Grammar competition and variational learning

Joel Wallenberg[1], Henri Kauhanen[2], George Walkden[2], and Caroline Heycock[3]

[1]*University of York*
[2]*University of Konstanz*
[3]*University of Edinburgh*

2026

NB: This is an Author Accepted Manuscript version reflecting changes made in the review process, but not the publisher's PDF. This contribution is published in the *Wiley-Blackwell Companion to Diachronic Linguistics* (doi: 10.1002/9781119898023.wbcdl029). When citing, please use the page numbers given there. The publisher should be contacted for permission to re-use or reprint the material in any form.

Abstract

This chapter describes the competing grammars / grammar competition paradigm for analyzing synchronic variation, diachronic change and the variation found during changes in progress, and its role as the foundation for Yang's variational learning model of child language acquisition. The chapter first sets out the essential components of the paradigm, and then discusses some common objections (with particular attention to syntactic theory), indicating how these can be resolved. Then, ramifications of the paradigm are discussed in more detail, including its relation to issues within linguistics (Yang's variational learning model of acquisition, the variable rules of Labovian sociolinguistics, and Kroch's Constant Rate Effect) and extending beyond (to evolutionary dynamics as a way of understanding language change). The chapter includes brief discussions of the mathematical underpinnings of the model, but is intended to be comprehensible also to a reader who wants a non-technical overview. We conclude that some notion of grammar competition is necessary for any coherent theory of language variation and change, and that grammar competition makes possible several important theoretical unifications in the above domains of linguistic theory.

Keywords: competing grammars, variational learning, evolutionary dynamics, Constant Rate Effect

Word Count: 11304

1 General introduction to the worldview

This chapter presents grammar competition (or "competing grammars"—we will use these two terms interchangeably) as a paradigm for analyzing variation in categorical aspects of linguistic varieties and change in these aspects over historical time. The grammar competition paradigm began in both the historical linguistics and child language acquisition literatures almost simultaneously, and seemingly independently, with the publication of Kroch (1989) and Fritzenschaft et al. (1990). We can conceive of this paradigm as having four main components, all of which are present in the founding publications, and all of which have been expanded on a great deal in subsequent work:

(1) A general hypothesis about linguistic competence: an individual can acquire and use more than one variant of some grammatical formative, where "variant" is understood

as one of the ways, out of multiple ways, of saying the same thing (in the Labovian sense).

(2) A consequent hypothesis about language use and linguistic variation: the linguistic behavior of individuals with variation of the above type will include surface exponents of multiple variants of the same grammatical formative, so categorical variation within language use can be explained as categorical variation within linguistic competence.

(3) A specific hypothesis about the representation of linguistic competence: an individual associates probabilities with the above type of variants, and draws on the variants in their inventory according to these probabilities. An acquirer must also learn the variants and their associated probabilities; this is "variational learning" (Yang 2002, ch. 2).

(4) A consequent hypothesis about language change: a change in a categorical aspect of a variety's grammar over historical time can be explained as the innovation of a new variant, which leads to the state of variation described above, followed by generational shifts in the probabilities associated with the variants. Thus, the linguistic output of a population over historical time is not only a mixture of individual language users, but also, simultaneously, a mixture of variants within individual competence. This also means that "stages of languages" which appear to be unusually variable do not represent unusually variable grammatical systems which only existed in the past, but rather mixtures of usual grammatical systems which can often be observed in the present.

The example in (5) is the type of data that grammar competition was originally proposed to account for:

(5) Mi feader & Mi moder for-þi þt ich nule þe forsaken; habbe
My father and my mother because that I not+would you forsake have
forsake me.
forsaken me
"Because I would not forsake you, my father and mother have forsaken me"
(*St. Juliana*, date: c1225; ID CMJULIA-M1,106.172 from *Penn Parsed Corpus of Middle English 2*, Kroch and Taylor 2000)

This example shows both a right-headed *v*P/VP (*þe forsaken*—OV) and a left-headed *v*P/VP (*forsake me*—VO) in the same sentence, with almost the same lexical items, in the same discourse context, produced by the same author. The pronominal objects are not subject to rightward (Pintzuk 1991; Kroch and Taylor 1997) or leftward (Wallenberg 2009, 2013) movement across the verb, and so can be taken as diagnostic for the underlying *v*P/VP structure. Finally, both objects are contrastive, possibly both contrastive topics, and so they have similar information-structural properties. These two variants are simply two ways of saying the same thing (in the Labovian sense), and one language user can alternate between them rapidly, with the "choice" of variant plausibly occurring below the level of consciousness.

One can see a similar instance of competing grammars in the phonological example below, originally from Glaser (1985), discussed in Fruehwald, Gress-Wright, and Wallenberg (2013):

(6) <tac> ~ <tag>
("day" Accusative singular, Early New High German)

Glaser finds both variants of this morphological form of the word *day* in an Early New High German manuscript of the *Augsburger Stadtbuch*, part of a wider pattern of variable final stop devoicing in Early New High German. Fruehwald, Gress-Wright, and Wallenberg (2013) build on these observations, showing that speakers alternated between devoiced stops

and voiced stops in final position, exhibiting competition between a devoicing phonological rule and the production of surface forms with voiced final stops identical with their underlying forms. Morphological alternations provide evidence that the observed variation went beyond spelling, and diachronic evidence shows that this was a change in progress. As expected, during the change, language users had more than one way of saying the same thing, and alternated between different versions of a rule when producing even the same lexical items in the same environments in the same manuscripts.

Finally, any introduction to this topic would be remiss not to include the empirical case which led to the idea of competing grammars in Kroch (1989): *do*-support in the history of English. The examples below show the two variants in question. Sentence (7) exhibits V-to-T movement of the finite lexical verb (with subsequent movement of T-to-C, or relevant head in the left-periphery, for question formation) and so has no *do*-support. (8), on the other hand, shows no V-to-T movement, but instead has *do*-support in T (which subsequently moves to C):

(7) Heard you that?
(*The Merry Wives of Windsor*, William Shakespeare, ID SHAKESP-E2-P2,51.C1.264 in *Penn Parsed Corpus of Early Modern English*, Kroch, Santorini, and Delfs 2004)

(8) Did you euer heare the like?
(*The Merry Wives of Windsor*, William Shakespeare, ID SHAKESP-E2-H,44.C1.68 in Kroch, Santorini, and Delfs 2004)

As with the other cases above, the variation is part of a change in progress, and it is impossible for an individual to utter both variants in a single token of the relevant linguistic structure (a matrix question, here). The variants are therefore in competition; the language user must "choose" between them in a given utterance of the relevant type, but may then "choose" a different variant in a subsequent similar utterance, apparently without consciousness.

The structure of the rest of the chapter is as follows. Section 2 addresses a number of arguments that have often been levelled against the concept of competing grammars, and shows that these arguments generally stem from a misunderstanding of what grammar competition means, or from perceived negative theoretical consequences of any analysis that posits competing grammars. In particular, we argue that grammar competition is not psychologically implausible and does not lead to conclusions which are psychologically implausible, nor is it incompatible with certain empirical results. In Section 3, we explore a number of theoretical consequences the competing grammars notion does have, if it is made more formal so that it is more of a theory than a general framework. Notably, grammar competition allows for a formal theory of language acquisition—variational learning (Yang 2002)—which has a number of specific predictions about child language acquisition and language change. Grammar competition also allows work on language change to make use of the mathematical framework of evolutionary dynamics, stemming originally from models of biological evolution. This theoretical synthesis also helps to explain and better test the Constant Rate Effect, one of the most important empirical results in the field, which became detectable because of the way Kroch (1989) stated the notion of competing grammars. Finally, we conclude and offer some directions for future research.

2 Frequently Raised Objections

In the previous section we set out what we take to be the core components of the grammar competition paradigm in (1)–(4) above, and gave brief illustrations of some of the types of data the paradigm was set up to address. Before going into more details of how the paradigm can be fleshed out, we first address some of the objections that have been raised in the literature. Some of these, we will argue, turn out to be unproblematic under our formulation of the key hypotheses.

2.1 What competes?

Henry (2002) contrasts the competing grammars paradigm with variation *within* grammars. She criticizes Kroch (1994) for failing to 'grasp the nettle' and allow for grammar-internal variation (2002, 272), maintaining that 'there is a considerable difference between a person having two grammars, and a single grammar which admits optionality' (2002, 273). Under our characterization of the paradigm, though, this difference is not a crucial one. All that is required is that acquirers be able to acquire multiple variants for a single grammatical variable, and associate these variants with probabilities. The nature of these variants will be given by one's grammatical theory. For Kroch (1989), the variants are settings of a single parameter. In Minimalist syntax, adopting the Borer-Chomsky Conjecture (Borer 1984; Chomsky 1995), variants would be different instantiations of a single functional lexical item (also the approach taken in Kroch 1994). One could also conceive of variants as rules, as in classical generative syntax and phonology (Chomsky 1965; Chomsky and Halle 1968; See section 3.1.5 on variable rules), or as constructions, as in Construction Grammar (e.g. Goldberg 1995; Boas and Sag 2012). In any of these cases, if we define a grammar as the full set of grammatical properties that determine the syntactic structure of a language, two (and only two) variants in competition is equivalent to two grammars in competition, and hence it has often been a useful idealization to treat probabilities as attached to grammars (e.g. Yang 2002, ch. 2). But it is not necessary to do so, and in this respect the 'grammars' in 'competing grammars' is a label that is there only for historical reasons.

2.2 A combinatorial explosion?

A related objection is made in Nevins and Parrott (2010). They note that, in varieties with more than one variable process operating independently, the number of grammars grows exponentially with the number of variable processes. With variables *a* and *b*, each with variants 0 and 1, there are four grammars: (0,0), (0,1), (1,0) and (1,1). More generally, there will be 2^n grammars, where n is the number of (binary) variables. Thus, to characterize a variety with five variable processes requires 32 grammars, for instance.

Is this 'combinatorial explosion' (Nevins and Parrott 2010, 1139) a problem? Only if we insist that variants must be grammars as just defined, and even then only if grammars are stored as such (i.e. with massive duplication of redundant properties). But neither of these is a necessary feature of the competing grammars paradigm as we have sketched it. Under the characterization in Section 1, the approach of Nevins and Parrott (2010) in terms of variable impoverishment rules in fact falls under the umbrella of the competing grammars paradigm.

2.3 Competing grammars and headedness

The competing grammars paradigm is often linked to a treatment of constituent order in which individual phrases (e.g. VP or IP) can be variably head-final or head-initial. This view—dubbed the *Double Base Hypothesis*—has its origins in important early work in the competing grammars paradigm by Santorini (1989, 1992, 1993) and Pintzuk (1991, 1995, 1999), with an empirical focus on historically attested West Germanic languages. Santorini (1992) proposed that variation and change in the history of Yiddish could be captured by positing two grammars in competition, one head-final in the IP, one head-initial in the IP, with the latter probabilistically gaining ground at the expense of the former. Similarly, Pintzuk (1999) argues against earlier treatments of Old English constituent order as derived via a head-final VP plus extraposition (Kemenade 1987), instead making the case that both VP and IP vary and change in headedness from Old to Middle English.

