared to the normal approximation, leading to redder colours in Figure 2.7.

The TSC appears to give a cleaner classification of verbs according to drift and selection, and greater consistency with different choices of bin size. This is likely due in part to the training data being subjected to the same binning protocol as the empirical time-series, but also because a strict threshold was applied to the neural network's output value to partition into the two classes. While the TSC neural networks produce a value between 0 and 1 as their output, making it more nuanced than this binary classification would suggest, this number is not a probability or a p-value like those produced by BwS or FIT. Thus, a (more) arbitrary threshold is necessary in order to classify time series as driven by drift or selection. A higher or lower threshold would put the boundary between the two classes in a different place. This hinders the interpretability of the result and the estimation of significance levels.

The BwS results further showcase variation in p-values under different binning strategies, previously observed within the FIT analysis [124]. As discussed in Section 2.1.5, this variability is an inherent feature of time series data. This motivates a more detailed investigation of the classifiability of individual time series. A time series that shows limited variation in parameter values under different temporal binnings is more classifiable than one that shows more variation. With our interest in selection, the two most relevant parameters are s, the selection strength, and the p-value associated with the drift hypothesis. We can visualise the variation in these parameters by performing Principal Components Analysis (PCA)[125] on combinations obtained through different binning strategies (in this case, bins of 10, 20 and 40 years). PCA takes a set of points in parameter space (in our case, the selection strength and p-value for each binning) and reduces it to an ellipse centred on the average of these points, and semi-major axes given by the square root of the eigenvalues of their covariance matrix. In other words, the interior of the resulting ellipses indicates the range of variation of the two parameters over different binning strategies. This way, they provide a visualization of not just the average, but the uncertainty and covariance of s and the p-value under different binning strategies. I show these ellipses for the COHA verbs in Figure 2.8. The upper panel contains the full range of p and s values obtained through the analysis, while the lower panel zooms in on the region

where the drift p-value is smaller than 0.05 (i.e., the conventional threshold for rejecting the null hypothesis). A correlation can be seen between the maximum likelihood value of s and the p-value (both through the positions and rotation angles of each ellipse).

As previously mentioned, the variant frequency x in this analysis is the fraction of irregular forms, and thus $s > 0$ in Figure 2.8 corresponds to selection favouring the irregular variant. The ellipses that lie entirely within the lower panel correspond to the verbs whose diachronic change is most likely to be driven by selection. We find four verbs with positive selection (*catch*, *light*, *wake* and *quit*), which corresponds to them becoming more irregular over time, and six verbs (*learn*, *lean*, *burn*, *smell*, *dwell* and *spill*) with negative selection, and thus regularising over time. Across the entire plane, there is evidence of both regularisation and irregularisation, although in most cases it is difficult to rule out drift as an explanation for the changes, as was observed by Newberry et al. (2017) [105].

To summarise, I have shown in this section that the BwS method can be readily applied to historical corpus data for changes in the frequencies of linguistic variants. It provides estimates of parameters in the Wright-Fisher model that do not rest on an assumption that frequency increments are drawn from a normal distribution, and there is broad consistency in the strength of support for a drift hypothesis with complementary methods.

2.2.2 Competing linguistic motivations in English verbs

In the previous section we observed a split between some verbs that were regularising and some that were irregularising. While the extension of regular inflection at the expense of irregulars seems to be the norm (e.g. [129, 130]), irregularisation is however an attested phenomenon. Cuskley et al. (2014) [127] found that the processes of regularisation and irregularisation tend to take place with similar frequency, something that is also perhaps suggested by Figure 2.8, which shows a similar density of verbs along the branch with positive s (towards irregularity) and negative s (towards regularity). Ringe and Yang (2022) [128] suggest that irregularisation may occur if the number of verbs within an irregularity class is high enough to surpass a productivity threshold. Following Bybee (2001) [131] and Prasada and Pinker (1993) [132], both Cuskley et al. (2014) [127] and Newberry et al. (2017) [105] propose phonological analogy as a potential mechanism for irregularisation. In the terms of this work, this would

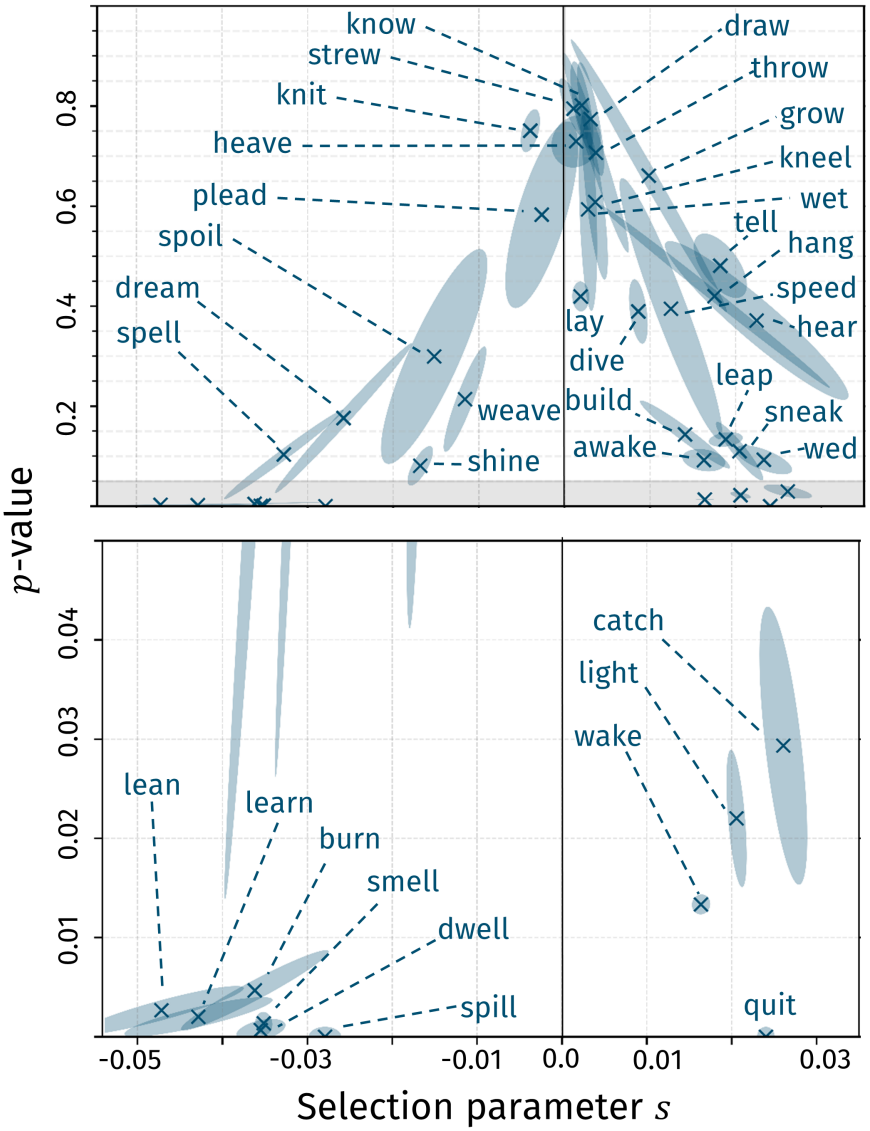

Figure 2.8 *Variability in the selection strengths s and p-values for the null hypothesis pure drift for the COHA verbs. Each cross shows the mean value of the two parameters for each verb obtained when aggregating frequencies into temporal bins of different lengths. Each ellipse indicates the variability in the parameters at the level of one standard deviation. The vertical axis is an indicator of selection, defined as one minus the p-value associated with the drift hypothesis. The lower panel shows those verbs that fall within the range of p-values that is conventionally used to reject the null hypothesis for a single observation. In this panel we see a clear split into those that are regularising (negative s) and are irregularising (positive s).*

correspond to the regular paradigm (adding ⟨-ed⟩) contributing a negative value to s whilst rules that apply only to a specific subset of verbs contribute a positive value to s. Note that we do not necessarily imply that these contributions are additive: for example, in optimality theory [133, 134], higher-ranked rules take precedence over lower-ranked rules. In general, we may regard opposing forces on linguistic variation as arising from competing motivations which have been discussed in a variety of language change contexts (e.g., [31, 75, 135–139]). By whatever mechanism this opposition is resolved, an overall positive s value here indicates that the irregularising rule is dominant.

In this section, I investigate a distinct motivation that may favour irregularisation, namely the phonological simplicity that is afforded by omitting a sound repetition that would occur under application of the regular rule. Specifically, I consider verbs whose infinitives end in alveolar stops (/d/ or /t/) and have an irregular past form where the regular ⟨-ed⟩ termination is omitted. Examples include *"I bled"* instead of *"I bleeded"* or *"She bet"* instead of *"She betted"*. Verbs where devoicing of final /d/ or changes in the root vowel take place on top of the omission of the termination are also considered. Thus, I hypothesise that the regular form is preferred from the point of view of inflectional simplicity (i.e. using the regular everywhere leads to a simpler inflectional system), while the irregular form is favoured by phonological simplicity. By applying the BwS algorithm to estimate the s parameter (and in particular, its sign), we can assess how these competing motivations play out.

For this investigation I switch to the 2019 English Google Books corpus [113], as the number of verbs falling into this category and whose past tense forms are both sufficiently frequent and can be reliably identified is relatively small. The larger size of Google Books relative to COHA allows more examples to be included. The validity of using frequency data from the Google Books corpus to draw conclusions on cultural evolution and language change has been questioned by Pechenick et al. (2015) [140] due to the over-representation of scientific literature in the English sub-corpus throughout the 20th and 21st centuries. While they propose restricting studies of cultural and language change to the fiction sub-corpus, I believe that using frequency data from the general English sub-corpus is justified for the purposes of our study. First, this work rests on the comparison between two data sets of English verbs differing only in their phonology. It is reasonable to assume that, if any bias exists in scientific texts regarding the use of irregular or regular forms of verbs, this bias will not be phonologically conditioned, thus

maintaining the validity of the comparison between both data sets. Secondly, I have chosen verbs that are reasonably present in both the general English corpus and the English Fiction corpus, so potential biases towards uncommon verbs in scientific literature should not be an issue. Thirdly, qualitative observations show no increased use of archaic verbal forms in scientific texts by virtue of their formality. Finally, the general English sub-corpus will contain more words than the restricted fiction sub-corpus, thus reducing the effect of sampling noise on results.

I identified 19 English verbs whose irregular and regular forms both show usage above 1% at least in one 5-year bin in the Google Books corpus in the considered time frame (1809 to 2009). These verbs are: *bend*, *bet*, *bite*, *blend*, *build*, *fit*, *glide*, *knit*, *light*, *pat*, *plead*, *quit*, *slide*, *speed*, *spit*, *thrust*, *tread*, *wed*, and *wet*. A difficulty in the analysis is that the irregular past-tense form can coincide with certain present-tense forms (e.g. *I hit* may be either present or past tense). A major exception occurs when the verb is preceded by a third-person singular pronoun (e.g., the present *he bets* versus the irregular past *he bet*), which can easily be distinguished in the bigram dataset. This separation is not perfect: for example, certain English varieties do not use the third person marker -s, but I consider the effect of these contributions to be negligible in the corpus. I also kept only those cases where the pronoun was judged to appear at the start of a sentence (by virtue of capitalisation), so as to exclude contexts where the pronoun is followed by the infinitive in a question or an inversion. Again, there are situations where capitalised pronouns can appear mid-sentence, but these are also rare. With this, total counts of usage for verbs with non distinct irregular past tense forms range roughly between $2,600$ (*knit*) and $120,000$ (*pat*), while counts for verbs whose irregular past tense is distinct from their base form range between $600,000$ (*glide*) and $40,000,000$ (*build*).

In order to formally test whether a potential bias towards irregularisation is significant, a similar analysis was carried out on a baseline set of 34 English verbs whose base form does not end in /d/ or /t/. Data was extracted from Google Books and all verbs satisfy the same conditions on minimal usage in the time frame of interest (1809-2009). The chosen verbs are: *awake*, *blow*, *burn*, *catch*, *cleave*, *creep*, *dive*, *dream*, *dwell*, *freeze*, *grow*, *hang*, *heave*, *hew*, *kneel*, *lean*, *leap*, *learn*, *shake*, *shear*, *shine*, *slay*, *slink*, *smell*, *sneak*, *spell*, *spill*, *spoil*, *strew*, *string*, *strive*, *swell*, *wake* and *weave*. Total usage for these verbs in the Google Books corpus for the specified period ranges between $211,000$ tokens (*slink*) and

$31,900,000$ (*learn*), in the same orders of magnitude as the /d/,/t/ set.

The maximum likelihood parameters for these 53 verbs are given in Appendix A.2. Here, results are visualised by plotting ellipses in the plane spanned by the selection strength and the indicator of selection, following the same procedure as previously described for the COHA verbs, albeit with the addition of a 5-year temporal binning strategy. With this, each ellipse in the *s*-*p* plane for each verb is produced with the results of the analyses of at most four temporal binnings.

The upper panel in Figure 2.9 shows the results for all 53 verbs. For the purpose of comparing the two sets of verbs, I partition the *s*-*p* plane into four regions: those with positive or negative selection strengths; and those where the p-value falls above or below 0.05. The lower panel of Figure 2.9 zooms in on this latter region, which we may regard as showing evidence of selection. In both panels, red crosses and ellipses correspond to verbs ending in alveolar stops, while blue crosses and ellipses correspond to verbs in the baseline set. Given our interest in irregularisation, three groups of verbs can be identified. 16 verbs (*awake*, *bend*, *bet*, *bite*, *catch*, *fit*, *hang*, *light*, *quit*, *shake*, *slide*, *sneak*, *spit*, *strew*, *wake*, *wed*) have their confidence regions (ellipses) completely contained in the region of likely selection of the irregular form ($p < 0.05$ and $s > 0$, lower-right panel). Of those, 9 are in the alveolar stop set and 7 are in the baseline set. Six verbs (*freeze*, *kneel*, *leap*, *plead*, *swell*, *thrust*) have confidence regions only partially contained in this region of the *s*-*p* plane, indicating that, while selective forces towards the irregular form are a plausible explanation to their dynamics, the pure drift hypothesis cannot be confidently ruled out. The remaining 31 verbs (8 in the alveolar stop set, 23 in the baseline set) have confidence regions contained entirely outside this region of likely irregularisation.

These results suggest that verbs in the alveolar stop set are more likely to be selected towards their phonologically simpler irregular form than their counterparts in the baseline set. To test the significance of these findings, I construct the 2×3 contingency table shown in Table 2.1, where one dimension expresses belonging to the alveolar stop or the baseline sets, while the other dimension expresses whether the verbs' ellipse falls in the irregularisation region in the bottom panels of Figure 2.9. The p-value for the null hypothesis that the baseline and alveolar stop verbs are drawn from the same distribution is 0.031, as obtained by applying the G-test of goodness-of-fit to the contingency table [141]. This indicates that the specific rule favouring phonological simplicity likely outcompetes a general tendency towards regularity.

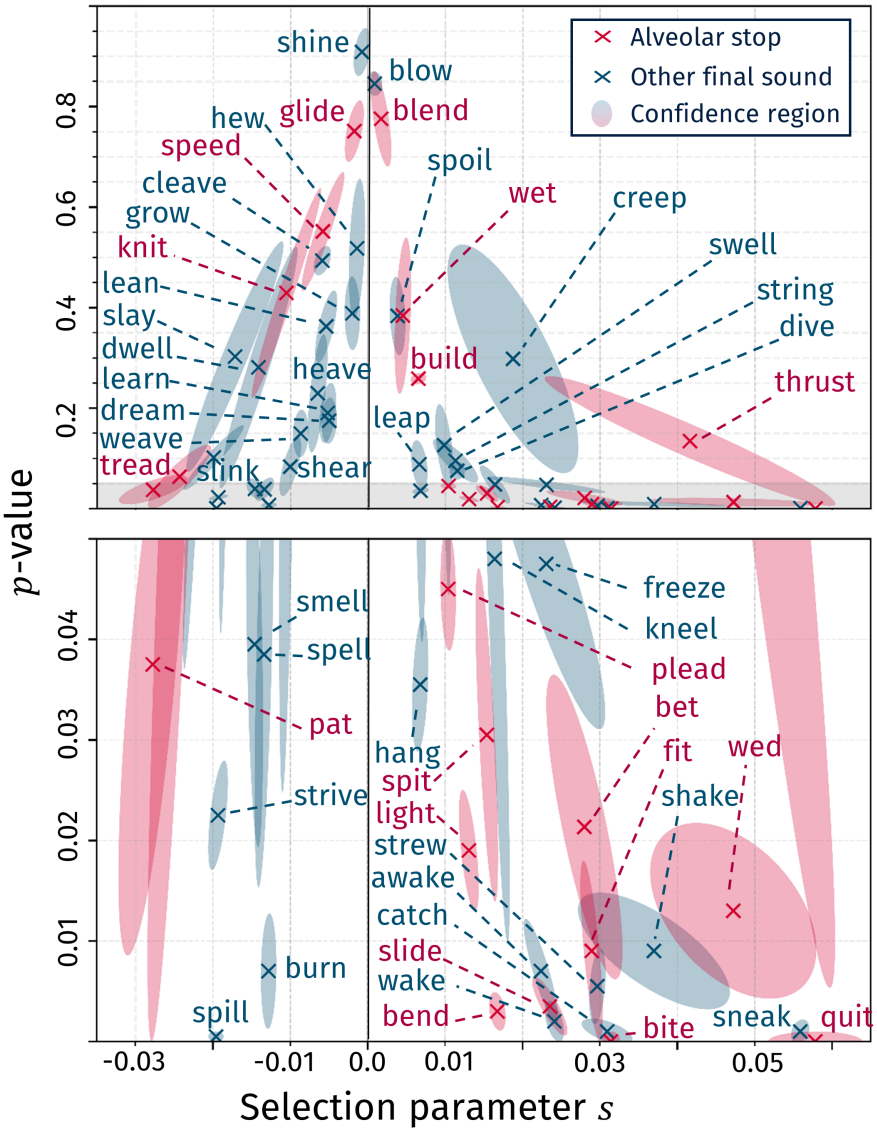

Figure 2.9 *Parameter estimates for verbs ending in alveolar stops (red) and verbs in the baseline set (blue) in the Google Books data set. The top panel shows the entire range of drift p-values and includes all 53 verbs. The bottom panel is restricted to $p < 0.05$, thus focusing on verbs that are likely to be undergoing directed selection. The distribution of verbs in the alveolar stop set seems to be skewed to the region where $s > 0$ and $p < 0.05$, suggesting they are more likely to be irregularising than the other verbs.*

	Irregularising	Inconclusive	Non-irregularising
Alveolar stop set	9	2	8
Baseline set	7	4	23

Table 2.1 *Contingency table for the comparison of irregularising behaviour between the set of verbs ending in alveolar stops and the baseline set. Irregularisation is significantly more common amongst verbs ending in alveolar stops, with a p-value of 0.031 as provided by the G-test.*

It is possible that other effects may be responsible for this subset of verbs tending to irregularise. For example, it is well understood that higher frequency items tend to tolerate greater irregularity [142]. Given the selection criteria imposed to arrive at the set of 12 verbs in this analysis, it is possible that the sample is skewed towards higher frequency and more irregular forms. However, as noted, the total token counts for both the baseline set and the alveolar stop set span similar ranges, and also have similar averages (of around 5 million for both sets). Therefore, I consider this alternative explanation unlikely.

To summarise, in this section I have shown that, by focussing on a subset of verbs that are subject to specific combination of competing motivations, the Wright-Fisher model combined with the BwS approximation can be used to determine the net effect of this competition. Specifically, I have presented evidence that phonological simplicity dominates inflectional simplicity in this competition, suggesting perhaps that this is an instance of an OCP constraint (Obligatory Contour Principle, [143, 144]). OCP constraints disfavour pairs of identical or near-identical consonants from being in close proximity to each other. In particular, the constraint here appears to be an OCP-place constraint [145–147], meaning that it does not just affect identical consonants, but all alveolar stops independently of voicing.

2.3 Discussion

In this chapter, I have introduced a method for obtaining reliable maximum-likelihood estimates of parameters within the Wright-Fisher model from time-series data of variant frequencies, and applied it to the analysis of evolutionary trajectories in language change. This approach is underpinned by the Beta-with-Spikes (BwS) distribution [96, 107] which captures the essential features of the Wright-Fisher model. These are the possibility of extinction or fixation of an

allele, accounted for by delta functions (spikes), and the continuous distribution governing the evolution of unfixed variants, which is well-characterised by its mean and variance.

The challenge in utilising the BwS approximation is accurately determining appropriate parameter values. Earlier works [95, 107] used Taylor series expansions to estimate how parameter values should change from one generation to the next. These have the benefit of being simple to evaluate, but the truncation of the Taylor series results in the approximation being unreliable when the selection coefficient is large. In particular, it can generate parameter values that cause the BwS distribution to be ill-defined. Here, I have developed a *self-contained* approximation, where the BwS distribution is used as the initial condition for one step of Wright-Fisher evolution, and a new BwS distribution is fit to the intermediate distribution that results. In this approach, the approximating distribution remains well-defined, and provides an adequate approximation to the Wright-Fisher model even when selection is strong. I have demonstrated the reliability of the method for the Wright-Fisher model with selection by comparing distributions directly, and by determining the error on the maximum-likelihood estimate of the selection coefficient for artificial time series where the true value is known.

When working with the Wright-Fisher model in a cultural evolutionary setting, another challenge arises in the indeterminacy of the generation time. The possibility of generations shorter than the temporal resolution of the data can be avoided altogether by alluding to scaling properties of the effective population size and selection strength. The converse possibility, that generations encompass multiple independent measures of variant frequency, needs to be acknowledged in any analysis. Here, I have done so by applying Principal Component Analysis [125] to results obtained under different choices of temporal binning. This allows for analyses that acknowledge the variation in estimates of evolutionary parameters and p-values that stem from this uncertainty in the generation time.

In Section 2.2.1, I benchmarked the self-contained BwS method through applications to previously-published data in language change. The method was applied to the set of 36 COHA verbs previously investigated using the Frequency Increment Test (FIT, [105, 124]) and Times Series Classification (TSC, [106]). FIT delivered unreliable results due to shortcomings of the normal approximation that it relies on, and TSC produced consistent results limited to a binary classification of time series into drift or selection. The self-contained BwS method

proved to combine the best features of both models, by being both consistent in a wide range of scenarios and producing easily interpretable results and maximum-likelihood estimates of the evolutionary parameters.

In Section 2.2.2, I applied this method to a novel data set, with the goal of showcasing its applicability to the testing of hypotheses on the mechanisms of historical language change. A comparison between a baseline set of verbs from the Google Books corpus and a set where the irregular past tense is formed by deletion of a repeated consonant reveals that they are distributed differently across the space of selection strengths s and drift p-values. Specifically, the phonological simplicity arising from coalescence or omission of the /ɪd/ termination tended to be favoured over the inflectional simplicity of the regular form. In principle, this method could be used to determine the relative importance of other pairs of constraints that correspond to opposing selective forces.

Despite these promising results, there are inevitably some limitations. A crucial one lies in the assumption that sample sizes are large enough that the uncertainty on allele frequency estimates drawn from them can be neglected. In reality, this may not be the case, which may warrant the use of a hidden Markov framework, as proposed by Bollback et al. (2008) [108] in the context of genetic time series data. A naïve numerical implementation of this scheme would involve integrating over each of the (now hidden) frequencies x_{t_i} in equation 2.30, which dramatically increases the computational demands. One way to circumvent the additional integrals is to employ an expectation-maximisation algorithm. However, I have found that if the effective population size is considered a free parameter, expectation-maximisation pushes this towards infinity due to deterministic trajectories being favored by the algorithm. Therefore, some further work is needed to develop tractable methods for jointly estimating effective population size and the selection coefficient when working with data subject to sampling uncertainty.

In summary, despite certain limitations, the method introduced here allows evolutionary parameters to be reliably estimated, and when combined with empirical likelihood ratio tests, can be used to test departure from a variety of null hypotheses. Although we have focused on the Wright-Fisher model with frequency-independent selection, it could be extended to models that involve other evolutionary processes. Here, I have shown that it is possible to draw inferences about contributions to selection from different sources, as exemplified in the analysis of competition between regular and irregular forms in English

verbs. By appealing to a wider range of corpora and instances of change, it may become possible to identify general mechanisms that are invariant over time and operate cross-linguistically, and are thus informative about language universals in general. Furthermore, the method is not specific to linguistic variation, and could be used to address similar questions in other instances of cultural evolution.

Chapter 3

Detection of changing social trends in historical data

In the previous chapter, we worked under the assumption that the Wright-Fisher evolutionary parameters (the effective population size N and the selection strength s) were constant over time. This is a reasonable assumption in the case of competition between regular and irregular verbs, due to the factors favouring one over the other likely being cognitive or linguistic in origin. By contrast, evolutionary forces originating in social dynamics are inherently time-dependent: the cultural environment in which language is embedded is itself adaptively evolving, often in timescales comparable or shorter than the typical timescales of fixation of lingueme variants. When this is the case, we may expect evolutionary forces to change over time.

Selection – effects biasing the competition process in favour of one variant – is particularly prone to being affected by changing societal attitudes. These may involve prestige, taboo, significance to a social group, or a desire to go against social norms [72, 148, 149]. Prestige, which refers to the social evaluations attached to specific lingueme variants, is usually rooted in class dynamics, with variants used by dominant classes being the ones receiving a more positive evaluation [150, 151]. Prestige variants may be enforced or encouraged in a speech community through institutional efforts, in what is usually termed *prescriptivism*. Regulating institutions may introduce changes to the prestige variety in the form of prescriptive grammar and spelling rules [152–155]. Taboo, whereby certain words may be considered unacceptable in some social contexts, is particularly

common among words referring to race and ethnic background, sex, disability, or disease. Taboo words tend to be replaced by socially-acceptable euphemisms, which may in turn be tainted by the connotations of the taboo words that they were originally replacing [156–158]. This rapid change in social perception is sometimes known as the *euphemism treadmill* [159]. Rapidly-changing attitudes towards linguistic variants may also appear due to cultural trends that pattern in complex ways with sociolinguistic variables not limited to class, such as gender, age, or ethnic background [72, 160, 161]. Anti-conformity biases have been argued to be behind the rise and fall of popularity of baby names, amongst other cultural trends [162]. Whatever their origin, changing social attitudes should be more accurately modelled using time-dependent selection. This chapter's goal is to develop a methodology able to achieve this in an empirically applicable way.

Some insight may again be gained by looking at existing developments in the context of evolutionary biology. Much attention has been paid to the ecology [163–165] and physiology [166–168] of microbial populations under fluctuating environments. Consequently, a variety of mathematical models have been developed for the description of population dynamics under variable environmental conditions [169–173]. Within the Wright-Fisher paradigm, Kimura (1954) [174] and Karlin and Levikson (1974) [175] approached the modelling of a stochastic environment by replacing the constant selection parameter s with a random iid sequence of selection parameters $(s_i)_{i\geq 1}$ changing at each generation i (see Huillet (2011) [176] for a thorough review). While interesting, these models are not practical for the empirical exploration of variable cultural dynamics, where we are interested in sustained changes, rather than random fluctuations at every generation.

Other well-established algorithms like change-point analysis [177] exist for the detection of change in time series, but they suffer from shortcomings that make them inadequate for a more nuanced analysis of change in linguistic datasets. First, change-point analysis is based on the assumption that the data is distributed around a constant average before and after a change, which changes the value of said average instantaneously. This makes this methodology only fit for the detection of rapidly occurring S-shaped curves of language change, where the usage frequency of a variant quickly changes and stabilises. Secondly, change-point analysis provides no extra linguistic information, as it does not assume a model of the underlying evolutionary dynamics.

Here, I propose a model of language change under variable cultural environments

in which evolutionary parameters in the Wright-Fisher paradigm change instantaneously at specific transition times in the trajectory of the time series of variant frequency data, remaining constant before and after the transition. I introduce this model, as well as maximum-likelihood methods for its application to the detection of changing social trends in historical data, in Section 3.1.1.

Selection is not the only evolutionary force that may change over time. Drift is related to the cultural relevance of a lingueme, as it increases its memory retention, making its usage more consistent and thus less fluctuating [4, 5]. Changes in lingueme relevance can be brought about by changes in technology and societal values. However, historical language data tends to be more scarce the further back in time one goes, leading to variations in sampling noise throughout time series that may be misconstrued as significant changes in N. In Section 3.1.2, I introduce a noise equalisation technique with the goal of alleviating this issue.

Finally, in Section 3.2, I apply this methodology to time series of language data. I focus on historical change in Spanish, with special focus on the detection of spelling reforms, which provide us with well-documented changes to evaluate the precision of the change-detection algorithm. This chapter is based on work published in Guerrero Montero and Blythe (2023) [98] and Guerrero Montero et al. [97].

3.1 Methods

3.1.1 Time-dependent evolutionary parameters

In Section 2.1.4, I discussed likelihood-maximisation methods for the estimation of evolutionary parameters in the Wright-Fisher model. It is straightforward to extend this maximum-likelihood approach to the case where different parameters apply over different time intervals. In particular, a single abrupt change can be modeled as two sets of parameters (N, s) that apply before and after a transition time T, such that the single-generation transition probability of the process becomes:

$$\mathrm{P}\left(x_{t+1} \,|\, x_t; N_1, s_1, N_2, s_2, T\right) = \begin{cases} \mathrm{P}_{\mathrm{WF}}\left(x_{t+1} \,|\, x_t; N_1, s_1\right) & \text{if } t < T \\ \mathrm{P}_{\mathrm{WF}}\left(x_{t+1} \,|\, x_t; N_2, s_2\right) & \text{if } t \geq T\,. \end{cases} \tag{3.1}$$

Here, $\mathrm{P_{WF}}$ represents the two-variant Wright-Fisher transition probability as defined in equation 2.3. For $t < T$, the parameters take values (N_1, s_1), and for $t > T$, they take values (N_2, s_2).

The optimal parameters $\hat{N}_1$, $\hat{s}_1$, $\hat{N}_2$, $\hat{s}_2$ and $\hat{T}$ for a time series of variant frequencies $X = \{(x_t, t)\} = \{(x_1, t_1), (x_2, t_2), ..., (x_m, t_m)\}$ can then be found by maximizing the likelihood function

$$\begin{aligned}\mathcal{L}\left(N_1, s_1, N_2, s_2, T \,|\, X\right) &= \prod_{i=1}^{m-1} \mathrm{P}\left(x_{i+1} \,|\, x_i; N_1, s_1, N_2, s_2, T\right) \\ &= \prod_{i=1}^{T-1} \mathrm{P_{WF}}\left(x_{i+1} \,|\, x_i; N_1, s_1\right) \prod_{i=T}^{m-1} \mathrm{P_{WF}}\left(x_{i+1} \,|\, x_i; N_2, s_2\right) .\end{aligned} \tag{3.2}$$

Again, we can use likelihood ratios to determine whether the time division provides a significantly better explanation of the data. Specifically, consider the test statistic:

$$\lambda = 2 \ln \left(\frac{\mathcal{L}\left(\hat{N}_1, \hat{s}_1, \hat{N}_2, \hat{s}_2, \hat{T} \,|\, X\right)}{\mathcal{L}\left(\hat{N}, \hat{s} \,|\, X\right)} \right) , \tag{3.3}$$

which compares the optimal likelihood of a model where both N and s change at a time T with one where N and s are fixed for the entire trajectory.

