This is a preprint - please use the following citation of the final print version, found at: https://doi.org/10.1016/j.cognition.2016.11.001

Cuskley, C., Castellano, C., Colaiori, F., Loreto, V., Pugliese, M. and Tria, F. (2017). The regularity game: investigating linguistic rule dynamics in a population of interacting agents. Cognition, 159, 25-32.

The regularity game: Investigating linguistic rule dynamics in a population of interacting agents

Christine Cuskley[*,1,3], Claudio Castellano[1,2], Francesca Colaiori,[1,2]
Vittorio Loreto[2,3,4], Martina Pugliese[2], Francesca Tria[3]

[1] Istituto dei Sistemi Complessi (ISC-CNR), via dei Taurini 19, I-00185 Rome, Italy
[2] Dipartimento di Fisica, Sapienza Università di Roma, P.le A. Moro 2, I-00185 Rome, Italy
[3] ISI Foundation, Via Alassio 11/C, Turin, Italy
[4] SONY Computer Science Lab, Paris, France
[*] Corresponding author, ccuskley@gmail.com

Abstract

Rules are an efficient feature of natural languages which allow speakers to use a finite set of instructions to generate a virtually infinite set of utterances. Yet, for many regular rules, there are irregular exceptions. There has been lively debate in cognitive science about how individual learners acquire rules and exceptions; for example, how they learn the past tense of *preach* is *preached*, but for *teach* it is *taught*. However, for most population or language-level models of language structure, particularly from the perspective of language evolution, the goal has generally been to examine how languages evolve stable structure, and neglects the fact that in many cases, languages exhibit exceptions to structural rules. We examine the dynamics of regularity and irregularity across a population of interacting agents to investigate how, for example, the irregular *teach* coexists beside the regular *preach* in a dynamic language system. Models show that in the absence of individual biases towards either regularity or irregularity, the outcome of a system is determined entirely by the initial condition. On the other hand, in the presence of individual biases, rule systems exhibit frequency dependent patterns in regularity reminiscent of patterns in natural language. We implement individual biases towards regularity in two ways: through 'child' agents who have a preference to generalise using the regular form, and through a memory constraint wherein an agent can only remember an irregular form for a finite time period. We provide theoretical arguments for the prediction of a critical frequency below which irregularity cannot persist in terms of the duration of the finite time period which constrains agent memory. However, within our framework we also find stable irregularity, arguably a feature of most natural languages not accounted for in many other cultural models of language structure. **Keywords: linguistic rules; morphology; language evolution; language development; memory**

1 Introduction

A striking feature of human language is its vast expressive power (Pinker & Jackendoff, 2005): human languages effortlessly convey everything from concrete, specific objects (e.g., *spacebar*) to broad, abstract concepts (e.g., *fairness*). The rule-based structure of human languages is key to this expressive power: word formation rules allow for new compounds (*spacebar*) or derivations (*fairness*) that speakers and hearers can readily parse. Rules allow speakers to use a finite set of instructions to generate scores of valid utterances, and allow new words to nestle into an existing language seamlessly. For example, knowing a suite of verb inflection rules even allows speakers to readily integrate entirely new (rather than derived or compounded) words into common usage (e.g., *Google* → *Googled*).

The apparent power of rules raises an interesting question: why are there irregular exceptions at all? Since rules are both productive and cognitively efficient (Trudgill, 2010; Gildea & Jurafsky, 1996; Thagard, 2005), why don't all aspects of a language obey the dominant rules? In fact, as much as rules are a universal of human languages, so are exceptions: from the phonological (e.g., disyllabic nouns are generally stressed on the first syllable as in **app**le and **ta**ble, but some nouns like ho**tel** exhibit stress on the second syllable[1]) to the morphological (e.g., in noun pluralisation, *goose-geese*, not **gooses*). Considerable study has been devoted to how language learners acquire a system with inconsistent structure (i.e., structure that may have highly frequent and idiosyncratic exceptions), but how languages support any irregularity at all, given that consistent structure is both easier to learn and more productive, remains a largely open question.

The question of how individual learners - particularly children - process and acquire exceptions has become central for both cognitive scientists and linguists (Pinker & Ullman, 2002). In particular, the formation of the simple past-tense in English has been a battleground for the debate. Previous investigations have focused on what specific cognitive mechanisms underlie the process of acquiring and producing rules and exceptions for individuals (see Pinker & Ullman, 2002; McClelland & Patterson, 2002 for a review). Extreme perspectives have argued on the one hand for a single cognitive mechanism underlying inflectional rules and exceptions (Rumelhart & McClelland, 1986; McClelland & Patterson, 2002), and on the other hand for qualitatively different levels of processing for rules and exceptions (Pinker & Ullman, 2002; Pinker, 1999). An emerging middle ground in this debate suggests that rather than rules and exceptions, the past tense is governed by

[1] c.f. (Robinson, Hank, Mike, & Gee, 1979) for a notable exception to this exception.

a set of ‘rules in competition’ (Yang, 2002; see also Albright & Hayes, 2003 for a ‘stochastic rules’ perspective). More recently, Tabor, Cho, and Dankowicz (2013) have shown using recurrent neural networks that a single learning mechanism can lead to the appearance of separate mechanisms for regularity and irregularity.

The goal of this paper will be to engage in this debate from a complex systems perspective, asking how the dynamics of rules function across a language as used by a population of individuals, rather than at the level of individual cognitive architecture. While a study of individual cognitive architecture is a necessary component to understanding regularity in language, the dynamics of regularity in a language system may be more than a simple function of extending what we know about individuals across a population. Indeed, given the cognitive efficiency of clean, consistent rules for an individual learner, the persistence of irregularity requires some explanation which goes beyond the individual learner. Rather than considering how an individual creates an internal rule-set which accommodates some exceptions (or multiple rules) in the language they speak, a dynamic systems perspective focuses on how a language system sustains irregularity despite the individual cognitive efficiency of a single regular rule. In other words, rather than seeking to explain how individuals learn and use a system with rules and exceptions, we aim to investigate why exceptions persist within the system at all.

Previous research in this area suggests a key relationship between irregularity and frequency (Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Bybee, 2007; Carrol, Svare, & Salmons, 2012; Cuskley et al., 2014): since frequency contributes to overall diachronic stability of linguistic variants (Pagel, Atkinson, & Meade, 2007; Pagel, Atkinson, Calude, & Meade, 2013), more frequent items are better able to sustain irregularity over time. From a system perspective, some corpus work has suggested that irregularity is unstable, in a state of constant ‘decay’ to the regular form (Lieberman et al., 2007). Some models of language evolution also suggest that learner biases function to reduce irregularity in language (e.g., ?, ?). However, using a larger historical corpus, a recent study suggests that much irregularity is largely stable over time (Cuskley et al., 2014), although there is some evidence of a transition in regularity dependent on frequency.