The debate about constituent order in these varieties is often framed in terms of a variable base word order analysis (the Double Base Hypothesis), in the sense of the GB-era context-free "base" in e.g. Chomsky (1986), versus a consistently head-initial analysis following Kayne's (1994) hypothesis of a universal head-initial base ('antisymmetry', e.g. Roberts 1997; Biberauer and Roberts 2005; see Pintzuk 2005 for a response). Crucially, though,

there is no incompatibility between a Kaynian head-initial analysis, with apparently head-final structures instead derived via movement, and the competing grammars paradigm. We refer the reader to Wallenberg (2009, ch. 5), in which antisymmetric assumptions about constituent order are unified with a competing grammars approach to variation and change.

More recently, Roberts (2021, 465) argues that the competing grammars paradigm is unable to handle the Final-over-Final Condition (FOFC; Sheehan et al. 2017), a condition which forbids a head-final projection from dominating a head-initial projection in the same spine. FOFC or something very like it clearly holds for historically attested Germanic languages in the verbal domain: VOAux sequences are not found (see e.g. Walkden 2014, 315–316). Roberts points out that, if Aux is the head of IP with VP as its complement, and V the head of VP with the object DP as its complement, a theory that allows the headedness of VP and IP to vary independently wrongly predicts the existence of VOAux during historical periods in which both IP and VP are changing headedness, assuming that subsentential code-switching between competing grammars is permitted (Roberts 2021, 460). This prediction, however, follows from the simple variable-headedness theory rather than the competing grammars paradigm as such. (Indeed, Pintzuk 1991 also makes the same observation with regard to the overlapping IP and VP changes in Old English.) Here we give a sketch of how FOFC can be derived while maintaining a competing grammars approach.

We follow Biberauer, Holmberg, and Roberts (2014) in assuming that constituent order is uniformly linearized as head-initial, with surface head-finality being derived by a feature ˆ that triggers comp-to-spec movement (e.g. of the DP object from the complement of VP to its specifier). For concreteness, we illustrate using the verbal extended projection. ˆ is associated with the categorial feature of an extended projection, e.g. [+V]. Functional heads in an extended projection inherit the categorial feature of the lexical head (e.g. V) of that extended projection by means of a process of upward feature percolation, and hence never inherently bear ˆ. These functional heads c-select for a categorial feature (e.g. [+V]); one possibility is that they specifically select the ˆ-bearing version (e.g. [+Vˆ]), in which case they will surface as head-final. If they select only [+V], without specifying the presence of ˆ, then they will surface as head-initial. ˆ only percolates upward if the ˆ-bearing version (e.g. [+Vˆ]) is selected by the higher head. If a ˆ-bearing V enters the derivation, then ˆ percolates up the extended projection, up to the point at which a functional head is merged which does not select [+Vˆ]; this yields either consistently head-final order or (if a functional head selecting [+V] rather than [+Vˆ] is merged) head-final order in the lower part of the extended projection and head-initial order in the higher part. Crucially, once a functional head that selects [+V] rather than [+Vˆ] is merged into an extended projection, no subsequent functional heads will be able to select [+Vˆ], as it has nowhere local to percolate from. If, however, the extended projection starts with a non-ˆ-bearing V, then merger of any subsequent functional heads that select [+Vˆ] will lead to crash due to feature mismatch, ensuring consistent head-initial order.

Assume that the point at which competition between lexical variants is resolved is the point of external Merge (from the lexicon/numeration). Assume further that in principle (i) two lexical items of the same category (e.g. V), one bearing ˆ and one without it, or (ii) two functional heads of the same category (e.g. I), one selecting a categorial feature with ˆ and one selecting a categorial feature without it, may constitute variants of the same grammatical formative, associated with probabilities. If a non-ˆ-bearing V is merged, no functional head selecting [+Vˆ] can subsequently be merged into the same extended projection and yield a well-formed outcome, regardless of the stored probability associated with that variant. If, on the other hand, a ˆ-bearing V is merged, competition higher in the extended projection may proceed as normal. Here we see FOFC holding both within and across 'grammars', in a theory within the competing grammars paradigm that takes competition to be inherently lexical in nature. Thus FOFC does not necessarily pose a challenge to the competing grammars paradigm, *pace* Roberts (2021).

2.4 *Competing grammars and stable variation*

Another objection raised by Henry (2002) to the competing grammars paradigm, again with reference to Kroch (1994), is 'that it does not allow for stable variation across long periods of time' (2002, 273), which she takes to be an undesirable consequence in view of the apparent existence of exactly such cases (cf. Hudson 1997, 97; Hale 2007, 173). It is true that Kroch (1994, 183) suggests that variation will not stabilize; however, this claim is not a necessary consequence of his conception of competing grammars, nor does he present a mechanistic model which rules out stable variation. Under the characterization of competing grammars in Section 1, too, there is nothing to rule out stable variation in principle: competition need not result in the ultimate victory of one variant.

The issue is, however, a little more complicated than this. First, within the competing grammars paradigm, models have been presented whose effect is to rule out stable variation. It can be shown, for instance, that in the case of two variants, assuming Yang's (2002) variational learner over intergenerational time and the usual simplifying assumptions, stable variation is not normally a possible outcome (we return to this in more detail in Section 3.1).

Second, if the difference in advantages between two competing variants is small enough (see again Section 3.1), very slowly changing probabilities at the population level may give the illusion of stable variation, if not enough temporal resolution is available in the empirical data to resolve the slowly changing probability vector. On this latter point, Wallenberg (2016) shows that over the *longue durée* the competition between intraposed and extraposed relative clauses in Indo-European languages is not stable, but rather exhibits a very gradual tendency over time in favor of intraposition. Wallenberg also argues that this *nearly* stable variation is the result of variants specializing for function along some continuous dimension, and that such specialization may be the key to unifying competing grammars with other cases of apparent stable variation.

Ultimately, whether stable variation exists at all is a—not yet resolved—empirical question. Establishing an answer to this question for any given case is not straightforward, since the standard methods of inferential statistics do not allow us to infer the absence of an effect from a failure to reject the null hypothesis. If clear cases of stable variation can be found, these pose a challenge to specific models of acquisition and change such as the two-variant variational learner mentioned above. They do not, however, challenge the guiding hypotheses of the competing grammars paradigm itself. On the contrary, we would maintain that questions of stable and unstable probabilistic variation only arise in the first place once the competing grammars paradigm, or something very like it, has been adopted.

2.5 *Variation in the individual and the population*

Rich corpus data such as that presented in Santorini (1989, 1992, 1993) and Pintzuk (1991, 1995, 1999) exhibits variation that must be taken seriously. The mere fact of variation in a diachronic corpus, however, does not force us to adopt the competing grammars hypotheses regarding human linguistic competence. It could in principle be the case that what appears to be probabilistic variation from a zoomed-out population perspective in fact results from the distribution of discrete options at an individual level. In other words, if 70% of our corpus data shows variant V_1 and 30% shows variant V_2, this could be because all writers in the population use V_1 70% of the time and V_2 30% of the time, or because 70% of writers use V_1 100% of the time and the remaining 30% use only V_2. Or it could be a mixture between these two extremes.

In practice it is clear that individuals sometimes show probabilistic variation in their written production, even in single-authored texts where the possibility of editorial interference can be ruled out (see e.g. Santorini 1992, 617–618). Establishing that this written performance in fact reflects the competence of a single individual is not a trivial task, but must be part of the standard philological methodology of anyone who takes competence as the primary object of study, as diachronic generative syntacticians do (see Whitman, Jonas,

and Garrett 2012). Once the possibility of probabilistic variation as part of linguistic competence is admitted, the question becomes a methodological one, of how to establish the extent of it.

Specific mechanistic models, such as variational learning, are useful in disentangling these competing hypotheses. When exposed to a particular linguistic environment, the variational learning model predicts the learner to arrive at a particular probability of using the competing variants, as we discuss in more detail in Section 3.1. To be more specific, it is possible to mathematically predict the expected value of this probability, as well as its variance. A cleverly designed experiment or fieldwork protocol could then be used to gather production data from a number of language users in order to assess to what extent these predictions of the model are borne out. While we are not aware of such a study in the existing literature, we again point out that it is only after the competing grammars framework has been formulated that the question can be rigorously formulated in the first place.

2.6 The proper place of competing grammars in linguistic theorizing

The final objection we will address in this section is a much broader one. It is clear that a theory that admits competing grammars opens up a much broader space of possibilities than one that does not. In fact, formally, the classical discrete generative conception of competence can be viewed as a special case of the competing grammars paradigm, one in which all probabilities are restricted to be either 1 or 0, with intermediate values disallowed. This vastly richer world of possibilities should give one pause for thought, especially given the traditional dictum in philosophy of science that more restrictive/predictive theories are to be preferred (Popper 1959).

The hypothesis that language acquirers can associate linguistic variants with probabilities thus faces a substantial burden of proof. We argue that this demand is in fact met. Space precludes a full treatment of all of the evidence for probabilistic acquisition, but Hudson Kam (2015, 907–913) reviews a wide range of studies that support this hypothesis. This evidence ranges from classical sociolinguistic studies of naturalistic acquisition (e.g. Labov 1989) to statistical learning studies (Saffran, Aslin, and Newport 1996), early work on probability learning (Estes 1976), and the artificial language learning and iterated learning paradigms (Hudson Kam and Newport 2005; Reali and Griffiths 2010). From Hudson Kam's review of the literature it emerges that children, though they frequently regularize their input, can also adroitly acquire sociolinguistic (probabilistic) variation if the circumstances are right.

That said, the full probabilistic apparatus of competing grammars should not be drawn on in every case. Early works in the competing grammars paradigm are very clear on this point, e.g. Santorini (1992, 620), responding to the claim that competing grammars 'illegitimately complicates the analysis of linguistic phenomena':

> [J]oint considerations of empirical adequacy and theoretical consistency may lead us to propose analyses of linguistic variation in terms of the interaction of more than one grammatical system, but unless forced to adopt such analyses by the linguistic evidence [...] we will prefer ones based on the assumption that a speaker's performance reflects a single grammatical system.

In other words, it is only when a traditional discrete grammatical analysis based on structural and distributional criteria has failed that an analysis in terms of competing grammars becomes admissible. Here, the critics of the competing grammars paradigm and its adherents are in agreement—at least in principle. There are very many linguistic phenomena for which a competing grammars analysis is unnecessary, hence undesirable. (Note that this does not exclude the possibility that there might be cases where a seemingly simple system actually contains competing grammars that are difficult to detect. See, for example, Fruehwald (2013, 2016)'s evidence that, at least at one period of time, Philadelphians represented two distinct allophones with no phonetic difference between them.)