There is a methodological caveat that needs to be discussed here. Wilks' theorem, which is usually applied in the estimation of the p-value for the null hypothesis in model comparison of this nature, states that for long enough time series the test statistic λ can be assumed to be χ^2-distributed. This theorem relies on the assumption that the model parameters take continuous values in a Euclidean space and that the likelihood function is continuous for all parameter values [121]. This is not true for the transition time T, which only takes on integer values measured in generations. Thus, λ cannot be assumed to be χ^2-distributed, no matter the length of the time series. Instead, a p-value must be obtained by constructing the empirical distribution of λ from trajectories in which there is no time division, similarly to the ELRT discussed in section 2.1.4. This approach can be iterated to include $n+1$ abrupt changes by further subdividing the trajectories, provided that n abrupt changes were already proven significant.

To test the ability of this algorithm to detect changes in selection strength, I first benchmark it here using artificially generated time series. I iteratively sample allele frequencies from equation 3.1 for T generations. At generation $T/2$, the

selection strength goes from $s = 0$ to $s = \Delta s$. For each set of values of $500 \leq N \leq 10000$, $6 \leq T \leq 50$ and $0.001 \leq \Delta s \leq 0.3$, I generate 2000 time series, compute their likelihood ratios using equation 3.3 and find the associated p-values. If a p-value is under the standard significance threshold, $p < 0.05$, the change in selection is considered detected.

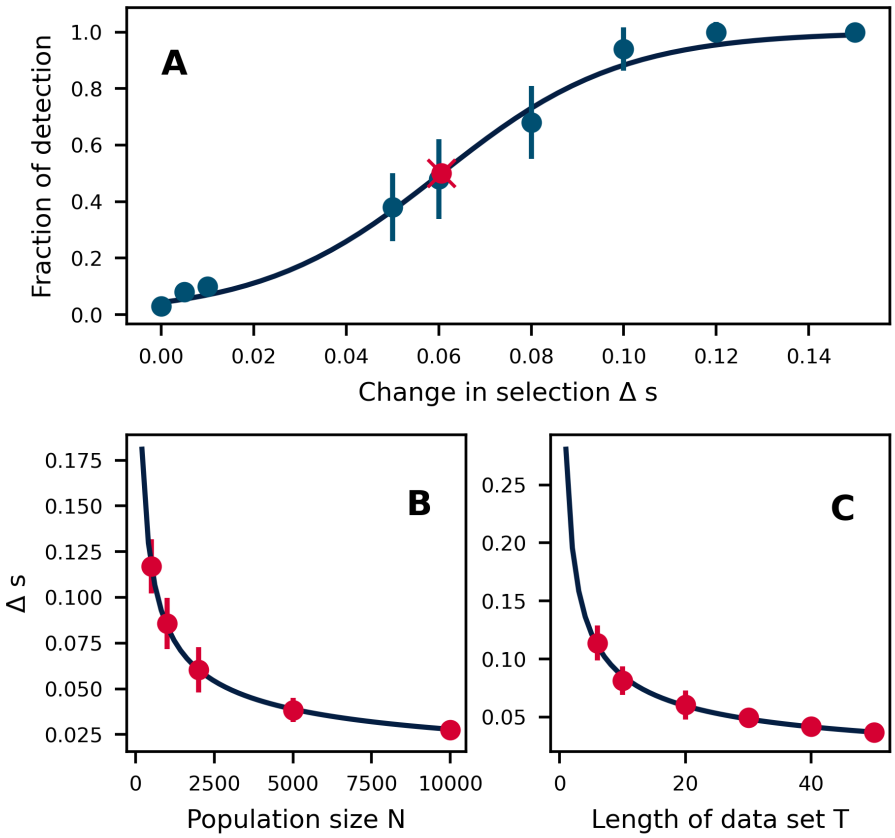

Figure 3.1 *(A) Fraction of significant selection as a function of Δs for $N = 2000$, $T = 20$, together with fitted logistic function and estimated characteristic value of Δs. The fitted logistic function (eq. 3.4) has parameters $a = -3.1 \pm 0.3$, $b = 51 \pm 5$, $r^2 = 0.992$. (B) Characteristic Δs as a function of N for fixed $T = 20$. The empirically fitted power law has proportionality constant $c = 2.33 \pm 0.12$, exponent $d = -0.481 \pm 0.008$, and $r^2 = 0.994$. (C) Characteristic Δs as a function of T for fixed $N = 2000$. Power law has parameters $c = 0.284 \pm 0.015$, $d = -0.52 \pm 0.02$, $r^2 = 0.999$. Blue dots represent empirical points obtained as the average of the detection of change in 2000 artificially generated time series. Red dots represent characteristic values of Δs, obtained from interpolated logistic functions.*

Figure 3.1A shows the fraction of time series for which changes in s are detected as a function of the change in selection strength Δs at fixed $N = 2000$ and $T = 20$. As expected, this fraction grows monotonically with Δs. Empirically, it

is well fit by the logistic function,

$$f(\Delta s) = \frac{1}{1 + \exp\left(-a - b\Delta s\right)}\,, \tag{3.4}$$

which allows us to identify a characteristic Δs through the value for which the fraction of detected changes equals one half. That is, above this characteristic Δs, we are more likely to detect the change than not.

Figures 3.1B and 3.1C show how this characteristic value varies with effective population size N and the number of generations T, respectively. They show that the larger the effective population size or the longer the time series, the smaller the change that can be detected. In both cases this is expected: fluctuations diminish as N increases, thereby increasing the signal-to-noise ratio; similarly, longer time series provide more information and allow stronger inferences to be drawn. Power laws can be empirically fitted to the data points in Figures 3.1B and 3.1C. Notably, both power laws have an exponent of roughly -0.5, and they can be combined under the assumption that their proportionality constants and exponents remain constant under changing N and T, leading to an empirical law for the characteristic Δs:

$$\Delta s \approx \frac{A}{\sqrt{NT}}\,, \tag{3.5}$$

where $A = 11.2 \pm 0.6$. This is useful to examine the effects of temporal binning on the detectability of change. In particular, consider a choice of generation time such that t data points in the time series are included in one generation. Following the arguments presented in Section 2.1.5, the maximum-likelihood population size N of this time series can be assumed to scale roughly as $N_t = tN$, where N is the maximum-likelihood value that would be obtained when analysing the time series under the assumption that each generation contains one data point only. Conversely, the transition time T_t would scale as $T_t = T/t$. Thus, the characteristic Δs remains unchanged under generation rescaling following equation 3.5. However, the maximum-likelihood s_t scales with temporal binning roughly as $s_t = ts$: while the threshold Δs would remain the same, the maximum-likelihood values of s would increase proportionally as greater generation times are chosen, thus making changes more easily detectable. This, of course, comes at a cost: the choice of generation time also determines the resolution of the search for the transition time T, thus decreasing the accuracy of its estimation.

Overall, these results indicate that, if present, temporal changes in selection strength can be detected using the proposed method, assuming trajectories are of

sufficient length and sufficiently low fluctuation level. When these are insufficient, coarser temporal binning may achieve significant results at the cost of the loss of accuracy.

3.1.2 Sampling error equalisation

When dealing with time-series of data, estimates of a variant's frequency may derive from samples of different sizes at different times, and therefore be subject to greater or lesser degrees of sampling error. Specifically, when working with corpus data, the general trend is for data to become scarcer as earlier time periods are examined. The Wright-Fisher model in general, and the likelihood-maximisation method laid out in the previous section in particular, are unable to differentiate sampling fluctuations from those arising from drift in the inter-generational transmission. To disentangle both effects, it is helpful to equalise the amplitude of the sampling error across the time series. Then, any detected changes in the drift parameter N can be ascribed to the process that generates the underlying frequencies, as changes in how samples are constructed have already been accounted for.

One way to do this is to create subsamples of the larger data sets in the time series, with the subsample size chosen in such a way that the contribution from sampling is of equal magnitude across the time series. That is, consider samples of corpus data of size n_t at some time t, within which a fraction x_t of tokens are of one particular lingueme variant of interest. These samples can be modelled as the result of binomial sampling from the corpus, with sample size n_t and an unknown 'true' success probability for the variant of interest p_t. Consider also the smallest original sample size across the entire time series, which we call $n_{\min}$. We then construct a binomial sample of size m_t and success probability x_t in which the new fraction of the variant of interest is given by y_t. The goal is then to choose m_t in such a way that the mean and variance of the final fraction y_t at time t are consistent with the variance of a single binomial sampling from the corpus of fixed size $n_{\min}$ and success probability p_t. The key point to note here is that the original sampling process already contributes some variance to y_t. Therefore, m_t will depend on the original sample size so that the additional variance arising from resampling gives the desired overall variance.

To determine the appropriate sample size m_t, we consider the first two moments of the random variable y_t. Given some value of x_t via the original sampling

process, the binomial resampling process implies that

$$\mathbb{E}(y_t \,|\, x_t) = x_t \tag{3.6}$$

$$\mathbb{E}(y_t^2|x_t) = \left(1 - \frac{1}{m_t}\right) x_t^2 + \frac{1}{m_t} x_t \,. \tag{3.7}$$

We now average over all possible realizations of the original sampling process to determine the first two moments of the resampled frequency y_t, finding

$$\mathbb{E}(y_t) = \mathbb{E}(x_t) \tag{3.8}$$

$$\mathbb{E}(y_t^2) = \left(1 - \frac{1}{m_t}\right) \mathbb{E}(x_t^2) + \frac{1}{m_t} \mathbb{E}(x_t) \,. \tag{3.9}$$

Although the true variant frequency p_t is unknown, we have that

$$\mathbb{E}(x_t) = p_t \tag{3.10}$$

$$\mathbb{E}(x_t^2) = \left(1 - \frac{1}{n_t}\right) p_t^2 + \frac{1}{n_t} p_t \,. \tag{3.11}$$

Substituting these expressions into equations 3.8 and 3.9, we find that

$$\begin{aligned} \mathrm{Var}(y_t) &= \mathbb{E}(y_t^2) - [\mathbb{E}(y_t)]^2 \\ &= \left[1 - \left(1 - \frac{1}{n_t}\right)\left(1 - \frac{1}{m_t}\right)\right] p_t(1 - p_t) \,. \end{aligned} \tag{3.12}$$

This is the variance that would be obtained if the original sampling process involved a sample of size:

$$\frac{1}{n_{\min}} = 1 - \left(1 - \frac{1}{n_t}\right)\left(1 - \frac{1}{m_t}\right) \,, \tag{3.13}$$

where, as previously mentioned, $n_{\min}$ is the fixed effective sample size of the smallest sample in the data set.

Rearranging, we find that the population size m_t of the resampling process should be

$$m_t = \frac{(n_t - 1)\, n_{\min}}{n_t - n_{\min}} \,. \tag{3.14}$$

When $n_t = n_{\min}$, the resulting m_t would diverge, indicating that no resampling is necessary – an infinite binomial sample will always contain a proportion of successes exactly equal to the success probability, thus maintaining the relative frequencies in the original sample.

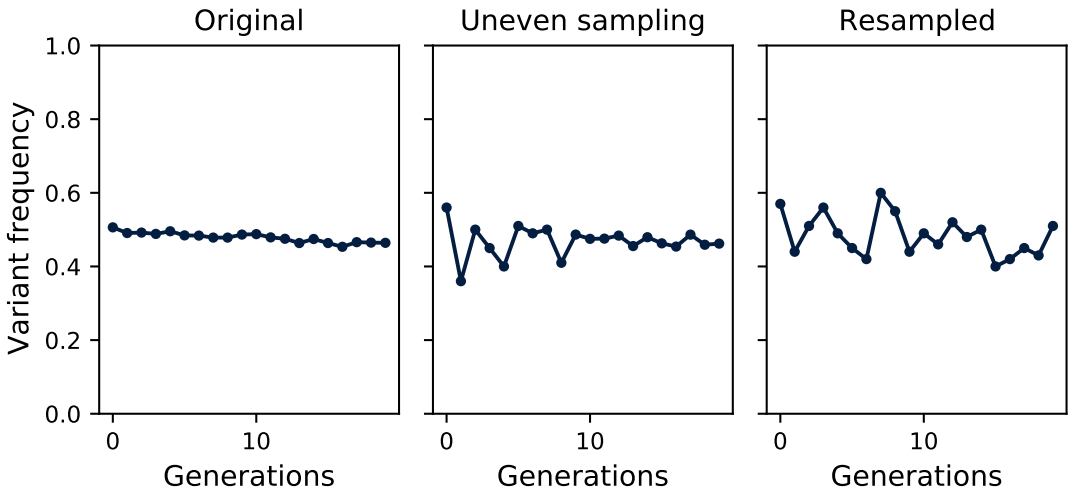

Figure 3.2 *Comparison of an artificially-generated time series before sampling (left), after uneven sampling typical of corpus data (centre) and after further applying sampling error equalization (right). Sampling error equalization homogenises the effects of sampling noise along the entire trajectory.*

Figure 3.2 illustrates the effects of sampling error equalisation through three time series. The one on the left is an artificially-generated trajectory evolving under constant evolutionary parameters, which represents the "true" underlying trajectory of change that is inaccessible to us. The central one represents the trajectory that would be available from corpus data extraction, where typical uneven sampling was reproduced by sampling the original time series with different sampling sizes in the first and second halves of the trajectory. Finally, the trajectory on the right represents the effects of noise equalisation, whereby the effects of sampling noise are homogenised along the trajectory.

I utilised this procedure to measure the effects of noise equalisation on the proportion of false positives in the detection of time-dependent evolutionary parameters, by applying the change-detection algorithm on artificially-generated time series after uneven sampling and after sampling error equalisation. All time series were generated using constant $N = 5000$ and $s = 0.0$ for 20 generations. Uneven sampling was simulated using sampling sizes of $n = 100$ during the first half of the trajectory, and $n = 4000$ after it. The change-detection algorithm produced false positives under the standard significance test in 16% of all analysed trajectories, much higher than the expected 5%. Conversely, only 2.5% of all error-equalised time series produced a false positive under the standard significance test. While this false positive rate being under 5% reveals

that noise equalisation may potentially lead to false negatives, it also proves that it is effective in minimising the false positive rate of detection of change arising from uneven sampling.

3.2 Applications

Having validated the change-detection method with artificially generated data in the previous section, I turn now to the analysis of time series of historical data. As previously discussed, we may expect the strength of selection to change over time due to social pressures like prestige, taboo, language contact, or identity [72, 148, 149]. Prescriptive grammar and spelling reforms introduced by regulating institutions are a particularly interesting case study, as they introduce changes in the prestige variant of the language [152, 178] and are thus expected to cause a sudden change in selection strength reflecting said changes. Furthermore, institutional reforms are usually well-documented, which makes them ideal to test the accuracy of the method in estimating the transition time T.

Here, I focus on historical change in the 2019 Spanish Google Books corpus [113]. First, I turn to an instance of unregulated change, which illustrates the necessity of noise sampling equalisation to avoid false positives in the detection of change. Then, I apply the method to six instances of regulated change introduced by the Real Academia Española (RAE), the central regulatory institution of the standard Spanish language. By choosing well-documented historical spelling reforms, the accuracy of the method in the detection of historical change can be tested.

3.2.1 Unregulated change in Spanish

Not all instances of historical language change are expected to involve strong and sudden fluctuations in the underlying evolutionary forces that cause them. Here, I first benchmark the methodology developed in the previous section by applying it to one such process. In particular, I analyse an instance of unregulated change, involving the competition dynamics between two completely equivalent forms of the past subjunctive tense (which has no equivalent in English), with verbal affixes ⟨-ra-⟩ and ⟨-se-⟩. Thus, the third person singular of the past subjunctive of the verb *escribir* (to write) could be either *escribiera* or *escribiese*. Both forms of the past subjunctive are considered completely equivalent in all contexts. In spite of

this, in the last 150 years, there has been a steady transition in the corpus, from a preference of the ⟨-se-⟩ form to a preference of the ⟨-ra-⟩ form. As of yet, there is no agreed upon explanation of this phenomenon, from either corpus-based or sociolinguistic perspectives [179, 180].

Rather than analysing instances of this change affecting individual verbs, I first identify a set of commonly used verbs and pool together their instances of use over time in the 2019 update to the Spanish Google Books corpus [113]. Then, we can find the total relative frequencies of usage of the ⟨-ra-⟩ and ⟨-se-⟩ forms over all verbs. 27 commonly used verbs were identified and used for this purpose, as listed in Appendix B. This procedure generates a single effective time series for the change, and has been found effective in related corpus analyses [181].

While this averaging over sets of words decreases the sampling noise in the data and increases the inferential power of the analysis, the data still suffers from uneven sampling error, particularly containing lower counts in early time periods. To remedy this, I apply the sampling error equalisation method as described in Section 3.1.2. To exemplify the necessity for this procedure, I apply the change-detection algorithm to the ⟨-ra-⟩ and ⟨-se-⟩ time series before and after sampling equalisation. A BwS approximation is used in the computation of all transition probabilities necessary to construct the likelihood of the model. To determine the p-value for the model with constant evolutionary parameters, I determine the maximum likelihood values of N and s without a change point, and generate 500 synthetic time series that match the length and starting frequency of each historical time series. For each of them, I then optimise the likelihood of the five-parameter model in equation 3.1. An empirical p-value can then be estimated as the fraction of artificial time series whose likelihood ratio (equation 3.3) exceeds that of the real trajectory. One can then apply a standard threshold of $p < 0.05$ to decide whether to accept the more complex model.

The results of this procedure are shown in Figure 3.3, where time series before (left) and after (right) noise equalisation are displayed. Both time series use a temporal binning of 1 year, equalling the temporal resolution of the Google Books corpus. The effects of noise equalisation are particularly apparent in the last 50 years of the time series. Furthermore, noise equalisation leads to significantly different results in the application of the change-detection algorithm. A change in model parameters is deemed significant by the algorithm if sampling error is not equalised first, as changes in fluctuations due to uneven sample sizes are misidentified as changes in drift in inter-generational transmission. This is

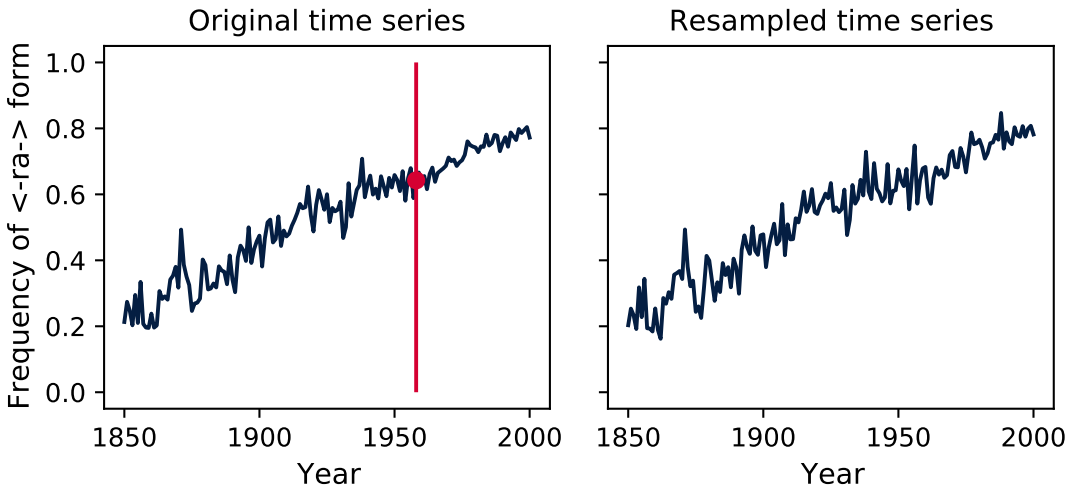

Figure 3.3 *Comparison of the frequency of usage of the ⟨-ra-⟩ form of the past subjunctive in the 2019 update to the Google Books Spanish corpus, before (left) and after (right) sampling error equalization. The effect of the sampling error equalization is particularly evident in the last 50 years of the time series. A significant change is detected before sampling error equalisation at $t = 1958$, due to the uneven sampling error in the time series.*

	T	N_1	N_2	s_1	s_2	p-value
Before noise equalisation	1958	63.7	369	0.016	0.011	<0.02
After noise equalisation	1961	46.9	124	0.016	0.026	0.15

Table 3.1 *Results for the analysis of unregulated change in the affixes of past subjunctive verbal forms in Spanish between the years 1850 and 2000, using time-divided models. The time division is found to not be significant after noise equalisation using a standard p-value threshold of 0.05, but it is significant before it.*

denoted in the left panel of Figure 3.3 with a red line and dot marking the detected transition time at $t = 1958$. Table 3.1 records results and p-values for this time series, both before and after noise equalisation.

In all, this analysis illustrates that not all change involves significant fluctuations in evolutionary forces, and thus a good change-detection algorithm should minimise the likelihood of false positives. The algorithm presented in this chapter is able to do so, provided it is supported with ways to minimise interference from uneven sampling such as the sampling error equalisation technique.

3.2.2 Regulated change in Spanish

Now, I focus on the detection of change in six instances of regulated historical change. In particular, I look at spelling reforms that were introduced by the Real Academia Española (RAE), the institution in charge of the standardisation of the Spanish language. Since its creation in 1713, the RAE has regulated Spanish orthography favouring the phonemic principle (i.e. the idea that there should be a clear mapping between spelling and pronunciation) over etymological or conservative approaches (i.e. those favouring spellings that closely reflect the history or origins of words) [182].

Here, I study words affected by one of six reforms. Reform A has to do with the simplification of the intervocalic ⟨ss⟩ digraph to a single ⟨s⟩ in 1763 [183], due to the different sounds that both spellings represented (/s/ and /z/, respectively) having merged in the 16th century [184]. Examples of words affected by this reform included *assunto* (affair, matter) and *esso* (that). Reform B, taking place in 1815, is concerned with the replacement of ⟨x⟩, with ⟨j⟩ in all non word-final contexts where it represented the phoneme /x/ [185]. This is again due to both letters having represented distinct sounds in Medieval Spanish (/ʃ/ and /ʒ/, respectively) that merged in the 15th century (and eventually drifted to their modern pronunciation /x/) [184]. This affected words like *baxo* (low, under) or *exército* (army). Reform C, also in 1815, replaced ⟨y⟩ with ⟨i⟩ in all non word-final closing diphthongs (i.e. /ai/, /ei/ and /oi/) [185]. Affected spellings included *ayre* (air) and *reynar* (to rule). Reforms D and E changed the accentuation rules for words ending in ⟨-n⟩ in 1881. These reforms stipulated that words ending in ⟨-n⟩ with a tonic last syllable had to be accentuated, while words ending in ⟨-n⟩ with a tonic penultimate syllable lost their previously prescribed accent [186]. Before these reforms, conjugated verbs followed different accentuation rules than other words; this rule change equalised accentuation rules in all grammatical contexts. Examples of words affected by this rule change included *álguien* (someone) and *cancion* (song). I treat words that gain an accent and words that lose an accent as independent sets (D and E, respectively). Finally, reform F in 1911 entailed the accentuated single vowel words *á* (to), *ó* (or), *é* (and), and *ú*, (alternative form of *ó* in certain phonological contexts) being replaced with their unaccented forms *a*, *o*, *e* and *u* [187].

I again commence by identifying a set of commonly used words affected by each of the reforms described above, and pool together their instances of use over

time in the 2019 update to the Spanish Google Books corpus [113] to find the total relative frequencies of usage of old spellings over all members of each set. Sampling error equalisation is subsequently applied to each of the resulting time series. The number of words in each set (other than reform F, which is limited to four words) ranges from 16 to 27. The exact sets are specified in Appendix B.

We are now ready to apply the change-detection method to the noise-equalised time series, with the goal of identifying the transition time T and evolutionary parameters that maximises the likelihood of the model in equation 3.1 for each of them. This is carried out in the same way as it was for the time series of unregulated change, by using 500 artificial time series to generated an empirical distribution of test statistics from which the p-value for the model with constant parameters can be estimated. Having split the time series once, the method can be iterated to each sub-series with the goal of identifying secondary change points. This procedure terminates when none of the sub-series admits a subdivision that yields a sufficiently improved description of the data according to the p-value threshold.

Results are shown in Figure 3.4. A 5-year generation time was chosen for the analysis of these time series, as it maximises the temporal resolution of the algorithm while reasonably reducing computational effort and increasing the detectability of changes in selection as explained in Section 3.1.1. In spite of this, the resulting trajectories are still subject to considerable fluctuations. For each of the RAE reforms, the frequency plotted is that of the old variant, which quickly approaches zero in all six cases–this highlights the acceptance and influence of the Real Academia Española amongst the literate population. Notably, the changes featured in reforms B and C seem to have already been in progress before the reform was introduced. It has been suggested that in many cases, language reforms tend to reflect pre-existing trends, as opposed to actuating the change [178].

I use a black line and dot to indicate the time at which the reform was introduced, and a red line and dot the first time T at which subdividing the time series improves the fit to the data, with a p-value threshold of $p = 0.05$ applied. In all six trajectories of regulated change, evidence is found that the selection s changed significantly over time. In each case, the first detected change point falls within twelve years of the reform being introduced, even when the trajectory is strongly fluctuating (see A) or does not have the typical S-shape of variant replacement in competition processes (see E). Notably, the first change point is detected in all

cases with p-values under 0.002.

By iterating the algorithm, time series can be further subdivided. In doing so, secondary time divisions are detected (dashed lines in Figure 3.4) whose p-values are below 0.05. In time series B and F, the earlier secondary point detects the beginning of the rapid decline in usage that was deemed less significant than the end by the first application of the algorithm. This is symptomatic of the algorithm's tendency to initially detect the reform after it has occurred, rather than at its inception. This is due to it not distinguishing past from present, making both the beginning and the end of the sharp decline following a reform equivalent. The later secondary point in B and the secondary point in A are not associated with documented reforms, and may reflect slight changes in social attitudes or simply be quirks of the data.

Table 3.2 records the primary detected years of change for each of the trajectories of RAE reforms, together with the actual year in which the reform was introduced. All reforms were detected with an error of at most 12 years, with the most important sources of discrepancy being strong fluctuations in the trajectory (see A) and the tendency of the algorithm to prioritise points after the change has taken place (B, C, D and F). When secondary change-points are also taken into account, the error in the detection of the reform is cut to 5 years for B and 1 year for F, as these are more likely to detect changes at the inception of change, rather than its completion.

All s, N and p-values for the most significant model (including all significant points of change) for each trajectory can be found in Table 3.3. With $N \simeq 100$ and $T \simeq 10$ (if measured from the beginning of the time series in units of generations), equation 3.5 predicts a characteristic $\Delta s \simeq 0.3$. Most detected changes involve a change in selection of the same order or greater than this characteristic Δs, lending credibility to the empirical power law. Also notably, N tends to increase after significant transition times, in spite of sampling effects on changing fluctuations having been accounted for through error equalisation. This is possibly reflective of an increase in size of the literate population and greater entrenchment of the standard variety of the language through universal access to education. These increases in N are possibly behind the tendency of the algorithm to initially detect change after it has taken effect: models with higher fluctuations (i.e. lower N) maximise the likelihood of the ongoing change, whereas those with lower fluctuations (i.e. higher N) are better at describing the stable regime following it.

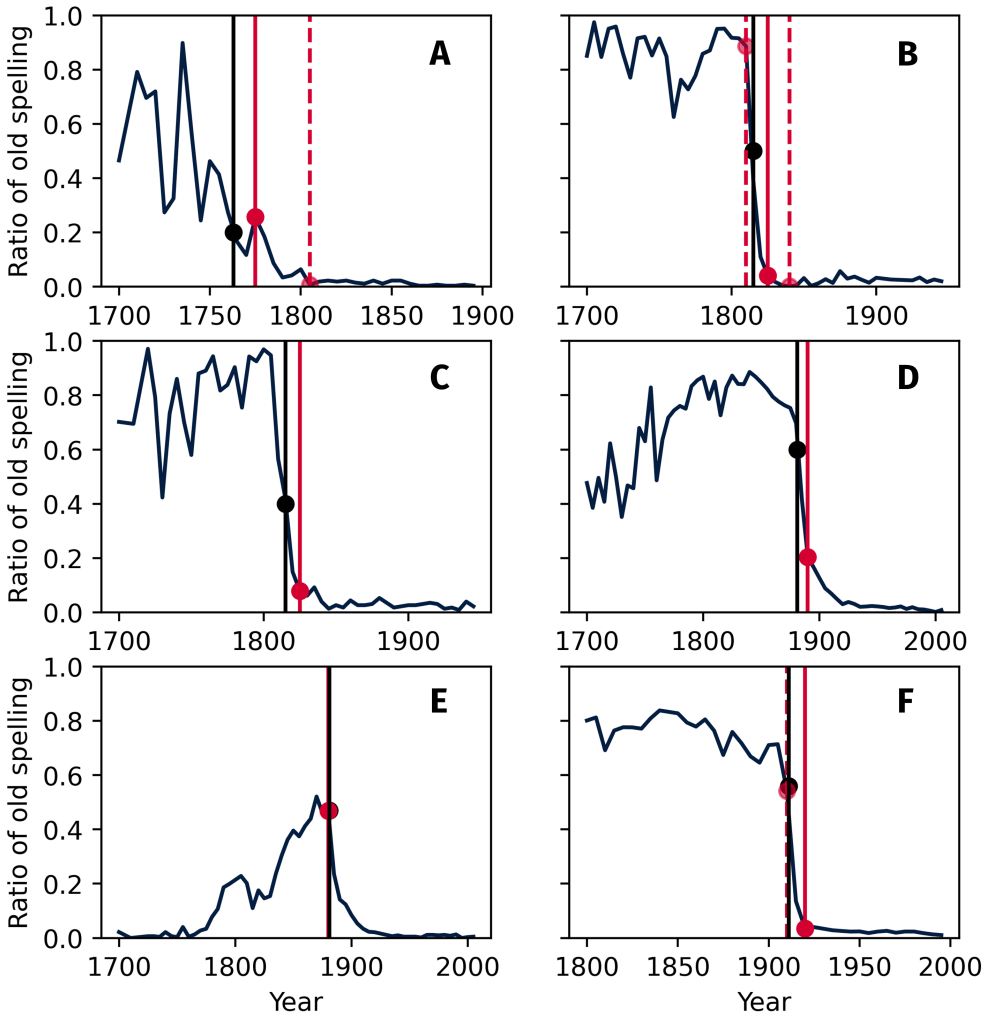

Figure 3.4 *Application of the change-detection algorithm to the data set of Spanish spelling reforms in the 2019 Spanish Google Books corpus, with temporal binning of the frequency data of 5 years. For each set of words that undergo a rule change, the ratio of usage of the old form is plotted over time. The ratio of usage of all old forms converges to zero after each reform. Black dots with solid vertical lines represent the year of publication of the RAE spelling reforms [183, 185–187]. Red dots with solid vertical lines represent the year at which selection strengths changed as first detected by the maximum-likelihood method with a p-value below* 0.05*. These fall within a period* ΔT *of 12 years or less relative to the date of the reform. Note that the temporal resolution of the time series is of 5 years, so an error of 10 years is equivalent to just two data points. Dashed vertical lines represent secondary points of change in evolutionary parameters, also detected with a p-value below 0.05 after iterating the change-detection model on the partial trajectories delimited by the original detected transition time. The number of such secondary points depends on the time series.*

	Reform	Reform year	Primary T	ΔT
(A)	⟨ss⟩ to ⟨s⟩	1763	1775	12 years
(B)	⟨x⟩ to ⟨j⟩	1815	1825	10 years
(C)	⟨y⟩ to ⟨i⟩	1815	1825	10 years
(D.1)	⟨-n⟩ accentuation	1881	1890	9 years
(D.2)	⟨-n⟩ accent loss	1881	1880	1 year
(E)	Vowel words accent loss	1911	1920	9 years

Table 3.2 *Years of introduction of reforms, years T detected in the first application of the change-detection algorithm, and differences between the two values ΔT for each of the RAE reforms.*

Reform	T	N before	s before	N after	s after	p-value
(A)	1775	5.0	-0.008	60	-0.59	<0.002
	1805	60	-0.59	292	-0.10	0.024
(B)	1810	17	0.003	33	-2.2	0.036
	1825	33	-2.2	2530	-1.0	<0.002
	1840	2530	-1.0	154	-0.02	0.026
(C)	1825	9.3	-0.30	121	-0.061	<0.002
(D1)	1890	16	-0.050	427	-0.21	<0.002
(D2)	1890	85	0.13	249	-0.53	<0.002
(E)	1910	82	-0.025	29	-1.3	0.004
	1920	29	-1.3	1460	-0.07	<0.002

Table 3.3 *List of all detected changes, transition times T, population sizes N before and after transition, selection parameters s before and after transitions, and p-values for each of the reforms under analysis with the change-detection algorithm.*

3.3 Discussion

In this chapter, I have introduced a method for the detection of abrupt changes in evolutionary forces in historical language change, chiefly associated with changes in the prestige language variety, but also possibly applicable to a wider variety of culturally motivated change. The method is based on an extension of the Wright-Fisher model, where evolutionary parameters N and s are time dependent. This time dependence takes on the simplest possible form, with evolutionary parameters instantaneously changing between otherwise constant values.