This paper will expand on the dynamic systems perspective on regularity by examining how competing rules function within a population of interacting artificial agents. The individual processing focus of previous research on the past tense has concentrated on models of individual learning, attempting to create single agents who, given some minimal linguistic input, learn a reasonable approximation of the past-tense much like a child learner would (e.g., see Pinker & Ullman, 2002; McClelland & Patterson, 2002 for a short overview). Instead, we apply a population and interaction-based approach to the problem, after a growing body of research which considers social processes as a key force in language dynamics (Kirby, Cornish, & Smith, 2008; Kirby & Hurford, 2002; Steels, 2011; Loreto & Steels, 2007). However, far from neglecting the importance of individual biases, this approach shows these individual cognitive biases are essential to recovering the dynamics we observe in natural language, and are magnified within the system through interaction and transmission (Kirby, Dowman, & Griffiths, 2007).

This approach fills a particular hole in the literature: while investigations of how individuals accommodate regularity and irregularity are fairly mature, detailed examinations of how and why language, as a large, generally rule-governed system, even has irregularity at all. While frequency has been suggested as a major player in this regard, there is considerable disagreement regarding the dynamics of regularity. Some studies have suggested that irregularity could decay or dissappear entirely; in other words, all verbs will eventually move towards the regular form, given sufficient time (Lieberman et al., 2007). However, recent data from a large historical corpus suggests that irregularity is fairly stable over time, rather than being in a process of inevitable decay (Cuskley et al., 2014). Irregularity is ubiquitous across different levels of a language, and also across diverse languages more generally, suggesting that irregularity is, in fact, a stable feature of language. While many cultural and dynamic systems approaches to language have sought to explain the emergence and sustainability of structure or rules across a population (Kirby et al., 2008; Kirby & Hurford, 2002; Steels, 2005; Kirby, Tamariz, Cornish, & Smith, 2015), few account for the fact that irregularity is also pervasive (for a notable exception, see Kirby, 2001).

We present a new model to examine linguistic rule dynamics - particularly the persistence of irregularity - modelled after a similar treatment of lexical dynamics known as the Naming Game (Steels, 1995; Loreto & Steels, 2007). In the Naming Game (NG), a population of agents interact over a pre-specified time scale measured by the number of pairwise games across the population (see also Centola & Baronchelli, 2015 for an experimental version of the game). In each "game", two agents are chosen to interact about a particular meaning, with one agent randomly assigned the role of speaker (S) and the other the role of hearer (H). The interaction consists of two core steps:

1. S sends H a string to represent the meaning; S chooses the string based on an inventory which has accumulated for the meaning over previous interactions (or, if S

has no inventory, a random string is invented). H guesses a meaning based on the transmitted string (if the string is absent from her inventory, she randomly guesses a meaning).

2. S and H update their inventories for the meaning according to predefined update rules, generally:

   - If H chooses the correct meaning, the interaction is considered a communicative success, and both agents keep the string-meaning pair.
   - If H chooses the incorrect meaning, communication is unsuccessful, S and/or H update their inventories based on the interaction; depending on the specific rules, they may disregard other form-meaning pairings or update weights to pairs in their inventory.

Using this simple model, populations which are initially unsuccessful at communication, having a broad range of random labels for a particular meaning, eventually converge on shared conventions to refer to meanings. The update rules for agents can be as simple as the H discarding their previous inventory and adopting the S's form (Baronchelli, Felici, Caglioti, Loreto, & Steels, 2006; Baronchelli, Loreto, & Steels, 2008), or can be more complex, involving different weights for forms over time depending on their communicative success in previous interactions (Wellens, Loetzch, & Steels, 2008).

While the original NG investigates how agent interaction leads to convergence on naming conventions, the current investigation adapts this general framework to focus on how a population of agents converges on shared rules or exceptions for a particular type. For example, how the word *walk* - and indeed, most verb types - inflect with the regular (roughly, add *-ed*) rule, while a verb type like *string* retains its strong umlaut form (*strung*). Our adaptation retains the basic speaker-hearer interaction at the core of the NG, but rather than inventories of strings applying to meanings, agents have inventories of inflections that can be applied to verb types. Given the established role of frequency in stability and regularity (Pagel et al., 2013; Cuskley et al., 2014; Bybee, 2007), verb types have different frequencies, wherein some verb types are used more in interaction than others.

We begin by outlining the basic structure of the models, and presenting previous findings showing how NG dynamics for rules function in the most basic case: where agents' inventories can be altered only through interaction, with no biases favouring either the regular or the irregular form. We then consider two more complex cases where agents have individual biases towards the regular form. First, we consider a child learner bias, implemented as a rate of replacement of "mature" agents with "child" agents who have a bias towards the regular form for verbs which they have not encountered (i.e., are Stage 2 learners as outlined in Rumelhart & McClelland, 1986). Second, we consider a more general memory constraint, wherein agents retain forms only for a particular temporal window, falling back on the regular form when this window has elapsed.

2 Method: The NG for rule dynamics

We use a minimal model adapted from the NG to investigate the dynamics of rules in competition over time, under the conditions of a fixed population size on a randomly connected network[2]. The model consists of N agents interacting over verb types defined by their frequency, f. For each agent, a lemma can potentially have one of three inflectional states: regular (R), irregular (I), or mixed (M).

In the mixed state, agents have a coexistent inventory of the R and I states, much like for some verbs (e.g., *sneak*) where English speakers find both regular (*sneaked*) and irregular (*snuck*) variants somewhat acceptable (Dale & Lupyan, 2012), and may even use them in seemingly free variation (Pinker & Prince, 1994). This implementation conceptualises regular and irregular inventories simply as different rules, allowing for the coexistence of competing rules within a single individual in the form of the M state (i.e., intraspeaker variation). The existence of the M state not only has psychological and linguistic validity, but analytical results show that it is crucial to recovering the type of frequency dependent transition observed in actual data (Colaiori et al., 2015).

Table 1 shows the interaction rules adapted from the basic NG (Baronchelli et al., 2006), and more broadly applicable to three-state dynamics in other realms (Colaiori et al., 2015). Although we adopt this set of rules throughout, see (Colaiori et al., 2015) for a detailed analytical treatment which allows for a prediction of the stable end-state dynamics of any three-state rule set with replacement.

At each interaction, a speaker (S) and a hearer (H) are randomly chosen from the population to engage in an interaction according to the rules outlined above. At any given interaction, the probability of interacting over a particular lemma is defined by its

[2]See e.g., Dall'Asta, Baronchelli, Barrat, & Loreto, 2006 for the minimal NG version on more complex, realistic networks; here we focus on the simplest population architecture in order to get a basic picture of rule dynamics.