On the other hand, there are situations in which mechanistic models in the competing grammars paradigm actually predict acquired probabilities of 0 and 1, with no allowance for intermediate values. This occurs, for example, when all of the learner's input is compatible with one grammatical analysis only (see Section 3.1). Thus, an equivalent way of looking at the epistemological situation is to maintain that learners are in fact inherently probabilistic, and that categorical behavior arises under specific circumstances. In a sense, our theory is now *more* falsifiable: we now predict fully categorical behavior only under a particular set of circumstances. Moreover, whether those circumstances apply in a given historical or present-day situation can be empirically investigated, not just assumed *ad hoc*.

Within the competing grammars paradigm, a major current focus of research is to develop predictive theories that narrow down which probabilities are assigned to which grammars under which (social, historical, and acquisitional) circumstances. These theories presuppose the hypotheses outlined in Section 1, and build on them to make links to mathematical models in evolutionary dynamics and probabilistic learning. In the following section, we outline some of these developments in more detail.

3 What this way of thinking buys you

The competing grammars approach to linguistic variation and change has two important advantages, both instances of theoretical unification. First, synchronic linguistic variation and language change over historical time can both be understood as instances of competing grammars. Secondly, linguists do not have to invent every aspect of a theory of diachronic change anew; rather, we can capitalize on the well-developed theory of selection and evolutionary dynamics from the biological sciences (with the required modifications for this case of cultural transmission). This includes both models of selection based on differential fitness (in fact, in Section 3.1 we will see that selection underlies the most fundamental applications of the variational learning model) as well as models of neutral evolution and the effect of stochastic fluctuations—the latter are crucial for linguistic cases given the limited size of human social networks and the limited duration of child language acquisition, both of which pose problems for purely deterministic accounts of language dynamics.

3.1 Variational learning

The *variational learning* (VL) model of Yang (2000, 2002) fits squarely within the grammar competition worldview outlined in Section 1, but extends the work of Kroch (1989, 1994, 2000) in two main ways. On the one hand, Yang gives an explicit mechanism for how acquisition—viewed fundamentally in terms of grammar competition within the individual—takes place in general, and in the context of variation in particular. On the other, he also aims to locate the motor for diachronic change within the heart of the acquisition system, rather than appealing to ideas such as preference for vernacular forms or other sociolinguistic pressures as in e.g. Kroch (1989, 1994). This is not to deny that types of sociolinguistic salience can play a role in driving some variants to replace others, particularly in cases of change from above (see Labov 1994; 2001, for an overview), but many cases of change (notably changes from below the level of consciousness) are not amenable to such analyses. (Note also how the notion of variants in competition is necessary even for a coherent description of changes from above in discrete linguistic features.)

Like Kroch, Yang aims to reconcile a categorical view of a grammar as a set of parametric choices with one of the key observations about the gradual nature of syntactic change referenced in Section 1, namely that—where we have enough evidence to be able to tell—syntactic change does not involve a gradual replacement of non-variable users of the old grammar by non-variable users of the new, but rather a series of changes in the probabilities associated with the new and old variant within the use—and competence—of individuals. But he also argues that the same view has to be taken of the course of acquisition in the individual. That is, the course of acquisition involves changes in the probabilities of use

of conflicting grammatical options rather than abrupt transitions between them. Further, if exposed to variable input—input that contains a mixture of sentences deriving from different grammars—Yang's model predicts that a learner will acquire both grammars, each one associated with a probability of use.

3.1.1 The learning algorithm(s)

Yang adapts a classical "instrumental conditioning" model of learning (Bush and Mosteller 1955) to syntactic acquisition. Informally, his model states that when a child hears a sentence, they pick a grammar from the space of possible grammars. Initially, each grammar is associated with the same weight—the same probability of being chosen. The process of acquisition is the process of updating these weights in the light of success or failure in parsing. If the parsing is successful, the selected grammar is rewarded: that is, the weight associated with it is increased, and all the alternative grammars are punished—the weights associated with them are decreased. If the parsing is unsuccessful, the reverse process takes place.

In principle, any number of learning algorithms could be used to update the grammar weights, and experimental research is needed to figure out the algorithm actually used by real-life language learners. In fact, in its widest sense, "variational learning" refers to any method of updating a probabilistic knowledge state; the foundational assumptions of the framework are independent of the specifics of concrete learning algorithms (cf. Yang 2000, ch. 2). In practical applications, however, some learning algorithm must be chosen, and here most studies have followed Yang in adopting, as a reasonable starting point, the linear reward–penalty procedure of Bush and Mosteller (1955), which we now outline. Let p_t stand for the current weight of grammar G_1; assuming competition between only two variants for simplicity, it then follows that the weight of G_2 is $1 - p_t$, since in this context G_2 simply means "not G_1". Upon receiving a sentence $s \in \mathcal{E}$ from their environment $\mathcal{E}$, the learner undergoes the following update, where $G_i \rightarrow s$ means that grammar G_i parses sentence s, and $G_i \nrightarrow s$ indicates parsing failure:

$$(9) \quad p_{t+1} = \begin{cases} p_t + \gamma(1 - p_t) & \text{if } G_1 \text{ chosen and } G_1 \rightarrow s \\ p_t - \gamma p_t & \text{if } G_1 \text{ chosen and } G_1 \nrightarrow s \\ p_t - \gamma p_t & \text{if } G_2 \text{ chosen and } G_2 \rightarrow s \\ p_t + \gamma(1 - p_t) & \text{if } G_2 \text{ chosen and } G_2 \nrightarrow s \end{cases}$$

Here, γ is a (typically small and constant) positive learning rate parameter that governs the size of each update. From the form of this learning rule it is easy to see that (i) the weight of G_1 gets increased whenever G_1 parses successfully or G_2 fails to parse, and that (ii) the weight of G_1 gets decreased whenever G_1 fails to parse or G_2 parses successfully—in line with the guiding assumptions outlined in prose in the previous paragraph.

If all the child's input is unambiguously the output of a single grammar, that grammar will never be punished, and the probability of it being used will therefore rise to 1; the child acquires a single, categorical grammar (although they will often exhibit variability during the course of acquisition). However, as we know from the historical record as well as from synchronic studies, in some cases the input to a child may be heterogeneous: some sentences being generated by one grammar, and some by another. This, of course, is the crucial case for the understanding of diachronic change.

To understand the learner's behavior in more complex environments, we look at how the *expected* value of p, $\mathrm{E}[p]$, evolves. This can be thought of as the average p one would observe across a large population of learners subject to the same linguistic environment, or alternatively as the average p of a single learner if that learner got to "relive" the entire course of their learning period anew multiple times.

Let c_i in (10) denote the *penalty probability* for grammar G_i; in other words, c_i is the probability of the learner encountering an input sentence such that G_i fails to parse it.

$$(10) \quad c_i = \Pr(G_i \nrightarrow s \mid s \in \mathcal{E})$$

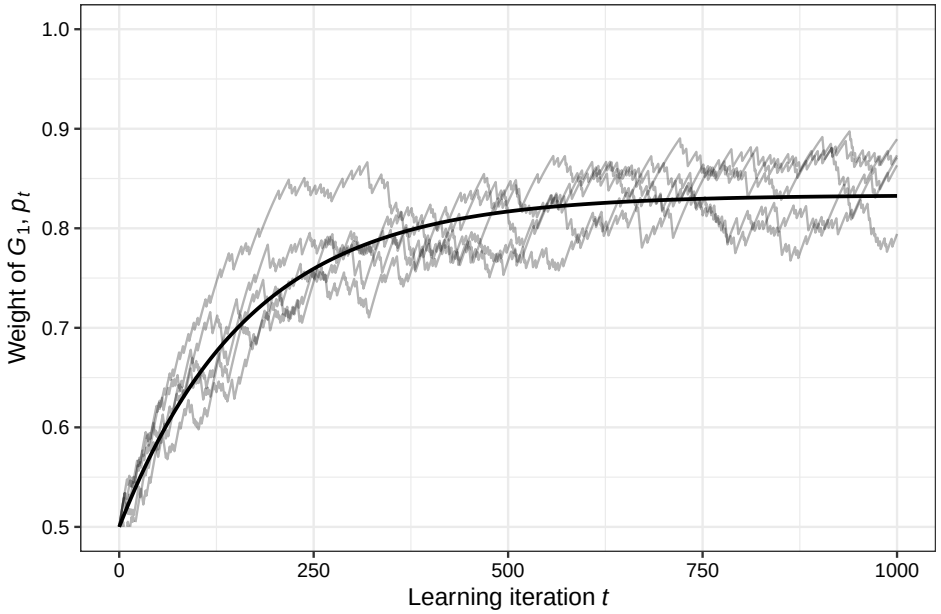

Figure 1: The trajectories of five variational learners subject to the same linguistic environment. The expected value of the grammar weight, $\mathrm{E}[p]_t$, is shown as the solid curve.

Bush and Mosteller (1955) then show that $\mathrm{E}[p]$ has the following solution, given that the learner starts with initial weight p_0:

$$\mathrm{E}[p]_t = \frac{c_2}{c_1 + c_2} - \Big(1 - \gamma(c_1 + c_2)\Big)^t \left(\frac{c_2}{c_1 + c_2} - p_0\right) \tag{11}$$

In particular, since $|1 - \gamma(c_1 + c_2)| < 1$, with every iteration t the prefactor of the second term on the right hand side, $\big(1 - \gamma(c_1 + c_2)\big)^t$, becomes smaller and smaller, i.e. closer to zero. In the limit $t \to \infty$, only the first term remains so that, eventually, the weight assigned by the learner to G_1 is closely approximated by (12).

$$\mathrm{E}[p]_\infty = \frac{c_2}{c_1 + c_2} \tag{12}$$

This result has two important consequences: firstly, it implies that the eventual grammar weights only depend on the penalty probabilities of the two grammars—for instance, the learner's initial state plays no role in this as it gets "forgotten" over iterated learning. Secondly, convergence to the limiting value $\mathrm{E}[p]_\infty$ is exponential, and thus occurs quickly.

To illustrate, in Figure 1 we show a number of realizations of this stochastic process: five independent learners exposed to the same learning environment. We set $c_1 = 0.1$ and $c_2 = 0.5$; these choices imply that grammar G_2 is punished more often than grammar G_1. (The learning rate parameter was set at $\gamma = 0.01$ and the initial weight for each of G_1 and G_2 at $p_0 = 0.5$ for all learners.) Each learner is expected to settle at a weight for G_1 of $0.5/(0.1+0.5) \approx 0.83$ in the limit, and this is exactly what the simulated learning trajectories show. Even though different learners take different learning paths due to the randomness inherent in sampling input from the environment, each learner's trajectory is well described by the evolution of the expected value $\mathrm{E}[p]_t$.