Likelihood maximisation methods can be applied then to find the optimal values of these parameters for a given time series of frequency data, as well as to test whether this time-divided model is necessary to explain its dynamics. As an unsupervised method, it does not rely on any prior knowledge as to when the change may have occurred. A thorough validation of the method using synthetic time series reveals an empirical power-law dependency between the characteristic change in selection Δs that can be significantly detected more than half the time, and the length of the time series and its level of fluctuation. This reveals a relationship between the choice of temporal binning and the probability of successful detection of change, with coarser binning increasing the probability of success at the expense of accuracy.

Significant changes in a trajectory of variant usage can be an artifact of sampling rather than a result of changes in the evolutionary forces driving its dynamics. To more confidently rule out this possibility, I have presented here a sampling error equalisation method that homogenises the effects of sampling throughout a time series. The application of the change-detection method to a time series of unregulated change in Spanish highlights the necessity of the error equalisation step: the method produces a (presumable) false positive when applied to the unprocessed time series. This change detection is no longer significant when the method is applied to the error-equalised trajectory.

The method has been further applied to the study of the dynamics of word spellings in Spanish before and after reforms introduced by the language's central regulatory institution, the Real Academia Española. Each of the changes was much better described by a time-divided model in which the selection strength changed at one or more points in time, the primary change point corresponded well with that at which the reform was introduced. This is despite the presence

of noise on the time series data, and despite some of the time series not following the S-shape trajectory typical of language change and necessary for other change-detection methods such as change-point analysis. A drawback of the method resides in its high computational demands, which were minimised in this application by choosing temporal binnings of 5 year at the expense of reducing the precision of the results.

In all, and in spite of these issues, the presented methodology provides a robust tool for the unsupervised detection of change in evolutionary forces rooted in the Wright-Fisher model of competition dynamics. Since changes in selection strength could derive from a variety of social and cultural factors, and apply to cultural evolutionary processes beyond language, this methodology could have broad applicability. Future steps may involve the development of more sophisticated functional forms for the time dependence that more accurately describe the underlying dynamics of cultural change, such as those involving frequency-dependent selection reflecting attitudes of conformity and anti-conformity.

Chapter 4

An iterated Bayesian learning model with nontrivial communication

In Chapter 1, I introduced linguemes as the basic units of linguistic function, and variants as the linguistic forms used to express them. Linguemes and variants exist on all levels of linguistic structure, from phonology to syntax. Thus, in English, the word *bird* (variant) is used to express the concept of a bird (lingueme). Similarly, the word order SVO (variant) is used to express the syntactic relation of transitivity (lingueme). According to Construction Grammar theories [52, 53, 188, 189], a speaker's mental representation of their language, also known as their grammar, is composed of a complex network of associations between linguemes and variants known as constructions. This network can be thought of as an adaptive system where change in one lingueme-variant pairing can only be fully understood in the context of its influence on and by closely related pairings [55–57]. However, the evolutionary paradigm presented so far oversimplifies these network dynamics – and with them, several contributing factors to language change. This oversimplification stems largely from two assumptions: that the competition dynamics of different linguemes are *isolated* from one another, and that the associations between variants and linguemes are clear-cut and *stable*. The goal of the following two chapters is to develop a stochastic paradigm of language change that goes beyond these assumptions, by modelling the interacting dynamics and many-to-many associations typical of diachronic processes affecting grammatical systems. The present chapter will focus on the development a model of cultural transmission that will serve as the foundation for the development of an empirically-applicable Wright-Fisher-like

paradigm in Chapter 5.

The isolation assumption models the competition processes between variants (e.g. irregular and regular variants of the past tense of a verb) as uninfluenced by the competition dynamics of related linguemes (e.g. irregular and regular variants of other verbs). However, a variety of linguistic scenarios have been identified where change is better characterised by considering the interaction and co-evolution of linguemes. In phonology, chain shifts are processes in which a change in pronunciation in one sound brings about changes in other sounds [190]. These changes are believed to take place in order to minimise ambiguity in the phonological system [191, 192]. There is thus a "repulsion" effect, whereby distinct phonemes (mental representations of sounds in the language's sound system) are preferentially expressed by maximally distinct sounds. Therefore, the evolutionary dynamics of all affected phonemes, and the interaction between them, have to be considered simultaneously if they are to be explained accurately. In morphosyntax, proportional processes such as analogy extend inflectional paradigms (i.e. sets of rules for inflecting words, such as adding ⟨-s⟩ to express possession) to words that they initially do not apply to [91]. An example of this is provided by the arisal of *octopi* as the plural of *octopus*, a word of Greek origin, in analogy to Latin plurals such as *cacti* from *cactus*. Regularisation processes, such as those explored in Chapter 2 in the context of English verbs, may also arise from analogy. Unlike chain shifts, which maximise the distinctiveness of every element in the phonological system, analogical processes seem to seek to simplify and homogenise inflectional paradigms. In either case, however, changes in individual linguemes cannot be understood without considering the effects that the linguistic system that they are embedded in has on their evolution. A mathematical framework aiming at explaining these phenomena should preferentially be able to model these lingueme interactions explicitly.

According to the stability assumption, a variant will always be limited to competing for usage of a single, unchanging lingueme, and will never come to realise a different one. The necessity for a model to account for change in a variant's meaning may not be obvious if considering only the evolutionary dynamics of content words, i.e. those which, like nouns and action verbs, are a key component of the meaning of the sentence that they appear in. Content words may be polysemous, e.g. *light* (not heavy, or not dark) or *school* (of Physics, or of fish). However, their meanings tend to be stable and easily inferred from context, and are usually robustly learned in the earliest stages of language acquisition

[193, 194]. The usefulness of moving beyond stability becomes clearer when one looks at function words, i.e. those which, like prepositions or pronouns, specify the grammatical relations within the sentence that they appear in. Functional vocabulary is highly polysemous [195]. Consider the following sentences:

(1) a. I left the oven on!
b. Your phone is on the floor.
c. Please carry on without me.

The word *on* has a different meaning in each of them, varying from specifying the state of an appliance to the location of an object or the direction of an action or process. This polysemy is a common feature of function words. The same grammatical function may also be realised by different function words, as in:

(2) a. I don't remember if we had to go left or right here.
b. I don't remember whether we had to go left or right here.

The words *if* and *whether* are full synonyms in this grammatical context. As these examples illustrate, functional vocabulary presents many-to-many associations between variants (words or structures) and linguemes (grammatical functions). A mathematical framework aiming at describing change in grammatical systems, and not just lexicon, should account for these networks of associations. For added clarity, table 4.1 summarises the key features of the isolation and stability assumptions as presented in this chapter, together with examples of processes and features that break each of them.

	Isolation	Stability
Key features	Evolution of linguemes unaffected by other linguemes	Only one lingueme per variant
Processes and features that break the assumption	Phonological chain shifts Analogy	Polysemy of content words Function words

Table 4.1 *Summary of the isolation and stability assumptions*

In this chapter, I create a mathematical paradigm of cultural transmission that describes grammatical change beyond the isolation and stability assumptions. In Section 4.1, I develop this model by building on models of iterated Bayesian learning (IBL) [85]. As previously discussed in Section 1.3.1, IBL is a modelling paradigm of cultural transmission that is able to implement the effects of cognitive

biases in the form of a prior distribution in a process of Bayesian learning. The model presented here breaks the stability assumption by incorporating any number of co-evolving linguemes and variants in a network where any of the linguemes can be realised, a priori, by any of the variants. On top of the production and learning phases of most IBL paradigms, an intermediate understanding phase is introduced. During understanding, imperfect communication and semantic effects provide us with the simplest nontrivial interactions between linguemes, thus breaking the isolation assumption.

This understanding phase has deep consequences on the statistics of the Markov chain generated by the iterated learning process. The most basic implementation of the IBL paradigm presents *convergence to the prior* [85], whereby the stationary distribution of grammars arising from the IBL chain of cultural transmission is identical to the prior distribution containing the learning biases of the speakers. A variety of extensions to the basic IBL paradigm have been developed in order to understand how different learning strategies [78], or more sophisticated social networks [79], semantic spaces [80], or pragmatics [81] may affect the emerging synchronic properties of the language system. In Section 4.2, I take an in-depth look at the effects of nontrivial understanding effects on the stationary distribution, and find analytical conditions for the breaking of convergence to the prior.

Another interesting statistical effect can arise in the stationary state: directionality in the form of detailed-balance breaking. Directionality is a well-documented feature of many processes of language change, which seem to almost always occur in a specific direction. The process of grammaticalisation, for example, – where content words lose lexical meaning to become function words – is a common mechanism of change in most human languages, but hardly ever occurs in the direction of loss of grammatical meaning [15, 16]. Examples of grammaticalisation include language change where body parts become prepositions [196], or the cyclic evolution of negation strategies [197]. However, directionality remains underexplored in stochastic models of language change. In Section 4.3, I find and discuss conditions for detailed-balance breaking in the model with nontrivial communication. In Section 4.4, I then proceed to quantify directionality in this and other IBL models by measuring their leading-order entropy production [198, 199], a concept borrowed from non-equilibrium statistical mechanics.

4.1 The model

In this section, I proceed to develop a stochastic model for the cultural transmission of a language, able to break the stability assumption by incorporating V variants and L linguemes that associate freely, and the isolation assumption by incorporating imperfect communication and semantic effects. I will further explore the consequences of these additions on the stationarity and reversibility of the cultural transmission process.

Consider a language possessing V variants and L linguemes. Here, a speaker's grammar (their mental representation of their language) will be specified by a set of frequencies $G = \{g_{lv} : v = 1, ..., V; l = 1, ..., L\}$ with which they expect variant v to be used to represent lingueme l. These frequencies are normalised for each lingueme, in such a way that

$$\sum_{v} g_{lv} = 1 \,. \tag{4.1}$$

Therefore, in principle, any variant in the model can be used to express any lingueme. This choice of normalisation highlights an interesting asymmetry of constructions: every lingueme should always be realised by at least one variant, but every variant need not realise at least one lingueme. Following the iterated Bayesian learning (IBL) paradigm [85], the evolution process leading from grammar G at generation t to grammar G' at generation $t + 1$ can be modelled as the concatenation of *production* and *learning* processes, where linguistic data produced at generation t is used in the learning process of generation $t + 1$. For simplicity, and following previous literature [4], I assume that each generation contains one speaker only, who can be understood as representing the average linguistic behaviour of a larger population. While some implementations of the IBL include several independent speakers per generation [see e.g. 79], this extra layer of social complexity would obscure the identification of effects arising from the layer of linguistic and communicative complexity that I aim at modelling here. In the production process, the speaker produces linguistic data as a set of M utterances, consisting of a set of pairs of variants and their associated linguemes, $U = \{(v_m, l_m) : m = 1, ..., M\}$. The probability of producing set U

given grammar G is then given by:

$$\begin{aligned} \mathrm{P}_{\mathrm{prod}}\left(U \mid G\right) &= \prod_{m=1}^{M} \mathrm{p}_{\mathrm{prod}}\left(v_m \mid l_m, G\right) \varphi\left(l_m\right) \\ &= \prod_{l=1}^{L} \prod_{v=1}^{V} \left[\mathrm{p}_{\mathrm{prod}}\left(v \mid l, G\right) \varphi\left(l\right)\right]^{u_{lv}} . \end{aligned} \tag{4.2}$$

In the second line, u_{lv} represents the total number of utterances in which the speaker chose to express lingueme l using variant v. These satisfy $\sum_l \sum_v u_{lv} = M$. φ is the lingueme probability, which represents the probability that the speaker will choose to communicate a given lingueme, i.e. how common or necessary a lingueme is to fulfil the communicative needs of a speaker. Unlike other IBL models implementing complex meaning spaces [see e.g. 80], I have chosen to assume that it is independent of the speaker's grammar. This implies that the likelihood of a speaker to want to express a given lingueme is fixed by their environment and does not vary from speaker to speaker, or from generation to generation. $\mathrm{p}_{\mathrm{prod}}\left(v \mid l, G\right)$ represents the probability that the speaker will choose variant v, given that they wish to communicate lingueme l with their grammar G. In a model with neutral production, the speaker will produce a variant associated to their lingueme of choice by sampling faithfully from their grammar:

$$\mathrm{p}_{\mathrm{prod}}\left(v \mid l, G\right) = g_{lv} . \tag{4.3}$$

However, a variety of production strategies can be implemented here, including production errors, selective biases, or analogy between different linguemes. These will be discussed individually in Section 5.2, in the context of the empirical application of the model.

In the learning process, the learner in generation $t+1$ infers the grammar of the language by putting together information from the utterances U of the previous generation and their own learning biases. Their learning strategy, borrowing from the general form of Bayes' theorem, is thus:

$$\mathrm{P}_{\mathrm{learn}}\left(G' \mid U\right) = \frac{\mathrm{P}_{\mathrm{L}}\left(U \mid G'\right) \Pi_0\left(G'\right)}{\mathrm{M}\left(U\right)} , \tag{4.4}$$

where Π_0 represents the prior distribution – the learning biases of the speaker favouring or disfavouring specific grammars – and the likelihood $\mathrm{P}_{\mathrm{L}}\left(U \mid G'\right)$ can be assumed to equal the production probability (eq. 4.2) under a neutral model

(eq. 4.3), i.e.

$$\mathrm{P_L}\left(U \mid G\right) = \prod_{l=1}^{L} \prod_{v=1}^{V} \left[g_{lv} \varphi\left(l\right)\right]^{u_{lv}} \,. \tag{4.5}$$

The marginal probability, $\mathrm{M}(U)$, is given by

$$\mathrm{M}\left(U\right) = \int \mathrm{P_L}\left(U \mid G\right) \Pi_0\left(G\right) \mathrm{d}G \,, \tag{4.6}$$

where $\int \mathrm{d}G$ represents an integral over all grammar frequencies g_{lv} for all values of l and v, restricted to the hypersurfaces defined by $\sum_v g_{lv} = 1$ for all l.

Here, following previous literature [4], I choose the prior to be a product of Dirichlet distributions of the form:

$$\Pi_0\left(G\right) = \prod_l \frac{\Gamma\left(\alpha_l\right)}{\prod_v \Gamma\left(\alpha_{lv}\right)} \prod_v g_{lv}^{\alpha_{lv}-1} \,, \tag{4.7}$$

whose shape is controlled by parameters α_{lv} with $\alpha_l = \sum_v \alpha_{lv}$. This choice of prior allows us to obtain a closed analytical form for the learning probability in equation 4.4 if we assume a neutral model of production (eq. 4.3), given by

$$\mathrm{P_{learn}}\left(G \mid U\right) = \prod_l \frac{\Gamma\left(u_l + \alpha_l\right)}{\prod_v \Gamma\left(u_{lv} + \alpha_{lv}\right)} \prod_v g_{lv}^{u_{lv}+\alpha_{lv}-1} \,. \tag{4.8}$$

This is also a product of Dirichlet distributions with shape parameters modified by the utterance counts u_{lv}, where $u_l = \sum_v u_{lv}$. As discussed in Section 1.3.3, this symmetry between the learning prior and posterior arise from the Dirichlet distribution being the conjugate prior of multinomial likelihoods [90].

Figure 4.1 exemplifies prior distributions with different shapes that can be obtained by tuning the shape parameters α_{lv}, for a language with one lingueme and two variants. The usual choice of $\alpha_{lv} < 1$ codifies a prior bias disfavouring variability, where the learner expects a priori that each lingueme is expressed by only one variant. Previous works introduced further invariability biases that work in the opposite direction in the mapping between variants and linguemes, such that each variant is expected to express only one lingueme [78, 79]. These models are usually grounded in lexical change, where language acquisition studies have shown that child learners strongly favour bijective mappings between linguemes and variants, in what is typically known as the *mutual exclusivity bias* [193, 194]. That is, children expect small quadruped furry animals that mew to always be referred to as *cats*, and the label *cat* to only ever be used for small quadruped

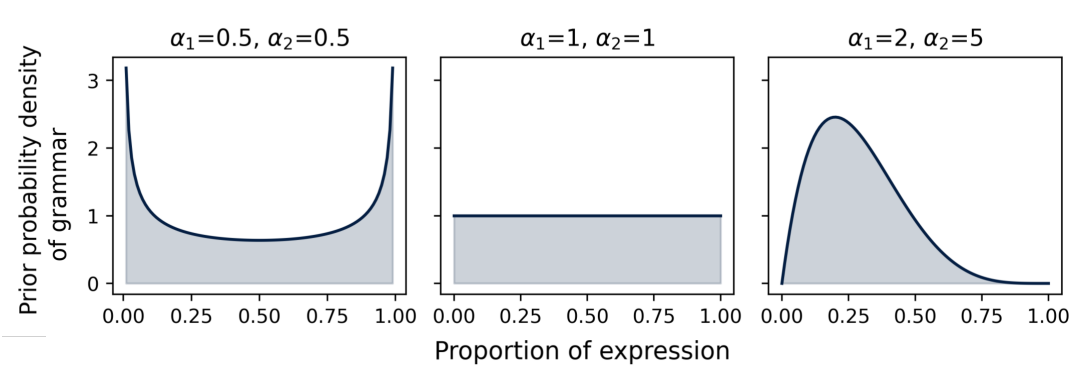

Figure 4.1 *Shapes of Dirichlet prior probabilities for systems with one lingueme and two variants ($L = 1$, $V = 2$). Left: typical choice of a symmetric prior probability disfavouring variability. Centre: a uniform prior expressing no preference for any grammar. Right: an example of a prior favouring grammars where both variants are available to express the lingueme, albeit with different frequencies.*

furry animals that mew. This does not necessarily hold when working with functional vocabulary and grammatical categories, where, if anything, child learners have been shown to overextend inflectional paradigms (saying e.g. *"I drinked"* instead of *"I drank"*), thus favouring the use of a single variant for every related grammatical lingueme [200–202]. For this reason, and for the sake of mathematical simplicity, I choose to not implement a bias in either direction at the level of the prior here.

Many previous implementations of IBL would only include production and learning phases. The key new component of the model I introduce here is an extra step between them, which I name the *understanding* phase. In it, the possibility of nontrivial communicative effects is introduced. Given a set of utterances U produced by generation t, generation $t+1$ will understand a potentially different set of utterances U'. This leads to an understanding probability given in its most general form by

$$\mathrm{P}_{\mathrm{und}}\left(U' \,|\, U\right) = \prod_{m=1}^{M} \mathrm{p}_{\mathrm{und}}\left(l'_m,\, v'_m \,|\, l_m,\, v_m\right) \,, \tag{4.9}$$

i.e. a product of the individual probabilities of understanding an utterance where variant v'_m is used to express lingueme l'_m, given an original utterance where v_m expressed l_m. This can account for effects ranging from communication noise to semantic effects like those behind the erosion of lexical meaning in grammaticalisation [15]. Imperfect communication provides us with a minimal

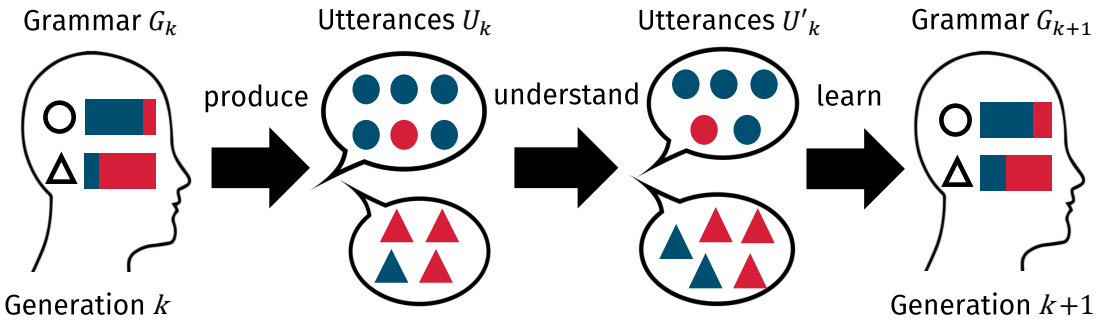

Figure 4.2 *Schematic representation of the iterated Bayesian learning model with imperfect understanding for a language containing two variants (represented by the colours blue and red) and two linguemes (represented by circles and triangles). A grammar consists of frequencies with which each variants is expected by the speaker to be used to express each lingueme. These frequencies are represented here by colour bars, with the length of each colour representing its expected relative frequency. The lingueme of an utterance is chosen according to some lingueme distribution, and its associated variant is chosen by sampling from the grammar. In a given generation k, a speaker uses their grammar G_k to produce utterances U_k, which are then understood as utterances U'_k by the speaker in generation $k+1$ and only then used to infer the grammar G_{k+1} through a learning process. This process is iterated indefinitely.*

model for the breaking of the isolation assumption, as it is one of the simplest nontrivial forms of interaction between linguemes during cultural transmission.

Figure 4.2 presents a schematic representation of the model for a language with two linguemes and two variants. Putting the production, understanding, and learning steps together, the transition probability distribution between grammar G and grammar G' after one generation can then be expressed as:

$$\mathrm{P}\left(G' \,|\, G\right) = \sum_{U,\,U'} \mathrm{P}_{\mathrm{learn}}\left(G' \,|\, U'\right) \mathrm{P}_{\mathrm{und}}\left(U' \,|\, U\right) \mathrm{P}_{\mathrm{prod}}\left(U \,|\, G\right) \,. \tag{4.10}$$

The Markov chain defined by this transition probability distribution reduces to the one presented in Section 1.3, equation 1.2, when understanding is perfect. That is:

$$\mathrm{P}_{\mathrm{und}}\left(U' \,|\, U\right) = \delta_{U',U} \,. \tag{4.11}$$

In particular, this is a sufficient condition for the prior distribution Π_0 to be the stationary distribution of the chain and for it to satisfy the detailed balance condition. As discussed in Section 1.3.1, these overly simplistic properties do not

adequately model key features of language change. It is thus key to understand what the stationary distribution of the model looks like, how and under which conditions it differs from the prior, and under which conditions it breaks detailed balance.

4.2 Stationarity

Let us first take a look at the stationary distribution of the process. As a reminder, the stationary distribution $\Pi(G)$ of a Markov process with transition probability distribution $\mathrm{P}(G' \mid G)$ is the probability distribution that satisfies:

$$\Pi(G') = \int \mathrm{P}(G' \mid G)\, \Pi(G)\, \mathrm{d}G\,. \tag{4.12}$$

Assuming that the Markov chain is irreducible (i.e. one where every grammar G' can be reached from every other grammar G with non-zero probability during inter-generational transmission) and aperiodic (sufficiently, one where $\mathrm{P}(G \mid G) \neq 0$), the stationary distribution exists and is unique. Furthermore, under these conditions, the Markov process converges to this distribution, meaning that the distribution of states $\mathrm{P}_k(G \mid G_0)$ expected k generations after an initial state G_0 converges to $\Pi(G)$ as $k \to \infty$, independently of the initial state G_0 [203].

In the study of language universals, it is often necessary to assume that the typological properties of human languages have already reached a stationary state [204]. Thus, understanding the properties of the stationary distributions of models of language change may help us understand how features of communication and learning affect language properties. The stationary distribution over grammars equalling the prior implies that, in the stationary state, the distribution of grammatical properties in human languages should be expected to be determined exclusively by our learning biases, and not our communicative needs or any other communicative or social effects. This feature of the simplest IBL models is commonly referred to as *convergence to the prior*.

As it turns out, most nontrivial choices of the understanding probability $\mathrm{P}_{\mathrm{und}}(U' \mid U)$ will lead to breaking of convergence to the prior, leading to stationary distributions that will depend both on it and on whatever understanding effects are incorporated in the model. As an example, consider a system with two linguemes and two variants ($L = 2$, $V = 2$) and a symmetric Dirichlet prior

against variability with $\alpha_{lv} = \alpha < 1$. Further consider a simple understanding probability for individual utterances given by

$$\mathrm{p}_{\mathrm{und}}\left(l', v' \,|\, l, v\right) = \delta_{v',v}\left[\delta_{l',l}\left(1-\eta\right) + \frac{\eta}{2}\right] , \tag{4.13}$$

representing an understanding process where variants are always perfectly understood, and the error rate of lingueme understanding is given by η. When error is indeed taking place, either lingueme will be understood with probability $\frac{1}{2}$. Thus, in this example model, what is said is always perfectly understood, but what is meant is not. Figure 4.3 features plots of the prior probability and numerically estimated stationary distribution of this process with $\eta = 0.1$. While the stationary distribution borrows some features from the prior, such as integrable divergences along boundary values of the grammar, its shape is modulated by the understanding process in a way that renders it distinct from the prior. In particular, in the stationary state, grammars where both linguemes are expressed by the same variant are more common than those where they are expressed by different variants, due to the effects of imperfect understanding. In the prior, conversely, any regular grammar is equally likely, independently of whether linguemes are expressed by different or equal variants. This example thus illustrates how a specific choice of understanding probability may lead to communicative effects in the stationary state.

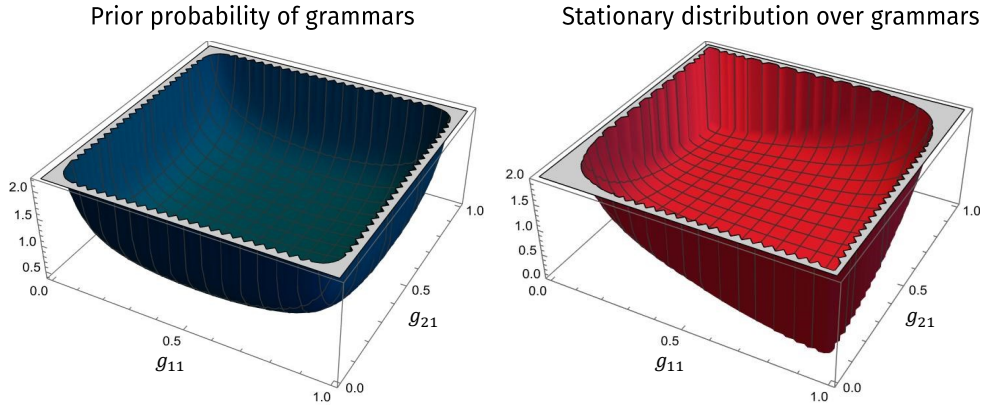

Figure 4.3 *Left: symmetric prior probability disfavouring variability for a system with $L = 2$, $V = 2$. Right: resulting stationary distribution emerging from an IBL process with imperfect understanding given by equation 4.13. Imperfect understanding breaks convergence to the prior, favouring grammars where both linguemes are expressed by the same variant.*

We do not know, however, what conditions exactly are necessary for this breaking of convergence to the prior to take place. We are thus interested in delimiting the

conditions under which the prior distribution Π_0 *is* the stationary distribution of our IBL model. For it, we have:

$$\begin{aligned}
\int \mathrm{P}\left(G' \,|\, G\right)&\Pi_0\left(G\right)\mathrm{d}G = \\
&= \sum_{U,U'} \mathrm{P}_{\text{learn}}\left(G' \,|\, U'\right)\mathrm{P}_{\text{und}}\left(U' \,|\, U\right)\int \mathrm{P}_{\text{prod}}\left(U \,|\, G\right)\Pi_0\left(G\right)\mathrm{d}G \\
&= \sum_{U,U'} \mathrm{P}_{\text{learn}}\left(G' \,|\, U'\right)\mathrm{P}_{\text{und}}\left(U' \,|\, U\right)\mathrm{M}\left(U\right) \\
&= \Pi_0\left(G'\right)\sum_{U'} \mathrm{P}_{\mathrm{L}}\left(U' \,|\, G'\right)\frac{\sum_U \mathrm{P}_{\text{und}}\left(U' \,|\, U\right)\mathrm{M}\left(U\right)}{\mathrm{M}\left(U'\right)} \\
&= \Pi_0\left(G'\right)\sum_{U'} \mathrm{P}_{\mathrm{L}}\left(U' \,|\, G'\right)\Omega\left(U'\right)\,. \qquad (4.14)
\end{aligned}$$

Here, in the first equality, the transition probability between grammars was expanded using equation 4.10. In the second equality, the definition of the marginal distribution (eq. 4.6) was used. In the third equality, the learning probability was expanded using its definition in terms of Bayesian learning (eq. 4.4). Finally, the following definition was introduced:

$$\Omega\left(U'\right) = \frac{\sum_U \mathrm{P}_{\text{und}}\left(U' \,|\, U\right)\mathrm{M}\left(U\right)}{\mathrm{M}\left(U'\right)} \qquad (4.15)$$

Comparing equations 4.14 and 4.12, we can infer that for the prior distribution to be the stationary distribution of the process, the following equality must hold for every value of G:

$$\sum_U \mathrm{P}_{\mathrm{L}}\left(U \,|\, G\right)\Omega\left(U\right) = 1\,. \qquad (4.16)$$

This is not true in general, but some conditions can be identified where it does hold (other than the previously identified condition of perfect understanding, eq. 4.11). One particularly interesting condition is obtained by choosing $\Omega\left(U\right) = 1$, which leads to:

$$\sum_U \mathrm{P}_{\text{und}}\left(U' \,|\, U\right)\mathrm{M}\left(U\right) = \mathrm{M}\left(U'\right)\,. \qquad (4.17)$$

Consider a game of *telephone*[1] where a set of utterances U is received and

[1]The game I am referring to was historically known as *Chinese whispers* in the United Kingdom. I am favouring its North American name here.

processed by a listener, who understands a set U' with probability $\mathrm{P}_{\mathrm{und}}(U' \,|\, U)$. The listener immediately communicates this set exactly as understood to another listener who processes U' with the same understanding probability, with the chain continuing indefinitely. Equation 4.17 then amounts to the marginal distribution $\mathrm{M}(U)$ being the stationary distribution of this Markov process. What exactly this means for the understanding process can be exemplified more clearly by considering a toy system where inter-generational transmission involves only one utterance, $U = \{(l, v)\}$. In that case, and given a Dirichlet prior with parameters α_{lv} and $\alpha_l = \sum_v \alpha_{lv}$, the marginal probability reduces to

$$\mathrm{M}(l, v) = \frac{\alpha_{lv}}{\alpha_l} \varphi(l) \ , \tag{4.18}$$

and condition 4.17 becomes

$$\sum_{l,v} \mathrm{p}_{\mathrm{und}}(l', v' \,|\, l, v) \frac{\alpha_{lv}}{\alpha_l} \varphi(l) = \frac{\alpha_{l'v'}}{\alpha_{l'}} \varphi(l') \ . \tag{4.19}$$

In this form, it is explicit that the understanding probability of a single utterance $\mathrm{p}_{\mathrm{und}}$ must be determined by cognitive biases (through α_{lv}) and the communicative environment of the speaker (through $\varphi(l)$) for this condition to be satisfied. Similarly nuanced relations between understanding, learning, and communicative needs remain key towards satisfying the more general equation 4.17. Finally, it is worth noting that the condition of perfect understanding, eq. 4.11, provides a trivial solution to equation 4.17 and is thus a particular case of it.