Before			After	
Speaker	Hearer		Speaker	Hearer
R	R	→	R	R
R	I	→	R	M
R	M	→	R	R
I	R	→	I	M
I	I	→	I	I
I	M	→	I	I
M(R)	R	→	R	R
M(I)	R	→	M	M
M(R)	I	→	M	M
M(I)	I	→	I	I
M(R)	M	→	R	R
M(I)	M	→	I	I

Table 1: Rules for interaction in the model. A speaker is in the mixed state chooses to utter the R or I inflection with equal probability. Throughout the paper, M(R) indicates an agent in the mixed state who choses a regular inflection for an utterance, while M(I) indicates a mixed agent who choses an irregular inflection for an utterance.

f. In other words, if a lemma's $f = 0.1$, the lemma will be the topic of one in every 10 interactions. We consider a total of N interactions to encompass a single time step, t, under the assumption that given the homogenous mixing there is the possibility for each agent to have the role of speaker and hearer after N interactions. For all agent-based simulations, we examine an $N = 1000$ and $t_{max} = 10,000$ (i.e., a total of Nt_{max} interaction events). We characterize the stable stationary end state of a system in terms of the proportion of agents in the population with an irregular inflection (ρ_I^s). The inclusion of the M as potential inflection in the starting condition for the model has little effect on the end-state outcome (discussed in further detail in Basic Dynamics, below). Thus, in the following models we consider different initial values of ρ_I^0 (and thus, ρ_R^0) in the population, with the M state arising only as a consequence of an agent encountering both R and I forms in interaction.

3 Results & Discussion

3.1 Basic dynamics

Prior to investigating mechanisms of replacement and memory constraints, it is important to understand the case where no such mechanisms operate. This is analogous to the basic Naming Game (NG) outlined in (Baronchelli et al., 2008), and covered in more detail with respect to regularity by Colaiori et al. (2015). We provide a brief treatment of this case here, in order to better understand the dynamics which result from implementing replacement and memory constraints.

Without any mechanisms to bias agents towards the regular or irregular form, and given that interaction rules favour no particular inflectional state (as outlined in Table 1), the end state of a rule system is dependent entirely on the initial condition of the population. In other words, the f of a lemma has no bearing on its regularity. Instead, the proportion of starting agents with a regular or irregular inflection determines the end regularity state (a process generally true of three-state systems of interaction with unbiased rules; Baronchelli et al., 2008, also found in Colaiori et al., 2015). Any given system eventually converges on a stable solution which is either entirely regular or entirely irregular, with no remaining agents in the M state.

If the population is very large, the relationship between the initial ρ_I and ρ_R gives a deterministic prediction of the end state: if $\rho_I^0 > \rho_R^0$ (or $\rho_R^0 > \rho_I^0$), the system unavoidably resolves to an irregular (or regular) stationary state (Colaiori et al., 2015). Notice that, since $\rho_M^0 = 0$ these conditions are equivalent to $\rho_I^0 > 1/2$ ($\rho_I^0 < 1/2$). As the population size N becomes smaller, the criterion becomes probabilistic. In other words, for a starting $\rho_I > \rho_R$, the system will have a higher probability of converging to an irregular state, while given a starting $\rho_I < \rho_R$, the system will have a higher probability of converging to an entirely regular state. Figure 1 shows how different starting ρ_I and ρ_R drift toward an end state that is entirely regular or irregular, with the outcome becoming more deterministic as the population size increases (see Colaiori et al., 2015 for further detail). In summary, this basic model shows that, under these simple conditions, the regularity of a given lemma is unrelated to its frequency, and dependent only on the relationship between initial ρ_I and ρ_R across the population.

This means that a population of agents with no implementation of individual cognitive biases towards regularity does not give rise to a system with a frequency dependent

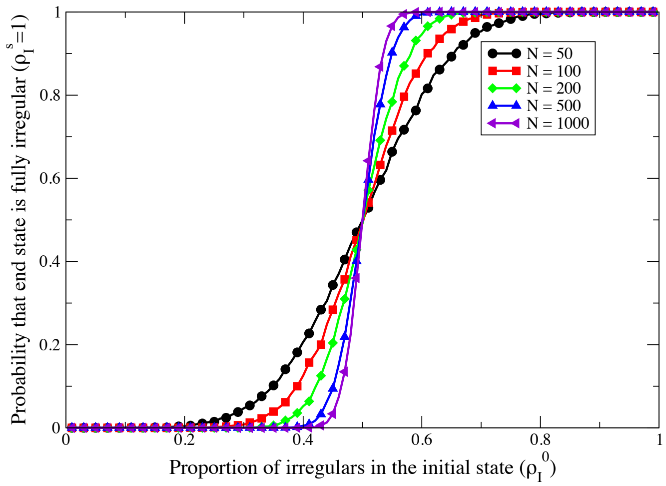

Figure 1: **Basic dynamics.** Plot of the probability to end in a fully irregular state as a function of the initial fraction of irregulars, for several population sizes N. Under conditions of simple interaction according to rules outlined in Table 1, a rule system resolves to either an entirely regular or irregular state. For a very large population, the end state is determined univocally by the the majority in the initial state: the system resolves to the R state if $\rho_I^0 < 1/2$, and to the I state if $\rho_I^0 > 1/2$. For smaller populations the transition is smoother.

transition. In other words, interaction and coordination among agents with no biases cannot recover the transition observed in rule dynamics in natural language (Cuskley et al., 2014; Lieberman et al., 2007; Bybee, 2007). In some sense, this result is intuitive: since linguistic rule systems are not only the product of simple interaction, but of interaction between "agents" with complex neural structures and biases. It is unsurprising that without any biases, the system behaves unrealistically; agents with an improbably blank slate give rise to a system qualitatively unlike what we observe in natural language. In fact, earlier adaptations of the NG have also shown that individual biases combine with interaction in non-trivial ways to produce features found in natural language systems (Puglisi, Baronchelli, & Loreto, 2008; Loreto, Mukherjee, & Tria, 2012). This makes it particularly important to investigate how biases combine with interaction to give rise to the sorts of frequency dependent dynamics observed in natural language.

3.2 Child learner bias

Child learners have a bias towards regular forms during early learning (i.e., are Stage 2 learners after Rumelhart & McClelland, 1986). In other words, children tend to follow a U-shaped learning curve (Gershkoff-Stowe & Thelen, 2004) wherein their inflection performance is at first very high as a result of rote learning a finite set of items, but as this set grows they begin to engage in rule generalisation and over-regularise some verbs in production (e.g., produce "goed" instead of "went"; see Maslen, Theakston, Lieven, & Tomasello, 2004).