If the learner's input is entirely composed of a mixture of sentences, each sentence either unambiguously the output of grammar G_1 or unambiguously the output of a different grammar G_2, the learning model predicts that the learner will arrive at a state of stable variation themselves, using each grammar with a probability that reflects the relative frequency of the sentences associated with that grammar in the learner's input (barring some selectional factor external to the learning process). To see this, let x and $1 - x$ be the frequencies of

sentences associated with G_1 and G_2, respectively, in the learner's environment. Then in this case the penalty probabilities are simply $c_1 = 1 - x$ and $c_2 = x$, so that

$$(13) \quad \mathrm{E}[p]_\infty = \frac{x}{x + 1 - x} = x$$

In other words, the frequency of the learner's use of the different variants will eventually faithfully match that in their input.

Crucially for the model, however, we also need to consider cases where at least some of the sentences that a learner hears are *ambiguous* as to which grammar they are associated with. For example, an SVO sentence is compatible with an English-style SVO grammar, but also with a Verb-Second grammar where the underlying order is SOV (e.g. right-headed VP with obligatory movement to left-headed C or I where possible). It is a property of the learning model that the "advantage", in an evolutionary sense, of a grammar is a function of the proportion of the outputs of that grammar heard by the learner that are *unambiguously* attributable to that grammar. This is what, in Yang's model, drives change. We now describe how the model works in this case; for simplicity, the discussion is framed in terms of two competing grammars, but the logic is similar for multi-way competition.

First, given a mixture of output from two grammars, G_1 and G_2, a learner is expected to acquire both grammars, as discussed above. If the learner hears an *unambiguous* sentence—e.g. only G_1 could have produced the sentence—then if the learner picked G_1 beforehand and used it to try to analyse the sentence, G_1 will be rewarded. If the learner picked G_2, on the other hand, G_2 would be punished, corresponding to a reward for G_1. Either way, G_1 ends up with an augmented weight. However, if the learner encounters an *ambiguous* input—i.e. either G_1 or G_2 can analyse the string the learner hears—then the learner will reward whichever grammar they happen to have picked. Ultimately, as the task is iterated many times, the learner will most frequently augment the grammar that was most successful in analyzing unambiguous inputs. In consequence, in this case the weights that the learner associates with each grammar will *not* mirror the respective weights for the previous generation, but will have shifted in favor of the grammar which signals itself most unambiguously.

Formally, let α_1 represent the probability with which grammar G_1 generates outputs which are *only* compatible with itself, and let α_2 denote the corresponding probability for G_2. Again, let x and $1 - x$ represent the frequencies of use of the two grammars in the learner's environment. The penalty probabilities then assume the forms

$$(14) \quad c_1 = \alpha_2(1 - x) \quad \text{and} \quad c_2 = \alpha_1 x$$

The learner is thus expected to acquire

$$(15) \quad \mathrm{E}[p]_\infty = \frac{c_2}{c_1 + c_2} = \frac{\alpha_1 x}{\alpha_1 x + \alpha_2(1 - x)}$$

Hence if $\alpha_1 \neq \alpha_2$, the expectation $\mathrm{E}[p]_\infty$ is no longer identical to x. For a simple illustration of this—how the presence of ambiguous input will affect the learner's probability matching—suppose that $x = 0.5$, so that each grammar is equally frequent in the learner's environment. Then these frequencies cancel out in the expression for $\mathrm{E}[p]_\infty$, and we obtain

$$(16) \quad \mathrm{E}[p]_\infty = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

It follows that $\mathrm{E}[p] > 0.5$ if and only if $\alpha_1 > \alpha_2$; in other words, the learner boosts the frequency of a grammar if and only if that grammar generates a greater proportion of unambiguous output than its competitor.

3.1.2 Intergenerational evolution

This process can be iterated over a sequence of generations of learners. To do this, we write x_n for the frequency of G_1 in the nth generation, and define the frequency for the $n+1$th

generation as follows.

$$x_{n+1} = \frac{\alpha_1 x_n}{\alpha_1 x_n + \alpha_2(1 - x_n)} \quad (17)$$

It follows that $x_{n+1} > x_n$ if and only if $\alpha_1 > \alpha_2$ (see Unique ID WBCDL030 for more details), so the winner of a diachronic competition between two grammars over a long period of history will also be the one that can analyse the most unambiguous sentences in the input to learning. This is the same thing as saying that the most successful grammar is the one which generates the highest frequency of unambiguous outputs. Thus α_1 is a measure of the fitness or "advantage" (Yang 2000) of G_1; a positive difference between α_1 and α_2 is a "fitness differential" in favor of G_1 over G_2. (Note: we here follow Yang in using the term "advantage" for α_i; cf. Heycock and Wallenberg 2013, where this term is reserved for the fitness differential.)

When the learning process is iterated over a number of generations, the first generation of learners becomes the second generation's parents (or adult speech community, more generally) and determines the composition of the linguistic input to the second generation's learning. Importantly, the proportion of unambiguous sentences a grammar generates (i.e. its fitness) is decisive independently of the initial weights of G_1 and G_2; that is, the initial frequencies of G_1 and G_2 in the linguistic environment when the competition begins are not relevant for determining the winner. This correctly allows for the overwhelmingly most commonly observed situation in syntactic change (and change in other discrete linguistic features), where a new variant begins as a minority variant in the speech community.

The special cases $\alpha_1 = \alpha_2 = 1$ (each grammar signals itself unambiguously in every sentence) and $\alpha_1 = \alpha_2 \neq 1$ (each grammar parses an identical proportion of sentences) merit further discussion. As mentioned in Section 2.4, and as is clear from the form of equation (17), stable variation ($x_{n+1} = x_n$) is predicted by the model if (and only if) $\alpha_1 = \alpha_2$. In other words, stable variation is a possible outcome of grammar competition under certain circumstances, even when only two variants compete. The situation for multi-way competition is considerably more interesting: it can be shown that when the number of competing variants is at least three, stable equilibria may arise even for unequal α values (Kauhanen 2019).

We further point out that equation (17) can be solved, i.e. it is possible to express x_n as a closed-form function of the initial condition x_0, i.e. the frequency of G_1 in the population at generation 0, and the advantage parameters. The solution is:

$$x_n = \frac{1}{1 + \left(\frac{\alpha_2}{\alpha_1}\right)^n \left(\frac{1}{x_0} - 1\right)} \quad (18)$$

By introducing $s = \log(\alpha_1/\alpha_2)$ and $k = \log(1/x_0 - 1)$, we can also express this as in (19).

$$x_n = \frac{1}{1 + e^{-sn+k}} \quad (19)$$

This is now the familiar logistic equation describing the S-shaped propagation of a single variant at the expense of another (see Appendix A). The slope s of the curve (actually, a discrete sequence of points, as our model describes evolution in terms of generations) is given by the logarithm of the ratio of the advantage parameters—the greater the difference between the advantages, the greater the rate of change, whereas $\alpha_1 = \alpha_2$ implies $s = 0$ and hence stasis—while the intercept k is related to the initial condition.

3.1.3 Empirical applications

Yang (2002) discusses a number of different diachronic changes that he argues can be explained as being driven by this learning algorithm and the relentlessness of change it predicts, including the loss of V2 in Old French and its (partial) loss in Old English. As he himself

notes (see e.g. Yang 2002, 143), his model ignores additional factors that might affect language learning and hence also result in change. His model is not, however, incompatible with additional sociological (or indeed cognitive) factors playing a role—in principle, *anything* that affects the penalty probability of a grammar may have an effect on the ensuing population dynamics (for an application of this in the context of sociolinguistic typology, see Kauhanen 2022). Thus the model is perfectly consistent with the possibility of a change taking place that is not driven by the bias inherent in the parsing/learning mechanism—as well as with the possibility of the dynamics emerging from the interplay of a number of factors.

On the other hand, the variational learning model does make the strong prediction that when two competing grammars differ in how unambiguously they are represented in the learner's input, then in the absence of other intervening factors, this will lead to the grammar that produces the highest proportion of unambiguous "signatures" overtaking and eliminating its competitor. Heycock and Wallenberg (2013) made use of this strong hypothesis to show that the loss of V-to-T movement in the mainland North Germanic languages is predictable from the fitness differential (i.e. difference in advantage) between the non-movement, V-in-situ grammar and the V-to-T movement grammar during child language acquisition. Both competing variants signal themselves in a number of finite subordinate clauses which also contain diagnostic adverbs, as in (9) and (10) for the older V-to-T grammar and innovative V-in-situ grammar, respectively (relevant adverb in boldface, relevant finite verb underlined).

(9) sum iak baer **nu** vitni til
which I bear **now** witness to
"which I now bear witness to"
(Old Swedish, from *Westgötalagen*, cited in Sundquist 2002a)

(10) han skulle göra them, som hans plägsedh **altijdh** war
he should do them, that his custom **always** was
"he should do for them as/what he usually did"
(Mark 15:8, Swedish *Gustav Vasa Bible*, date: 1526/1541, *Fornsvenska Textbanken 1.0*)

There are also many clauses in these languages which optionally involve verb movement to a higher functional head ("embedded verb-second clauses"), regardless of which of the two competing variants is used. These sentences obscure some instances of the V-to-T variant in the input, as in (11) below. However, the newer V-in-situ variant need not be obscured in every such clause, and can still sometimes signal itself, as in (12).

(11) så at the hadhe **icke** tijdh til at äta.
so that they had **not** time for to eat.
"such that they didn't have time to eat"
(Mark 6:31, Swedish *Gustav Vasa Bible*, date: 1526/1541)

(12) at the ock **icke** komma vthi thetta pijno rwmet.
that they also **not** come into this torment room-the
"So that they should not also come into this place of torment."
(Luke 16:28, Swedish *Gustav Vasa Bible*, date: 1526/1541)

Heycock and Wallenberg (2013) show that this asymmetry in unambiguous signalling between the grammars results in a fitness differential in favor of the innovative, V-in-situ grammar compared with the older V-to-T variant. In this way, the starting conditions for acquisition predict that V-in-situ should gradually replace the V-to-T variant over generational time, once it was innovated. Such a change may in fact still be ongoing in Faroese (Heycock et al. 2012). In evolutionary terms, one could say that the fitness differential makes the V-to-T variant susceptible to invasion by a V-in-situ variant (though V-to-T can remain stably the only variant in a population until V-in-situ is innovated and escapes into usage at a learnable

level). This level of predictive power arguably did not exist in the field of language change prior to the frameworks of grammar competition and variational learning.