4.3 Reversibility

We now look at an even more restrictive condition, that of detailed balance. As a reminder, a distribution $\Pi(G)$ satisfies detailed balance if

$$\mathrm{P}(G' \,|\, G)\, \Pi(G) = \mathrm{P}(G \,|\, G')\, \Pi(G') \tag{4.20}$$

for every G and G' [203]. Any distribution that satisfies detailed balance automatically satisfies equation 4.12 and is therefore also the stationary distribution of the process. A Markov process whose stationary distribution satisfies detailed balance is said to be *reversible*, which means that, in the stationary state, any trajectory of the process is equally likely in the forwards and backwards directions.

The stationary distribution of a model of language change satisfying detailed balance implies that, in the stationary state, diachronic processes as modelled in this paradigm are reversible, with the probability flow between any two types of grammars being perfectly balanced. However, as previously discussed, many processes in language change are directional, and thus cannot be captured by previous IBL models. The multi-lingueme model in this chapter may provide a basis for their description.

Thus, we are interested in the conditions under which a stationary distribution $\Pi(G)$ of the IBL process satisfies detailed balance, and whether any of those conditions are satisfied by stationary distributions distinct from the prior Π_0. Consider the following:

$$\begin{aligned}
&\mathrm{P}(G' \mid G)\,\Pi(G) - \mathrm{P}(G \mid G')\,\Pi(G') = \\
&= \sum_{U,U'} \mathrm{P}_{\text{learn}}(G' \mid U')\,\mathrm{P}_{\text{und}}(U' \mid U)\,\mathrm{P}_{\text{prod}}(U \mid G)\,\Pi(G)\,. \\
&\quad - \sum_{U,U'} \mathrm{P}_{\text{learn}}(G \mid U)\,\mathrm{P}_{\text{und}}(U \mid U')\,\mathrm{P}_{\text{prod}}(U' \mid G')\,\Pi(G') \\
&= \Pi_0(G')\,\Pi(G) \sum_{U,U'} \mathrm{P}_{\mathrm{L}}(U \mid G)\,\mathrm{P}_{\mathrm{L}}(U' \mid G')\,\frac{\mathrm{P}_{\text{und}}(U' \mid U)}{\mathrm{M}(U')} \\
&\quad - \Pi_0(G)\,\Pi(G') \sum_{U,U'} \mathrm{P}_{\mathrm{L}}(U \mid G)\,\mathrm{P}_{\mathrm{L}}(U' \mid G')\,\frac{\mathrm{P}_{\text{und}}(U \mid U')}{\mathrm{M}(U)}\,.
\end{aligned} \tag{4.21}$$

Here, the transition probability between grammars was expanded using equation 4.10, and Bayesian learning (eq. 4.4) was applied. Note that P_{prod} and P_{L} are identical under the neutral production model presented here, which was used in the second equality. Detailed balance requires this quantity to equal 0 for every G and G'. This, in turn, leads to the following condition:

$$\begin{aligned}
\frac{\Pi(G')}{\Pi(G)} &= \frac{\Pi_0(G')}{\Pi_0(G)} \frac{\sum_{U,U'} \mathrm{P}_{\mathrm{L}}(U \mid G)\,\mathrm{P}_{\mathrm{L}}(U' \mid G')\,\mathrm{P}_{\text{und}}(U' \mid U)/\mathrm{M}(U')}{\sum_{U,U'} \mathrm{P}_{\mathrm{L}}(U \mid G)\,\mathrm{P}_{\mathrm{L}}(U' \mid G')\,\mathrm{P}_{\text{und}}(U \mid U')/\mathrm{M}(U)} \\
&= \frac{\Pi_0(G')}{\Pi_0(G)} \frac{\Lambda(G', G)}{\Lambda(G, G')}\,,
\end{aligned} \tag{4.22}$$

which, again, has to hold for any choice of the grammars, and where the following definition was introduced:

$$\Lambda(G', G) = \sum_{U,U'} \mathrm{P}_{\mathrm{L}}(U \mid G)\,\mathrm{P}_{\mathrm{L}}(U' \mid G')\,\frac{\mathrm{P}_{\text{und}}(U' \mid U)}{\mathrm{M}(U')}\,. \tag{4.23}$$

From equation 4.22, several conditions for detailed balance can be identified. First, conditions under which the prior distribution satisfies detailed balance can be identified as those where the function $\Lambda(G', G)$ is symmetric in its arguments, i.e.

$$\Lambda(G', G) = \Lambda(G, G') \ . \tag{4.24}$$

Notably, this equation is satisfied by the condition of perfect understanding, equation 4.11. Thus, perfect understanding represents a condition where the process both converges to the prior and satisfies detailed balance. Note that equation 4.24 represents a sufficient, not necessary, condition for detailed balance. In particular, detailed balance where the stationary distribution does not equal the prior will not satisfy this equation.

A further condition for detailed balance can be identified that is a particular case of a previously identified condition for convergence to the prior, rather than identical to it. It is given by:

$$\mathrm{P}_{\mathrm{und}}(U' \,|\, U)\,\mathrm{M}(U) = \mathrm{P}_{\mathrm{und}}(U \,|\, U')\,\mathrm{M}(U') \ , \tag{4.25}$$

whereby

$$\begin{aligned}
\Lambda(G', G) =& \sum_{U, U'} \mathrm{P_L}(U \,|\, G)\,\mathrm{P_L}(U' \,|\, G')\,\frac{\mathrm{P}_{\mathrm{und}}(U' \,|\, U)}{\mathrm{M}(U')} \\
=& \sum_{U, U'} \mathrm{P_L}(U \,|\, G)\,\mathrm{P_L}(U' \,|\, G')\,\frac{\mathrm{P}_{\mathrm{und}}(U \,|\, U')}{\mathrm{M}(U')} \\
=& \sum_{U, U'} \mathrm{P_L}(U' \,|\, G)\,\mathrm{P_L}(U \,|\, G')\,\frac{\mathrm{P}_{\mathrm{und}}(U' \,|\, U)}{\mathrm{M}(U)} \\
=& \Lambda(G, G') \ .
\end{aligned} \tag{4.26}$$

Here, relabelling of the summation variables was used to demonstrate symmetry in G and G'. Equation 4.25 is a particular case of equation 4.17, where the marginal distribution $\mathrm{M}(U)$ is not only the stationary distribution of the game of telephone defined by the transition probability $\mathrm{P}_{\mathrm{und}}(U' \,|\, U)$, but it further satisfies detailed balance for the same process, denoting an even more restrictive relationship between understanding, cognitive biases and communicative needs. Remarkably, this exemplifies that, when nontrivial understanding is included in the IBL paradigm, convergence to the prior is not synonymous with detailed balance. Directionality can arise in the stationary state even when that stationary state is – at the synchronic level – identical to the prior.

There is one further condition that can be identified which satisfies equation 4.22 for every choice of G and G'. Most notably, this condition is generally not satisfied by $\Pi_0(G)$, meaning that it leads to detailed balance of a stationary distribution distinct from the prior. The condition is given by

$$\mathrm{P}_{\mathrm{und}}(U' \,|\, U) = \mathrm{P}_{\mathrm{und}}(U') \ . \tag{4.27}$$

With this, equation 4.22 becomes

$$\begin{aligned}\frac{\Pi(G')}{\Pi(G)} =& \frac{\Pi_0(G')}{\Pi_0(G)} \frac{\sum_{U,U'} \mathrm{P_L}(U \,|\, G)\, \mathrm{P_L}(U' \,|\, G')\, \mathrm{P}_{\mathrm{und}}(U') \,/\mathrm{M}(U')}{\sum_{U,U'} \mathrm{P_L}(U \,|\, G)\, \mathrm{P_L}(U' \,|\, G')\, \mathrm{P}_{\mathrm{und}}(U) \,/\mathrm{M}(U)} = \\ =& \frac{\sum_{U,U'} \mathrm{P_L}(U \,|\, G)\, \mathrm{P}_{\mathrm{learn}}(G' \,|\, U')\, \mathrm{P}_{\mathrm{und}}(U')}{\sum_{U,U'} \mathrm{P_L}(U' \,|\, G')\, \mathrm{P}_{\mathrm{learn}}(G \,|\, U)\, \mathrm{P}_{\mathrm{und}}(U)} \\ =& \frac{\sum_{U'} \mathrm{P}_{\mathrm{learn}}(G' \,|\, U')\, \mathrm{P}_{\mathrm{und}}(U')}{\sum_{U} \mathrm{P}_{\mathrm{learn}}(G \,|\, U)\, \mathrm{P}_{\mathrm{und}}(U)} \ .\end{aligned} \tag{4.28}$$

Here, Bayesian learning (eq. 4.4) was used in the first equality. This equation is satisfied for every value of G and G' for a stationary distribution defined as

$$\Pi(G) = \sum_{U} \mathrm{P}_{\mathrm{learn}}(G \,|\, U)\, \mathrm{P}_{\mathrm{und}}(U) \ . \tag{4.29}$$

This equals $\Pi_0(G)$ only if $\mathrm{P}_{\mathrm{und}}(U)$ equals the marginal distribution $\mathrm{M}(U)$. Equation 4.27 describes an unrealistic scenario of data-independent understanding: it presents an cultural transmission process where the input data is completely ignored. It is, however, interesting as a mathematical demonstration that detailed balance and convergence to the prior are decoupled in this model: detailed balance is not always satisfied by a stationary distribution identical to the prior, and stationary distributions distinct from the prior may satisfy detailed balance.

All conditions for convergence to the prior and detailed balance presented so far are summarised in table 4.2. In general, however, most choices of the understanding probability will lead to neither reversibility nor convergence to the prior. This proves that nontrivial communication affects the emerging distributional properties and begets directionality in the process of cultural transmission. Exactly how directionality is generated will depend on the specific choice of understanding probability $\mathrm{P}_{\mathrm{und}}(U' \,|\, U)$. In order to explore this formally, the following section is devoted to the quantification of directionality in this and a few other IBL models through the application of entropy production, a tool from non-equilibrium statistical mechanics.

4.4 Measuring directionality through entropy production

As we have seen, the condition of perfect understanding (eq. 4.11) leads to a simple IBL paradigm satisfying both convergence to the prior and detailed balance. This IBL model is the multi-lingueme version of the one introduced back in Section 1.3.1, whose transition probability distribution is given by

$$\mathrm{P}_0\left(G' \,|\, G\right) = \sum_U \mathrm{P}_{0,\mathrm{learn}}\left(G' \,|\, U\right) \mathrm{P}_{0,\mathrm{prod}}\left(U \,|\, G\right), \tag{4.30}$$

where production is unbiased and equal to the likelihood that a learner assigns to data, i.e.

$$\mathrm{P}_{0,\mathrm{prod}}\left(U \,|\, G\right) = \prod_l \prod_v \left[g_{lv}\varphi\left(l\right)\right]^{u_{lv}} = \mathrm{P}_{0,\mathrm{L}}\left(U \,|\, G\right) , \tag{4.31}$$

and learning takes place by sampling from the Bayesian posterior, i.e.

$$\mathrm{P}_{0,\mathrm{learn}}\left(G' \,|\, U\right) = \frac{\mathrm{P}_{0,\mathrm{L}}\left(U \,|\, G'\right) \Pi_0\left(G'\right)}{\mathrm{M}\left(U\right)} . \tag{4.32}$$

In this basic formulation of IBL, detailed balance of the prior is satisfied by virtue of the symmetry between the learning and production processes, due to the likelihood and the production probabilities being equal. I will refer to this as the *rational learner* condition. It assumes that the learner is perfectly aware of the mechanisms giving rise to the data that they received. A variety of mechanisms that break this symmetry have been explored in the literature [78–81]. These works tend to focus on the consequences of these mechanisms on the stationary

Condition	**Convergence to the prior**	**Detailed balance**
Perfect understanding (eq. 4.11)	Yes	Yes
Stationary game of telephone (eq. 4.17)	Yes	Not generally
Data-independent understanding (eq. 4.27)	Not generally	Yes

Table 4.2 *Summary of the conditions for convergence to the prior and detailed balance in the IBL model with nontrivial understanding. While convergence to the prior and detailed balance co-occur in certain scenarios, there are conditions where the stationary distribution of a reversible chain does not equal the prior, and where the stationary distribution does equal the prior but the chain is not reversible.*

distribution of the system, but not on the reversibility of the process in the stationary state and the directionality of change that arises from it.

Entropy production, introduced in the study of non-equilibrium thermodynamic systems, is able to quantify how far from equilibrium a process is by measuring its deviation from detailed balance [198, 199]. Here, I aim at applying it to quantify deviations from reversibility in the stationary states of iterated Bayesian models of the cultural transmission of language. In this context, the entropy production Σ will be given by [198]:

$$\Sigma = \int \mathrm{P}\left(G'|G\right)\Pi\left(G\right)\ln\left(\frac{\mathrm{P}\left(G'|G\right)\Pi\left(G\right)}{\mathrm{P}\left(G|G'\right)\Pi\left(G'\right)}\right)\mathrm{d}G\mathrm{d}G' \,, \tag{4.33}$$

thus quantifying differences in likelihood between forward ($\mathrm{P}\left(G'|G\right)\Pi\left(G\right)$, from G to G') and backward ($\mathrm{P}\left(G|G'\right)\Pi\left(G'\right)$, from G' to G) trajectories connecting grammars G and G' over the entire grammar space in the stationary state. This is generally not trivial, as stationary distributions $\Pi\left(G\right)$ will generally not have closed analytical forms for arbitrary IBL models, and for the model with nontrivial understanding presented in this chapter in particular. Here, I overcome this issue by considering only differentiable models that are close to equilibrium, and finding computationally tractable leading-order contributions to entropy production. I further apply this result to the quantification and comparison of directionality in a variety of out-of-equilibrium IBL models.

4.4.1 Leading-order entropy production

Consider a general IBL model with stationary distribution $\Pi_\epsilon\left(G\right)$ and transition probability $\mathrm{P}_\epsilon\left(G' \,|\, G\right)$ that depend differentiably on a model parameter ϵ, such that at $\epsilon = 0$, the transition probability becomes $\mathrm{P}_0\left(G' \,|\, G\right)$ as introduced above, and its stationary distribution becomes its prior $\Pi_0\left(G\right)$. P_0 and Π_0 satisfy detailed balance,

$$\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right) = \mathrm{P}_0\left(G|G'\right)\Pi_0\left(G'\right). \tag{4.34}$$

Generally, and in particular when $\epsilon \neq 0$, we have instead:

$$\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right) = \mathrm{P}_\epsilon\left(G \,|\, G'\right)\Pi_\epsilon\left(G'\right) + \sum_{n=1}^{\infty}\epsilon^n\Delta_n\left(G' \,|\, G\right) \,, \tag{4.35}$$

where

$$\Delta_n\left(G' \,|\, G\right) = \frac{\partial^n}{\partial \epsilon^n}\left[\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right) - \mathrm{P}_\epsilon\left(G \,|\, G'\right)\Pi_\epsilon\left(G'\right)\right]\Big|_{\epsilon=0} . \tag{4.36}$$

This expansion can be introduced in equation 4.33 to find the leading order contribution in ϵ, as laid out in Appendix C. This results in:

$$\Sigma = \frac{\epsilon^2}{2}\int \frac{\Delta_1\left(G'|G\right)^2}{\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G' + \mathcal{O}(\epsilon^3)\,. \tag{4.37}$$

Thus, the leading-order contribution in ϵ to the entropy production of the system is quadratic, and depends only on Δ_1, P_0 and Π_0. While analytical forms will usually be available for the latter two, that will not generally be the case for the former. In order to estimate it, let us further define:

$$\Phi\left(G' \,|\, G\right) \equiv \frac{\partial}{\partial \epsilon}\,\mathrm{P}_\epsilon\left(G' \,|\, G\right)|_{\epsilon=0} \tag{4.38}$$

$$\delta\left(G\right) \equiv \frac{\partial}{\partial \epsilon}\,\Pi_\epsilon\left(G\right)|_{\epsilon=0}\ . \tag{4.39}$$

With that, we have

$$\begin{aligned}\Delta_1\left(G'|G\right) =& \Phi\left(G'|G\right)\Pi_0\left(G\right) - \Phi\left(G|G'\right)\Pi_0\left(G'\right) \\ &+ \mathrm{P}_0\left(G'|G\right)\delta\left(G\right) - \mathrm{P}_0\left(G|G'\right)\delta\left(G'\right)\ .\end{aligned} \tag{4.40}$$

Generally speaking, the first order correction to the transition probability distribution (eq. 4.38) will be more easily computable than the first order correction to the stationary distribution (eq. 4.39), as the stationary distribution will generally have no analytical form for arbitrary ϵ. $\delta\left(G\right)$ can be estimated, however, by assuming quick convergence to stationarity from an initial state equalling the prior. This assumption leads to:

$$\delta\left(G\right) \approx \int \Phi\left(G \,|\, G'\right)\Pi_0\left(G'\right)\mathrm{d}G'\,. \tag{4.41}$$

By introducing equation 4.40 into equation 4.37 together with this approximation, we arrive at

$$\Sigma = \frac{\epsilon^2}{2}\int \frac{\Delta_\Phi\left(G' \,|\, G\right)^2}{\mathrm{P}_0\left(G'|G\right)\Pi_0(G)}\mathrm{d}G\mathrm{d}G' - \epsilon^2\int \frac{\left(\int \Delta_\Phi\left(G' \,|\, G\right)\mathrm{d}G'\right)^2}{\Pi_0(G)}\mathrm{d}G\,, \tag{4.42}$$

where I have defined:

$$\Delta_{\Phi}\left(G' \,|\, G\right) \equiv \Phi\left(G'|G\right)\Pi_0(G) - \Phi\left(G|G'\right)\Pi_0(G')\,. \tag{4.43}$$

Analytical steps to arrive at this result are again spelled out in Appendix C. This reduces the computation of the integrand of the entropy production to the known P_0 and Π_0, as well as the anti-symmetric quantity $\Delta_{\Phi}\left(G' \,|\, G\right)$, which can be computed for each choice of model.

4.4.2 Entropy production in IBL models

We are now set for the application of equation 4.42 to the quantification of directionality in IBL models of language change. As previously discussed, directionality is often generated in this paradigm through the breakage of the *rational speaker* condition, whereby a learner has full access to the processes that generated the data that they received. I will now explore three different ways in which this condition can be broken, and compare the extent to which they generate directionality by measuring entropy production for each of them. In general, and to facilitate comparison, I will be considering systems with three linguemes and two variants ($L = 3$, $V = 2$), with a symmetric Dirichlet prior,

$$\Pi_0\left(G\right) = \prod_l \frac{\Gamma\left(2\alpha\right)}{\Gamma\left(\alpha\right)^V} \prod_v g_{lv}^{\alpha-1}\,, \tag{4.44}$$

where $\alpha < 1$ to encode a bias against variability.

Nontrivial understanding

The condition of nontrivial understanding introduced in this chapter breaks the rational speaker condition by introducing a layer of uncertainty between the data generated by the speaker and received by the listener through an understanding phase, as shown in equation 4.10. Here, I will be considering a single-utterance understanding probability given by

$$\mathrm{p}_{\mathrm{und}}\left(l', v' \,|\, l, v\right) = \delta_{v',v}\left[\delta_{l',l}\left(1-\eta\right) + \eta\psi\left(l' \,|\, l\right)\right]\,, \tag{4.45}$$

where variants are always perfectly understood, and the error rate in lingueme understanding is given by η. When error in understanding is taking place, lingueme

l' will be understood when l was meant with probability $\psi\left(l' \,|\, l\right)$. Note that $\eta = 0$ leads to perfect understanding, which reduces the transition probability distribution to $\mathrm{P}_0\left(G' \,|\, G\right)$. When $\eta \neq 0$, the leading-order contribution to entropy production can be computed by first finding Δ_Φ, which is given by

$$\Delta_\Phi^{\mathrm{und}}\left(G' \,|\, G\right) = \sum_{\substack{U \\ \text{differ only by} \\ \text{utterance } m}} \sum_{U'} \mathrm{P}_{0,\mathrm{learn}}\left(G \,|\, U\right) \mathrm{P}_{0,\mathrm{learn}}\left(G' \,|\, U'\right) \times \\ \times \left[\psi\left(l'_m \,|\, l_m\right) \mathrm{M}\left(U\right) - \psi\left(l_m \,|\, l'_m\right) \mathrm{M}\left(U'\right)\right] . \tag{4.46}$$

Thus, we are summing over all U and U' that differ only by one utterance. Directionality is thus generated through an anti-symmetric term that depends on the understanding process through ψ. It is then to be expected that different ψ's will beget directionality in the stationary state to different degrees. To explore this, I will consider two choices of ψ. The first one introduces explicitly directional misunderstanding, whereby each lingueme can only ever be understood as one other lingueme in the language if imperfect understanding is taking place. This represents a system where semantic change happens overwhelmingly in a specific direction, as is typical of many processes of language change [197].

$$\begin{aligned} \psi_{\mathrm{dir}}\left(l' \,|\, l_1\right) &= \delta_{l',l_2} , \\ \psi_{\mathrm{dir}}\left(l' \,|\, l_2\right) &= \delta_{l',l_3} , \\ \psi_{\mathrm{dir}}\left(l' \,|\, l_3\right) &= \delta_{l',l_1} . \end{aligned} \tag{4.47}$$

The second choice represents a neutral model of misunderstanding, whereby linguemes are sampled from the lingueme distribution φ if imperfect understanding is taking place:

$$\psi_{\mathrm{neutral}}\left(l' \,|\, l\right) = \varphi\left(l'\right) . \tag{4.48}$$

Biased learning: maximum a posteriori estimation

Maximum *a posteriori* (MAP) estimation is a learning strategy that differs from the unbiased sampling from the posterior used in this chapter and other IBL models in that it biases learning in favour of the maximum of the posterior distribution. This has been shown to give rise to stationary distributions that accentuate features of the prior, rather than being identical to it [78, 85]. This

effect can be implemented differentiably through an exponent $\gamma > 0$ that makes the posterior more peaked around its maximum:

$$\mathrm{P}_{\mathrm{learn}}\left(G \,|\, U\right) = \frac{\left[\mathrm{P}_{0,\mathrm{L}}\left(U \,|\, G\right)\Pi_0\left(G\right)\right]^{1+\gamma}}{\int\left[0, \mathrm{P}_{\mathrm{L}}\left(U \,|\, G'\right)\Pi_0\left(G'\right)\right]^{1+\gamma}\mathrm{d}G'}\,. \tag{4.49}$$

When $\gamma = 0$, this simplifies to the unbiased learning probability $\mathrm{P}_{0,\mathrm{learn}}\left(G \,|\, U\right)$. In this form, this posterior can be expanded up to first order in γ to find the Δ_Φ of the model, and through it the leading-order contribution to entropy production:

$$\Delta_\Phi^{\mathrm{MAP}}\left(G' \,|\, G\right) = \sum_U \mathrm{M}\left(U\right)\mathrm{P}_{0,\mathrm{learn}}\left(G \,|\, U\right)\mathrm{P}_{0,\mathrm{learn}}\left(G' \,|\, U\right)\times \\ \times\left[\ln\left(\mathrm{P}_{0,\mathrm{learn}}\left(G \,|\, U\right)\right) - \ln\left(\mathrm{P}_{0,\mathrm{learn}}\left(G' \,|\, U\right)\right)\right]\,. \tag{4.50}$$

Biased production: selection with non-rational learners

Biased production does not have to break detailed balance, as long as the learners are able to identify the biases that gave rise to the data and account for them in their learning, i.e. $\mathrm{P}_{\mathrm{L}}\left(U \,|\, G\right) = \mathrm{P}_{\mathrm{prod}}\left(U \,|\, G\right)$. When this is *not* the case, convergence to the prior is broken, and it is thus worth exploring whether directionality is present. The single-utterance production strategy I will be considering here is:

$$\mathrm{p}_{\mathrm{prod}}\left(v, \,|\, l, G\right) = \frac{g_{lv}\mathrm{e}^{s_v}}{g_{lv}\mathrm{e}^{s_v} + 1 - g_{lv}}\,. \tag{4.51}$$

This is a case of lingueme-independent selection acting on variant v with force s_v. In the case with only two variants, we have $s \equiv s_1 = -s_2$. This production strategy reduces to unbiased production $\mathrm{P}_{0,\mathrm{prod}}\left(U \,|\, G\right)$ when $s = 0$. The Δ_Φ of this models is then:

$$\Delta_\Phi^{\mathrm{sel}}\left(G' \,|\, G\right) = \sum_U \mathrm{M}\left(U\right)\mathrm{P}_{0,\mathrm{learn}}\left(G \,|\, U\right)\mathrm{P}_{0,\mathrm{learn}}\left(G' \,|\, U\right)\times \\ \times\sum_l\left[\left(g_{lv_2} - g'_{lv_2}\right)u_{lv_1} - \left(g_{lv_1} - g'_{lv_1}\right)u_{lv_2}\right]\,. \tag{4.52}$$

As a reminder, u_{lv} represents the number of utterances using variant v to express lingueme l in a data set.

Results for the entropy production of the models

Results are shown in figure 4.4. The integral of the entropy production (eq. 4.42) becomes computationally intractable as the sets of utterances in inter-generational transmission U increase in size, as the Δ_Φ and P_0 terms involve larger sums. Thus, Monte Carlo integration [205] could only produce reasonably precise results within reasonable computation times for small utterance set sizes between 1 and 5.

All models have non-zero leading-order contributions to the entropy production, reflecting that they all generate directional behaviour in the stationary state to some extent. The model with MAP learning starts off as the one with the highest contribution but quickly decreases as the number of utterances increase. This is possibly reflective of the MAP and sampling strategies becoming increasingly similar as the posterior distribution becomes peaked around its maximum as the utterance sample size increases. Every other model increases in entropy production as the utterance set size increases. The two models with nontrivial understanding produce drastically different results. The model with neutral misunderstanding consistently produces the least entropy, reflecting a lower degree of directional behaviour in the stationary state. The model with explicitly directional misunderstanding, conversely, produces the most entropy out of all models when the number of utterances is higher. This exemplifies the power of communicative and semantic processes in generating directionality in language change, and demonstrates the potential of simple models with nontrivial understanding in being able to reproduce some statistical features of these diachronic processes. Further work would be needed to explore whether this can be extended to empirical studies.

4.5 Discussion

In this chapter, I have presented a novel iterated Bayesian learning paradigm for grammatical change in which L linguemes (word meanings or grammatical relations) can be expressed by any of V variants (words or structures). Between the production and learning phase typical of most iterated learning models [76, 85], a novel understanding phase is introduced, which is able to codify imperfect understanding, semantic effects, and nontrivial communicative effects.

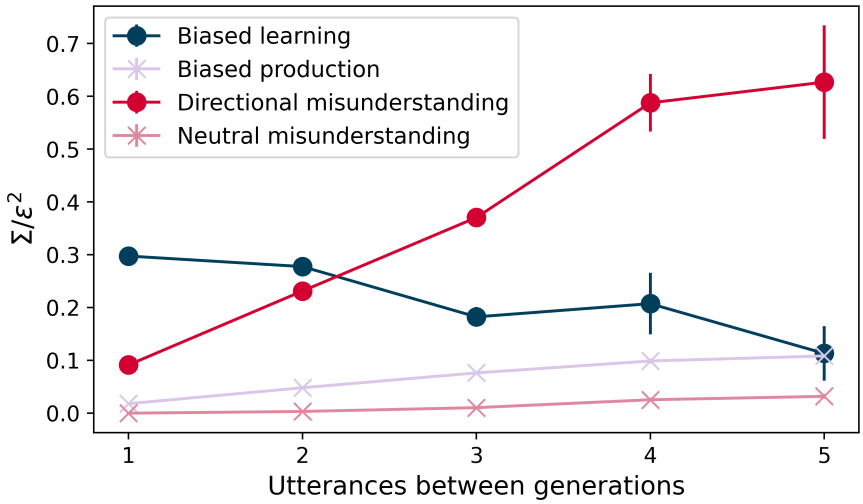

Figure 4.4 *Leading order contribution to the entropy production from a variety of out-of-equilibrium iterated Bayesian learning models as a function of the number of utterances transmitted between generations.*

It thus provides a minimal model for the breaking of the isolation assumption, allowing for nontrivial interactions between the evolutionary dynamics of different linguemes.

This has deep implications for the statistical properties of the stationary state of the Markov chain generated by the model. As explored in Section 4.2, nontrivial understanding effects are able to generate stationary distributions distinct from the prior, aligning with the idea that distributional universals are moulded by communication, and not just cognition. Some nontrivial understanding effects can be identified, however, that maintain convergence to the prior. In Section 4.3, I went a step further and studied the diachronic properties of the arising stationary state. Analytical conditions for the breaking of detailed balance were found, concluding that most nontrivial understanding effects lead to detailed balance breaking, and therefore directional change, even in the stationary state. The model is thus a first step towards being able to mathematically capture the directional behaviour typical of processes of language change such as grammaticalisation.

Section 4.4 aimed at understanding how exactly directional behaviour is generated close to equilibrium. A general formula for the leading-order contribution to entropy production was found and applied to four different IBL paradigms, including two different models with nontrivial understanding effects. When a neutral model of misunderstanding is used, whereby understanding depends only

on the relative frequency of linguemes in the environment, very little directional behaviour is achieved, reflected by comparatively low entropy production. Conversely, a great degree of directionality is achieved in the model with explicitly directional misunderstanding which mimics the cyclic behaviour of some grammaticalisation processes [197], surpassing the entropy production of models with biased learning and production. In all, directionality is an underexplored topic in the modelling of language change. This chapter demonstrates that techniques from non-equilibrium statistical mechanics may allow for the systematic and testable quantification of directionality, not only in stochastic models but potentially also in empirical applications. This, however, remains to be tested in future research.