This bias itself is undoubtedly internally complex in real language processing, and much debate has centered around whether it arises while processing is still refining divisions between the word and rule level (Pinker & Ullman, 2002), or whether it is the result of general statistical learning mechanisms which are refined with input (McClelland & Patterson, 2002). For this model, we take the fact of over-regularisation as an important factor in linguistic outcomes (Sankoff, 2008), but do not speculate on its internal nature. In other words, we do not aim to address exactly how language users acquire or implement this sort of bias (a topic covered in significant detail by earlier work, e.g.,Pinker, 1999; Pinker & Ullman, 2002; Plunkett & Juola, 1999; Bybee & Slobin, 1982, among others). Rather, in our model, "child" agents exhibit a uniform "built-in" bias towards regularity: they assume a regular inflection for all lemmas in their inventory in initial production, only acquiring irregular forms through interaction.

Biased "child" agents enter the model through simple *replacement*: "adult" agents are replaced at a probabilistic rate, r. Practically, this is implemented by choosing a single agent from the population randomly at each interaction, and replacing them with a probability r. Analytical results show that the relationship between r and f (frequency) is most relevant (Colaiori et al., 2015), therefore we explore a single value of r ($r = 0.01$) over a range of frequencies. Practically, this value means that at a given interaction there is a 1% chance that a random "adult" agent will be replaced with a child; accordingly, over the course of a single unit of t, approximately 1% of the total population will have been replaced. In order to keep the model minimal, we do not consider growth of the population; rather, r is envisioned more usefully as a constant rate of turnover in a population with a fixed size.

As with the basic model outlined in the previous section, the end state of a system is at least partially dependent on the starting condition. However, introducing replacement also introduces frequency dependence, giving different outcomes in regularity for different lemmas as a function of their f. Figure 2 shows the probability that a given run will end in a state with a positive fraction of irregularity (ρ_I^s), as well as the average value of ρ_I^s for several values of ρ_I^0.

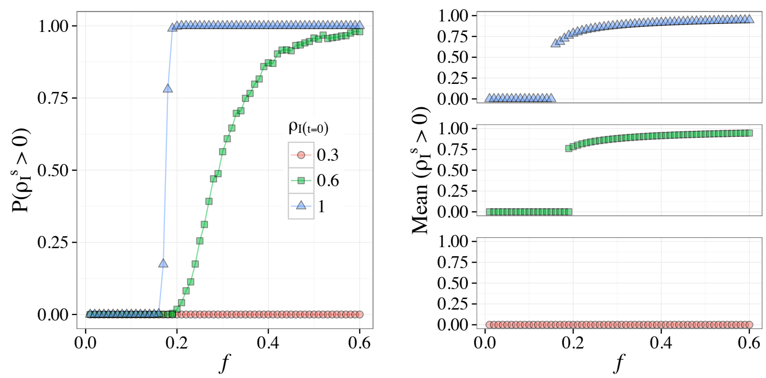

Figure 2: **Naming Game model with replacement.** The graph on the left shows the probability that the system will end up in a state with some positive fraction of irregularity (ρ_I^s) plotted against frequency, f. Results for three different initial fractions of irregularity are shown (ρ_I^0). The graph on the right shows the average value of ρ_I^s as a function of f, again for three different initial values of ρ_I^0. These show that a certain level of irregularity is necessary in order for it to stabilise and persist within the population, demonstrating that the initial condition has some bearing on the final state. However, systems with sufficient initial irregularity, $>\approx 0.38$ display clear frequency dependent transitions.

The case $\rho_I^0 = 1$ is particularly interesting to test what happens to irregular verbs over time, particularly given previous claims that irregular verbs decay slowly to the regular form (Lieberman et al., 2007). In other words, what happens to a completely irregular verb over time given the pressure from incoming child learners to conform to the regular rule? The behaviour of the case where $\rho_I^0 = 1$ displays a clear discontinuous change in regularity in agreement with analytical models (Colaiori et al., 2015) and more reminiscent of the patterns found more recently in natural language data (Cuskley et al., 2014). In other words, under some conditions, verbs stabilise in a predominantly irregular state. All verbs below a certain frequency $f \approx 0.16$ become completely regular. Above this value, with a probability close to 1, verbs remain predominantly irregular, although a sizeable fraction of the agents adopt the regular inflection. Even for highly frequent forms, no lemma exhibits complete, consistent irregularity, even at the highest values of f (this could be considered analogous to the roughly 4% over-regularisation rate found in corpora of child speech, Marcus, 1996, which would indicate that a totally comprehensive corpus would include regularisations even of verbs uncontroversially considered to be irregular). As the fraction of irregulars in the initial condition becomes smaller, the change of behaviour occurs for larger frequencies and is less abrupt. If the initial fraction of irregulars gets smaller than a threshold ≈ 0.38 all verbs become fully regular, regardless of their frequency.

3.3 Memory constraints

In this model, we implement constraints on individual agents' memory: accurate recall of an inflection relies both on the cumulative number of encounters with a lemma as well as time elapsed since last encounter (Rodi, Loreto, Servedio, & Tria, 2015; Novikoff, Kleinberg, & Strogatz, 2012). Earlier work on individual learning models of the past tense in English have shown memory constraints, particularly as they relate to frequency, to be an important factor in over-regularisation errors in children (Marcus, 1996). However, this memory constraint can be considered domain general, applying not only for linguistic rules, but also, for example, to visual memory (Logie, 2014), as well as more complex learning tasks (e.g., studying an academic subject, Rodi et al., 2015).

The memory constraint is implemented in terms of deterministic loss of the irregular form (and reversion to the regular form) after a certain amount of time, unless the agent is involved in interactions about the lemma, thus refreshing her memory. In other words, each agent has a time window, W, for each lemma within which they can recall the I form. Each time an agent encounters a lemma, there is a refresh event: the time of last encounter, t_l, is re-set to the current time, and total elapsed time since the last irregular encounter, τ, is updated: $\tau = t - t_l$. At every interaction, if $\tau > W$, then the temporal window has elapsed and the agent will revert to the R form (although I can be re-acquired through interaction; see Table 1). As with the previous model, we assume a fixed population size.

Under these conditions, the initial value of the window for each lemma largely determines the behaviour of different frequencies, much like the value for r in the replacement model. For the following models we consider a $W_{t0} = 100$, and provide a more general theoretical treatment which can account for other values of W_{t0} in Section 3.3.3. First, we examine the case where the window is fixed (Section 3.3.1), resulting in transitional outcomes reminiscent of replacement. Second, we allow the window to grow linearly, dependent on the total number of encounters with a verb, k (Section 3.3.2). Finally, we provide a brief theoretical treatment of the relationship between frequency and memory constraints.

3.3.1 Fixed window

Figure 3 shows results for a fixed $W = W_{(t=0)} = 100$. The resulting dynamics look much like replacement, although the location of the frequency dependent transition is lower, given that the effective rate of reversion to the R form is lower than for $r = 0.01$, an issue which is covered in more detail below (Section 3.3.3). For a system which starts in the completely irregular state, the transition occurs between $0.015 < f_c < 0.02$. This transition is slightly shifted with a lower $\rho_I^0 = 0.6$, while for a $\rho_I^0 = 0.3$ no irregularity remains in the system at all.