Finally, we note that the Heycock et al. (2012) investigation into Faroese provides indirect evidence that speakers probabilistically select from a set of internal grammars, as Yang's model proposes. That study argued that Faroese production data reflected the presence of competing V-to-T and V-in-situ grammars in the population. Interestingly, acceptability judgment experiments showed that Faroese speakers give intermediate acceptabilty judgments to the same types of sentences that display the mixture of two grammars in production. Such a result is predicted if both production and judgments involved speakers selecting from their internal set of grammars based on complementary probability weights that they learned for them in acquisition. The authors show that the experiments' results cannot be explained under a 1-invariant-grammar-per-speaker framework.

3.1.4 A challenge for future research: the subset problem

The precision with which the variational learning model allows us to state hypotheses about the course of change also allows us to frame an important problem for future research. Under the right circumstances, the strong variational learning hypothesis gives rise to the subset problem: if grammar G_1 generates a language that is a proper subset of the language generated by G_2, G_1 will always be out-competed by G_2. Truswell (2021) discusses this issue in the context of variation in Early Middle English: he discusses an attested dialect which allows for more different positions for phrases at the left edge of the clause than is possible in other contemporaneous dialects. He notes that as this permissive syntactic variant comes into competition with less permissive variants (as it is in some texts), the permissive grammar should have a greater advantage than its competitors under the hypothesis of variational learning. It should therefore rise over time and eventually exclude the other, more restrictive systems. But this is not what happened: after a period of time the evidence in texts for this more "flexible" grammar declined, and it was eventually lost entirely.

The historical case described by Truswell evidently poses a challenge to the model, therefore. But it should be noted that this is a general issue for the model, which arises even if there is no evidence in the input for the superset grammar. Given the assumption that at the beginning of the process of acquisition children entertain all possible grammars, superset grammars will never be penalized and so always remain as an option. In consequence, the model predicts that—all things being equal—grammars with relatively free word order will constantly be introduced by learners, regardless of their input—and as the productions of these learners constitutes the input for the next generation, such grammars should then outcompete more restrictive competitors. It may therefore be necessary to explore extensions of the basic model that incorporate some sort of economy factors. Alternatively, it may be that some apparent permissive variants are not single variants at all, but rather multiple variants in competition already, and that these states of variation should be reanalyzed.

3.1.5 Note on variable rules

It is worth mentioning that the sociolinguistics literature has proposed (and to some extent, assumed) the existence of grammatical rules which apply probabilistically, at least as far back as Labov (1972). These "variable rules", often allophonic rules in this literature, have some probability of application, which may be modulated by language-internal and sociolinguistic conditioning factors. Though not usually described just in this way, each variable rule also implies some complementary probability of non-application, which one can think of as an identity rule in e.g. an allophonic string-rewrite rule. For example, the existence of a stochastically applying rule that deletes final /t/ or /d/ in certain environments, as in the seminal application of variable rules to lexical phonology in Guy (1991), implies the existence of a complementary rule which stochastically realizes the /t/ or /d/ as [t] or [d], respectively, in the same environments (we thank Tony Kroch, p.c., for pointing this out). Thus, variable

rules are another case of competing grammars.

In a development somewhat parallel to variational learning in the sociolinguistics literature, Smith, Durham, and Fortune (2007) show convincingly that children estimate the probabilities of use of linguistic variants from their caregivers' linguistic behavior, presumably encoding these as base probabilities in variable rules (though this wording is ours). They also show that children match the modulation of application probabilities for different styles from their caregivers' speech in different social settings. This kind of data provides a strong argument that children represent probabilities along with each variant when they acquire linguistic variation, a result that would not be statable without a notion of competing grammars.

3.1.6 Note on reinforcement learning

We also point out that adopting the competing grammars framework immediately connects the study of language acquisition and language change with the rich literature on reinforcement learning. The linear reward–penalty learning procedure at the heart of variational learning originated in the context of mathematical psychology, as a simple model of instrumental conditioning (see Bush and Mosteller 1955). Importantly, it can also be viewed as a special case of more sophisticated learning algorithms studied in reinforcement learning theory (see Sutton and Barto 2018 for an introduction). The general reinforcement learning problem is normally formulated as one of maximizing accumulated reward; if reward is understood as parsing success, then this is precisely what a variational learner is engaged in. In view of the considerable success of reinforcement-based explanations of human psychology in general, we view this theoretical unification as a promising opportunity for further research into more sophisticated probabilistic learning algorithms and their diachronic consequences.

3.2 Evolutionary dynamics

In addition to allowing us to derive patterns and processes of language acquisition and change from underlying mechanistic principles, the competing grammars approach also allows us to view language change as a particular type of process of cultural evolution, leading to a form of theoretical unification in the direction of general mathematical models of evolutionary dynamics. *Evolution* means change through the differential replication of competing variants (whether biological or cultural); this differential replication may be the consequence of a fitness difference or may simply result from random drift in a pool of equally fit variants (in which case the "competition" may not be resolved). Under this conception, grammars constitute *replicators*; in a steady state of population dynamics, the frequency with which a particular grammar is used in a speech community reflects, in general, this grammar's success in replicating itself over time. Language users, in turn, are *interactors* (Hull 1988): replicators replicate through interactions between interactors. For instance, the entire period of language acquisition of a single individual may be conceptualized as a long sequence of interactions with other language users.

It is conceptually enlightening to contrast this process—change through replication—with another notion of change, change through deformation. Whereas ecological or population-dynamic change (the historically primary notion of evolution) consists of changes to the distribution of replicators, physical objects change in a different way, through material deformation (cf. Lewontin 1985). We believe it is significant that much traditional work in historical linguistics has proceeded on the analogy of change through deformation: languages—conceptualized simultaneously as atomic and Platonic entities with an existence which is independent of the minds of individuals—are said to change by proceeding from one state to another through a series of intermediate stages, not unlike the way in which a physical object may be deformed into a new shape through the application of the right sequence of forces (see e.g. Lass 1997, 277–281 on this view). Although this way of stating things en-

joys a certain economy of expression, ontologically it is seriously misguided: a language has no existence above and beyond a set of mental representations in an individual's mind, or the minds of a community, or perhaps several communities, of language users (see Walkden 2021 for discussion). Importantly, it follows that instances of language change cannot be explained by reference to processes or "laws" at the level of Platonic language-objects; by contrast, every observed regularity in language variation and change must, in principle, be reductively explainable as a consequence of the interactions of grammatical replicators in populations of language users. The point is more than conceptual: we maintain that *scientific progress* in explaining (as opposed to describing) processes of language change hinges crucially on taking this difference seriously. Grammar competition, and its extension in variational learning, are frameworks which explicitly distinguish replicators from interactors, and take these rather than "a language" as the objects of study.

To illustrate this in a bit more detail, we will briefly point out two ways in which variational learning and grammar competition connect with the theory of general evolutionary dynamics. This has to do with the shape of inter-generational evolution predicted by the learning model, as embodied in equation (17). Writing equation (20), we note that $\varphi(x_n)$ can be interpreted as the *average advantage* obtaining in the speech community: it is simply the frequency-weighted average of the two advantage quantities α_1 and α_2.

$$\varphi(x_n) = \alpha_1 x_n + \alpha_2(1 - x_n) \tag{20}$$

This allows us to rewrite equation (17) as

$$x_{n+1} = \frac{\alpha_1}{\varphi(x_n)} x_n \tag{21}$$

It is now evident that $x_{n+1} > x_n$, and hence the frequency of grammar G_1 increases, if and only if the advantage of G_1 is greater than average, i.e.:

$$\alpha_1 > \varphi(x_n) \tag{22}$$

(In fact, a little algebra shows that in this constant-fitness case, the condition $\alpha_1 > \varphi(x_n)$ is equivalent to $\alpha_1 > \alpha_2$.) Conversely, the frequency of the grammar decreases if and only if its advantage is below average.

Equations of the form (21) are known as *replicator equations* in the literature on evolutionary game theory (e.g. Sandholm 2010). While it is in and of itself significant that a linguistically relevant learning model gives rise to such an equation over inter-generational iteration, it is equally important to note that the simple equation (21) admits of straightforward generalization. By replacing the constant advantage parameters α_1 and α_2 by advantage functions which are sensitive to the current frequencies of the two grammars in the population, it becomes possible to model *frequency-dependent* selection (for more details, see Unique ID WBCDL030), a well-established effect in some biological systems. (A linguistic example might be a sociolinguistically salient variant whose probability of adoption increases as its frequency in the population increases.) Such equations allow for more complex behavior, such as the simultaneous existence of multiple equilibria, which are only locally stable (as opposed to the globally stable equilibrium of equation 17). Among other things, this makes it possible to derive stable variation within the competing grammars framework—even when only two variants compete. For linguistic applications of this line of thinking (one in phonology, one in sociolinguistics), see Baumann and Ritt (2017) and Kauhanen (2020).

Much work in linguistics employing variational learning has focused on the "hydrodynamic" limit in which the model's predictions become deterministic—indeed, the original applications in Yang (2000), as well as a number of studies inspired by those original applications (Heycock and Wallenberg 2013; Danckaert 2017; Simonenko, Crabbé, and Prévost 2019; Kauhanen 2022), do this by assuming that the noise inherent in learning disappears in large populations. While there are good reasons for accepting this approximation, especially when advantage differences between competing grammars are large and hence dominate the

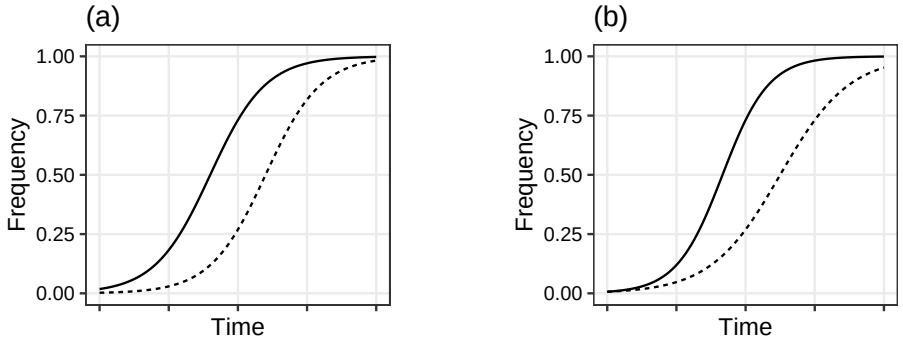

Figure 2: Constant (a) and inconstant (b) rates.

random fluctuations, there remain cases in which it will be useful to study the stochastic, finite dynamics of small populations. We point out that the variational learning and competing grammars frameworks—even though often discussed in the context of deterministic approximations—are fully compatible with this. Above, we have mentioned that a closed-form solution exists for the expected value of the grammar weights when learners employ the linear reward–penalty procedure. However, a closed-form solution also exists for the variance (see Bush and Mosteller 1955); hence it becomes possible to analytically study (and not just simulate), for instance, the effect of introducing finite learning periods instead of the infinite limits assumed in deterministic approximations.