Chapter 5

A Wright-Fisher paradigm of grammar change

In the previous chapter, I introduced the notion that the evolutionary paradigms applied in Chapters 2 and 3 of this thesis are limited by assumptions of *isolation* and *stability* of the linguistic competition process. According to the isolation assumption, the evolutionary dynamics of linguemes are isolated from one another, which is patently untrue when examining processes such as analogical change or phonological chain shifts [91]. According to the stability assumption, variants and linguemes may only associate in very restrictive ways, with variants only ever realising one lingueme. As previously established, the prevalence of polysemy [195] – especially when considering functional vocabulary – means this does not hold generally. Thus, a great portion of processes of historical change, and particularly those involving change in functional vocabulary and grammatical relations, are currently inaccessible to evolutionary models such as Wright-Fisher. The goal of this chapter is to develop an evolutionary paradigm addressing these issues, and proving its applicability to the formulation and testing of hypotheses using time series of historical corpus data.

The IBL model developed in the previous chapter set the foundation for the quantitative exploration of change beyond the isolation and stability assumptions. By modelling arbitrarily associating linguemes and variants, it was able to move beyond the stability assumption. By further introducing nontrivial communicative effects interlinking different linguemes, it provided us with a minimal but flexible model breaking the isolation assumption. It is, however, not

ideal for data-analytic efforts in this form. In Section 5.1, inspired by Florencia Reali and Tom Griffiths's 2010 work *Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift* [4], I work towards deriving a Wright-Fisher-like model from this IBL paradigm with nontrivial communication. In the biological analogy, the resulting model treats different linguemes as different subpopulations of an organism. On top of the selection and variation effects present in the Wright-Fisher model introduced in Section 1.3.2, this allows for the introduction of evolutionary interactions between different linguemes that are akin to migration effects [107, 206] in the biological analogy.

In Section 5.2, I extend the methodological advancements introduced in Chapter 2 to make them applicable to historical data involving the co-evolution of several linguemes under the presented mathematical paradigm. The evolutionary dynamics of functional vocabulary are a particularly interesting case study for this model, as they usually involve many-to-many associations between linguemes and variants. In Section 5.3, I apply the model to two such systems: English relativisers throughout Middle and Modern English, and the rise of do-support in Early Modern English. These applications demonstrate the potential of the model presented in this chapter towards the quantitative formulation and testing of hypotheses in language change.

5.1 Equivalence between IBL and WF paradigms

In Section 4.1, a Markov transition probability distribution ($\mathrm{P}\left(G' \,|\, G\right)$, eq. 4.10) was derived for the model with nontrivial understanding effects. However, when working with time series of corpus data, we do not have access to the grammars of the speakers that produced the texts – only to their utterances. Thus, from an empirical perspective, the transition probability between sets of utterances at different generations $\mathrm{P}\left(U_1 \,|\, U_0\right)$ is more interesting than that between grammars. The goal of this section is thus to derive this utterance transition probability for a model with L linguemes and V variants subject to nontrivial understanding effects. Generally, this is given by

$$\mathrm{P}\left(U_1 \,|\, U_0\right) = \prod_l \frac{u_l^1!}{\prod_v u_{lv}^1!} \prod_v \mathrm{p}_{\mathrm{prod}}\left(v \,|\, l,\, U_0\right)^{u_{lv}^1} \,, \tag{5.1}$$

where $u_l^1 = \sum_v u_{lv}^1$ is the total usage of lingueme l in the set of utterances U_1, and $\mathrm{p}_{\mathrm{prod}}(v \mid l, U_0)$ is the average population-level production probability of variant v given that speakers want to express lingueme l and were exposed to utterances U_0. Generally, the production strategy of a single speaker will be a function of their grammar:

$$\mathrm{p}_{\mathrm{prod}}^1(v \mid l, G) = f(g_{lv}) \ . \tag{5.2}$$

Here, g_{lv} is the speaker's grammar. The function f represents the production biases of the speaker, such that in the case of neutral, non biased production one has $f(g_{lv}) = g_{lv}$. That is, a neutral speaker will produce a lingueme-variant pair as frequently as their grammar says the pair appears in the language. Following Reali and Griffiths (2010) [4], I approximate the production strategy of the entire population as the mean of this production strategy over the posterior over grammars after exposure to utterances U_0, i.e.

$$\begin{aligned} \mathrm{p}_{\mathrm{prod}}(v \mid l, U_0) =& \mathbb{E}\left[\mathrm{p}_{\mathrm{prod}}^1(v \mid l, G) \mid U_0\right] \\ =& \mathbb{E}\left[f(g_{lv}) \mid U_0\right] \\ \approx& f(\hat{g}_{lv}(U_0)) \ , \end{aligned} \tag{5.3}$$

where $\hat{g}_{lv}(U_0)$ represents the mean of g_{lv} over the posterior over grammars of a speaker, and the average production strategy was approximated as the production strategy applied to the average grammar, which is given by:

$$\begin{aligned} \hat{g}_{fe}(U_0) =& \mathbb{E}_{\mathrm{und}}\left[\mathbb{E}_{\mathrm{learn}}\left[g_{lv} \mid U\right] \mid U_0\right] \\ =& \sum_U \left[\int g_{lv} \mathrm{P}_{\mathrm{learn}}(G \mid U)\, \mathrm{d}G\right] \mathrm{P}_{\mathrm{und}}(U \mid U_0) \ . \end{aligned} \tag{5.4}$$

Unlike Reali and Griffiths, who assumed perfect communication, we now take nontrivial understanding into account, as reflected in equation 4.10: in this model, the utterances that speakers were exposed to are not the same ones that they understood and subsequently learned from. Therefore, we have to average not only over the grammar frequencies g_{lv}, but also over the sets of understood utterances U that they would have been inferred from.

Assuming a Dirichlet prior (eq. 4.7), the learning probability takes on a tractable form (eq. 4.8) that leads to a closed analytical form of the average of the learning

process:

$$\begin{aligned}
\mathbb{E}_{\text{learn}}\left[g_{lv} \,|\, U\right] =& \prod_{l'} \frac{\Gamma\left(u_{l'} + \alpha_{l'}\right)}{\prod_{v'} \Gamma\left(u_{l'v'} + \alpha_{l'v'}\right)} \int g_{lv} \prod_{v'} g_{l'v'}^{u_{l'v'} + \alpha_{l'v'} - 1} \mathrm{d}G \\
=& \frac{\Gamma\left(u_{l} + \alpha_{l}\right)}{\prod_{v'} \Gamma\left(u_{lv'} + \alpha_{lv'}\right)} \int_{\sum_{v'} g_{lv'} = 1} g_{lv} \prod_{v'} g_{lv'}^{u_{lv'} + \alpha_{lv'} - 1} \prod_{v'} \mathrm{d}g_{lv'} \\
=& \frac{\Gamma\left(u_{l} + \alpha_{l}\right)}{\prod_{e'} \Gamma\left(u_{lv'} + \alpha_{lv'}\right)} \frac{\Gamma\left(u_{lv} + \alpha_{lv} + 1\right) \prod_{v' \neq v} \Gamma\left(u_{lv'} + \alpha_{lv'}\right)}{\Gamma\left(u_{l} + \alpha_{l} + 1\right)} \\
=& \frac{\Gamma\left(u_{lv} + \alpha_{lv} + 1\right)}{\Gamma\left(u_{lv} + \alpha_{lv}\right)} \frac{\Gamma\left(u_{l} + \alpha_{l} + 1\right)}{\Gamma\left(u_{l} + \alpha_{l}\right)} \\
=& \frac{u_{lv} + \alpha_{lv}}{u_{l} + \alpha_{l}} \,.
\end{aligned} \tag{5.5}$$

In the second equality, all integrals not performed over grammar frequencies corresponding to lingueme l cancelled out the normalising factors of the prior probabilities, resulting in an integral that equals the normalising factor of a Dirichlet distribution. In the fifth equation, the identity $\Gamma(x+1) = x\Gamma(x)$ was used to simplify the resulting expression. This is the multi-lingueme equivalent of the result obtained by Reali and Griffiths [4], as summarised in Section 1.3.3. If assuming nontrivial communication, however, the token counts in u_{lv} need to be further averaged over the understanding process, as laid out in equation 5.4. If the set of utterances is large and thus the understanding probability is sharply peaked around its average, we can approximate:

$$\hat{g}_{lv}\left(U_0\right) = \mathbb{E}_{\text{und}}\left[\frac{u_{lv} + \alpha_{lv}}{u_{l} + \alpha_{l}} \,|\, U_0\right] \approx \frac{\mathbb{E}_{\text{und}}\left[u_{lv} \,|\, U_0\right] + \alpha_{lv}}{\mathbb{E}_{\text{und}}\left[u_{l} \,|\, U_0\right] + \alpha_{l}} \,. \tag{5.6}$$

In general, the average token count post-communication will be given by:

$$\mathbb{E}_{\text{und}}\left[u_{lv} \,|\, U_0\right] = u_{lv}^0 - \sum_{l',v'} \mathbb{E}_{\text{und}}\left[u_{lv \to l'v'} \,|\, U_0\right] + \sum_{l',v'} \mathbb{E}_{\text{und}}\left[u_{l'v' \to lv} \,|\, U_0\right] \,. \tag{5.7}$$

Here, $u_{lv \to l'v'}$ represents the number of tokens originally intended to express lingueme l using variant v that ended up being understood by the following generation as expressing lingueme l' using v'. These are, of course, unknown quantities. From here, we can only go further if we choose a particular form of the understanding probability. For the purpose of most practical applications, I will be assuming a simple understanding process given for individual utterances (l, v) by:

$$\mathrm{p}_{\text{und}}\left(l',\, v' \,|\, l, v\right) = \delta_{v,v'}\left[\delta_{l,l'}\left(1 - \eta\right) + \eta\psi\left(l' \,|\, l\right)\right] \,, \tag{5.8}$$

where the understanding of variants is perfect, and imperfect understanding of linguemes occurs with rate η. When a lingueme l was meant and imperfect understanding is indeed taking place, a lingueme l' is understood instead with probability $\psi(l' \mid l)$. With this understanding probability, $u_{lv \to l'v'}$ are independent binomially distributed variables with number of trials u^0_{lv} and success probability $[\delta_{l',l}(1-\eta) + \eta\psi(l' \mid l)]$. This allows for the explicit computation of their average:

$$\mathbb{E}_{\text{und}}\left[u_{lv \to l'v} \mid U_0\right] = u^0_{lv}\left[\delta_{l,l'}(1-\eta) + \eta\psi(l' \mid l)\right] , \tag{5.9}$$

which in turn simplifies equation 5.7 to

$$\mathbb{E}_{\text{und}}\left[u_{lv} \mid U_0\right] = u^0_{lv}(1-\eta) + \eta \sum_{l'} u^0_{l'v}\psi(l \mid l') . \tag{5.10}$$

With this, the average speaker grammar (eq. 5.5) takes the form

$$\hat{g}_{lv}(U_0) = \frac{u^0_{lv}(1-\eta) + \eta\sum_{l'} u^0_{l'v}\psi(l \mid l') + \alpha_{lv}}{u^0_{l}(1-\eta) + \eta\sum_{l'} u^0_{l'}\psi(l \mid l') + \alpha_{l}} . \tag{5.11}$$

As a sanity check, it can be easily proved that these satisfy the normalisation condition $\sum_v \hat{g}_{lv}(U_0) = 1$. With this, equation 5.1 finally takes the form:

$$\begin{aligned} \mathrm{P}(U_1 \mid U_0) &= \prod_l \frac{u^1_l!}{\prod_v u^1_{lv}!} \prod_v f\left(\hat{g}_{lv}(U_0)\right)^{u^1_{lv}} \\ &= \prod_l \mathrm{P}_{\text{WF}}\left(\{u^1_{lv} : v = 1, ..., V\} \mid \{f\left(\hat{g}_{lv}(U_0)\right) : v = 1, ..., V\}\right) , \end{aligned} \tag{5.12}$$

that is, it is a product of a Wright-Fisher process for each lingueme, where the role of the fitness functions is now taken by the average production strategy composed with the average speaker grammars, $f\left(\hat{g}_{fe}(U_0)\right)$. This facilitates applying the model to historical data immensely, as we can inherit insight and computational techniques from previous chapters. There are some key differences, however, that must be taken into account when exporting methodologies from Chapters 2 and 3. Firstly, new evolutionary forces can be encoded in fitness functions when the isolation and stability assumptions are abandoned. Secondly, the Beta-with-Spikes approximation may no longer be viable when working with systems with more than two expressions of interest.

5.2 Methodology for application to historical data

In previous chapters, we discussed how causal effects in language change can be mapped onto evolutionary forces under the Wright-Fisher paradigm, which in turns allows for the formulation of hypotheses that are empirically testable through model fitting and comparison using historical data. When the isolation and stability hypotheses are broken, a variety of new evolutionary effects can be incorporated into the fitness functions in equation 5.12. However, in order for them to be adequately measured, the methodologies discussed in previous chapters need to be adjusted, and several additional considerations need to be taken into account.

5.2.1 Evolutionary forces in grammar change

The first step when formulating a hypothesis of the evolutionary dynamics of a historical process is to translate that hypothesis into explicitly parametrisable evolutionary forces in the fitness function in equation 5.12. Notably, this fitness function explicitly distinguishes between evolutionary forces that enter the process during learning and understanding – which appear in the average grammar $\hat{g}_{lv}$ – and those introduced during production – which are parametrised in the production function f. As I will present shortly, production can include effects ranging from social biases to analogy. Notably, equation 5.3 explicitly determines the order in which the composition of production and learning or understanding effects has to take place. In the following, I will indulge in abuse of notation and continue referring to all fitness functions as g_{lv} only, independently of whether they incorporate production effects or not.

In order to test any hypothesis, it must be compared to a null model. As in previous chapters, I will be taking this to be a model of pure drift, where any changes in the relative frequencies of variants are due to unbiased stochasticity in inter-generational transmission. So far, the effects of drift were quantified through a drift parameter N, whose value was taken as constant throughout the evolutionary trajectory of the system. Here, additional considerations need to be taken into account. In particular, I am taking the N parameter to correspond to the entire population of utterances. The drift parameter of the evolutionary dynamics of an individual lingueme l can then be estimated as a proportion of N

corresponding to the relative usage $\varphi(l)$ of lingueme l in the population, i.e.:

$$N_l = N\varphi(l), \tag{5.13}$$

which may be approximated as constant throughout the trajectory. With that, the Wright-Fisher process in equation 5.12 for lingueme l can be reformulated as

$$\mathrm{P_{WF}}\left(X = \{x_{fe}\} \,|\, \{g_{lv}\}\right) = \frac{(N\varphi(f))!}{\prod_e (N\varphi(f)\, x_{fe})!} \prod_e g_{lv}^{N\varphi(f)x_{fe}}. \tag{5.14}$$

Note that, in order to keep methodologies consistent with previous chapters and approximate frequency of usage as a continuous variable, I am no longer using the total token counts u_{lv} of variant v expressing lingueme l in a given time window, but its normalised usage $x_{lv} = u_{lv}/u_l$. This satisfies $\sum_e x_{lv} = 1$ for every l. Both $\varphi(l)$ and $\{x_{lv}\}$ can be estimated from historical data.

As we established in previous chapters, selective forces are commonly defaulted to when modelling directional evolution, and may include a variety of effects ranging from the social to the cognitive. Generally, a fitness function incorporating selection will be given by

$$g_{lv}^{\mathrm{sel}}(X) = \frac{x_{lv}\mathrm{e}^{s_{lv}}}{\sum_{v'} x_{lv'}\mathrm{e}^{s_{lv'}}}. \tag{5.15}$$

This form assumes that selection parameters s_{lv} depend on both the lingueme and the variant of the utterances. This may be a sensible choice in certain scenarios, e.g. when dealing with social biases favouring specific variants only in specific contexts. Conversely, some scenarios may be better described using lingueme-independent selective forces, i.e. $s_{lv} = s_v$. For example, articulatory or economy biases may favour short or easily-pronounced variants, independently of the lingueme that they express.

Variation, equivalent to mutation in biology, encompasses processes where new variants arise and transform from one to another. A fitness function for variation has the general form:

$$g_{lv}^{\mathrm{var}}(X) = x_{lv}\left(1 - \sum_{l'} \epsilon_{vv'}^{l}\right) + \sum_{v'} \epsilon_{v'v}^{l} x_{fv'}, \tag{5.16}$$

where $\epsilon_{vv'}^{l}$ represents the inter-generational variation rate between variants v and v' for lingueme l. A variety of production effects can be understood as variation

in this form, including systematic errors and innovation. As Reali and Griffiths [4] already noted, the average speaker grammar in equation 5.11 with $\eta = 0$ is equivalent to variation when taking $\epsilon^{l}_{v'v} = \frac{\alpha_{lv}}{N\varphi(l)+\alpha_l}$ for all v'. Learning biases can therefore also be subsumed under the umbrella of variation.

Notably, variation and selection effects do not challenge the isolation assumption explicitly: fitness functions 5.15 and 5.16 are fully decoupled from those of other linguemes in the system and only depend on utterances expressing lingueme l. One may wonder, then, what kind of evolutionary effects can account for the explicit coupling of linguemes appearing in equation 5.11 as a result of understanding effects, and for effects breaking the isolation assumption in general. For that, it may be useful to consider an evolutionary force that we have ignored so far: migration. In population genetics, migration has to do with the movement of individuals between independent subpopulations. While migration has been parametrised in a variety of ways, it often relies on the assumption that a subset of the population can be approximated as infinitely large [107, 206]. Here, I opt for a similar characterisation where all populations are assumed to be finite. Given migration parameters m_{ij} denoting the rate of population transfer from subpopulation i to subpopulation j in one generation, the fitness function of allele a in population i can be expressed as:

$$g_{ia}(X) = x_{ia}\left(1 - \sum_j m_{ij}\right) + \sum_j m_{ji}x_{ja}\,, \tag{5.17}$$

where, in order for subpopulations to maintain a stable size, their incoming and outgoing migratory fluxes have to be equal, i.e.:

$$\sum_j m_{ij} = \sum_j m_{ji}\,. \tag{5.18}$$

Equation 5.8 is equivalent to a migration process as described by 5.17, with migration parameters $m_{ll'} = \eta\psi(l' \mid l)$. The balancing of incoming and outgoing migration in equation 5.18 thus reduces to the condition that the understanding process does not change the relative proportion of linguemes, i.e.:

$$\varphi(l') = \sum_l \psi(l'|l)\,\varphi(l)\ , \tag{5.19}$$

Thus, different linguemes are, in the genetic analogy, akin to different subpopulations of an organism, with imperfect communication and other semantic processes

Effect	Parameters	Example
Selection	s_{lv}, selection strength of v	Social biases
Variation	$\epsilon^{l}_{vv'}$, variation rate from v to v'	Systematic errors
Migration	η, migration rate between linguemes	Imperfect communication
Analogy	$\gamma_{ll'}$, analogy rate between l and l'	Analogical extension

Table 5.1 *Non-exhaustive summary of parametrisable evolutionary forces in the presented model.*

being equivalent to migration between them. I will generally refer to these processes as *lingueme migration*. The simplest of these processes is one where $\psi\left(l' \mid l\right) = \varphi\left(l'\right)$, in which case the fitness function reduces to:

$$g^{\text{mig}}_{lv}\left(X\right) = x_{lv}\left(1-\eta\right) + \eta \sum_{l'} x_{l'v}\varphi\left(l'\right) \ . \tag{5.20}$$

This represents an understanding process where, in case of imperfect communication, the listener defaults to sampling from the distribution of linguemes.

Effects breaking the isolation assumption may also occur in production: analogy is a prominent example of this. For it, I will be positing the following fitness function:

$$g^{\text{ana}}_{lv}\left(X\right) = \frac{x_{lv} + \sum_{l'} x_{l'v}\gamma_{l'l}}{1 + \sum_{l'} \gamma_{l'l}} \, , \tag{5.21}$$

where $\gamma_{l'l}$ represents the rate of inter-generational analogical extension from lingueme l' to lingueme l. Table 5.1 summarises all evolutionary forces presented thus far, together with their associated parameters and linguistic effects.

Note that the model is not limited to one evolutionary effect per fitness function: several effects can be included in a single fitness function through composition. When this is the case, fitness functions accounting for effects in production must be composed last, as reflected in equation 5.3. For example, evolutionary dynamics presenting both selection in production and variation due to learning biases would be given by

$$g_{lv}\left(X\right) = g^{\text{sel}}_{lv}\left(g^{\text{var}}\left(X\right)\right) = \frac{\left[x_{lv}\left(1 - \sum_{v''}\epsilon^{l}_{v''}\right) + \epsilon^{l}_{v}\right]\mathrm{e}^{s_{lv}}}{\sum_{v'}\left[x_{lv'}\left(1 - \sum_{v''}\epsilon^{l}_{v''}\right) + \epsilon^{l}_{v}\right]\mathrm{e}^{s_{lv'}}} \, , \tag{5.22}$$

The deterministic trajectories (those with $N \to \infty$) generated by fitness functions incorporating different evolutionary forces are plotted in Figure 5.1 for selection (left, eq. 5.15), variation (centre, eq. 5.16) and lingueme migration (right, eq. 5.20) for a system with two linguemes and two variants, and two different starting

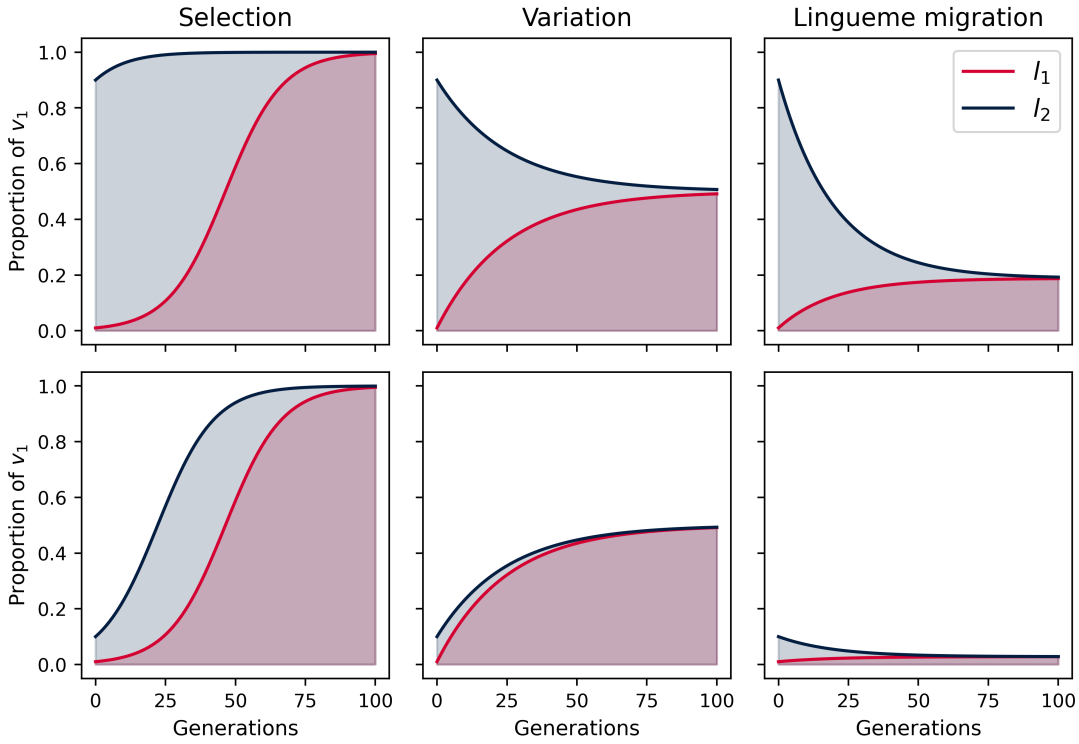

Figure 5.1 *Deterministic trajectories of different evolutionary forces for a system with $L = 2$, $V = 2$. Upper panels: trajectories with starting frequencies $x_{l_1 v_1} = 0.01$, $x_{l_2 v_1} = 0.9$. Lower panels: trajectories with starting frequencies $x_{l_1 v_1} = 0.01$, $x_{l_2 v_1} = 0.1$. Left: trajectories with selection favouring variant v_1, with $s_{v_1} = 0.1$. Centre: trajectories with variation, with $\epsilon = 0.02$. Right: trajectories with lingueme migration, with $\eta = 0.05$, $\varphi(l_1) = 0.8$, $\varphi(l_2) = 0.2$.*

variant frequencies for the second lingueme. Selection creates typical S-curves of change for the favoured variant. Variation tends to homogenise variant usage in all linguemes, independently of starting frequencies. Lingueme migration, on the other hand, homogenises variant use across linguemes in a way that depends on both starting frequencies and relative usage of each of the linguemes. Lingueme l_1 was taken to be four times more frequent than lingueme l_2 in this example, which is reflected in its slower trajectory in the example of lingueme migration. In stochastic processes and those where evolutionary parameters are small, however, it may not be as easy to distinguish the effects of variation and migration. This problem will be empirically explored in Section 5.3.

5.2.2 Maximum-likelihood methods in grammar change

Given a relevant set of linguemes and variants, we are again interested in constructing frequency time series of the form $X = \{(X^i, t_i) \,:\, i = 1, ..., m\}$, where $X^i = \{x^i_{lv} \,:\, l = 1, ..., L,\, v = 1, ..., V\}$ and x^i_{lv} is the proportion of utterances in the data corresponding to usage of variant v to express lingueme l in temporal bin i. The distribution of linguemes $\varphi(l)$ also needs to be estimated from the data as the proportion of total usage of each of the linguemes in the data set. This distribution will be assumed to stay constant throughout the trajectory of the data.

Once all relevant hypotheses for the evolutionary dynamics of a historical data set have been parametrised as evolutionary forces in the fitness functions, statistical testing can be applied to determine whether these forces are significant in the dynamics of a given data set. I again use maximum-likelihood methods to this end. Thus, the likelihood of a hypothesis formulated in terms of evolutionary parameters Θ given data X will be given by

$$\mathcal{L}(\Theta \,|\, X) = \prod_{i=1}^{m-1} \prod_{l=1}^{L} \mathrm{P_{WF}}\left(\left\{x^{i+1}_{lv} \,:\, v = 1, ..., V\right\} | X^i; \Theta\right) . \tag{5.23}$$

Unlike previously, the likelihood now involves a product of L co-evolving Wright-Fisher processes, one per lingueme of interest. Just like in Chapters 2 and 3, empirical estimations of the parameters, $\hat{\Theta}$ can be obtained as those that maximise the likelihood. Model comparison is also possible in the same way, by obtaining a test statistic λ as the ratio of the maximal log-likelihoods of the model of interest and a null-model with parameter space Θ_0, i.e.

$$\lambda = 2\ln\left(\frac{\mathcal{L}\left(\hat{\Theta} \,|\, X\right)}{\mathcal{L}\left(\hat{\Theta}_0 \,|\, X\right)}\right) . \tag{5.24}$$

Just like previously, a p-value can then be obtained by comparing λ to a numerically-generated likelihood ratio distribution [104].

When working with hypotheses involving several evolutionary effects, this process can be iterated. Assume two independent evolutionary effects A and B with parameter spaces Θ_A and Θ_B both produce significant p-values when tested against the null hypothesis. Further assume that A's likelihood is higher than B's. Then, the goodness-of-fit of a model incorporating both effects, with parameter

space Θ_{AB}, can be tested by finding the test statistic in equation 5.24 with $\hat{\Theta} = \hat{\Theta}_{\mathrm{AB}}$ and $\hat{\Theta}_0 = \hat{\Theta}_{\mathrm{A}}$ and similarly comparing it to an artificially generated distribution.

5.2.3 Temporal binning and parameter scaling

Just like in Chapter 2, the choice of Wright-Fisher generation time and temporal binning of the data may have varying effects on the significance level of models and numerical value of the estimated parameters. Following Section 2.1.5, I will be mitigating these effects by reanalysing time series under several choices of binning, and reporting the average and standard deviation of evolutionary parameters and p-values. In order to do so, evolutionary parameters must be properly normalised, as their values scale with the choice of generation time. Scaling laws for N and s were already found in Chapter 2, leading to

$$N_1 = tN_t \qquad s_1 = s_t/t\,, \tag{5.25}$$

where the subscript 1 represents the value of the parameter for a generation choice of 1 time unit, and the subscript t represents the value of the parameter for a generation choice of t time units.

Further approximate scaling laws can be found for the parameters in equations 5.16, 5.20 and 5.21 by composing their associated fitness functions and dropping all quadratic terms in evolutionary parameters. As an example, for equation 5.20, we have:

$$\begin{aligned} g_{lv}^{\mathrm{mig}}\left(g^{\mathrm{mig}}(X)\right) =& \left[x_{lv}(1-\eta) + \eta\sum_{l'} x_{l'v}\varphi(l')\right](1-\eta) \\ &+ \eta\sum_{l'}\left[x_{lv}(1-\eta) + \eta\sum_{f'} x_{l'v}\varphi(l')\right]\varphi(l') \\ =& x_{lv}\left(1-2\eta+\mathcal{O}\left(\eta^2\right)\right) + \eta\sum_{l'} x_{l'v}\varphi(l') + \mathcal{O}\left(\eta^2\right) \\ &+ \eta\sum_{l'} x_{l'v}\varphi(l') + \mathcal{O}\left(\eta^2\right) \\ =& x_{lv}(1-2\eta) + 2\eta\sum_{l'} x_{l'v}\varphi(l') + \mathcal{O}\left(\eta^2\right)\,. \end{aligned} \tag{5.26}$$

This composition describes the deterministic ($N \to \infty$) trajectory of a system

evolving under lingueme migration with parameter η for two generations. As can be seen in the final result, by comparing to equation 5.20, this equals up to first order in η the evolution of the system for a single generation with parameter 2η. Similar scaling properties can be found for equations 5.16 and 5.21, leading to:

$$\epsilon_1 = \epsilon_t/t \qquad \eta_1 = \eta_t/t \qquad \gamma_1 = \gamma_t/t \tag{5.27}$$

where again, the subscript 1 represents the value of the parameter for a generation choice of 1 time unit, and the subscript t represents the value of the parameter for a generation choice of t time units.