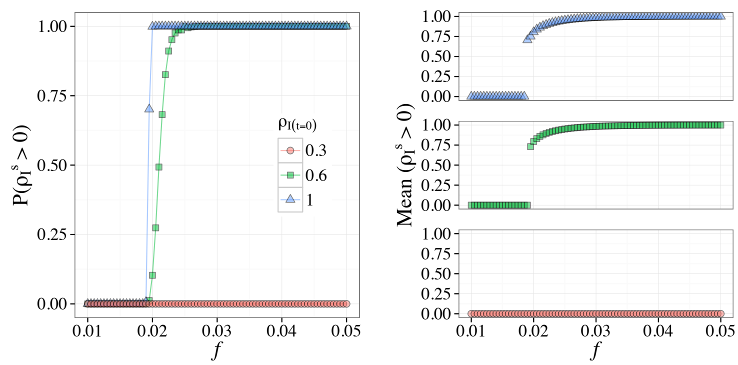

Figure 3: **Naming game with a fixed forgetting window.** The graph on the left shows the probability that the system will end up in a state with some positive fraction of irregularity (ρ_I^s) plotted against frequency, f. Results for three different initial fractions of irregularity are shown (ρ_I^0). The graph on the right shows the average value of ρ_I^s as a function of f, again for three different initial values of ρ_I^0. Results for a fixed forgetting window are almost identical to replacement.

3.3.2 Expanding window

In the basic case of forgetting presented above, agents have a static value of W determined at the start of the simulation and constant across all lemmas. Here, we test the condition where the window for a given lemma in an agent expands each time the lemma is encountered. This is reminiscent of expanded retrieval and spacing effects in memory, wherein information is better retained when the intervals at which it is reinforced are optimally spaced and/or expand with each reinforcement (Baddeley, 1997). Such effects are not only domain general, but have also been shown to hold for learning in other animals (e.g., rats and pigeons; Balota, Duchek, & Logan, 2007), and have been confirmed in theoretical models (see e.g., Novikoff et al., 2012 for the spacing effect and e.g., Ebbinghaus, 1885 for expanded retrieval) as well as artificial learning networks (Rodi et al., 2015). Here we implement an increase in W linearly as a function of k, defined as the total number of irregular interactions an agent has had with a lemma (such that $W_t = W_{(t=0)} + k$). Even this moderate expansion of W shifts the location of the transition: lower frequencies are able to remain irregular where they eventually regularised given a static value of W (Figure 4).

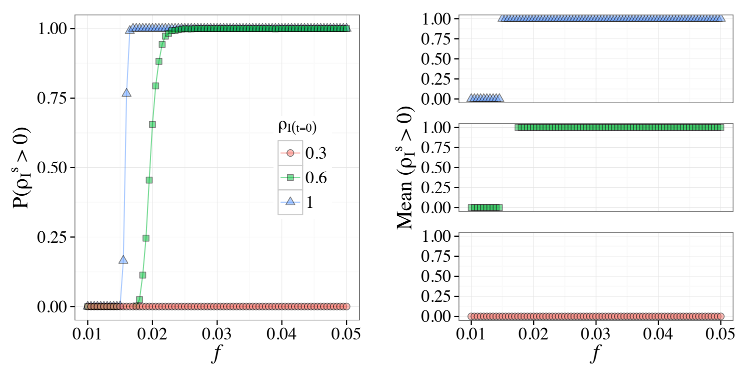

Figure 4: **Naming Game with linear expansion of the forgetting window.** The graph on the left shows the probability that the system will end up in a state with some positive fraction of irregularity (ρ_I^s) plotted against frequency, f. Results for three different initial fractions of irregularity are shown (ρ_I^0). The graph on the right shows the average value of ρ_I^s as a function of f, again for three different initial values of ρ_I^0. In this case, where the forgetting window is expanded, the frequency at which lemmas can remain irregular given an entirely irregular start state ($\rho_I^0 = 0$) reduces considerably, from $f \approx 0.02$ in 3 to $f \approx 0.016$. The stable end state with an expanding window also exhibits an important qualitative difference: verbs resolve to either entirely regular or irregular states.

More importantly, each f in a given system with an expanding window resolves to a completely regular or irregular state, with no agents remaining in the M state. Therefore, in the case of an expansion of W, $\mathrm{P}(\rho_I^s > 0)$ (in Figure 4, left) can be conceptualised as the probability that a given system will resolve to a completely irregular state. Accordingly, the mean value of $\rho_I^s > 0$ is either 1 or 0 in all cases (Figure 4, right). The discontinuous transition is more abrupt with expansion (with $\rho_I = 0$ or 1 for all verbs) than for no expansion or for replacement, since some verbs remain entirely irregular given a high enough frequency.

A comparison of no expansion and linear expansion shows that lemmas which would regularise without expansion remain irregular if W expands. Figure 5 shows a time series of linear expansion. The stabilisation of the irregular state for $f = 0.016$ is particularly evident in a time series, which shows a dip indicating that agents begin to revert to the regular form, but in re-encountering the irregular form in interaction, their windows expand and the lemma recovers to the fully irregular form across the population[3]. While this frequency best illustrates important differences between a static W and a value of W which grows linearly, the specific value of f which illustrates this is dependent primarily on the initial value of W across the population. Section 3.3.3 will provide a framework which allows for more detailed consideration of alternative values of a static W and how these change the transition frequency.

3.3.3 Theoretical treatment

In order to fully understand the dynamics of regularity, it is important to consider how models might behave with slightly different parameters than the ones presented in the simulations above. Colaiori et al. (2015) provide a detailed theoretical treatment of the replacement case, showing that given a specific set of interaction rules and a constant replacement rate, the nature and location of a frequency dependent transition can be predicted. In particular, when interactions take place according to the rules laid out in Table 1, a discontinuous transition between a fully regular state (for low frequencies) and a largely irregular state (for high frequencies) occurs at a critical frequency

$$f_c = \frac{r}{n_c}, \tag{1}$$

where r is the replacement rate and $n_c \approx 0.058$ is an analytically derived numerical constant.

Results above show that implementing memory constraints yields a very similar pattern

[3]More extreme expansion of the window (e.g., quadratic expansion, $W_t = W_{t(last)} + k$, or exponential $W_t = W + 2^k$) leads to dynamics similar to linear expansion, although effects are more extreme. Since the value of W grows more drastically, the critical frequency of f_c is lower, and quickly hits ceiling effects (such that, for example, there is little difference between quadratic and exponential expansion).