3.3 The Constant Rate Effect

The competing grammars framework also allows us to make sense of rates of change in language dynamics. Specifically, the ideas that (i) grammatical variants are the replicators in language change, (ii) multiple variants may exist in a single language user (interactor), and (iii) one variant might replace another in the population over generational time, allow us to state and evaluate hypotheses concerning where one expects and where one does not expect the rates of change among multiple replacement processes to be identical.

Kroch (1989), studying the emergence and spread of English *do*-support based on data from Ellegård (1953), observed that different contexts (e.g. questions, negative declaratives) seemed to be characterized by different S-curves, but that these S-curves were linked by a crucial property: they all had the same slope. This led to the formulation of the Constant Rate Hypothesis (CRH):

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

As we have seen, the shape and position of logistic curves is governed by an equation with two parameters: the slope (s), which determines how steep or shallow the S-curve is, and the intercept (k), which determines where the S-curve meets the vertical axis (see equation 19). The "rate of replacement" in the above quote from Kroch is understood as the slope parameter s (which, as we have seen above, can be related to the advantage differential obtaining between the competing variants). So a slightly more explicit way of stating the Constant Rate Hypothesis is that, when fitting logistic curves to different linguistic contexts, if the different contexts exhibit reflexes of a single underlying grammatical alternation then the slope parameters of these curves should not differ. A hypothetical Constant-Rate-compatible scenario is illustrated in Figure 2(a); a state of affairs that is incompatible with a Constant Rate Effect is illustrated in Figure 2(b).

The Constant Rate Hypothesis was developed at the height of early Principles & Parameters, when the search was on for "the controlling effect of abstract grammatical analyses

Table 1: Studies investigating the Constant Rate Hypothesis.

Study	Language	Phenomenon	CRE claimed?
Kroch (1989)	English	*do*-support	Y
Kroch (1989)	English	British *have* vs. *have got*	Y
Santorini (1993)	Yiddish	IP headedness	Y
Ball (1994)	English	*wh*-complementizer	Y
Taylor (1994)	Greek	IP/VP headedness	Y
Pintzuk (1995)	English	VP headedness	Y
Wagner (1996)	Portuguese	*haver* vs. *ter*	Y
Frisch (1997)	English	Jespersen's Cycle	Y/N
Wallage (2008, 2013)	English	Jespersen's Cycle	Y
Cukor-Avila (2002)	English	AAE quotatives	Y/N
Sundquist (2002b, 2006)	Norwegian	VP headedness	Y
Breitbarth (2005)	German	Afinite construction	Y/N
Kallel (2005, 2007)	English	Negative concord	Y
Pintzuk and Taylor (2006)	English	VP headedness	N
Sundquist (2007)	German	Jespersen's Cycle	N
Tagliamonte and D'Arcy (2007)	English	Toronto quotatives	Y/N
Heusinger (2008)	Spanish	Differential Object Marking	Y
Durham et al. (2012)	English	British quotatives	Y
Fruehwald, Gress-Wright, and Wallenberg (2013)	German	Final fortition	Y
Wolk et al. (2013)	English	Genitives	N
Gardiner (2015)	Egyptian	Possessives	Y
Hundt (2015)	English	*do*-support (Aus, NZ)	N
Wallenberg (2013)	Yiddish	IP headedness and object position	Y
Wallenberg (2016)	Various	Relative clause extraposition	Y
Willis (2017)	Welsh	Pronoun *chdi*	Y
Simonenko, Crabbé, and Prévost (2018)	French	VP headedness	Y
Simonenko, Crabbé, and Prévost (2019)	French	Null subjects and agreement	N
Zimmermann (2022)	English	Possessive *have*	Y

on patterns in usage data" (Kroch 1989, 239). It has stimulated a wide variety of studies on different linguistic variables in different languages: Table 1 provides a listing of the studies we are aware of. In addition to the studies in Table 1, we make an interesting historical note: Lindgren (1953), in his study of the loss of final schwa in Middle High German, derived a number of similar S-shaped curves for that change in different environments from hand-collected manuscript data. His plots appear to show a CRE across word classes, and he seemed aware that the curves could be modelled with the logistic function (or a similar function). Lindgren also anticipates the concept of grammar competition in his description that S-shaped curves are produced as older forms and newer forms "kämpfen" (*fight*) (Lindgren 1953, 186). He stopped just short, however, of explicitly pursuing the hypothesis that the curves he derived for the different word classes have the same slope. He also may well have found a CRE in the diphthongization of Early New High Bavarian long vowels in Lindgren (1961); see also discussion and modelling of his data in Best (2008).

The Constant Rate Hypothesis has led to a large amount of theoretically informed empirical work, but at the same time has not been without controversy. First, it is not clear what the operationalization of the hypothesis stated above—in terms of "same slope, different intercept"—actually follows from, beyond the intuition that different surface realizations of one underlying parameter should share something in their trajectories of change. (Hypothesizing same slopes was also a conscious, direct challenge to Bailey 1973's different rates of change hypothesis.) Kauhanen and Walkden (2018) address this point. By extending Yang's (2002) variational learning model (discussed above in Section 3.1) so that it incorporates constant, context-specific biases that affect production by favoring one variant over

another, Kauhanen and Walkden (2018, §2) are able to derive behavior that is almost indistinguishable from the classical operationalization of the Constant Rate Effect in Kroch (1989). An added bonus of this model is that it derives an upper bound on the time separation between contexts (i.e. approximately the difference between intercepts, k, in the classical operationalization). In this respect, Kauhanen and Walkden's model is more restrictive than the classical operationalization of Kroch (1989); see Kauhanen and Walkden (2018, §4) for details. For a different view of the Constant Rate Effect, related to the notion of families of solutions of differential equations, see Postma (2017).

The second point of controversy relates to how "same slope" is actually established for a given dataset. There are two main methods of investigating this: likelihood ratios, as in Kroch (1989), and logistic regression, as in Fruehwald, Gress-Wright, and Wallenberg (2013). Standard practice in null hypothesis significance testing involves investigating a null hypothesis of identity across different conditions; the p-value provided by inferential statistical tests is a measure of how likely the observed state of affairs is under the assumption that the null hypothesis is true. A low p-value for an effect can lead one to reject the null hypothesis. But in the case of the Constant Rate Hypothesis it is exactly the null hypothesis that we want to establish, and since we are assuming its truth, simply failing to reject the null hypothesis "invites type 2 error of unknown probability" (Paolillo 2011, 266), i.e. a false negative.

This thorny problem can in principle be overcome if we have an independent way of estimating the probability of type 2 error, β. This probability is linked to the statistical power of the test, $1 - \beta$. Hence, power analysis based on simulated datasets can be used for this purpose (Ecay 2015, 36–40). Since statistical power is affected by sample size, and since the temporal resolution of studies pursuing the CRH is not typically very high (the median across the studies listed in Table 1, wherever we have access to this information, is 4 time periods), many existing studies of the CRH may in fact be underpowered for the purpose—including the original study of *do*-support by Kroch (1989) (Ecay 2015, 63–66).

To illustrate the challenge, we present the results of a Monte Carlo power analysis. For this, we simulate two contexts of a logistic change and ask how much data needs to be observed in order for a slope difference of a given size to be reliably detected using the customary statistical method (logistic regression with an interaction between the time and context variables). Without loss of generality, we set the slope of one context to $s_1 = 1$ and the slope of the other context to $s_2 = (1 - E)s_1 = 1 - E$, where E quantifies the effect size. To translate the slopes into quantities with a more natural interpretation, we follow Ecay (2015) and focus on their inverses, which relate to the time it takes for the change to propagate. If Δt_1 and Δt_2 are the propagation times for the two contexts, we then have

$$\frac{\Delta t_2}{\Delta t_1} = \frac{s_1}{s_2} = \frac{1}{1 - E} \tag{23}$$

(see Appendix B). Setting $E = 0.2$ then corresponds to a 25% increase in propagation time ($1/0.8 = 1.25$). We maintain that no one would seriously consider two contexts to show a CRE if their propagation times diverged by such an amount, and hence concentrate on effect sizes in the range from 0 to 0.2.

We then pick different values of E from this range, as well as overall sample sizes N between 100 and 100,000. We assume the data points to be equally distributed between the two contexts, over a total of four time periods. We then simulate synthetic data based on these assumptions, repeat the exercise 1,000 times for each combination of E and N, and calculate the proportion of regressions in which the interaction effect is detected at level $\alpha = 0.05$. This proportion is the power of the test, i.e. probability of correctly detecting an interaction effect when one exists. Figure 3 shows the results—the main takeaway being that considerably large sample sizes—larger than are normally possible in historical corpus studies—are required to reach reasonable levels of statistical power.

The above observations lend weight to Zimmermann's (2022, 5) claim that "all previous examinations of the CRH have serious shortcomings". Zimmermann (2022) himself

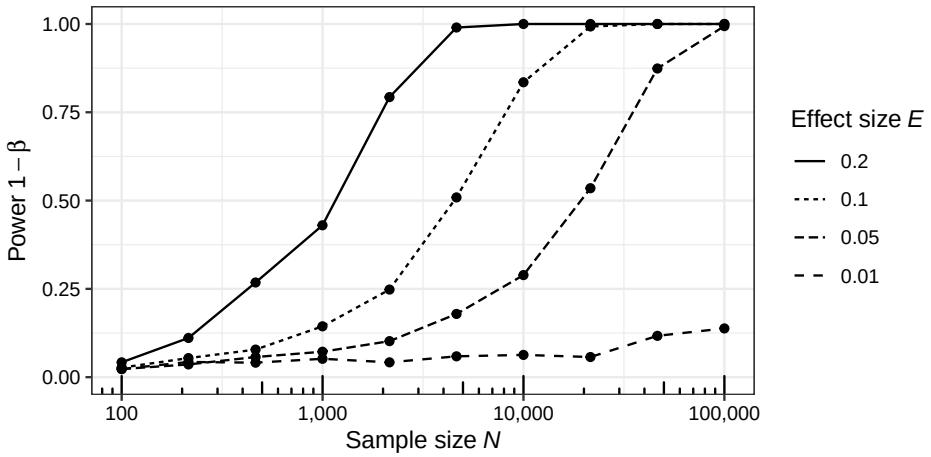

Figure 3: Dependence of statistical power on sample size and effect size in the detection of CRH-violating changes. See text for simulation assumptions.

presents an improved test of the CRH based on data from the Corpus of Historical American English (COHA; Davies 2010) at a very high resolution (with a total sample size of over 50,000), investigating possessive *have*. The high *have* variant with V-to-T movement (e.g. *I have not the power*) declines at the expense of a lower, in-situ variant (e.g. *I don't have the power*) across four contexts—negation, inversion, VP-adjuncts and VP-ellipsis—in a way that is entirely consistent with the CRH. More studies of this quality are needed in order to investigate whether the CRH holds more generally; it is also possible to explore other methods of inferential statistics that do not rely on the notion of null hypothesis testing at all (see Ecay 2015; Bacovcin 2017; Wallenberg et al. 2021; Kauhanen, in press).