5.2.4 Approximation of the Wright-Fisher transition probability

There is then just one key feature of the methodologies developed in previous chapters that cannot be carried over here. When working with systems with multiple variants, the highly accurate Beta-with-Spikes approximation to the Wright-Fisher model cannot be applied, as it was developed for systems with only two variants (or only one variant of interest). Here, I will be using a Dirichlet approximation, arguably the next best approximation that does not rely on the diffusion limit, and a natural extension of the Beta approximation [95, 111] to the multivariate Wright-Fisher process. It is given for a single time step and a single lingueme by:

$$\mathrm{P_B}\left(x_{lv}^{t+1} : v = 1, ..., V \,|\, X^t; \Theta\right) = \frac{1}{B\left(\beta_l\right)} \prod_v \left(x_{lv}^{t+1}\right)^{\beta_{lv}-1} \tag{5.28}$$

where $B\left(\beta\right)$ is the normalisation of the distribution, given by:

$$B\left(\beta_l\right) = \frac{\prod_v \Gamma\left(\beta_{lv}\right)}{\Gamma\left(\sum_v \beta_{lv}\right)}, \tag{5.29}$$

and β_{lv} are parameters found by fitting the moments and variances of the exact Wright-Fisher distribution to those of the Dirichlet approximation. Given exact means $\mathbb{E}\left(x_{lv}^{t+1}\right)$ and variances $\mathrm{Var}\left(x_{lv}^{t+1}\right)$ of the Wright-Fisher process after one generation, β_{lv} can be found as:

$$\beta_{lv} = \left[\frac{\mathbb{E}\left(x_{lv}^{t+1}\right)\left(1 - \mathbb{E}\left(x_{lv}^{t+1}\right)\right)}{\mathrm{Var}\left(x_{lv}^{t+1}\right)}\right] \mathbb{E}\left(x_{lv}^{t+1}\right) . \tag{5.30}$$

$\mathbb{E}\left(x_{lv}^{t+1}\right)$ and $\mathrm{Var}\left(x_{lv}^{t+1}\right)$ can in turn be found as:

$$\mathbb{E}\left(x_{lv}^{t+1}\right) = g_{lv}\left(X^t;\Theta\right) \tag{5.31}$$

$$\mathrm{Var}\left(x_{lv}^{t+1}\right) = \frac{1}{N\varphi\left(l\right)} g_{lv}\left(X^t;\Theta\right)\left(1 - g_{lv}\left(X^t;\Theta\right)\right), \tag{5.32}$$

and thus β_{lv} depend implicitly on X^t, $\varphi\left(l\right)$ and the model parameters N and Θ. Whenever there is a gap in the time series due to no frequency of usage data being available for an intermediate generation, the self-contained scheme presented in Section 2.1 can be naturally extended to a system with L linguemes and V variants in order to estimate β_{lv} after any number of generations.

5.3 Applications to historical grammar change in English

Having developed a Wright-Fisher-based methodology for the analysis of grammatical change in the previous section, I proceed now to apply it to historical data sets. Here, I focus on the historical change of grammatical structures in English. All data analysed in this section was extracted from the corpora by my collaborator Prof Robert Truswell from the School of Philosophy, Psychology and Language Sciences at the University of Edinburgh.

A key feature of the model lies in its applicability to phenomena involving many-to-many associations in networks of linguemes and variants by virtue of its relaxation of the stability assumption. I thus focus first on the description of one such data set from the Penn Parsed Corpora of Historical English [207] and the Parsed Corpus of Middle English Poetry [208], involving the evolutionary dynamics of relative pronouns (also known as relativisers) in Middle and Modern English. The application of the model reveals that evolutionary forces of both variation and lingueme migration are significant in explaining the behaviour of the data. This highlights the importance of models moving past stability and isolation in the empirical study of grammatical change.

I further apply the model to yet another historical data set, relating to the rise of do-support in Early Modern English. As one of the most widely-studied phenomena of grammatical change in English historical linguistics, a plethora of hypotheses have been put forward for its causal mechanisms [209–212]. However,

most efforts so far have involved only regression models, which do not account for the inherent stochasticity of language change. Here, I explore a variety of evolutionary forces whose presence can be statistically tested in time series from the Penn-York Computer-annotated Corpus of a Large amount of English based on the TCP (PYCCLE-TCP) [213]. I find that only selection affecting negative declarative sentences remains significant under a variety of temporal binnings of the data. Upon further inspection, significant changes in selection are found to be affecting the trajectory of interrogative sentences. While this does not necessarily rule out that other evolutionary forces are at play, higher-quality data would be needed to achieve the necessary inferential power to detect them.

5.3.1 Relative pronouns in Middle and Modern English

As previously established, the grammatical structures of human languages involve many-to-many associations between linguemes and variants, where one lingueme may be fulfilled by several variants, and one variant may fulfil several linguemes. Here, I focus on a family of functional vocabulary known as *relative pronouns* or *relativisers*, where these patterns of many-to-many association are particularly prevalent [214].

Relativisers are words or syntactic devices used to introduce relative clauses. For example, consider the following sentences:

(1) a. Rob and Dan are the linguists who we are collaborating with.
b. Rob and Dan are the linguists that we are collaborating with.
c. Rob and Dan are the linguists we are collaborating with.

In (1-a), the relativiser *who* introduces the relative clause *who we are collaborating with*, which modifies the noun phrase *the linguists*, usually known as the antecedent of the relative clause. Notably, *who* could be replaced with *that*, as in (1-b), or omitted altogether, as in (1-c), and the sentence would remain grammatical in most English varieties. Note that in linguistics, the term *grammatical* is not used in a prescriptive sense (i.e. how language "should" be used), but rather in a descriptive sense (i.e. use of language that is perceived as intelligible and natural by speakers). Compare that to the following:

(2) a. Rob and Dan, who we are collaborating with, are based in the Central

Campus.

b. *Rob and Dan, that we are collaborating with, are based in the Central Campus.

c. *Rob and Dan, we are collaborating with, are based in the Central Campus.

Here, *who* works as a relativiser introducing the relative clause *who we are collaborating with* that modifies the antecedent *Rob and Dan*, while *that* and the omission of the relativiser would be ungrammatical, which is marked in the example sentences by the preceding asterisks (*). These examples illustrate that relative clauses can fulfill different pragmatic functions with respect to the noun phrase they modify. In (1), the relative clause is helping identify who Rob and Dan are, while in (2), it is simply providing additional information about them. In the terms of the present chapter, these different types of relative clause are different linguemes in this grammatical system, while the relativisers used to introduce them are their associated variants. We can thus see the complex ways in which the use of variants is patterned across different linguemes here, with some variants being able fulfill several linguemes whereas others are limited to a smaller range. This makes this a suitable case study for testing whether explicitly breaking the isolation assumption through lingueme migration (eq. 5.20) may be necessary to explain the behaviour of grammatical systems when many-to-many associations of linguemes and variants are present.

In order to do so, we must first identify an appropriate set of linguemes and variants. A variety of types of relative clauses can be identified in English depending on the function that the modified noun fulfills in the clause, and whether the antecedent is human or non-human, amongst other variables. Furthermore, a variety of strategies for relative clause formation beyond the use of relativisers exist in English, including reduced and infinitival relatives [215]. Here, for simplicity, I focus on the evolutionary dynamics of three broad types of relative clauses, and five types of relativisers.

The three types of relative clauses (linguemes) are ordinary, free, and clause-adjoined relative clauses. These may differ in their pragmatics – notably, the pragmatic differences between (1) and (2) are conveyed through ordinary and clause-adjoined relatives, respectively. However, they are more easily distinguished in terms of their syntactic features.

Ordinary relative clauses are those that modify a noun phrase. As such, they can

often be replaced with an adjective while preserving the meaning of the sentence, as exemplified by these two sentences:

(3) a. The restaurant that everyone's talking about is not that good.
b. The popular restaurant is not that good.

Free relative clauses are those that replace a noun phrase, rather than modifying it. Thus, they can often be replaced with a noun or noun phrase. As an example, consider:

(4) a. I was shocked by what he said.
b. I was shocked by his words.

Finally, *clause-adjoined relatives* are those that are adjacent to the main clause, rather than being subordinate to it by forming a constituent with their antecedent. They are usually identifiable in English spelling by being delimited by commas. The most natural way to replace them is with a separate sentence. For example:

(5) a. The Alhambra, which receives 3 million visitors every year, is in the city of Granada.
b. The Alhambra is in the city of Granada. It receives 3 million visitors every year.

For each one of these three linguemes, I consider five possible variants. They are: no relativiser, *that*, *which*, *what*, and other HW-words such as *where* or *how*, usually referred to in this way because of their Old English spelling. Time series of the relative usage of each one of these variants for each one of the three linguemes under consideration were extracted for the period between the years 1300 and 1900, comprising much of Middle English and Modern English, from the Penn Parsed Corpora of Historical English (PPCHE, [207]) and the Parsed Corpus of Middle English Poetry (PCMEP, [208]). The resulting time series with temporal binning of 50 years are presented in figure 5.2. In it, each lingueme under consideration is presented in an independent plot. The relative use of each of the variants for each of the linguemes is represented as the relative proportion of their associated colours. A variety of variants are used for each of the linguemes throughout the considered time periods. Several features of the data stand out.

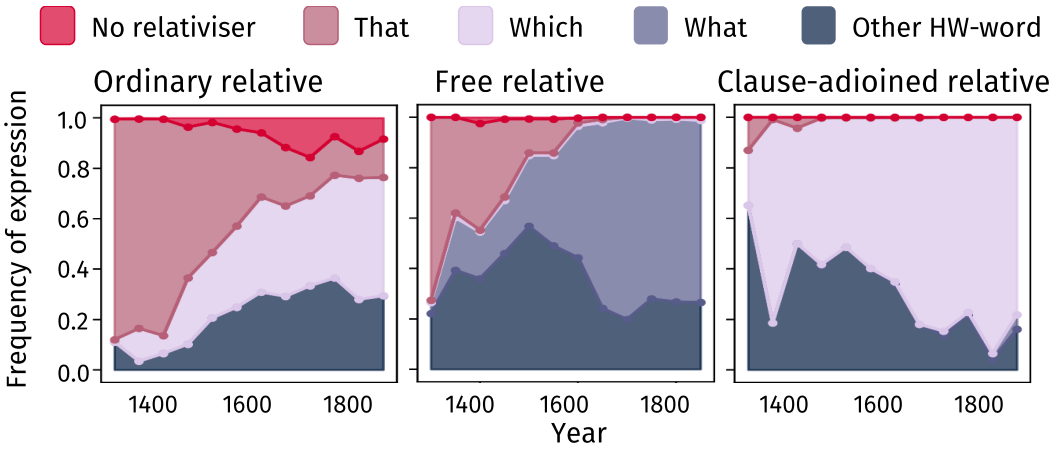

Figure 5.2 *Time series of the evolution of the relative frequency of usage of variants for each of the linguemes in the relativisers data set with temporal binning of 50 years. The linguemes under consideration are ordinary relatives (left), free relatives (centre), and clause-adjoined relatives (right). The variants under consideration are the use of no relativiser, that, which, what, and other HW-words, each codified using a different colour. The relative proportion of a colour in the plot for any given year represents the relative frequency of usage of its associated variant.*

First, *that* has seen a steady decrease in usage in all linguemes, with it going from being almost categorically used in ordinary relatives in 1300 to only marginally used in 1900. It furthermore dropped to zero usage in free and clause-adjoined relatives. HW-words other than *which* and *what* are reasonably present in all three linguemes, while *which* and *what* show a complementary distribution, with the former being present in ordinary and clause-adjoined relatives, while the latter is only found heading free relatives. Both see a steady increase through the considered time period, at the expense of *that*, which is only represented in ordinary relatives at the end of the trajectory. The use of no relativiser appeared initially in the 1400s as an option for ordinary and free relatives, but only stayed relevant for the former.

A variety of evolutionary forces could be tested for here. However, I am primarily interested in testing whether a model including lingueme migration is necessary to explain the behaviour of the data. In order to do so confidently, it will not suffice to check whether such a model is significant, but it is also necessary to compare its fit to those provided by other evolutionary models. In particular, the effects of variation and lingueme migration may be confounded in noisy data, potentially leading to them not being distinguishable. To maximise the inferential power of the method, and to account for variability under different binning strategies as

discussed in Section 5.2, I analyse the data under five different temporal binnings of 20, 30, 40, 50 and 60 years. For each model, p-values against a null model are obtained by comparing the test statistic to a distribution constructed using 500 artificially-generated time series evolving under the null model.

Results of the application of the model with different parametrisations of evolutionary forces are shown in Table 5.2. Both variation and lingueme migration produce significant p-values when tested against a null hypothesis of pure drift. This allows us to test for the significance of a model with both evolutionary forces by comparing its fit to that of the model with the highest likelihood, as described in Section 5.2. This also leads to an overall significant p-value, as well as significant p-values for all temporal binnings, suggesting that both variation and lingueme migration are necessary to explain the behaviour of the data. Variation, in particular, is a feasible explanation for the appearance of no relativiser as a possible variant in this data set from an initial state where it was not used to express any of the linguemes. The efficiency it provides by effectively expressing a lingueme using one fewer word may be the reason behind its innovation, as an instance of the economy principle in action [71]. Lingueme migration, on the other hand, may be a likely explanation of the introduction and increase of use of *which* as a relativiser in ordinary relatives, possibly resulting from influence from clause-adjoined relatives. For completeness, the results of the fits of further models with lingueme-independent selection affecting each of the variants are also included in Table 5.2. Selection produces overall significant p-values for the no relativiser and that expressions, and may be a plausible explanation for the increase in usage of no relativiser, as well as the decrease of that throughout the considered time period. Unlike the model with both types of variation, these selection models do not produce significant fits for all individual temporal binning.

In all, these results show that processes interlinking the evolution of distinct but related grammatical linguemes, grouped here under the label of *lingueme migration*, may be relevant to the evolutionary dynamics of grammatical systems. Not only this, but they may be significant and produce effects that are detectable even in the presence of variation of variants. As the main evolutionary force breaking the isolation assumption, this lends credibility to our efforts to create models with this feature. How exactly to interpret these effects in terms of causal linguistic phenomena, however, requires a more in-depth qualitative analysis in collaboration with historical linguists.

Model	N	**Parameters**($\times 10^{-3}$)	p**-value**
Pure drift	7000 ± 4000	N/A	N/A
Variations	4600 ± 2700	$\epsilon = 1.9 \pm 1.3$	< 0.005
Lingueme migration	4600 ± 2700	$\eta = 5 \pm 3$	< 0.005
Both	4400 ± 2600	$\epsilon = 2.4 \pm 1.9$ $\eta = 1.2 \pm 0.7$	0.01 ± 0.01
Selection	7000 ± 4000	$s_{\emptyset} = 8 \pm 5$	0.03 ± 0.2
	7000 ± 4000	$s_{\text{that}} = -0.5 \pm 0.6$	0.01 ± 0.03
	7000 ± 4000	$s_{\text{which}} = 2.4 \pm 1.2$	0.3 ± 0.3
	8000 ± 6000	$s_{\text{what}} = 7.2 \pm 1.1$	0.10 ± 0.09
	6000 ± 4000	$s_{\text{HW}} = -3 \pm 7$	0.45 ± 0.17

Table 5.2 *Results of the application of the multi-lingueme model with different parametric evolutionary forces to the relativisers data. Estimations of the evolutionary parameters shown are the average of the results obtained for different binning strategies, normalised following the procedure laid out in Section 2.1.5. Errors are obtained as the standard deviation of the same sample of results. All models with one or both of variation and lingueme mutation produce significant p-values. Only the models with selection affecting the no relativiser and that expressions present overall significant p-values, but they do not produce p-values for all binning strategies independently.*

5.3.2 The rise of the periphrastic do

I now proceed to analyse one of the most thoroughly studied instances of grammatical change in English historical linguistics: the rise of do-support. Do-support refers to the use of the verb *do* as an auxiliary, commonly found in sentences such as:

(6) a. Believe me, I do like it.
b. I don't think they're coming in the end.
c. How do spiders make webs?

In affirmative sentences such as (6-a), *do* is optionally used for emphasis. Conversely, in negative declarative sentences such as (6-b), and interrogative sentences such as (6-c), do-support is compulsory in most modern English varieties, with the verb *do* serving a purely grammatical function and the main verb providing the lexical meaning. This was not always the case: for much of English history, negative declarative sentences were formed by simply adding *not* after the main verb, whereas questions would use inversion of the main verb and

the subject. Thus, one would have:

(7) a. Ne I know not þe, knyȝt, by cort ne þi name. [216, page 12]
I know thee not, Sir Knight, thy court, nor yet thy name.
b. Why cried'st thou? Who hath thee done offence? [217, line 1083]

The first in-depth quantitative exploration of the diachronic processes leading to the rise of do-support was carried out in Alvar Ellegård's influential 1955 work *The auxiliary do: the establishment and regulation of its use in English* [209]. In it, Ellegård noticed that do-support had increased in usage at different times in different syntactic contexts, with negative interrogative sentences being the first ones to show a significant proportion of it. He also highlighted that affirmative declarative sentences seemed to provide an example of failed change: do-support increased to as much as 10% in the 1500s in this type of sentence, only to again drop in usage and be relegated to the emphatic contexts in which it is utilised in present-day English. Sociolinguistic factors [211] and several steps of semantic change [212] have been put forward as explanations of these phenomena. These accounts are generally based on sophisticated qualitative analyses and regression models, which do not entertain the effects of drift inherent to language change.

Here, I aim at performing a preliminary exploration of the rise of do-support using the Wright-Fisher paradigm for grammar change. First, as always, a set of linguemes and variants needs to be determined. The two possible variants here are easily delimited: they are the use of do-support, or lack thereof. Choosing a set of linguemes may be more difficult – Ellegård noted that historical do-support usage depends on several variables, such as the argument structure of the sentence and the semantic class of the main verb. However, Kroch (1989) [210] noted that, in spite of these variables, the rate of change was fully uniform before 1575 and only dependent on sentence type thereafter. Thus, following Kroch's assertion, three types of sentences can be identified as our linguemes in this preliminary study: affirmative declarative sentences such as (6-a), negative declarative sentences such as (6-b), and interrogative sentences such as (6-c).

Time series of the relative usage of do-support for each of these three linguemes were extracted for the period between the years 1500 and 1800 from the Penn-York Computer-annotated Corpus of a Large amount of English based on the TCP (PYCCLE-TCP) [213]. The resulting time series with temporal binning of 15 years are presented in figure 5.3. For each of the linguemes, the proportion

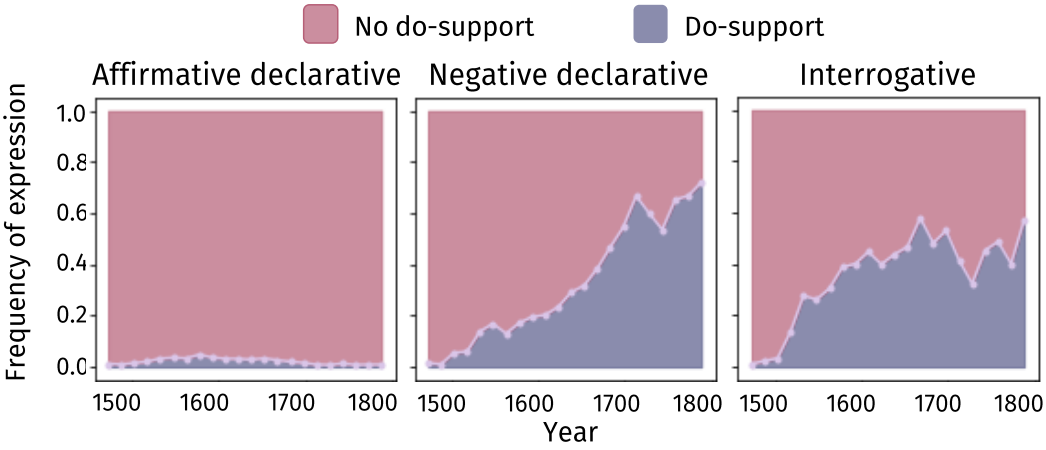

Figure 5.3 *Time series of the evolution of the relative frequency of usage of variants for each of the linguemes in the data set pertaining to the rise of do support, with a temporal binning of 15 years. The linguemes under consideration are affirmative declarative sentences, negative declarative sentences, and interrogative sentences. The use or absence of do support define the two variants under consideration.*

of usage of do-support is presented as the proportion of the plot shaded blue. The slight rise in use in affirmative declarative sentences remarked by Ellegård is visible here. Do-support in negative declarative sentences steadily increases in usage during the considered time period. In the case of interrogative sentences, this increase is only indisputable in the 16th centure, after which its trajectory seems to be dominated by unbiased drift. It is difficult to tell at this stage whether this drift is a true feature of the underlying dynamics of the data, or a consequence of sampling noise.

Based on these observations, I test for a variety of evolutionary forces in this data, including variation, lingueme migration, analogy, and lingueme-dependent selection. Again, to maximise inferential power, I do so for temporal binnings of 10, 15, 20, 25 and 30 years, and report the average rescaled maximum-likelihood parameters and p-values. Results are shown in Table 5.3. Out of all evolutionary forces considered, only lingueme-independent selection acting on negative sentences produces significant p-values for all temporal binnings, and has an average p-value under the threshold.

Lingueme-dependent selection acting on negative sentences is able to capture the steady increase of do-support throughout the considered time span. However, its increase in interrogative sentences fails to be captured by the model in this data, whether it be through selection or through analogy to the trajectory of negative declarative sentences. This may be a consequence of sampling noise decreasing

Model	N	**Parameters**($\times 10^{-3}$)	p**-value**
Pure drift	5100 ± 500	N/A	N/A
Variation	5000 ± 500	$\epsilon = 0.37 \pm 0.06$	0.08 ± 0.10
Lingueme migration	5200 ± 600	$\eta = -0.13 \pm 0.15$	0.26 ± 0.19
Analogy	5200 ± 700	$\gamma_{+-} = -0.18 \pm 0.15$	0.24 ± 0.15
	5200 ± 500	$\gamma_{-?} = 32 \pm 19$	0.24 ± 0.10
	5100 ± 500	$\gamma_{?+} = 2.0 \pm 1.7$	0.28 ± 0.16
Lingueme-dependent selection	5100 ± 500	$s_+ = 2.4 \pm 0.9$	0.24 ± 0.14
	7100 ± 1600	$s_- = 12.6 \pm 0.3$	0.01 ± 0.03
	5200 ± 500	$s_? = 1.5 \pm 1.5$	0.28 ± 0.15

Table 5.3 *Results of the application of the multi-lingueme model with different parametric evolutionary forces to the do-support data. Estimations of the evolutionary parameters shown are the average of the results obtained for different binning strategies, normalised following the procedure laid out in Section 2.1.5. Errors are obtained as the standard deviation of the same sample of results. One p-values. Only the model with lingueme-dependent selection affecting negative declaratives produces a significant p-value.*

the statistical significance of results, a ubiquitous challenge when working with historical data. It may, however, be a consequence of changing underlying dynamics not being adequately captured by a model with constant parameters. In Table 5.4, I report results obtained by fitting the time-dependent model developed in Chapter 3 to each of the time series. Doing so independently for each of them is consistent with results obtained thus far, as lingueme-dependent selection does not break the isolation condition. The model with changing parameters produces a significant fit to the time series of interrogative data, with a detected change in selection strength in 1690. Varying social dynamics, including changes in social status and differential usage across age groups, have been discussed as a possible driving factor in the rise of do-support in negative sentences throughout the 1500s [211], and it is thus possible that similar changes in social attitudes taking place in the late 1600s are being detected here. Affirmative declarative sentences are thus the only lingueme which remains well described by drift in inter-generational transmission. Higher quality data would be necessary to increase the inferential power of the model in order to test more nuanced hypotheses.

Function	**Time division**	s **before**	s after	p**-value**
Affirmative	1540	0.034	-0.018	0.232
Negative	1630	0.028	0.004	0.14
Interrogative	1690	0.0013	0.0015	0.034

Table 5.4 *Results of the application of the change-detection algorithm to each lingueme in the do-support data set, with temporal binning of 5 years. Significant change in selection is detected in 1690 in interrogative sentences.*

5.4 Discussion

In this chapter, by following steps similar to those developed by Reali and Griffiths [4], I was able to prove the equivalence of the IBL paradigm with nontrivial communication developed in Chapter 4 to a Wright-Fisher-like model of language change. Being able to account for any number of linguemes expressed by any number of variants, this model breaks the stability assumption typical of evolutionary models of lexical change, whereby expressions are assumed to only represent one immovable linguemes. Being able to further introduce evolutionary forces of lingueme migration – akin to migration between subpopulations in biological settings – it further breaks the isolation assumption, which posits that different linguemes have no effect on the evolutionary dynamics of one another. This allows it to model time series of grammatical systems typically out of reach of simpler models of competition dynamics. Crucially, many methodological advancements introduced in previous chapters remain relevant to the application of the new paradigm.

In Section 5.3, the model was applied to two data sets of historical change in English grammar. The first one was concerned with the evolutionary dynamics of relative pronouns in Middle and Modern English, which present many-to-many associations between linguemes and variants. Most notably, the newly introduced evolutionary force of lingueme migration was shown to be significant even in the presence of variation, demonstrating its potential in modelling competition dynamics in functional vocabulary. The second data set involved the rise of do-support in Early Modern English, a highly-discussed instance of diachronic change involving a wide variety of theorised social and structural factors [209–212]. The statistical testing of these factors, however, has so far ignored the stochastic nature of language change. The preliminary study presented here has found significant selective forces acting on negative declarative sentences,

as well as changing selective forces on interrogative sentences. Affirmative declarative sentences, which present a puzzling instance of failed change, seem to be adequately described by models of pure drift.

A wide variety of diachronic phenomena may be at play in the rise of do-support in particular, and processes of grammatical change in general. However, stochastic models such as the one presented here are limited by the quality of the data they analyse. This challenge is further compounded when classifying utterances not only in terms of their variants, but also the linguemes that they express. Not only does this further limit the number of tokens per time point and trajectory, but it also hinders their automated extraction, as the instances of a grammatical lingueme that can be detected algorithmically are only a subset of all occurrences of said lingueme. As the quality of parsed corpora increases, potentially through the use of AI-assisted annotation, evolutionary models of grammar change can be expected to reach their full potential in assisting historical linguists in the formulation and testing of hypotheses.

Chapter 6

Conclusion

At the start of this thesis, I discussed how language, like many other biological, social, and cultural systems, can be understood as a complex adaptive system, both at the social and at the cognitive levels [1, 44]. As such, many properties and phenomena of language emerge at the collective level, in ways that cannot be predicted from the individual behaviour of speakers. Language change is the collection of adaptive mechanisms of the language system, the evolutionary process connecting the cognitive processes and ever-changing social interactions of individuals to the emergent structure and conventionalisation of language. Understanding language change is thus not only interesting from the perspective of historical linguistics, but also language evolution, language universals, and the emergence of structure from human systems as a whole. The contribution of this thesis stems from the application of insight from Statistical Mechanics to the modelling and empirical study of language change.

Digital corpora provide us with a wealth of time-organised language use data containing key information on the nature of language change. Crucially, language change can be understood as an evolutionary process equivalent to those in population genetics [64], where the role of genes is played by *linguemes* (meanings, grammatical relations or phonemes) and that of alleles is played by *variants* (words, structures or sounds). This formulation allows us to inherit empirically-applicable models that codify a wide variety of dynamics as population-level evolutionary forces such as stochastic drift, selection (social of linguistic effects biasing dynamics in favour or against specific variants), and variation (innovation effects giving rise to new variants). Among all of them, the Wright-Fisher model

is most relevant to the evolutionary process in language change, as it can be proved equivalent to a variety of models of cultural transmission of language [4, 5]. However, the Wright-Fisher paradigm presents a series of issues that hinder its empirical applicability to historical language data, and oversimplifies some features of language change in ways that limit the range of diachronic phenomena that it can accurately model. Much of this thesis has been dedicated to describing these limitations and oversimplifications, proposing improvements that address these issues by building on well-established Wright-Fisher methodologies, and demonstrating their empirical applicability through analyses of historical corpus data sets.

Chapter 2 was devoted to the development of methodological advancements addressing the numerical instability of usual approximations to the Wright-Fisher process, as well as the uncertainty arising from the inaccessibility of the Wright-Fisher generation time in cultural and linguistic contexts. The former issue was addressed by developing a self-contained parameter estimation scheme based on the Beta-with-Spikes [96] approximation to the Wright-Fisher transition probability. The latter required the application of Principal Component Analysis [125] to find confidence regions for p-values and estimated model parameters over a range of generation times. These methodologies were successfully applied to the detection of a phonologically-conditioned bias in the competition dynamics between past tense forms of English verbs.

In Chapter 3, I addressed the oversimplification whereby model parameters remain constant throughout evolutionary trajectories as modelled in the Wright-Fisher paradigm. This oversimplification fails to capture the changes in selection strength arising from changing social attitudes towards specific linguistic variants. I therefore proposed a minimal model with time-dependent parameters, where evolutionary forces change abruptly at a specified transition time but remain constant otherwise. Maximum-likelihood estimation methods can then be applied to find transition times reflecting shifts in social attitudes in historical data. Accurate estimations were obtained when benchmarking this model using data on historical spelling reforms in Spanish.

Chapters 4 and 5 addressed the assumptions of stability and isolation, which limit the evolutionary models presented thus far to the competition dynamics of individual linguemes that are assumed to be unaffected by the dynamics of other linguemes, no matter how closely related. I proposed a model of cultural transmission of a language with arbitrarily-associating linguemes and

variants, based on the iterated Bayesian learning paradigm [85]. This model introduced nontrivial communicative effects that were able to interrelate the evolutionary dynamics of distinct linguemes. I was further able to demonstrate the equivalence of this model of cultural transmission to a Wright-Fisher-like model of competition. This model incorporated a new evolutionary force akin to migration between linguemes, which subsumed a variety of communicative and semantic effects, as well as analogy. The model was further applied to two data sets of grammatical change in English, demonstrating the significance of this newly formulated evolutionary force in explaining the dynamics of historical data.

Also in Chapter 4, in a detour from the thesis's overarching focus on empirical analysis, the consequences of nontrivial communication on the stationary properties of language were examined. In particular, directional behaviour in this and other iterated Bayesian learning paradigms was quantified and compared using entropy production. This preliminary analysis shows the potential of this concept from non-equilibrium statistical mechanics in exploring directed change in linguistic contexts. The development of techniques for the detection and quantification of directional behaviour in empirical data, both from corpora and from typological data bases, may allow us to formulate and test hypotheses on the directionality of language change and deepen our understanding of its underlying mechanisms.

In all, this thesis demonstrates the potential of the Wright-Fisher paradigm in the exploration of empirical questions in linguistics. There are some limitations, however, to the methodological advances introduced here. Even after the relaxation of the stability and isolation assumptions, there are still some linguistic phenomena that cannot be captured by Wright-Fisher models. Chiefly, the model still characterises linguemes and variants as categorical units. In phonology, however, variants (sounds) are determined by continuous physical properties such as the formant frequencies of vowels [192, 218]. When modelling semantic change in processes such as lexicalisation and grammaticalisation, changes in linguemes (meaning) are often gradual [15]. Advances in quantitative semantics allow for the modelling of meaning as vectors in a multi-dimensional space [219]. Thus, future developments in evolutionary models should aim at capturing both linguemes and variants that exist in a continuum, rather than discrete immovable categories.

At the methodological level, all analyses in this thesis have been carried out within the framework of frequentist statistics. This does not denote a personal

preference, but rather the continuation of a trend in the field of evolutionary modelling, where previous efforts have aimed at quantifying significance levels through p-value thresholds [96, 104, 105, 124]. Future developments may benefit from further validating results through the application of Bayesian statistics, allowing for the introduction of prior biases in model fitting, and more nuanced characterisations of uncertainty in parameter estimations [220].

Another limitation of evolutionary models stems from the quality of the data available to them. Historically, there has been a trade-off between the quantity of data available in a corpus, and the quality of its annotation, as this usually relies on the work of human annotators. Empirical studies such as those carried out using the multi-lingueme Wright-Fisher paradigm presented in Chapter 5 may often rely on detailed corpus annotation to identify features of interest of utterances. This thus limits the inferential power of models by having them depend on smaller corpora subject to higher sampling noise. Automated annotation relying on artificial intelligence may prove to be key towards eliminating this quantity-quality trade-off in historical data.