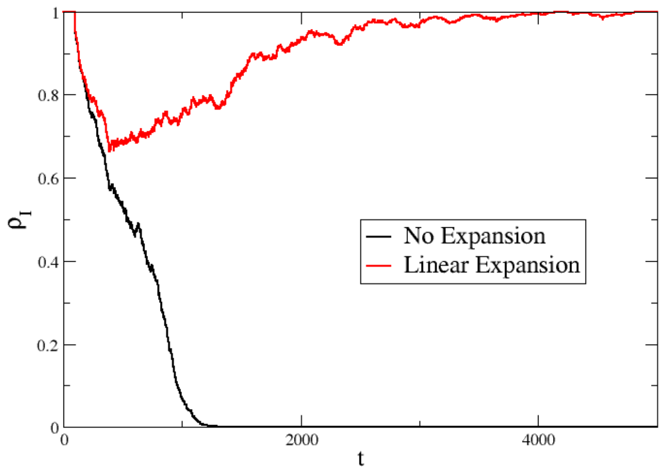

Figure 5: **Time series of** ρ_I for $f = 0.016$ for the Naming Game models with no expansion and linear expansion. In the case of expansion, the lemma begins to regularise and then recovers to the irregular form around $t = 150$ as agents' windows start to expand.

of behaviour. However, we present a more detailed theoretical treatment below in order to both understand important subtle differences between replacement and a static window, and also to consider the more complex case of an expanding window. This treatment allows for some predictions of the behaviour of a system, particularly in terms of the transition frequency at which verbs regularise, for different values of W other than the somewhat arbitrary value used in the simulations presented above.

A reasonable justification of the similarity between replacement and forgetting, in particular in the case of static value of W, is that W plays a role analogous to the inverse of the replacement rate $(1/r)$ in the dynamics: in a time interval $1/r$, on average, one agent is replaced by a new agent in the regular state. Likewise in a time interval W, on average, one agent forgets the irregular form and switches to the regular state. According to this argument one should expect a transition at a frequency:

$$f_c = \frac{1}{W n_c}. \tag{2}$$

In other words, the higher the value of W, the lower the critical frequency (f_c) at which irregular verbs can remain stable over time and across the population. Put differently, if the memory of agents is improved, increasingly lower frequency items can remain stably irregular within their shared language. A comparison with simulation results (see Fig. 6) shows that the prediction of Eq. 2 correctly captures the dependence of the transition frequency on the forgetting time W. However, there is also a considerable mismatch: the theoretical prediction is approximately 10 times larger than the value obtained in the simulations.

This discrepancy is due to the fact that encountering a lemma "refreshes" the memory and resets the time of the forgetting event to zero. As Figure 7 illustrates, even though W has a constant value, due to recurring refresh events, the total time spent by an agent in the irregular state before forgetting is larger than W.

We define as W_{eff} the typical *effective* time for an agent to forget the irregular form of a lemma. Inserting its value in Eq. 2 provides a more accurate estimate of the transition frequency:

$$f_c = \frac{1}{W_{\text{eff}}(f_c)n_c}, \tag{3}$$

where we have made explicit the dependence of W_{eff} on f. Since $W_{\text{eff}} \geq W$, Eq. 3 predicts a critical frequency smaller than Eq. 2.

In the Appendix, we report a brief analytical treatment, which provides a formula (Eq. 10) for the value of W_{eff} as a function of W and f. Inserting this expression into Eq. 3 one obtains a nonlinear equation for the frequency f_c, which can be easily solved graphically (plotting the left and right hand sides of the equation separately as a function of f and determining the intersection point) for any value of W. The values obtained in this way are compared with simulation results in Fig. 6, displaying a much better agreement than the naive theory[4] (Eq. 2).

In the case where W is not fixed, a theoretical approach taking into account the expansion would be considerably more complicated. However, simulations (Fig. 6) show that the difference in f_c between no expansion and linear expansion is not large and tends to decrease as W increases.

[4]The analytical estimates are off by a factor ≈ 1.5, which is acceptable in this case, given that the theory includes no fitting parameters.

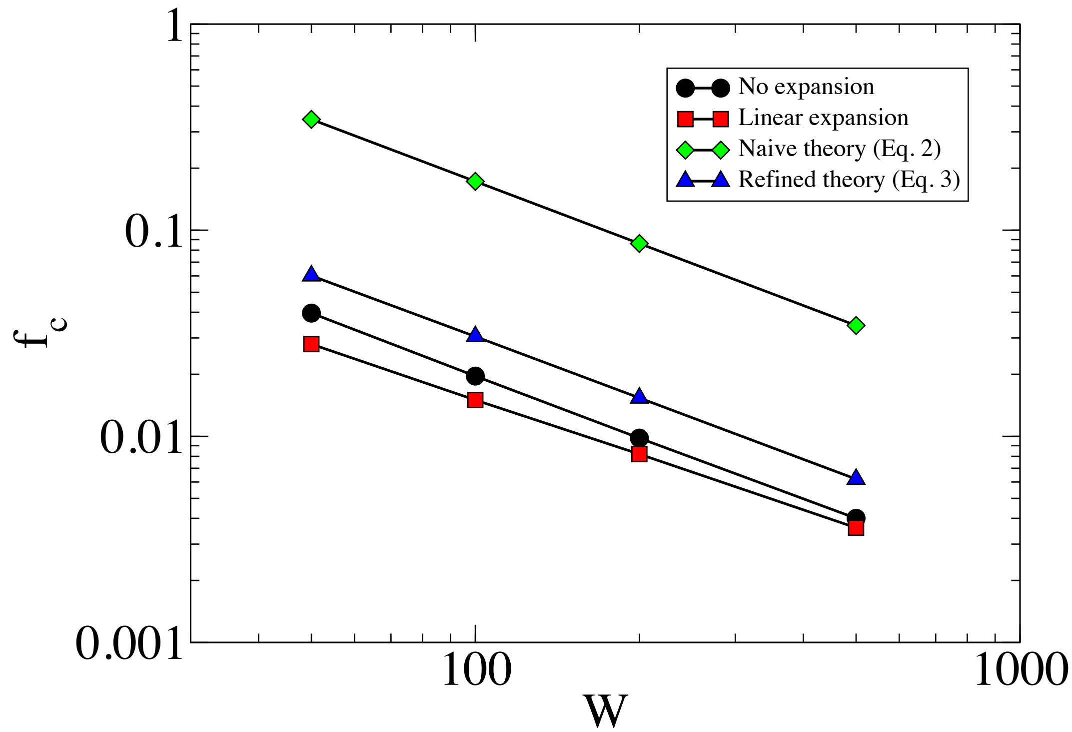

Figure 6: **Behaviour of f_c vs. W.** Comparison of the transition frequency f_c (in the case of no expansion, black circles) determined in simulations as a function of W, with theoretical estimates obtained with a naive theory (Eq. 2, green diamonds) and a more refined theory (blue triangles) . For completeness also the value of f_c in the case of linear expansion is displayed (Eq. 3, red squares).