These methodological challenges notwithstanding, the CRH remains an important, empirically testable hypothesis about the relationship between corpus (E-language) frequencies and cognitive (I-language) representations. Importantly, the CRH only appears to be statable within the competing grammars framework—while the framework as outlined in Section 1 does not entail the CRH, the CRH, for all intents and purposes, entails the competing grammars paradigm (or something so similar as to be indistinguishable). This is simply because, without the notion of variable, probabilistic representations at the individual level, the context-sensitive modulation of empirical frequencies the CRH talks about remains mysterious: if language users were fully categorical, then the only way they could boost the use of a variant in a given context would be by flipping the probability from 0 to 1. If all the language users in the relevant speech community were to do this, then what we would observe at the level of the E-language would be total (i.e. maximal) divergence between two contexts: one would have the ceiling frequency of 1, the other the floor frequency of 0. The only way to escape this conclusion with categorical users would appear to be to suggest that only some portion of users are sensitive to the context-sensitive modulation of their linguistic behavior. The CRH then becomes a hypothesis about sociolinguistics rather than one about competence at the individual level. While this is a possible stance to take, we submit it not only goes against the original intention behind the CRH, but also encounters significant empirical difficulties: for instance, why would the sociolinguistically motivated contextual modulation remain constant over time? For reasons such as this, the competing grammars framework is an important part of the diachronic linguist's toolkit—that the framework gives us the wherewithal to explore hypotheses such as the CRH is a point in its favor.

4 Conclusion

The competing grammars framework maintains that language acquirers and users acquire probabilities which they attach to a number of competing variants of linguistic variables—

ordinarily (but as we have seen, perhaps unhelpfully) referred to loosely as "grammars" in the literature—and that language change is nothing but change in these probabilities over time. Here we have shown how some common objections to this paradigm can be dissolved, as well as what benefits can be reaped by working within the paradigm. These benefits include theoretical unification with mathematical models of learning and evolutionary dynamics, as well as the ability to state, examine and test specific hypotheses thanks to quantitative predictions that the paradigm makes available. These include predictions about which grammar will win out in a given diachronic situation, what the propagation time of a change is likely to be, what learning trajectories at the individual level should look like, how child language acquisition trajectories relate to the innovation and propagation time of new variants, how language-external factors can be expected to interact with language-internal factors such as parsing advantage, and specific quantitative hypotheses such as the Constant Rate Hypothesis. This way of thinking also helps the analyst avoid the pitfalls of reifying "a language" as the object of study, and the consequent positing of mysteriously transient hybrid grammatical objects to account for how "a language" appears in the middle of a change in progress in variant frequencies. All of this serves to increase falsifiability, a welcome outcome for any theory.

Data and code availability

R (R Core Team 2021) code for replicating the learning simulations in Section 3.1 and the power analysis in Section 3.3 can be obtained from https://doi.org/10.5281/zenodo.8356506.

See also

WBCDL001
WBCDL030

A The S-curve

In this appendix, we prove that (19) is the solution to (17), and hence that under the simple intergenerational dynamics derived from variational learning, change has the shape of an S-curve, more specifically, that of the logistic equation. Although the result is by no means new (in fact, it could be called "trivial"), the proof is not easy to find in existing literature and hence we feel it pays to present the argument here in full mathematical detail.

Dividing both the numerator and denominator of the righthand side of (17) by $\alpha_1 x_n$ (on the assumption that $\alpha_1 \neq 0 \neq x_n$), we obtain

$$x_{n+1} = \frac{1}{1 + \frac{\alpha_2}{\alpha_1}\frac{1-x_n}{x_n}} \tag{24}$$

or

$$x_{n+1} = \frac{1}{1 + \rho\left(\frac{1}{x_n} - 1\right)} \tag{25}$$

where we have introduced $\rho = \alpha_2/\alpha_1$. Let $z_n = 1/x_n$. Then

$$z_{n+1} = 1 + \rho(z_n - 1) \tag{26}$$

We now claim that

$$z_n = 1 + \rho^n(z_0 - 1) \tag{27}$$

for all $n \geq 0$ and prove this by induction. The claim is obviously true for $n = 0$.

Then assume the claim holds for $n = m$. We then have

$$\begin{aligned} z_{m+1} &= 1 + \rho(z_m - 1) \\ &= 1 + \rho(1 + \rho^m(z_0 - 1) - 1) \\ &= 1 + \rho\rho^m(z_0 - 1) \\ &= 1 + \rho^{m+1}(z_0 - 1) \end{aligned} \tag{28}$$

and so the claim holds for $n = m + 1$ as well, which completes the induction.

Changing back to the original x-variables, we thus have

$$x_n = \frac{1}{1 + \rho^n \left(\frac{1}{x_0} - 1\right)} \tag{29}$$

To see that this is an S-curve, we note that

$$\rho^n \left(\frac{1}{x_0} - 1\right) = e^{\log\left(\rho^n\left(\frac{1}{x_0}-1\right)\right)} = e^{\log(\rho^n)+\log\left(\frac{1}{x_0}-1\right)} = e^{-n\log\left(\frac{1}{\rho}\right)+\log\left(\frac{1}{x_0}-1\right)} \tag{30}$$

Thus by choosing $s = \log(1/\rho) = \log(\alpha_1/\alpha_2)$ and $k = \log(1/x_0 - 1)$ we find that

$$x_n = \frac{1}{1 + e^{-sn+k}} \tag{31}$$

This is the equation for the logistic function (with time measured in generations n).

B CRH power analysis

The following procedure was adopted for the power analysis reported in Section 3.3. We assume a logistic change in two contexts, with slopes s_1 and s_2 and intercepts k_1 and k_2; more formally, the probability of grammar G_1 in context c at time t is written

$$p_c(t) = \frac{1}{1 + e^{-s_c t + k_c}} \tag{32}$$

Without loss of generality, we choose $s_1 = 1$ and $k_1 = 0$, and consider four equally spaced time points between the extremes $t = -3$ and $t = 3$. This implies that the probabilities of G_1 at the extremal time points in the first context are approximately $p_1(-3) \approx 0.05$ and $p_1(3) \approx 0.95$. This means we track most of the change as it propagates through the speech community.

For the second context, we set $s_2 = (1 - E)s_1 = 1 - E$ where E is the *effect size* parameter. E directly measures the difference in the slopes between the two contexts, and hence evidence for a non-zero E constitutes evidence against the Constant Rate Hypothesis. We then set $k_2 = -3E$. This choice implies that $p_1(-3) = p_2(-3)$, i.e. the probability of G_1 is the same in the two contexts initially. As E is increased, the slope of the second context decreases, leading to a slower change in that context.

The time it takes for p to increase from $p = \delta$ to $p = 1 - \delta$ is

$$\Delta t = \frac{2\log(1/\delta - 1)}{s} \tag{33}$$

as can be directly verified through equation (32). We will call this the *propagation time*. The ratio of the propagation times for the two contexts is

$$\frac{\Delta t_2}{\Delta t_1} = \frac{s_1}{s_2} \tag{34}$$

i.e. the dependency on δ cancels out. With our choices of s_1 and s_2, we have

$$\frac{\Delta t_2}{\Delta t_1} = \frac{1}{1 - E} \tag{35}$$

For $0 < E < 1$, this measures how much longer the change takes in context 2, compared to context 1.

To generate synthetic data for purposes of the power analysis, we then sample, for each time point t and each context c, $N/8$ times from the Bernoulli distribution with success probability $p_c(t)$, where N is the total sample size. (We sample $N/8$ data points since we have two contexts and four time points.) The synthetic data are then analysed using ordinary logistic regression with time and context as independent variables, including their interaction. If the interaction coefficient is significant at the customary $\alpha = 0.05$ level, we record this outcome as a correct rejection of the null hypothesis of a CRH; otherwise, we have a false negative. We generate 100 datasets and regressions in this way for each choice of total sample size N and effect size E, and record the proportion of correct decisions in the regression. This proportion is our estimate of the power $1 - \beta$ of the test.

References

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

Bailey, Charles-James N. 1973. *Variation and linguistic theory.* Arlington, VA: Center for Applied Linguistics.

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

Baumann, A., and N. Ritt. 2017. On the replicator dynamics of lexical stress: accounting for stress-pattern diversity in terms of evolutionary game theory. *Phonology* 34 (3): 439–471.

Best, Karl-Heinz. 2008. XXXIV. Kaj Brynolf Lindgren (1922-2007). *Glottometrics* 16:127–131.

Biberauer, Theresa, Anders Holmberg, and Ian Roberts. 2014. A syntactic universal and its consequences. *Linguistic Inquiry* 45 (2): 169–225.

Biberauer, Theresa, and Ian Roberts. 2005. Changing EPP-parameters in the history of English: accounting for variation and change. *English Language and Linguistics* 9:5–46.

Boas, Hans C., and Ivan Sag. 2012. *Sign-based Construction Grammar.* Stanford, CA: CSLI Press.

Borer, Hagit. 1984. *Parametric syntax.* Dordrecht: Foris.

Breitbarth, Anne. 2005. Live fast, die young: the short life of Early Modern German auxiliary ellipsis. PhD diss., Universiteit van Tilburg.

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

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

———. 1986. *Barriers.* Cambridge, MA: MIT Press.

———. 1995. *The minimalist program.* Cambridge, MA: MIT Press.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English.* New York, NY: Harper & Row.

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

Danckaert, Lieven. 2017. The loss of Latin OV: steps towards an analysis. In *Elements of comparative syntax: theory and description,* edited by Enoch Aboh, Eric Haeberli, Genoveva Puskás, and Manuela Schönenberger, 401–446. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9781501504037-015.

Davies, Mark. 2010. Corpus of Historical American English (COHA). https://www.english-corpora.org/coha/.

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

Ecay, Aaron W. 2015. A multi-step analysis of the evolution of English *do*-support. PhD diss., University of Pennsylvania. http://repository.upenn.edu/edissertations/1049/.

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

Estes, William K. 1976. The cognitive side of probability learning. *Psychological Review* 83:37–64.

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

Fritzenschaft, Agnes, Ira Gawlitzek-Maiwald, Rosmarie Tracy, and Susanne Winkler. 1990. Wege zur komplexen Syntax. *Zeitschrift für Sprachwissenschaft* 9 (1 and 2): 52–134.

Fruehwald, Josef. 2013. Phonological involvement in phonetic change. PhD diss., University of Pennsylvania.

———. 2016. The early influence of phonology on a phonetic change. *Language* 92 (2): 376–410.