Evolutionary forces, and how they map onto different linguistic and cultural effects, have been a central topic throughout this thesis. Selection, variation, and the newly introduced lingueme migration all encompass a wide variety of phenomena, including social preferences, cognitive biases, analogy, semantic effects, and innovation. In the case that a phenomenon is not covered by these forces, this paradigm can be extended to include evolutionary forces that account for it. For example, frequency-dependent selection is able to model conformity and anti-conformity biases that are not adequately subsumed by the form of selection presented in this thesis [221]. However, the generality of these evolutionary forces also presents some shortcomings: selection can be caused by such a wide range of diachronic phenomena that it is impossible to know which one is actually responsible for the significant selective forces driving the evolutionary dynamics of a given data set. This issue arises from the fact that the Wright-Fisher paradigm operates at the macroscopic level of the language system, and thus is only able to model emergent phenomena arising from a variety of microscopic social, cognitive and spatial effects with similar population-level effects. In order to narrow down the causal mechanisms of an empirical trajectory of change, the quantitative results of evolutionary models need to be complemented with the sophisticated qualitative insight of historical linguists. However, in order to more generally understand how microscopic behaviours beget

evolutionary forces at the macroscopic level, more complex stochastic models are necessary. The effects of social complexity have been explored by Utterance Selection models [5, 92], demonstrating that a richer variety of diachronic phenomena may arise when complex social networks are accounted for. Models implementing spatial complexity have been able to demonstrate the effects of this variable on language diversification and dialect formation [222, 223], and even questioned the validity of neutral models as drivers of change [224]. Cognitive forces like biases for efficiency and expressivity have been shown to drastically impact language structure [48], and may be mathematically formulated using information theory [225]. Future developments in the evolutionary modelling of language should aim at integrating these social, spatial, and cognitive accounts towards a holistic characterisation of language change and how it shapes the properties of human languages.

Appendix A

Detailed results of the analysis of English verbs using the self-contained Beta-with-Spikes method

A.1 Maximum likelihood parameters for the COHA verbs

In the following tables I quote the maximum-likelihood estimates of the parameters in the Wright-Fisher model obtained by applying the Beta-with-Spikes method outlined in the main text to frequency counts derived from the COHA corpus. Each table corresponds to a different binning strategy: for example, in the first table, frequency counts from each period of 10 consecutive years are aggregated to form a single frequency estimate for the corresponding time period.

Two different effective population sizes N are quoted: one ('for drift') under the assumption that $s = 0$, and the other ('for selection') that is obtained when both N and s are optimised via the maximum likelihood analysis. The p-value is the empirical p-value for the drift hypothesis, obtained as described in Section 2.1.4 of the main text. The maximum likelihood values are all quoted to three significant figures, and the p-values to two significant figures.

10-year bins

Verb	N for drift	N for selection	s	p-value
awake	1820	1990	0.021	0.11
build	2820	3130	0.026	0
burn	410	542	-0.05	0
catch	3690	4170	0.031	0.028
dive	145	147	0.0092	0.43
draw	2880	2880	0.00067	0.96
dream	219	233	-0.036	0.002
dwell	568	554	-0.029	0.002
grow	4510	4770	0.031	0.032
hang	1520	1830	0.048	0
hear	6160	6920	0.047	0.044
heave	147	145	0.0069	0.68
kneel	360	362	0.0028	0.82
knit	121	122	-0.0043	0.77
know	4240	4350	0.0035	0.65
lay	4070	4250	0.002	0.47
lean	753	926	-0.053	0
leap	313	346	0.022	0.16
learn	652	847	-0.052	0
light	294	368	0.021	0.03
plead	1500	1590	-0.012	0.19
quit	790	927	0.022	0
shine	1330	1320	-0.014	0.17
smell	319	399	-0.036	0.004
sneak	415	453	0.023	0.068
speed	299	327	0.024	0.052
spell	288	307	-0.017	0.31
spill	304	357	-0.032	0
spoil	223	226	-0.0036	0.79
strew	127	127	-0.0023	0.86
tell	3760	3960	0.018	0.37
throw	2090	2180	0.012	0.25
wake	815	928	0.015	0.012
weave	460	457	-0.007	0.45
wed	116	120	0.026	0.03
wet	201	201	0.0024	0.88

20-year bins

Verb	N for drift	N for selection	s	p-value
awake	5130	5730	0.011	0.13
build	4850	5290	0.0064	0.27
burn	615	909	-0.042	0
catch	9890	12000	0.02	0.058
dive	483	546	0.0095	0.24
draw	6690	6660	0.0028	0.84
dream	296	326	-0.034	0.004
dwell	1080	1430	-0.04	0
grow	17600	17600	-4.2e-05	1.0
hang	4260	4310	0.0052	0.58
hear	16400	17500	0.013	0.39
heave	353	350	0.0018	0.85
kneel	1580	1640	0.0018	0.76
knit	244	251	-0.0049	0.66
know	8890	8900	4.1e-05	1.0
lay	9110	10200	0.0016	0.43
lean	1060	1990	-0.063	0
leap	498	616	0.019	0.1
learn	937	1910	-0.054	0
light	360	834	0.022	0.006
plead	2100	2110	-0.00072	0.94
quit	1140	1690	0.025	0
shine	2500	2740	-0.018	0.066
smell	500	1440	-0.036	0
sneak	674	871	0.024	0.02
speed	807	941	0.014	0.14
spell	498	1310	-0.042	0
spill	688	1220	-0.025	0
spoil	565	747	-0.028	0.018
strew	334	334	0.0012	0.87
tell	7620	7570	0.012	0.62
throw	9380	9420	0.00046	0.94
wake	1160	1650	0.015	0.014
weave	1020	1000	-0.011	0.14
wed	147	216	0.028	0.1
wet	1100	1140	0.0021	0.78

40-year bins

Verb	N for drift	N for selection	s	p-value
awake	9510	20300	0.016	0.042
build	5780	8090	0.0098	0.16
burn	1770	15100	-0.017	0.014
catch	15000	40800	0.027	0.002
dive	697	879	0.0072	0.49
draw	30800	48600	0.0057	0.52
dream	824	868	-0.0075	0.52
dwell	1430	2060	-0.037	0
grow	31200	31100	-0.0012	0.95
hang	146000	148000	-0.00099	0.68
hear	20100	20300	0.0073	0.68
heave	743	835	-0.0046	0.66
kneel	3090	6490	0.0062	0.25
knit	252	261	-0.0027	0.83
know	8310	8720	0.0025	0.75
lay	9020	13400	0.0021	0.35
lean	2810	5530	-0.025	0.008
leap	775	968	0.015	0.14
learn	2680	9430	-0.023	0.006
light	318	1110	0.019	0.03
plead	3770	3690	0.0047	0.62
quit	659	1610	0.026	0
shine	3910	8610	-0.019	0.01
smell	477	2600	-0.034	0
sneak	924	920	0.014	0.24
speed	1050	1040	-0.00025	1.0
spell	509	2180	-0.039	0
spill	712	2060	-0.026	0
spoil	1350	2490	-0.014	0.086
strew	569	623	0.0043	0.65
tell	11000	11700	0.025	0.46
throw	6510	6630	-0.0017	0.93
wake	908	2810	0.019	0.014
weave	1160	1690	-0.017	0.052
wed	248	364	0.016	0.14
wet	2570	3220	0.0037	0.13

A.2 Maximum likelihood parameters for verbs in the study of competing motivations

In this section, we provide the corresponding tables for the set of verbs ending in alveolar stops from drawn from the Google Books corpus. Dashes mean that the corresponding time series did not have enough data points per time bin in the corresponding binning for it to be included in the study.

5-year bins

Verb	N for drift	N for selection	s	p-value
bend	7450	9980	0.019	0.004
bet	-	-	-	-
bite	4600	6960	0.029	0
blend	5110	5040	0.0037	0.56
build	21000	22200	0.0058	0.28
fit	586	596	0.026	0.036
glide	6520	6520	-0.0026	0.84
knit	-	-	-	-
light	823	908	0.011	0.036
pat	1420	1420	-0.02	0.12
plead	5040	5160	0.01	0.054
quit	376	377	0.076	0
slide	2050	2210	0.023	0.002
speed	628	626	-0.003	0.61
spit	757	847	0.014	0.064
thrust	2510	2880	0.074	0.002
tread	2980	3000	-0.014	0.25
wed	93.6	83.6	0.079	0
wet	-	-	-	-

Verb	N for drift	N for selection	s	p-value
awake	1390	2230	0.025	0
blow	11100	11200	0.0012	0.8
burn	1110	1350	-0.012	0
catch	5070	6640	0.039	0
cleave	442	447	-0.0042	0.54
creep	3820	4190	0.042	0.014
dive	757	766	0.011	0.088
dream	1830	1900	-0.0041	0.36
dwell	1300	1300	0	1.0
freeze	2290	2510	0.045	0.008
grow	95400	95300	-0.0012	0.5
hang	2070	2350	0.0077	0.052
heave	507	516	-0.0042	0.58
hew	1320	1320	-0.00037	0.93
kneel	986	1390	0.02	0.002
lean	1350	1380	-0.0037	0.47
leap	1460	1500	0.0056	0.22
learn	2250	2360	-0.0047	0.3
shake	3910	4520	0.065	0
shear	545	546	-0.0073	0.26
shine	1380	1380	0.00025	0.98
slay	8430	8430	-0.00048	0.98
slink	1310	1320	-0.011	0.28
smell	710	799	-0.013	0.018
sneak	2050	3130	0.055	0
spell	542	584	-0.011	0.1
spill	863	1260	-0.02	0.002
spoil	1370	1380	0.0046	0.26
strew	646	1010	0.028	0
string	2180	2650	0.019	0.038
strive	3520	3830	-0.017	0.026
swell	1070	1240	0.011	0.01
wake	775	1290	0.025	0
weave	994	988	-0.0086	0.16

10-year bins

Verb	N for drift	N for selection	s	p-value
bend	12400	28100	0.017	0
bet	491	685	0.039	0
bite	5330	16600	0.03	0
blend	8290	8170	0.0024	0.68
build	23900	27000	0.0057	0.23
fit	1340	1750	0.03	0
glide	26300	25900	0.0014	0.85
knit	-	-	-	-
light	840	1120	0.012	0.018
pat	2680	3090	-0.022	0.028
plead	6390	7030	0.011	0.028
quit	522	670	0.052	0
slide	3240	3240	0.018	0.012
speed	444	445	-0.0024	0.75
spit	1210	1590	0.013	0.054
thrust	3760	4630	0.051	0.006
tread	7190	8840	-0.023	0
wed	159	163	0.038	0.038
wet	-	-	-	-

Verb	N for drift	N for selection	s	p-value
awake	1310	3710	0.025	0
blow	24500	24700	0.00079	0.81
burn	1260	2260	-0.013	0
catch	9740	28500	0.033	0
cleave	461	472	-0.0042	0.51
creep	12600	14200	0.017	0.066
dive	1050	1140	0.012	0.048
dream	2650	3100	-0.0053	0.16
dwell	1790	1970	-0.014	0.096
freeze	6130	7090	0.02	0.038
grow	184000	187000	-0.0018	0.19
hang	3010	4150	0.0071	0.04
heave	1130	1320	-0.007	0.14
hew	7420	7640	-0.0015	0.43
kneel	895	1560	0.019	0.008
lean	1350	1420	-0.0038	0.49
leap	1660	1830	0.0066	0.1
learn	3130	3660	-0.0054	0.17
shake	9620	11900	0.037	0.004
shear	983	1100	-0.0098	0.06
shine	3570	3550	-0.00087	0.9
slay	17200	20000	-0.015	0.21
slink	2120	2230	-0.015	0.12
smell	607	785	-0.014	0.022
sneak	2190	3740	0.055	0
spell	998	1630	-0.014	0.004
spill	789	1890	-0.02	0
spoil	1710	1730	0.0039	0.26
strew	453	1130	0.03	0
string	8410	12200	0.012	0.028
strive	2980	3440	-0.018	0.034
swell	844	1070	0.011	0.03
wake	631	1710	0.025	0
weave	2110	2280	-0.0093	0.044

20-year bins

Verb	N for drift	N for selection	s	p-value
bend	13800	126000	0.016	0
bet	677	913	0.02	0.056
bite	5670	18200	0.033	0
blend	5580	5580	0.0013	0.88
build	22600	28500	0.0064	0.26
fit	1390	2490	0.032	0
glide	102000	116000	-0.0035	0.58
knit	1460	1480	-0.0043	0.72
light	746	1850	0.014	0.008
pat	2660	6670	-0.039	0
plead	5860	7540	0.011	0.044
quit	590	1020	0.054	0
slide	5160	23200	0.026	0
speed	434	430	-0.0042	0.63
spit	1570	9000	0.017	0
thrust	4700	4250	-0.013	0.52
tread	8800	21600	-0.028	0
wed	310	353	0.037	0.012
wet	800	898	0.0051	0.6

Verb	N for drift	N for selection	s	p-value
awake	1150	3460	0.023	0.006
blow	26200	26700	0.00064	0.87
burn	1040	3790	-0.013	0.004
catch	12500	47200	0.03	0
cleave	374	399	-0.0074	0.41
creep	11200	13400	0.016	0.14
dive	1120	1550	0.012	0.046
dream	3030	5060	-0.0061	0.074
dwell	2570	4140	-0.02	0.01
freeze	14800	24300	0.013	0.086
grow	190000	190000	-0.0022	0.36
hang	3850	8570	0.0064	0.022
heave	1530	2210	-0.0064	0.12
hew	9390	10200	-0.0016	0.4
kneel	978	2630	0.015	0.02
lean	1840	2300	-0.0059	0.28
leap	3010	5730	0.007	0.02
learn	5100	8270	-0.0052	0.072
shake	22800	32600	0.021	0.026
shear	1550	2980	-0.011	0.008
shine	2570	2540	-0.002	0.78
slay	12100	162000	-0.026	0.01
slink	3280	5200	-0.027	0.004
smell	487	863	-0.015	0.018
sneak	2750	4340	0.056	0
spell	890	3350	-0.014	0.004
spill	624	7300	-0.02	0
spoil	1860	1910	0.0035	0.43
strew	295	1110	0.031	0.006
string	12100	20400	0.0086	0.1
strive	3730	4800	-0.021	0.008
swell	618	828	0.01	0.1
wake	476	1350	0.025	0
weave	2450	2850	-0.0081	0.12

40-year bins

Verb	N for drift	N for selection	s	p-value
bend	12900	196000	0.014	0.008
bet	748	3900	0.025	0.008
bite	7370	38000	0.034	0
blend	5020	5030	-0.00054	0.98
build	18600	26800	0.0082	0.27
fit	1880	2450	0.027	0
glide	70600	77800	-0.0024	0.74
knit	2120	10400	-0.017	0.13
light	366	2180	0.016	0.014
pat	3470	67700	-0.03	0
plead	6110	16700	0.0099	0.054
quit	966	1460	0.049	0
sit	200	200	0.17	0.99
slide	5730	36500	0.028	0
speed	487	443	-0.014	0.21
spit	1580	27000	0.017	0.004
thrust	17100	26400	0.054	0.008
tread	9620	26800	-0.033	0
wed	513	702	0.035	0.002
wet	1340	1620	0.0039	0.16

Verb	N for drift	N for selection	s	p-value
awake	1810	13500	0.017	0.022
blow	14500	15200	0.00087	0.9
burn	758	5850	-0.012	0.024
catch	18500	101000	0.023	0.004
cleave	353	408	-0.0077	0.52
creep	16100	16300	0.00068	0.98
dive	700	1130	0.012	0.12
dream	2780	5950	-0.0052	0.17
dwell	2670	3730	-0.022	0.018
freeze	20200	86700	0.014	0.058
grow	112000	110000	-0.0028	0.5
hang	3720	28400	0.0059	0.028
heave	1210	4750	-0.0083	0.076
hew	8230	11300	-0.0021	0.31
kneel	965	3050	0.012	0.16
lean	1640	3080	-0.0082	0.21
leap	2410	20700	0.0074	0.012
learn	4610	10100	-0.0046	0.16
shake	24700	69000	0.025	0.006
shear	727	8120	-0.013	0.01
shine	1920	1920	-0.00042	0.98
slay	16000	461000	-0.028	0.014
slink	3860	4260	-0.026	0.004
smell	290	599	-0.016	0.1
sneak	3790	4200	0.059	0.004
spell	497	2340	-0.014	0.044
spill	385	10600	-0.019	0
spoil	1260	1230	0.0034	0.58
strew	300	656	0.03	0.016
string	25700	59300	0.0053	0.22
strive	3990	4020	-0.022	0.022
swell	565	804	0.0081	0.36
wake	506	2030	0.022	0.008
weave	1350	1830	-0.0087	0.27

Appendix B

Sets of words used in analyses of change in Spanish Google Books data

Infinitive forms of verbs in set of unregulated change: Alternative ⟨-ra-⟩ and ⟨-se-⟩ forms of the past subjuntive

creer, *dar*, *deber*, *decir*, *dejar*, *encontrar*, *estar*, *hablar*, *hacer*, *llamar*, *llegar*, *llevar*, *parecer*, *pasar*, *pedir*, *poder*, *poner*, *quedar*, *querer*, *saber*, *seguir*, *tener*, *tomar*, *traer*, *venir*, *ver*, and *volver*.

Old spellings of words in set (A): ⟨ss⟩ to ⟨s⟩

assegurar, *assentar*, *assentir*, *assunto*, *confessar*, *diesse*, *essa*, *esse*, *essencia*, *esso*, *estuviesse*, *fuesse*, *gustasse*, *hiciesse*, *passar*, *pudiesse*, *quisiesse*, *tassar*, *tuviesse*, and *usasse*.

Old spellings of words in set (B): ⟨x⟩ to ⟨j⟩

abaxo, *baxar*, *baxo*, *bruxa*, *bruxería*, *caxa*, *conduxo*, *debaxo*, *dexar*, *dibuxar*, *dibuxo*, *dixo*, *enxuto*, *exe*, *exemplo*, *exercer*, *exercicio*, *exército*, *floxo*, *fluxo*, *fixar*, *fixo*, *quexa*, *roxo*, *texa*, and *traxo*.

Old spellings of words in set (C): ⟨y⟩ to i

aceyte, *aceytuna*, *afeyte*, *amaynar*, *ayre*, *bayle*, *deleyte*, *deydad*, *estoyco*, *frayle*, *gayta*, *heroyco*, *layco*, *oyga*, *peyne*, and *reyna*.

Old spellings of words in set (D.1): Word-final tonic syllable becoming accentuated

accion, *alacran*, *algun*, *almacen*, *atencion*, *bailarin*, *cancion*, *capitan*, *comun*, *corazon*, *estacion*, *jardin*, *latin*, *nacion*, *ningun*, *opcion razon*, *recien*, *region*, *relacion*, *segun*, *Serafin*, *situacion*, *tambien*, and *union*.

Old spellings of words in set (D.2): Non word-final tonic syllable losing its accent

abdómen, *álguien*, *Cármen*, *certámen*, *cólon*, *crímen*, *desórden*, *dictámen*, *exámen*, *gérmen*, *jóven*, *márgen*, *órden*, *orígen*, *resúmen*, and *volúmen*.

Old spellings of words in set (E): Single-vowel words losing their accent

á, *é*, *ó*, and *ú*.

Appendix C

Derivation of leading-order entropy production

This appendix expands on the details of the derivation of the leading-order entropy production presented in Section 4.4.1. The entropy production for an IBL model with transition probability distribution $\mathrm{P}\left(G'|G\right)$ and stationary distribution $\Pi\left(G\right)$ is given by:

$$\Sigma = \int \mathrm{P}\left(G'|G\right)\Pi\left(G\right)\ln\left(\frac{\mathrm{P}\left(G'|G\right)\Pi\left(G\right)}{\mathrm{P}\left(G|G'\right)\Pi\left(G'\right)}\right)\mathrm{d}G\mathrm{d}G' \,. \tag{C.1}$$

The key issue with the direct application of this equation is that the stationary distribution is often not easily computationally accessible. This motivates finding reasonably approximated, leading-order contributions.

As laid out in the main text, we are considering an IBL model with stationary distribution $\Pi_\epsilon\left(G\right)$ and transition probability distribution $\mathrm{P}_\epsilon\left(G' \,|\, G\right)$. These depend differentiably on a model parameter ϵ in such a way that at $\epsilon = 0$, the stationary distribution becomes the prior $\Pi_0\left(G\right)$ and the transition probability becomes $\mathrm{P}_0\left(G' \,|\, G\right)$, such that they satisfy detailed balance,

$$\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right) = \mathrm{P}_0\left(G|G'\right)\Pi_0\left(G'\right) . \tag{C.2}$$

As further introduced in Section 4.4.1, when $\epsilon \neq 0$, we have:

$$\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right) = \mathrm{P}_\epsilon\left(G \,|\, G'\right)\Pi_\epsilon\left(G'\right) + \sum_{n=1}^{\infty}\epsilon^n\Delta_n\left(G' \,|\, G\right) \,, \tag{C.3}$$

where

$$\Delta_n\left(G' \,|\, G\right) = \frac{\partial^n}{\partial \epsilon^n}\left[\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right) - \mathrm{P}_\epsilon\left(G \,|\, G'\right)\Pi_\epsilon\left(G'\right)\right]\Big|_{\epsilon=0}. \tag{C.4}$$

Notably, the anti-symmetric nature of these quantities implies that

$$\int \Delta_n\left(G' \,|\, G\right)\mathrm{d}G\mathrm{d}G' = 0 \qquad \forall n\,. \tag{C.5}$$

We can now introduce this series expansion into equation C.1 to find:

$$\begin{aligned}
\Sigma &= \int \mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)\ln\left(\frac{1}{1-\epsilon\frac{\Delta_1(G' \,|\, G)}{\mathrm{P}_\epsilon(G' \,|\, G)\Pi_\epsilon(G)} - \epsilon^2\frac{\Delta_2(G' \,|\, G)}{\mathrm{P}_\epsilon(G' \,|\, G)\Pi_\epsilon(G)} + \mathcal{O}(\epsilon^3)}\right)\mathrm{d}G\mathrm{d}G' \\
&= \int \mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)\ln\left(1 + \epsilon\frac{\Delta_1\left(G' \,|\, G\right)}{\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)} + \epsilon^2\frac{\Delta_2\left(G' \,|\, G\right)}{\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)} + \right. \\
&\qquad\qquad \left. + \epsilon^2\left(\frac{\Delta_1\left(G' \,|\, G\right)}{\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)}\right)^2 + \mathcal{O}(\epsilon^3)\right)\mathrm{d}G\mathrm{d}G' \\
&= \int \mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)\left[\epsilon\frac{\Delta_1\left(G' \,|\, G\right)}{\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)} + \epsilon^2\frac{\Delta_2\left(G' \,|\, G\right)}{\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)} + \right. \\
&\qquad\qquad \left. + \frac{1}{2}\epsilon^2\left(\frac{\Delta_1\left(G' \,|\, G\right)}{\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)}\right)^2 + \mathcal{O}(\epsilon^3)\right]\mathrm{d}G\mathrm{d}G' \\
&= \epsilon\int \Delta_1\left(G' \,|\, G\right)\mathrm{d}G\mathrm{d}G' + \epsilon^2\int \Delta_2\left(G' \,|\, G\right)\mathrm{d}G\mathrm{d}G' \\
&\quad + \frac{\epsilon^2}{2}\int \frac{\Delta_1\left(G' \,|\, G\right)^2}{\mathrm{P}_\epsilon\left(G' \,|\, G\right)\Pi_\epsilon\left(G\right)}\mathrm{d}G\mathrm{d}G' + \mathcal{O}(\epsilon^3) \\
&= \frac{\epsilon^2}{2}\int \frac{\Delta_1\left(G' \,|\, G\right)^2}{\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G' + \mathcal{O}(\epsilon^3)\,.
\end{aligned} \tag{C.6}$$

In the second and third equalities, the Taylor expansions of $\frac{1}{1-x}$ and $\ln\left(1+x\right)$ where utilised. In the final equality, equation C.5 eliminated the first two terms. In particular, it eliminated the $\mathcal{O}\left(\epsilon\right)$ contribution to the entropy production, meaning that the leading order is quadratic. This leads to equation 4.37 in the main text.

By taking

$$\mathrm{P}\left(G'|G\right) = \mathrm{P}_0\left(G'|G\right) + \epsilon\Phi\left(G'|G\right) + \mathcal{O}(\epsilon^2) \tag{C.7}$$

and

$$\Pi\left(G\right) = \Pi_0\left(G\right) + \epsilon\delta\left(G\right) + \mathcal{O}(\epsilon^2)\,, \tag{C.8}$$

we have:

$$\begin{aligned}\Delta_1\left(G'\,|\,G\right) =& \Phi\left(G'|G\right)\Pi_0\left(G\right) - \Phi\left(G|G'\right)\Pi_0\left(G'\right) \\ &+ \mathrm{P}_0\left(G'|G\right)\delta\left(G\right) - \mathrm{P}_0\left(G|G'\right)\delta\left(G'\right) \\ =& \Delta_\Phi\left(G'\,|\,G\right) + \Delta_\delta\left(G'\,|\,G\right)\,, \end{aligned} \tag{C.9}$$

where I have defined:

$$\Delta_\Phi\left(G'\,|\,G\right) = \Phi\left(G'|G\right)\Pi_0\left(G\right) - \Phi\left(G|G'\right)\Pi_0\left(G'\right)\,, \tag{C.10}$$

$$\Delta_\delta\left(G'\,|\,G\right) = \mathrm{P}_0\left(G'|G\right)\delta\left(G\right) - \mathrm{P}_0\left(G|G'\right)\delta\left(G'\right) \tag{C.11}$$

Note that normalisation imposes:

$$\int \Phi\left(G'|G\right)\mathrm{d}G' = 0 \tag{C.12}$$

$$\int \delta\left(G\right)\mathrm{d}G = 0 \tag{C.13}$$

Further note that from the stationarity condition for $\Pi(G)$:

$$\Pi(G) = \int \mathrm{P}\left(G|G'\right)\Pi(G')\mathrm{d}G' \tag{C.14}$$

we obtain the following for $\delta(G)$:

$$\delta(G) = \int \Phi\left(G|G'\right)\Pi_0(G')\mathrm{d}G' + \int \mathrm{P}_0\left(G|G'\right)\delta(G')\mathrm{d}G'\,. \tag{C.15}$$

$\delta\left(G\right)$ will generally not be known. By assuming quick convergence to the stationary state from a prior-distributed initial state, we can approximate it as:

$$\delta(G) \approx \int \Phi\left(G|G'\right)\Pi_0(G')\mathrm{d}G'\,, \tag{C.16}$$

which in turn implies

$$\int \mathrm{P}_0\left(G|G'\right)\delta(G')\mathrm{d}G' \approx 0\,. \tag{C.17}$$

By introducing equation C.9 into C.6, we find:

$$\Sigma = \frac{\epsilon^2}{2}\int \frac{\Delta_\Phi\left(G'\,|\,G\right)^2 + 2\Delta_\Phi\left(G'\,|\,G\right)\Delta_\delta\left(G'\,|\,G\right) + \Delta_\delta\left(G'\,|\,G\right)^2}{\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G'\,. \tag{C.18}$$

Let us look at each term independently:

$$\begin{aligned}\int \frac{\Delta_\delta\left(G' \mid G\right)^2}{\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G' = \\ &= 2\int \frac{\mathrm{P}_0\left(G'|G\right)^2\delta(G)^2 - \mathrm{P}_0\left(G'|G\right)\delta(G)\mathrm{P}_0\left(G|G'\right)\delta(G')}{\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G' \\ &= 2\int \frac{\delta\left(G\right)^2}{\Pi_0\left(G\right)}\mathrm{d}G - 2\int \frac{\mathrm{P}_0\left(G|G'\right)\delta(G)\delta(G')}{\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G' \\ &\approx 2\int \frac{\Phi\left(G|G'\right)\Pi_0(G')\Phi\left(G|G''\right)\Pi_0(G'')}{\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G'\mathrm{d}G''. \end{aligned} \tag{C.19}$$

In the first equality, Δ_δ was expanded and terms grouped capitalising on their symmetry in G and G' under integration. In the second equality, terms were cancelled using the denominator. In the last equality, the first term was expanded using equation C.16 and the second one was eliminated using equation C.17.

Let us look at the central term in equation C.18:

$$\begin{aligned}2\int \frac{\Delta_\Phi\left(G' \mid G\right)\Delta_\delta\left(G' \mid G\right)}{\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G' = \\ &= 4\int \frac{\Phi\left(G'|G\right)\Pi_0(G)\mathrm{P}_0\left(G'|G\right)\delta(G) - \Phi\left(G'|G\right)\Pi_0(G)\mathrm{P}_0\left(G|G'\right)\delta(G')}{\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G' \\ &= 4\int \Phi\left(G'|G\right)\delta(G)\mathrm{d}G\mathrm{d}G' - 4\int \frac{\Phi\left(G'|G\right)\Pi_0\left(G\right)\delta(G')}{\Pi_0\left(G'\right)}\mathrm{d}G\mathrm{d}G' \\ &\approx -4\int \frac{\Phi\left(G|G'\right)\Pi_0(G')\Phi\left(G|G''\right)\Pi_0(G'')}{\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G'\mathrm{d}G''\,. \end{aligned} \tag{C.20}$$

In the first equality, terms were again expanded and grouped based on invariance under replacement of G and G'. The second equality again cancelled out terms using the denominator, first using detailed balance in the second term. Finally, equation C.16 was used and variables renamed to highlight the similarity of this

term to equation C.19. These can now be put together as

$$\begin{aligned}
&\int \frac{\Delta_\delta\left(G' \,|\, G\right)^2 + 2\Delta_\Phi\left(G' \,|\, G\right)\Delta_\delta\left(G' \,|\, G\right)}{\mathrm{P}_0\left(G'|G\right)\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G' = \\
&\quad = -2\int \frac{\Phi\left(G|G'\right)\Pi_0(G')\Phi\left(G|G''\right)\Pi_0(G'')}{\Pi_0\left(G\right)}\mathrm{d}G\mathrm{d}G'\mathrm{d}G'' \\
&\quad = -2\int \frac{\left[\int \Phi\left(G|G'\right)\Pi_0(G')\mathrm{d}G'\right]^2}{\Pi_0\left(G\right)}\mathrm{d}G \\
&\quad = -2\int \frac{\left[\int \left(\Phi\left(G|G'\right)\Pi_0(G') - \Phi\left(G'|G\right)\Pi_0(G)\right)\mathrm{d}G'\right]^2}{\Pi_0\left(G\right)}\mathrm{d}G \\
&\quad = -2\int \frac{\left[\int \Delta_\Phi\left(G' \,|\, G\right)\mathrm{d}G'\right]^2}{\Pi_0\left(G\right)}\mathrm{d}G\,.
\end{aligned} \tag{C.21}$$

Here, I have utilised equation C.12 to express the result in terms of Δ_Φ. With that, we finally have for the entropy production:

$$\Sigma = \frac{\epsilon^2}{2}\int \frac{\Delta_\Phi\left(G' \,|\, G\right)^2}{\mathrm{P}_0\left(G'|G\right)\Pi_0(G)}\mathrm{d}G\mathrm{d}G' - \epsilon^2\int \frac{\left(\int \Delta_\Phi\left(G' \,|\, G\right)\mathrm{d}G'\right)^2}{\Pi_0(G)}\mathrm{d}G\,, \tag{C.22}$$

which is the expression reported in the main text.