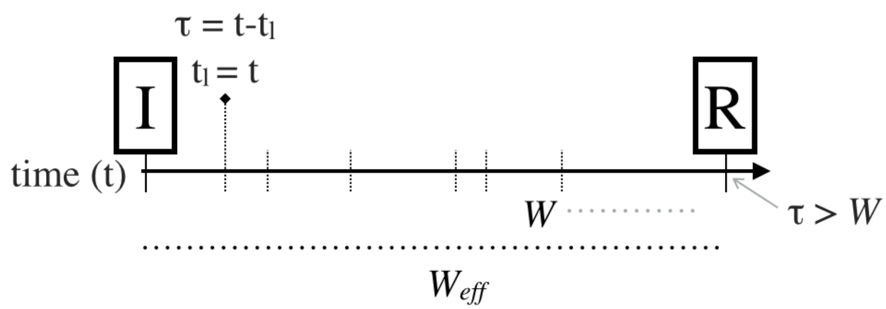

Figure 7: **Representation of refresh events**

4 Discussion & Conclusions

Using the mechanisms of replacement and general memory constraints, our models have shown that individual biases combined with interaction among a population lead to system-wide rule dynamics where highly frequent items can remain stably irregular. These results indicate that the sort of frequency dependent decay predicted by Lieberman et al. (2007) only occurs under a certain frequency threshold. Moreover, the patterns observed echo those found in a larger diachronic sample of English (Cuskley et al., 2014). Both a constant influx of child learners in a population and individual constraints on agent memory lead to a discontinuous transition in regularity, with more frequent verbs retaining a stable irregular form while less frequent verbs tend to regularise. In accordance with results for child learner bias, we found that memory constraints lead to a system where different initial conditions and specific frequencies resolve to (ir)regularity with a probability, rather than deterministically. In other words, two separate evolutions with the same initial conditions may resolve to completely different outcomes for the same frequency. Finally, we presented a theoretical framework which allows for the estimate of the critical transition frequency given particular memory constraints.

These models represent an initial step in understanding the dynamics of linguistic rules which function across complex populations. Our goal was to make this first step simple by considering a small, closed population which is homogenously mixed. However, in the future, this approach could be used to examine how different social network structures may lead to divergent linguistic rules (e.g., *burnt* in British English and *burned* in American English; Michel et al., 2011), how different types of learners might effect rule dynamics differently (Cuskley et al., 2015), and how linguistic rules evolve and spread over growing or shrinking populations. This framework could be expanded to examine more general cases of contact dynamics in language (Weinreich, 1963; Thomason, 2001; Bakker & Matras, 2013), with the potential to address specific quantitative questions in sociolinguistics: for instance whether an influx of non-native adult speakers leads to decreased morphological complexity (Lupyan & Dale, 2010), or indeed, how linguistic rules and systems diverge to the point of creating entirely new languages (e.g., Creoles and Pidgins; Michaelis, Maurer, Haspelmath, & Huber, 2013).

The individual mechanisms at work could also be further specified, by giving "child" agents more nuanced biases refined by learning, or refining the memory window to be more commensurate with actual memory systems. Finally, these two biases could be combined to investigate the differences between child and adult learners, with different memory constraints to account for differences in child and adult language acquisition (Gathercole & Baddeley, 2014; Cuskley et al., 2015). More generally, while our models sought to examine the dynamics of existing rule sets over time, a further step would be to extend work examining how rules and exceptions emerge in the first place (Kirby, 2001), a question with particular relevance for language evolution (Michel et al., 2011). Our application of the NG framework to linguistic rules highlights broadly how agent-based models of interaction, coordination, and cultural transmission can be applied to a diverse array of collective linguistic, cultural, and cognitive phenomena.

5 Acknowledgements

This work was supported by the European Science Foundation as part of the DRUST project, a EUROCORES EuroUnderstanding programme: http://www.eurounderstanding.eu/. V.L and F.T. acknowledge support from the KREYON project funded by the Templeton Foundation under contract N. 51663.

6 Appendix

The effective time W_{eff} necessary for an agent to forget the irregular form of a lemma can roughly be estimated (see Fig. 7) as:

$$W_{\text{eff}} \simeq \langle t_e \rangle \overline{n_r} \tag{4}$$

Here $\langle t_e \rangle$ is the average number of time steps separating two successive refresh events, while $\overline{n_r}$ is the average number of such refresh events before a forgetting event.

To compute these two quantities we start by defining p_{not}, the probability not to have a refresh event at a given time; p_{not}^W is then the probability to forget before a refresh event occurs. Hence the probability p_r that a given agent will experience a refresh event before forgetting is:

$$p_r = 1 - p_{not}^W. \tag{5}$$

The probability to have a refresh event at a given time is proportional to the interaction frequency f and to the effective density of irregulars in the population. This effective density is best captured as $\rho_I + \rho_M/2$, since agents in the M state have an equal probability of using the R or I form in interaction. Given this, we can estimate p_{not} as $p_{not} \simeq 1 - f(\rho_I + \rho_M/2)$. This allows us to calculate the average number of refresh events before a forgetting event occurs:

$$\overline{n_r} = \sum_{k=0}^{\infty} k p_r^k (1 - p_r)\,, \tag{6}$$

and the average time between two successive refresh events as:

$$\langle t_e \rangle = \sum_{k=0}^{W-1} k p_{not}^k (1 - p_{not}) \tag{7}$$

that after some algebra turn out to be

$$\overline{n_r} = p_r/(1 - p_r) = (1 - p_{not}^W)/p_{not}^W\,, \tag{8}$$

and

$$\langle t_e \rangle = (1 - p_{not}^W)/(1 - p_{not}) - W p_{not}^{W-1}\,. \tag{9}$$

Inserting the results for $\overline{n_r}$ and for $\langle t_e \rangle$ into Eq. 4 we get

$$W_{\text{eff}} \simeq \frac{(1 - p_{not}^W)^2}{p_{not}^W (1 - p_{not})} - \frac{W(1 - p_{not}^W)}{p_{not}}, \tag{10}$$

with

$$p_{not} \simeq 1 - f\left(\rho_I + \frac{\rho_M}{2}\right). \tag{11}$$

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*, 119-161.

Baddeley, A. D. (1997). *Human memory: Theory and practice*. Psychology Press.

Bakker, P., & Matras, Y. (2013). *Contact languages: A comprehensive guide*. De Gruyter Mouton.

Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). The foundations of remembering: Essays in honor of henry l. roediger, iii. In J. S. Nairne (Ed.), (p. 83-105). New York, NY: Psychology Press.

Baronchelli, A., Felici, M., Caglioti, E., Loreto, V., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics*, *P06014*.

Baronchelli, A., Loreto, V., & Steels, L. (2008). In-depth analysis of the naming game dynamics: the homogenous mixing case. *International Journal of Modern Physics C*, *19*(5), 785-812.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford, UK: Oxford University Press.