Fruehwald, Josef, Jonathan Gress-Wright, and Joel C. Wallenberg. 2013. Phonological rule change: the Constant Rate Effect. In *NELS 40: Proceedings of the 40th Annual Meeting of the North East Linguistic Society,* edited by Seda Kan, Claire Moore-Cantwell, and Robert Staubs, 219–230. Amherst, MA: GLSA Publications.

Gardiner, Shayna. 2015. Taking possession of the Constant Rate Hypothesis: variation and change in Ancient Egyptian possessive constructions. *University of Pennsylvania Working Papers in Linguistics* 21 (2): 69–78. https://repository.upenn.edu/pwpl/vol21/iss2/9.

Glaser, Elvira. 1985. *Graphische Studien zum Schreibsprachwandel vom 13. bis 16. Jahrhundert.* Heidelberg: Carl Winter.

Goldberg, Adele E. 1995. *Constructions: a Construction Grammar approach to argument structure.* Cambridge, MA: University of Chicago Press.

Guy, Gregory R. 1991. Explanation in variable phonology: an exponential model of morphological constraints. *Language Variation and Change* 3 (1): 1–22.

Hale, Mark. 2007. *Historical linguistics: theory and method.* Malden, MA: Blackwell.

Henry, Alison. 2002. Variation and syntactic theory. In *The handbook of language variation and change,* edited by Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 267–282. Oxford: Blackwell.

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

Heycock, Caroline, Antonella Sorace, Zakaris Svabo Hansen, Frances Wilson, and Sten Vikner. 2012. Detecting the late stages of syntactic change: The loss of V-to-T in Faroese. *Language.*

Heycock, Caroline, and Joel Wallenberg. 2013. How variational acquisition drives syntactic change: the loss of verb movement in Scandinavian. *Journal of Comparative Germanic Linguistics* 16:127–157.

Hudson, Richard K. 1997. Inherent variability and linguistic theory. *Cognitive Linguistics* 8 (1): 73–108.

Hudson Kam, Carla L. 2015. The impact of conditioning variables on the acquisition of variation in adult and child learners. *Language* 91 (4): 906–937.

Hudson Kam, Carla L., and Elissa Newport. 2005. Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development* 1:151–195.

Hull, David L. 1988. *Science as a process: an evolutionary account of the social and conceptual development of science.* Chicago, IL: University of Chicago Press.

Hundt, Marianne. 2015. *Do*-support in early New Zealand and Australian English. In *Grammatical change in English world-wide,* edited by Peter Collins, 65–86. Amsterdam: John Benjamins.

Kallel, Amel. 2005. The loss of negative concord and the Constant Rate Hypothesis. *University of Pennsylvania Working Papers in Linguistics* 10 (2): 128–142. https://repository.upenn.edu/pwpl/vol10/iss2/11/.

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

Kauhanen, Henri. 2019. Stable variation in multidimensional competition. In *The determinants of diachronic stability,* edited by Anne Breitbarth, Miriam Bouzouita, Lieven Danckaert, and Melissa Farasyn, 263–290. Amsterdam: Benjamins.

———. 2020. Replicator–mutator dynamics of linguistic convergence and divergence. *Royal Society Open Science* 7:201682. https://doi.org/10.1098/rsos.201682.

———. 2022. A bifurcation threshold for contact-induced language change. *Glossa: a journal of general linguistics* 7 (1): 1–32. https://doi.org/10.16995/glossa.8211.

———. in press. Grammar competition, speaker models and rates of change: a critical reappraisal of the Constant Rate Hypothesis. *Journal of Historical Syntax* 7.

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

Kayne, Richard. 1994. *The antisymmetry of syntax.* Cambridge, MA: MIT Press.

Kemenade, Ans van. 1987. *Syntactic case and morphological case in the history of English.* Dordrecht: Foris.

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

Kroch, Anthony S. 1994. Morphosyntactic variation. In *Proceedings of the 30th annual meeting of the Chicago Linguistic Society,* edited by Katherine Beals, Jeannette Denton, Robert Knippen, Lynette Melnar, Hisami Suzuki, and Erica Zeinfeld, 180–201. Chicago, IL: Chicago Linguistic Society.

———. 2000. Syntactic change. In *The handbook of contemporary syntactic theory,* edited by Mark Baltin and Chris Collins, 629–739. Oxford: Blackwell.

Kroch, Anthony S., Beatrice Santorini, and Lauren Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. University of Pennsylvania. https://www.ling.upenn.edu/histcorpora/PPCEME-RELEASE-3/.

Kroch, Anthony S., and Ann Taylor. 1997. Verb movement in Old and Middle English: dialect variation and language contact. In *Parameters of morphosyntactic change,* edited by Ans van Kemenade and Nigel Vincent, 297–325. Cambridge: Cambridge University Press.

———. 2000. Penn-Helsinki Parsed Corpus of Middle English. University of Pennsylvania. https://www.ling.upenn.edu/histcorpora/PPCME2-RELEASE-4/.

Labov, William. 1972. *Language in the inner city: studies in the Black English vernacular.* Philadelphia, PA: University of Pennsylvania Press.

———. 1989. The child as linguistic historian. *Language Variation and Change* 1 (1): 85–97.

———. 1994. *Principles of linguistic change.* Vol. 1: Internal factors. Oxford: Blackwell.

———. 2001. *Principles of linguistic change.* Vol. 2: Social factors. Oxford: Blackwell.

Lass, Roger. 1997. *Historical linguistics and language change.* Cambridge: Cambridge University Press.

Lewontin, Richard C. 1985. The organism as the subject and object of evolution. In *The dialectical biologist,* edited by Richard Levins and Richard Lewontin. Cambridge, MA: Harvard University Press.

Lindgren, Kaj B. 1953. *Die Apokope des mittelhochdeutschen –e in seinen verschiedenen Funktionen.* Helsinki: Suomalainen Tiedeakatemia.

———. 1961. *Die Ausbreitung der nhd. Diphthongierung bis 1500.* Suomalaisen tiedeakatemian toimituksia/ Annales academiae scientiarum fennicae; Sarja/Ser. B, Nide/ Tom. 123,2). Helsinki.

Nevins, Andrew, and Jeffrey Parrott. 2010. Variable rules meet impoverishment theory. *Lingua* 120:1135–1159.

Paolillo, John C. 2011. Independence claims in linguistics. *Language Variation and Change* 23:257–274. https://doi.org/10.1017/S0954394511000081.

Pintzuk, Susan. 1991. Phrase structures in competition: variation and change in Old English word order. PhD diss., University of Pennsylvania.

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

———. 1999. *Phrase structures in competition: variation and change in Old English word order.* New York: Garland.

———. 2005. Arguments against a universal base: evidence from Old English. *English Language and Linguistics* 9:115–138.

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

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

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

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

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

Roberts, Ian. 1997. *Comparative syntax.* London: Arnold.

———. 2021. *Diachronic syntax.* 2nd ed. Oxford: Oxford University Press.

Saffran, Jenny R., Richard N. Aslin, and Elissa Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274:1926–1928.

Sandholm, William H. 2010. *Population games and evolutionary dynamics.* Cambridge, MA: MIT Press.

Santorini, Beatrice. 1989. The generalization of the verb-second constraint in the history of Yiddish. PhD diss., University of Pennsylvania.

———. 1992. Variation and change in Yiddish subordinate clause word order. *Natural Language and Linguistic Theory* 10:595–640.

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

Sheehan, Michelle, Theresa Biberauer, Anders Holmberg, and Ian Roberts. 2017. *The Final-over-Final Condition: a syntactic universal.* Cambridge, MA: MIT Press.

Simonenko, Alexandra, Benoît Crabbé, and Sophie Prévost. 2018. Text form and grammatical changes in Medieval French: a treebank-based diachronic study. *Diachronica* 35 (3): 393–428.

———. 2019. Agreement syncretization and the loss of null subjects: quantificational models for Medieval French. *Language Variation and Change* 31:275–301.

Smith, Jennifer, Mercedes Durham, and Liane Fortune. 2007. "Mam, my trousers is fa'in doon!": community, caregiver, and child in the acquisition of variation in a Scottish dialect. *Language Variation and Change* 19 (1): 63.

Sundquist, John D. 2002a. Morphosyntactic change in the history of the mainland Scandinavian languages. PhD diss., Indiana University.

———. 2002b. Object shift and Holmberg's Generalization in the history of Norwegian. In *Syntactic effects of morphological change,* edited by David W. Lightfoot, 326–349. Oxford: Oxford University Press.

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

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

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement learning: an introduction.* 2nd ed. Cambridge, MA: MIT Press.

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

Taylor, Ann. 1994. Variation in past tense formation in the history of English. *University of Pennsylvania Working Papers in Linguistics* 1:143–159.

Truswell, Robert. 2021. Grammar competition and word order in a Northern Early Middle English text. *Languages* 5 (59). https://doi.org/https://doi.org/10.3390/languages6020059.

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

Walkden, George. 2014. Object position and Heavy NP Shift in Old Saxon and beyond. In *Information structure and syntactic change in Germanic and Romance languages,* edited by Kristin Bech and Kristine Gunn Eide, 313–340. Amsterdam: John Benjamins.

———. 2021. Against mechanisms: towards a minimal theory of change. Special issue: Whither Reanalysis?, *Journal of Historical Syntax* 5 (33): 1–27.

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

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

Wallenberg, Joel C. 2009. Antisymmetry and the conservation of c-command: scrambling and phrase structure in synchronic and diachronic perspective. PhD diss., University of Pennsylvania.

———. 2013. Scrambling, LF, and phrase structure change in Yiddish. *Lingua* 133:289–318.

———. 2016. Extraposition is disappearing. *Language* 92 (4): e237–e256.

Wallenberg, Joel C., Rachael Bailes, Christine Cuskley, and Anton Karl Ingason. 2021. Smooth signals and syntactic change. *Languages* 6 (2): 60. https://doi.org/10.3390/languages6020060.

Whitman, John, Dianne Jonas, and Andrew Garrett. 2012. Introduction. In *Grammatical change: origins, nature, outcomes,* edited by Dianne Jonas, John Whitman, and Andrew Garrett, 1–12. Oxford: Oxford University Press.

Willis, David W. E. 2017. Endogenous and exogenous theories of syntactic change. In *The Cambridge handbook of historical syntax,* edited by Adam Ledgeway and Ian Roberts, 491–514. Cambridge: Cambridge University Press.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: exploring cross-constructional variation and change. *Diachronica* 30 (3): 382–419. https://doi.org/https://doi.org/10.1075/dia.30.3.04wol.

Yang, Charles D. 2000. Internal and external forces in language change. *Language Variation and Change* 12 (3): 231–250.

———. 2002. *Knowledge and learning in natural language.* Oxford: Oxford University Press.

Zimmermann, Richard. 2022. An improved test of the Constant Rate Hypothesis: late Modern American English possessive *have*. *Corpus Linguistics and Linguistic Theory,* 1–30. https://doi.org/10.1515/cllt-2021-0038.