Glossary

Analogy

Processes whereby the rules of inflection of a set of words are extended to words that they did not originally apply to.

Compositionality

The property of human languages whereby the meaning of a sentence is generated by putting together the meanings of its individual words and the rules that are used to create the sentence.

Construction

An association of linguistic meaning and expression in a speaker's grammar.

Construction Grammar

A family of theories according to which speakers' grammars are made up of a complex reactive network of constructions.

Content word

A word that is a key contributor to the meaning of the sentence it appears in. Nouns, adjective, and action verbs are content words.

Diachronic

Referring to properties or processes along a language or languages' history.

Distributional universal

Property that, while not present in all languages, is shared by a majority of them.

Domain-general

Referring to cognitive processes that are applied to a variety of social, learning, and processing tasks, and not just language.

Domain-specific

Referring to cognitive processes that are specialised in dealing with language.

Function word

A word that determines the grammatical or topical relations in a sentence, or specifies the attitude or mood of the speaker. Adpositions, conjunctions, pronouns, auxiliary verbs and articles arev function words.

Grammar

The mental representation of a language in a speaker's mind.

Grammaticalisation

Process of language change where content words lose lexical meaning and become function words.

Inflectional paradigm

A set of rules used to generate new words for existing ones in a systematic way, like the use of ⟨-s⟩ to form plurals or ⟨-ed⟩ to form past tenses.

Lingueme

A unit of linguistic function or meaning, such as a phoneme, a concept, or a grammatical relation.

Morphology

The study of words and their rules of formation.

Nativist theories

A family of theories according to which the properties of human languages are, to the greatest extent, genetically determined and encoded in strong, domain-specific cognitive constraints.

Phonology

The study of phonemes – the basic sounds that act as the building blocks that make up words in spoken languages – and the rules that organise them.

Pragmatics

The study how context, implication and intention contribute to meaning.

Semantics

The study meaning, both of individual words and of how their individual meanings create those of the larger structures that they are a part of.

Synchronic

Referring to language properties at a specific moment in time.

Syntax

The study of the formation of sentences from their constituent words.

Transitivity

Property of verbs that take a direct object, such as *like* or *bring.*.

Universal

Property that is shared by all human languages.

Usage-based theories

A family of theories according to which the properties of human languages are greatly shaped by usage in communication.

Utterance

A unit of empirically-measurable language use, such as a phrase or a sentence.

Variant

A unit of linguistic expression, such as a sound, a word, or a syntactic structure, that is used to express a lingueme.

Bibliography

[1] Clay Beckner, Richard Blythe, Joan Bybee, Morten H Christiansen, William Croft, Nick C Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. Language is a complex adaptive system: Position paper. *Language Learning*, 59(s1):1–26, 2009.

[2] Joan Bybee. *Language, usage and cognition*. Cambridge University Press, 2010.

[3] Michael Pleyer and Stefan Hartmann. *Cognitive Linguistics and Language Evolution*. Elements in Cognitive Linguistics. Cambridge University Press, 2024.

[4] Florencia Reali and Thomas L Griffiths. Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, 277(1680):429–436, 2010.

[5] Gareth J Baxter, Richard A Blythe, William Croft, and Alan J McKane. Utterance selection model of language change. *Phys. Rev. E*, 73:046118, 2006.

[6] Charles F Hockett. Animal "languages" and human language. *Human Biology*, 31(1):32–39, 1959.

[7] Derek Bickerton. *Language and species*. University of Chicago Press, 1990.

[8] Robert C Berwick and Noam Chomsky. *Why only us: Language and evolution*. MIT press, 2016.

[9] Edward Sapir. *Language: An introduction to the study of speech*. Courier Corporation, 1921.

[10] Charles F Hockett. The problem of universals in language. *Universals of language*, 2:1–29, 1963.

[11] Joseph H Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113, 1963.

[12] Bernard Comrie. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press, 1989.

[13] William Croft. *Typology and universals*. Cambridge University Press, 2003.

[14] Matthew S Dryer. Order of subject, object and verb (v2020.3). In Matthew S Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo, 2013.

[15] Paul J Hopper and Elizabeth C Traugott. *Grammaticalization*. Cambridge University Press, 2003.

[16] Joan Bybee. Cognitive processes in grammaticalization. In *The new psychology of language*, pages 145–167. Psychology Press, 2014.

[17] Noam Chomsky. *Aspects of the Theory of Syntax*. MIT press, 1965.

[18] Noam Chomsky. *Reflections on language*. Temple Smith London, 1976.

[19] Steven Pinker and Paul Bloom. Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4):707–727, 1990.

[20] Noam Chomsky. *The minimalist program*. MIT press, 2014.

[21] Michael Tomasello. *First verbs: A case study of early grammatical development*. Cambridge University Press, 1992.

[22] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

[23] Rebecca L Gómez and LouAnn Gerken. Infant artificial language learning and language acquisition. *Trends in cognitive sciences*, 4(5):178–186, 2000.

[24] Geoffrey K Pullum and Barbara C Scholz. Empirical assessment of stimulus poverty arguments. *The linguistic review*, 19(1-2):9–50, 2002.

[25] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[26] Jeffrey L Elman. *Rethinking innateness: A connectionist perspective on development*, volume 10. MIT press, 1996.

[27] George Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press, 1987.

[28] Morten H Christiansen, Rick AC Dale, Michelle R Ellefson, and Christopher M Conway. The role of sequential learning in language evolution: Computational and experimental studies. In *Simulating the evolution of language*, pages 165–187. Springer, 2002.

[29] Prahlad Gupta and Jamie Tisdale. Word learning, phonological short-term memory, phonotactic probability and long-term memory: towards an integrated framework. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3755–3771, 2009.

[30] Alexa R Romberg and Jenny R Saffran. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):906–914, 2010.

[31] Elizabeth Bates and Brian MacWhinney. Competition, variation and language learning. In Brian MacWhinney, editor, *Mechanisms of language acquisition*, pages 157–93. Lawrence Erlbaum, Hillsdale, NJ, 1987.

[32] Joan Bybee. From usage to grammar: The mind's response to repetition. *Language*, pages 711–733, 2006.

[33] Michael Tomasello. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press, 2005.

[34] Michael Tomasello. *Origins of human communication*. MIT press, 2010.

[35] Michael Tomasello. *A natural history of human thinking*. Harvard University Press, 2014.

[36] John H Holland. Complex adaptive systems. *Daedalus*, 121(1):17–30, 1992.

[37] Stephen Lansing. Complex adaptive systems. *Annual review of anthropology*, 32(1):183–204, 2003.

[38] Simon A Levin. Ecosystems and the biosphere as complex adaptive systems. *Ecosystems*, 1:431–436, 1998.

[39] Brian Arthur. *The economy as an evolving complex system II*. CRC Press, 2018.

[40] Walter Buckley. Society as a complex adaptive system. In *Systems research for behavioral science*, pages 490–513. Routledge, 2017.

[41] Rika Preiser, Reinette Biggs, Alta De Vos, and Carl Folke. Social-ecological systems as complex adaptive systems. *Ecology and Society*, 23(4), 2018.

[42] Murray Gell-Mann. Complex adaptive systems. In *Santa Fe Institute Studies in the Sciences of Complexity - Proceedings*, volume 19, page 17. Addison-Wesley Publishing Co, 1994.

[43] Serena Chan. Complex adaptive systems. In *ESD. 83 research seminar in engineering systems*, volume 31, pages 1–9. MIT Cambridge, MA, USA, 2001.

[44] Luc Steels. Language as a complex adaptive system. In *International Conference on Parallel Problem Solving from Nature*, pages 17–26. Springer, 2000.

[45] Bart De Boer. Self-organization in vowel systems. *Journal of phonetics*, 28 (4):441–465, 2000.

[46] Luc Steels. Self-organizing vocabularies. In *Artificial Life V*, pages 179–184, 1997.

[47] John Batali. Computational simulations of the emergence of grammar. In James Hurford, Chris Knight, and Michael Studdert-Kennedy, editors, *Approaches to the Evolution of Language: Social and Cognitive bases*, pages 405–426. Cambridge University Press, Cambridge, 1998.

[48] Simon Kirby. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In Chris Knight, Michael Studdert-Kennedy, and James Hurford, editors, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, page 303–323. Cambridge University Press, 2000.

[49] Namhee Lee, Lisa Mikesell, Anna Dina L Joaquin, Andrea W Mates, and John H Schumann. 111 grammar as a complex adaptive system. In *The Interactional Instinct: The Evolution and Acquisition of Language*. Oxford University Press, 08 2009.

[50] Simon Kirby. Language is an adaptive system: the role of cultural evolution in the origins of structure. In *The Oxford Handbook of Language Evolution*. Oxford University Press, 11 2011.

[51] Luc Steels. Modeling the cultural evolution of language. *Physics of life reviews*, 8(4):339–356, 2011.

[52] Charles J Fillmore, Paul Kay, and Mary Catherine O'Connor. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538, 1988.

[53] Adele E Goldberg. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.

[54] William Croft. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, USA, 2001.

[55] Holger Diessel. *The acquisition of complex sentences*, volume 105. Cambridge University Press, 2004.

[56] Adele E Goldberg. Constructions at work. *Cognitive Linguistics*, 2009.

[57] Holger Diessel. *The grammar network*. Cambridge University Press, 2019.

[58] Christian Walloth. *Emergent Nested Systems: A Theory of Understanding and Influencing Complex Systems as Well as Case Studies in Urban Systems*. Springer, 2016.

[59] Albert Baugh and Thomas Cable. *A history of the English language*. Routledge, 1993.

[60] William Labov. *Principles of linguistic change: Vol. 1.* Blackwell, Oxford, 1994.

[61] Jean Aitchison. Language change. In *The Routledge Companion to Semiotics and Linguistics*, pages 111–120. Routledge, 2005.

[62] August Schleicher. *Die Darwinsche Theorie und Sprachwissenschaft: offenes Sendschreiben an Herrn Dr. Ernst Häckel*. Böhlau, 1873.

[63] Charles Darwin and Tom Griffith. *The descent of man*, volume 4. Prometheus Books New York, 1874.

[64] William Croft. *Explaining language change: An evolutionary approach*. Pearson Education, 2000.

[65] Richard Dawkins. *The selfish gene*. Oxford University Press, 1976.

[66] David L Hull. *Science as a process: an evolutionary account of the social and conceptual development of science*. University of Chicago Press, 1988.

[67] John C Wells. *Accents of English*, volume 1. Cambridge University Press, 1982.

[68] Lyle Campbell. *Historical linguistics*. Edinburgh University Press, 2013.

[69] Paul T Roberge. The creation of pidgins as a possible window on language evolution. *LOT Occasional Series*, 10:101–138, 2008.

[70] Rudi Keller. *On Language Change: The Invisible Hand in Language*. Routledge, 1994.

[71] George K Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

[72] William Labov. *Principles of linguistic change: Vol. 2.* Blackwell, Oxford, 2001.

[73] William Labov. *Principles of linguistic change: Vol. 3.* Blackwell, Oxford, 2010.

[74] Nikolaus Ritt. *Selfish sounds and linguistic evolution: A Darwinian approach to language change*. Cambridge University Press, 2004.

[75] Eva Zehentner. *Competition in Language Change: The Rise of the English Dative Alternation*. De Gruyter Mouton, Berlin, Boston, 2019.

[76] Simon Kirby and James R Hurford. The emergence of linguistic structure: An overview of the iterated learning model. *Simulating the evolution of language*, pages 121–147, 2002.

[77] Simon Kirby, Kenny Smith, and Henry Brighton. From UG to universals: Linguistic adaptation through iterated learning. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 28(3):587–607, 2004.

[78] Simon Kirby, Mike Dowman, and Thomas L Griffiths. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245, 2007.

[79] Kenny Smith. Iterated learning in populations of Bayesian agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31, 2009.

[80] Andrew Perfors and Daniel J Navarro. Language evolution can be shaped by the structure of the world. *Cognitive science*, 38(4):775–793, 2014.

[81] Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015.

[82] Joshua B Tenenbaum, Thomas L Griffiths, and Charles Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318, 2006.

[83] Thomas L Griffiths and Joshua B Tenenbaum. Optimal predictions in everyday cognition. *Psychological science*, 17(9):767–773, 2006.

[84] James Joyce. Bayes' theorem. In Edward N Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2003.

[85] Thomas L Griffiths and Michael L Kalish. Language evolution by iterated learning with Bayesian agents. *Cognitive science*, 31(3):441–480, 2007.

[86] James R Norris. *Markov chains*. Cambridge university press, 1998.

[87] Sewall Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97–159, 1931.

[88] Ronald A Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930.

[89] James F Crow and Motō Kimura. *An introduction in Population Genetics Theory*. Harper and Row, New York, 1970.

[90] Daniel Fink. A compendium of conjugate priors, 1997.

[91] Hans H Hock. *Principles of historical linguistics*. Mouton de Gruyter, 1991.

[92] Richard A Blythe and William Croft. How individuals change language. *Plos one*, 16(6):e0252582, 2021.

[93] Lindell Bromham, Xia Hua, Thomas G Fitzpatrick, and Simon J Greenhill. Rate of language evolution is affected by population size. *PNAS*, 112:2097–102, 2015.

[94] Søren Wichmann, Dietrich Stauffer, Christian Schulze, and Eric W Holman. Do language change rates depend on population size? *Advances in Complex Systems*, 11:357–369, 2008.

[95] Cyriel Paris, Bertrand Servin, and Simon Boitard. Inference of selection from genetic time series using various parametric approximations to the Wright-Fisher model. *G3 Genes—Genomes—Genetics*, 9(12):4073–4086, 2019.

[96] Paula Tataru, Thomas Bataillon, and Asger Hobolth. Inference under a Wright-Fisher model using an accurate beta approximation. *Genetics*, 201: 1133–1151, 2015. doi: 10.1534/genetics.115.179606.

[97] Juan Guerrero Montero, Andres Karjus, Kenny Smith, and Richard A Blythe. Reliable detection and quantification of selective forces in language change. *Corpus Linguistics and Linguistic Theory*, 2023.

[98] Juan Guerrero Montero and Richard A Blythe. Self-contained beta-with-spikes approximation for inference under a Wright-Fisher model. *Genetics*, 225(2), 2023.

[99] Richard A Blythe and William Croft. S-curves and the mechanisms of propagation in language change. *Language*, pages 269–304, 2012.

[100] Motō Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, 1983.

[101] Martin Kreitmann. Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.*, 1:539–59, 2000.

[102] Richard E Lenski, Michael R. Rose, Suzanne C Simpson, and Scott C Tadler. Long-term experimental evolution in Escherichia Coli. i. adaptation and divergence during 2,000 generations. *Am. Nat.*, 138:1315–41, 1991.

[103] Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular Cell*, 58:586–97, 2015.

[104] Alison F Feder, Sergey Kryazhimskiy, and Joshua B Plotkin. Identifying signatures of selection in genetic time series. *Genetics*, 196(2):509–522, 2014. doi: 10.1534/genetics.113.158220.

[105] Mitchell Newberry, Christopher Ahern, Robin Clark, and Joshua B Plotkin. Detecting evolutionary forces in language change. *Nature*, 551:223–226, 2017. doi: 10.1038/nature24455.

[106] Folgert Karsdorp, Enrique Manjavacas, Lauren Fonteyn, and Mike Kestemont. Classifying evolutionary forces in language change using neural networks. *Evolutionary Human Sciences*, 2, 2020.

[107] Paula Tataru, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. Statistical inference in the Wright–Fisher Model using allele frequency data. *Systematic Biology*, 66(1):e30–e46, 2017.

[108] Jonathan P Bollback, Thomas L York, and Rasmus Nielsen. Estimation of 2 N es from temporal allele frequency data. *Genetics*, 179(1):497–502, 2008.

[109] S. Lukić and J Hey. Demographic inference using spectral methods on SNP data, with an analysis of the human Out-of-Africa expansion. *Genetics*, 192 (2):619–39, 2012.

[110] Miguel Lacerda and Cathal Seoighe. Population genetics inference for longitudinally-sampled mutants under strong selection. *Genetics*, 198(3): 1237–1250, 2014.

[111] Tin-Yu J Hui and Austin Burt. Estimating effective population size from temporally spaced samples with a novel, efficient maximum-likelihood algorithm. *Genetics*, 200(1):285–93, 2015.

[112] Mark Davies. The Corpus of Historical American English, 2010.

[113] Jean-Baptiste Michel, Yuan K Shen, Aviva P Aiden, Adrian Veres, Matthew K Gray, the Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A Nowak, and Erez L Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[114] Henri Kauhanen and George Walkden. Deriving the constant rate effect. *Nat Lang Linguistic Theory*, 36:483–521, 2018.

[115] Charles Yang. Internal and external forces in language change. *Language Variation and Change*, 12(3):231–250, 2000.

[116] Brian Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.*, 10:195–205, 2009.

[117] Raydonal Ospina and Silvia LP Ferrari. Inflated beta distributions. *Statistical papers*, 51:111–126, 2010.

[118] Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions, volume 2*, volume 289, chapter Chapter 25: Beta Distributions. John Wiley & Sons, 1995.

[119] SS Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4): 784–786, 1974.

[120] Stephen J Wright. Coordinate descent algorithms. *Mathematical programming*, 151(1):3–34, 2015.

[121] Samuel Stanley Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9: 60–62, 1938.

[122] George Casella and Roger L Berger. *Statistical Inference*. Cengage Learning, 2 edition, 2001.

[123] Michael T Madigan, John M Martinko, Jack Parker, et al. *Brock biology of microorganisms*, volume 11. Prentice hall Upper Saddle River, NJ, 1997.

[124] Andres Karjus, Richard A Blythe, Simon Kirby, and Kenny Smith. Challenges in detecting evolutionary forces in language change using diachronic corpora. *Glossa*, 5, 2020.

[125] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[126] Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716, 2007.

[127] Christine F Cuskley, Martina Pugliese, Claudio Castellano, Francesca Colaiori, Vittorio Loreto, and Francesca Tria. Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English. *PloS ONE*, 9(8):e102882, 2014.

[128] Don Ringe and Charles Yang. The threshold of productivity and the 'irregularization' of verbs in Early Modern English. In Bettelou Los, Claire Cowie, Patrick Honeybone, and Graeme Trousdale, editors, *English Historical Linguistics: Change in structure and meaning*. John Benjamins, Amsterdam, 2022.

[129] Joan Bybee. Regular morphology and the lexicon. *Language and Cognitive Processes*, pages 425–455, 1995.

[130] Helen Sims-Williams. Analogical levelling and optimisation: The treatment of pointless lexical allomorphy in Greek. *Transactions of the Philological Society*, 114(3):315–338, 2016.

[131] Joan Bybee. *Phonology and Language Use*. Cambridge University Press, Cambridge, 2001.

[132] Sandeep Prasada and Steven Pinker. Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8:1–56, 1993.

[133] Alan Prince and Paul Smolensky. Optimality: From neural networks to universal grammar. *Science*, 275(5306):1604–1610, 1997.

[134] Bruce P Hayes. Phonetically driven phonology. *Functionalism and formalism in linguistics*, 1:243–285, 1999.

[135] John W DuBois. Competing motivations. In John Haiman, editor, *Iconicity in Syntax*, pages 343–366. John Benjamins, Amsterdam, 1985.

[136] Elizabeth Bates and Brian MacWhinney. Functionalism and the competition model. In Brian MacWhinney and Elizabeth Bates, editors, *The crosslinguistic study of sentence processing*, pages 3–73. Cambridge University Press, Cambridge, 1989.

[137] John Haiman. Iconic and economic motivation. *Language*, 53:781–819, 1983.

[138] Simon Kirby. Competing motivations and emergence: explaining implicational hierarchies. *Linguistic Typology*, 1:995–1026, 1997.

[139] John A Hawkins. *Efficiency and complexity in grammars*. OUP Oxford, 2004.

[140] Eitan A Pechenick, Christopher M Danforth, and Peter S Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*, 10(10):1–24, 10 2015.

[141] John H McDonald. *Handbook of Bological Statistics*, chapter G-test of goodness-of-fit. Sparky House Publishing, 2014.

[142] Joan Bybee. *Frequency of use and the organization of language*. Oxford University Press, 2006.

[143] William R Leben. *Suprasegmental Phonology*. PhD thesis, Massachusetts Institute of Technology, 1973.

[144] Joseph P Stemberger. Morphological haplology. *Language*, 57(4):791–817, 1981.

[145] John J. McCarthy. OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17:207–264, 1986.

[146] Stephan A Frisch, Janet B Pierrehumbert, and Michael B Broe. Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, 22:179–228, 2004.

[147] Konstantin Pozdniakov and Guillaume Segerer. Similar place avoidance: A statistical universal. *Linguistic Typology*, 2007.

[148] Juan M Hernández-Campoy and Juan C Conde-Silvestre. *The handbook of historical sociolinguistics*, volume 68. John Wiley & Sons, 2012.

[149] April MS McMahon. *Understanding Language Change*. Cambridge University Press, 1994.

[150] Peter Garrett. *Attitudes to language*. Cambridge University Press, 2010.

[151] Anni Sairio and Minna Palander-Collin. The reconstruction of prestige patterns in language history. *The handbook of historical sociolinguistics*, pages 626–638, 2012.

[152] Joan Rubin, Björn H Jernudd, Jyotirindra DasGupta, Joshua A Fishman, and Charles A Ferguson, editors. *Language Planning Processes*. De Gruyter Mouton, 2013.

[153] Lieselotte Anderwald. Variable past-tense forms in nineteenth-century American English: Linking Normative Grammars and language change. *American Speech*, 87(3):257–293, 2012.

[154] R Anthony Lodge. Authority, prescriptivism and the French standard language. *Journal of French Language Studies*, 1(1), 1991.

[155] Anne Curzan. *Fixing English: Prescriptivism and language history*. Cambridge University Press, 2014.

[156] Kate Burridge. Euphemism and language change: The sixth and seventh ages. *Lexis. Journal in English Lexicology*, (7), 2012.

[157] Edwin Battistella, Keith Allan, and Kate Burridge. Euphemism & dysphemism: Language used as shield and weapon. *Language*, 69:406, 06 1993.

[158] Kerry Linfoot-Ham. The linguistics of euphemism: A diachronic study of euphemism formation. *Journal of language and linguistics*, 4(2):227–263, 2005.

[159] Steven Pinker. *The Blank Slate: The Modern Denial of Human Nature*. Viking, 2002.

[160] William Labov. The social motivation of a sound change. *Word*, 19(3): 273–309, 1963.

[161] William Labov. The intersection of sex and social class in the course of linguistic change. *Language variation and change*, 2(2):205–254, 1990.

[162] Alberto Acerbi and R Alexander Bentley. Biases in cultural transmission shape the turnover of popular traits. *Evolution and Human Behavior*, 35 (3):228–236, 2014.

[163] Katrina J Edwards, Thomas M Gihring, and Jillian F Banfield. Seasonal variations in microbial populations and environmental conditions in an extreme acid mine drainage environment. *Applied and environmental microbiology*, 65(8):3627–3632, 1999.

[164] Serdar Turkarslan, Arjun V Raman, Anne W Thompson, Christina E Arens, Mark A Gillespie, Frederick von Netzer, Kristina L Hillesland, Sergey Stolyar, Adrian López García de Lomana, David J Reiss, Drew Gorman-Lewis, Grant M Zane, Jeffrey A Ranish, Judy D Wall, David A Stahl, and Nitin S Baliga. Mechanism for microbial population collapse in a fluctuating resource environment. *Molecular systems biology*, 13(3):919, 2017.

[165] Jen Nguyen, Juanita Lara-Gutiérrez, and Roman Stocker. Environmental fluctuations and their effects on microbial communities, populations and individuals. *FEMS microbiology reviews*, 45(4), 2021.

[166] Joshua Schimel, Teri C Balser, and Matthew Wallenstein. Microbial stress-response physiology and its implications for ecosystem function. *Ecology*, 88(6):1386–1394, 2007.

[167] Merav Parter, Nadav Kashtan, and Uri Alon. Environmental variability and modularity of bacterial metabolic networks. *BMC evolutionary biology*, 7: 1–8, 2007.

[168] Nico Geisel, Jose MG Vilar, and J Miguel Rubi. Optimal resting-growth strategies of microbial populations in fluctuating environments. *PLoS One*, 6(4), 2011.

[169] Samuel Karlin and Uri Liberman. Random temporal variation in selection intensities: one-locus two-allele model. *Journal of Mathematical Biology*, 2 (1):1–17, 1975.

[170] Naoyuki Takahata, Kazushige Ishii, and Hirotsugu Matsuda. Effect of temporal fluctuation of selection coefficient on gene frequency in a population. *Proceedings of the National Academy of Sciences*, 72(11):4541–4545, 1975.

[171] John H Gillespie. The effects of stochastic environments on allele frequencies in natural populations. *Theoretical population biology*, 3(3):241–248, 1972.

[172] Edo Kussell and Stanislas Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309(5743):2075–2078, 2005.

[173] Ville Mustonen and Michael Lässig. Molecular evolution under fitness fluctuations. *Physical review letters*, 100(10):108101, 2008.

[174] Motoo Kimura. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. *Genetics*, 39(3):280, 1954.

[175] Samuel Karlin and Benny Levikson. Temporal fluctuations in selection intensities: Case of small population size. *Theoretical Population Biology*, 6(3):383–412, 1974.

[176] Thierry Huillet. On the Karlin–Kimura approaches to the Wright–Fisher diffusion with fluctuating selection. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02016, 2011.

[177] Wayne A Taylor. Change-point analysis: a powerful new tool for detecting changes, 2000.

[178] Gijsbert Rutten and Rik Vosters. Language standardization 'from above'. In *The Cambridge Handbook of Language Standardization*, pages 65–92. Cambridge University Press, 2021.

[179] Matías Guzmán Naranjo. The se-ra alternation in spanish subjunctive. *Corpus Linguistics Linguistic Theory*, 13:97–134, 2017.

[180] Ilpo Kempas. Sobre la variación en el marco de la libre elección entre cantara y cantase en el español peninsular. *Moenia*, 17:243–264, 2011.

[181] Roberta Amato, Lucas Lacasa, Albert Díaz-Guilera, and Andrea Baronchelli. The dynamics of norm change in the cultural evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 115(33):8260–8265, 2018.

[182] Susan Baddeley and Anja Voeste. *Orthographies in early modern Europe*. De Gruyter Mouton, 2012.

[183] Real Academia Española. *Ortografía de la lengua castellana*. Imprenta de la Real Academia Española, Madrid, Spain, 3 edition, 1763.

[184] Ralph John Penny. *A history of the Spanish language*. Cambridge University Press, 2002.

[185] Real Academia Española. *Ortografía de la lengua castellana*. Imprenta de la Real Academia Española, Madrid, Spain, 8 edition, 1815.

[186] Real Academia Española. *Prontuario de ortografía castellana en preguntas y respuestas*. Gregorio Hernando, Madrid, 7 edition, 1881.

[187] Real Academia Española. *Compendio de la Gramática de la lengua castellana*. Perlado, Páez y Cía, Madrid, 27 edition, 1911.

[188] Charles J Fillmore. The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, pages 35–55, 1988.

[189] Adele E Goldberg. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224, 2003.

[190] Anna Łubowicz. Chain shifts. *The Blackwell companion to phonology*, pages 1–19, 2011.

[191] André Martinet. Function, structure, and sound change. *Word*, 8(1):1–32, 1952.

[192] James Burridge and Bert Vaux. Brownian dynamics for the vowel sounds of human language. *Physical Review Research*, 2(1):013274, 2020.

[193] Ellen M Markman and Gwyn F Wachtel. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20 (2):121–157, 1988.

[194] Ellen M Markman. Constraints children place on word meanings. *Cognitive science*, 14(1):57–77, 1990.

[195] Brigitte Nerlich. *Polysemy: Flexible patterns of meaning in mind and language*, volume 142. Walter de Gruyter, 2003.

[196] Bernd Heine. *The handbook of historical linguistics*, chapter Grammaticalization, pages 573–601. Wiley Online Library, 2017.

[197] Elly Van Gelderen. *The linguistic cycle: Language change and the language faculty*. Oxford University Press, 2011.

[198] Udo Seifert. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Physical review letters*, 95(4):040602, 2005.

[199] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on progress in physics*, 75(12):126001, 2012.

[200] Gary F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. Overregularization in language acquisition. *Monographs of the society for research in child development*, pages i–178, 1992.

[201] Virginia A Marchman. Children's productivity in the english past tense: The role of frequency, phonology, and neighborhood structure. *Cognitive Science*, 21(3):283–303, 1997.

[202] James L McClelland and Karalyn Patterson. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in cognitive sciences*, 6(11):465–472, 2002.

[203] Richard Serfozo. *Basics of applied stochastic processes*. Springer Science & Business Media, 2009.

[204] Elena Maslova. A dynamic approach to the verification of distributional universals. *Linguistic Typology*, 2000.

[205] Malvin H Kalos and Paula A Whitlock. *Monte Carlo methods*. John Wiley & Sons, 2009.

[206] Motō Kimura and George H Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561, 1964.

[207] Anthony Kroch. Penn Parsed Corpora of Historical English LDC2020T16, 2020.

[208] R. Zimmermann. The Parsed Corpus of Middle English Poetry (PCMEP), 2015.

[209] Alvar Ellegård. *The auxiliary do : the establishment and regulation of its use in English*. Almqvist I& Wiksell, 1955.

[210] Anthony S Kroch. Reflexes of grammar in patterns of language change. *Language variation and change*, 1(3):199–244, 1989.

[211] Anthony Warner. Why do dove: Evidence for register variation in Early Modern English negatives. *Language Variation and Change*, 17(3):257–280, 2005.

[212] Aaron Ecay. *A multi-step analysis of the evolution of English do-support*. PhD thesis, University of Pennsylvania, 2015.

[213] Aaron Ecay. The Penn-York Computer-annotated Corpus of a Large amount of English based on the TCP (PYCCLE-TCP), 2015.

[214] Bernard Comrie. Noun phrase accessibility and universal grammar 1. In *Universal Grammar (RLE Linguistics A: General Linguistics)*, pages 3–45. Routledge, 2014.

[215] Ivan A Sag. English relative clause constructions. *Journal of linguistics*, 33 (2):431–483, 1997.

[216] Sir Gawain and the Green Knight.

[217] Geoffrey Chaucer. The Canterbury Tales, 1400.

[218] Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

[219] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[220] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[221] Mark Pagel, Mark Beaumont, Andrew Meade, Annemarie Verkerk, and Andreea Calude. Dominant words rise to the top by positive frequency-dependent selection. *Proceedings of the National Academy of Sciences*, 116 (15):7397–7402, 2019.

[222] James Burridge. Spatial evolution of human dialects. *Physical Review X*, 7(3):031008, 2017.

[223] James Burridge. Unifying models of dialect spread and extinction using surface tension dynamics. *Royal Society open science*, 5(1):171446, 2018.

[224] James Burridge and Tamsin Blaxter. Spatial evidence that language change is not neutral. *arXiv preprint arXiv:2005.07553*, 2020.

[225] Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389–407, 2019.