Bybee, J., & Slobin, D. (1982). Rules and schemas in the development and use of the english past tense. *Language*, 265-269.

Carrol, R., Svare, R., & Salmons, J. (2012). Quantifying the evolutionary dynamics of German verbs. *Journal of Historical Linguistics*, *2*(2), 153-172.

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, *112*(7), 1989-1994. Retrieved from `http://www.pnas.org/content/112/7/1989.abstract` doi: 10.1073/pnas.1418838112

Colaiori, F., Castellano, C., Cuskley, C., Loreto, V., Pugliese, M., & Tria, F. (2015, Jan). General three-state model with biased population replacement: Analytical solution and application to language dynamics. *Phys. Rev. E*, *91*, 012808. Retrieved from `http://link.aps.org/doi/10.1103/PhysRevE.91.012808` doi: 10.1103/PhysRevE.91.012808

Cuskley, C., Colaiori, F., Castellano, C., Loreto, V., Pugliese, M., & Tria, F. (2015). The adoption of linguistic rules in native and non-native speakers: Evidence from a wug task. *Journal of Memory and Language*, *84*, 205 - 223. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0749596X15000790` doi: http://dx.doi.org/10.1016/j.jml.2015.06.005

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., & Tria, F. (2014, 08). Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of english. *PLoS ONE*, *9*(8), e102882. doi: 10.1371/journal.pone.0102882

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, *15*(03n04), 1150017.

Dall'Asta, L., Baronchelli, A., Barrat, A., & Loreto, V. (2006). Nonequilibrium dynamics of language games on complex networks. *Physical Review E*, *74*(3), 036105.

Ebbinghaus, H. (1885). *Memory: a contribution to experimental psychology.* (trans. H. A. Ruger and C. E. Bussenius, Teachers College at Columbia University, 1913)

Gathercole, S., & Baddeley, A. (2014). *Working memory and language processing.* New York, NY: Psychology Press.

Gershkoff-Stowe, L., & Thelen, E. (2004). U-shaped changes in behaviour: A dynamic systems perspective. *Journal of Cognition and Develeopment*, *5*(1), 11-36.

Gildea, D., & Jurafsky, D. (1996). Learning bias and phonlogical-rule induction. *Computational Linguistics*, *22*(4), 497-530.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102-110.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681-10686.

Kirby, S., Dowman, M., & Griffiths, T. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241-5245.

Kirby, S., & Hurford, J. R. (2002). Simulating the evolution of language. In (p. 121-147). Springer.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. doi: 10.1016/j.cognition.2015.03.016

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(2007).

Logie, R. H. (2014). *Visuo-spatial working memory.* New York, NY: Psychology Press.

Loreto, V., Mukherjee, A., & Tria, F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, *109*(18), 6819-6824. Retrieved from `http://www.pnas.org/content/109/18/6819.abstract` doi: 10.1073/pnas.1113347109

Loreto, V., & Steels, L. (2007). Social dynamics: emergence of language. *Nature Physics*, *3*(11), 758-760.

Lupyan, G., & Dale, R. (2010, 01). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1), e8559.

Marcus, G. F. (1996). Why do children say "breaked"? *Current Directions in Psychological Science*, *5*, 81-85.

Maslen, R. J., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in english. *Journal of Speech, Language, and Hearing Research*, *47*, 1319-1333.

McClelland, J. L., & Patterson, K. (2002). 'words *or* rules' cannot exploit the regularity in exceptions. *Trends in Cognitive Sciences*, *6*(11), 464-645.

Michaelis, S. M., Maurer, P., Haspelmath, M., & Huber, M. (Eds.). (2013). *Apics online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://apics-online.info/`

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182. Retrieved from `http://www.sciencemag.org/content/331/6014/176.abstract` doi: 10.1126/science.1199644

Novikoff, T. P., Kleinberg, J. M., & Strogatz, S. H. (2012). Education of a model student. *PNAS*, *109*(6), 1868–73.

Pagel, M., Atkinson, Q. D., Calude, A., & Meade, A. (2013). Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*, *110*(21), 8471-8476.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, *449*, 717-720.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Harper Collins: New York.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, *95*(2), 201-236.

Pinker, S., & Prince, A. (1994). Regular and irregular morphology and the psychological status of rules of grammar. In S. D. Lima & R. L. (Eds.), *The reality of linguistic rules* (p. 321-352). Philadelphia: John Benjamins.

Pinker, S., & Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456-463.

Plunkett, K., & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, *23*(4), 463-490.

Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, *105*(23), 7936-7940. Retrieved from `http://www.pnas.org/content/105/23/7936.abstract` doi: 10.1073/pnas.0802485105

Reali, F., & Griffiths, T. (2009). The evolution of frequency distributions: Relating regularisation to inductive biases through iterated learning. *Cognition*, *111*(3), 217-328.

Robinson, S., Hank, B. B., Mike, W., & Gee, M. (1979). *Rapper's delight.* The Sugar Hill Gang (Sugar Hill Studios).

Rodi, G. C., Loreto, V., Servedio, V. D. P., & Tria, F. (2015, 06 01). Optimal learning paths in information networks. *Scientific Reports*, *5*, 10286 EP -. Retrieved from `http://dx.doi.org/10.1038/srep10286`

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing (vol 2): Psychological and biological models* (p. 216-271). Cambridge: MIT Press.

Sankoff, G. (2008). Linguistic outcomes of language contact. In J. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (chap. 25). Blackwell.

Steels, L. (1995). A self-orgnaizing spatial vocabulary. *Artificial Life*, *2*(3), 319-332.

Steels, L. (2005). The emergence and evolution of linguistic structure: from lexical to grammatical communication systems. *Connection Science*, *17*(3-4), 213-230. doi: 10.1080/09540090500269088

Steels, L. (2011). Modelling the cultural evolution of language. *Physicis of Life Reviews*, *8*(4), 339-356.

Tabor, W., Cho, P., & Dankowicz, H. (2013). Birth of an abstraction: A dynamical systems account of the discovery of an elsewhere principle in category learning. *Cognitive Science*, *37*(7), 1193-1227.

Thagard, P. (2005). *Mind: Introduction to cognitive science.* Cambridge, MA: MIT Press.

Thomason, S. (2001). *Language contact.* Edinburgh University Press.

Trudgill, P. (2010). *Investigations in sociohistorical linguistics.* Cambridge, UK: Cambridge University Press.

Weinreich, U. (1963). *Languages in contact: findings and problems.* Mouton.

Wellens, P., Loetzch, M., & Steels, L. (2008). Flexible word meaning in embodied agents. *Connection Science*, *20*(2-3), 173-191.

Yang, C. (2002). *Knowledge and learning in natural language.* Oxford University Press, Oxford UK.