THE UNIVERSITY *of* EDINBURGH
School of Physics
and Astronomy

MPhys Project Report

An Agent-based Approach To Language Formation and Evolution

Investigating optimal communication and language turnover in a speech community

Casimir Fisch

Abstract

An agent-based model of language is developed as an extension to the Utterance Selection Model, in which speakers adjust their mental conception of language, or grammar, over successive conversations with a nonlinear update rule. The convergence to shared optimal grammars which maximise intelligibility and minimise ambiguity is observed in Monte Carlo simulations. Using a mean-field prescription of the model and linear stability analysis, the preference for these optimal grammars is investigated. We identify the convergence to a first mapping between a signal and a meaning as an important process by which an initial consensus is attained across the speakers. Subsequently, a mechanism of language change is implemented in the model, by allowing speakers to generate innovations that can propagate in the community.

Supervisor	Project duration
Prof. R. A. Blythe	September 2022 – March 2023

Personal statement

The first few weeks of the project were spent familiarising myself with some of the literature on language modelling and linguistics more broadly, to reflect on the possible aims of this open-ended project. This was complemented with insightful conversations with my supervisor, whose experience with the field was greatly beneficial to approach this topic. Simultaneously, I started implementing the model and its update rule into Python code so that its predictions could be tested and visualised. Shortly after, the expected convergence to shared mappings across the speech community was observed, which was a result noted in the literature.

Following this initial period, and after running quite a few simulations, I started noticing a preference of the system to what we would later call 'optimal' configurations, in which speakers would naturally minimise ambiguity—this was a new result that we chose to investigate. To probe it, the multiplicity of these systems was calculated and compared with that of an unbiased model. Much of the mathematical derivation is to the credit of my supervisor. I focused on implementing it in Monte Carlo simulations with the help of the Numba library, which greatly increases the computation speed by translating Python programs into machine code fed directly to the CPU. This tool was subsequently used in the majority of the computational work done in this project.

The remainder of semester 1 was spent on the stability analysis of the model, which required the derivation of a mean-field prescription of the model and the application of linear stability analysis. Prof. Blythe contributed heavily to the mathematical derivations, which took me a few weeks to assimilate. Most of my work was concentrated on the implementation of these methods into numerical routines, and their interpretation in the context of language formation. Work was also done on the dependence of convergence on the different parameters and configurations of the system.

As semester 1 finished, I was unfortunately confronted with a personal situation that required me to take several weeks off university work, extending into the start of semester 2. During that time, I was nevertheless able to finish the work done on linear stability analysis, which required a careful interpretation of its (oftentimes confusing) results.

With help from my supervisor, we were able to complete this part of the project and move on to the next one, which was concerned with an implementation of language change in the model. This occupied the rest of semester 2 and had a much more exploratory nature—the aim being to determine whether a plausible mechanism of turnover in the language could be developed. Initially, our approach was to consider the phenomenon of stochastic resonance, in which an external forcing is able to amplify fluctuations to result in global change in the system. However, the formulation of the model made it so that speakers were increasingly less likely to alter their communication system once established, with optimal systems being identified as absorbing states.

This made the implementation of stochastic resonance in the model more difficult, and led me to investigate alternative approaches to seek language turnover in the system. Multiple mechanisms were considered, with variable and mitigated results. The one showing the most promise is presented in this report, which introduces a 'trendsetter' among the speakers, able to generate innovations that catch on in the community. The remaining weeks were spent assembling the different results of this project into a sensible narrative, and writing this report.

Acknowledgements

I would like to thank my supervisor, Prof. Richard Blythe, for introducing me to the stimulating topic of language modelling over the many insightful conversations we've shared. His contributions to the project were essential to its progression, and his helpful comments and guidance always ensured that I was on track.

Contents

1 Introduction

Language is often regarded as humanity's defining characteristic, by which we attribute words to thoughts and ideas to convey information to others. Its formation and subsequent evolution in society constitute some of the most fundamental questions that linguistics, psychology, cognitive science and anthropology attempt to answer.

Remarkably, some of the most significant contributions to the topic in the past decades have come from the fields of complex system physics and computational modelling. Indeed, the increasing power of computational methods has led to a wide range of applications reaching beyond the traditional domain of statistical physics, for instance in the field of social dynamics [1]. As will be described below, these complex, socially-driven systems are aptly described by *agent-based models*, which probe the collective behaviour of interacting agents.

Agent-based approaches naturally apply to language, which is driven by the social interactions of its speakers. Many such models has been formulated over the years, comprising many different paradigms and conceptual frameworks to study language formation and evolution. As described in Section 2.3, a general finding from these models is the observed convergence to a shared, common communicative system among agents, or speakers.

In this project we develop a stochastic agent-based model of language use, formulated as an extension to the Utterance Selection Model (outlined in Section 2.4), in which speakers adjust their understanding of language over successive conversations. Section 3 describes its mathematical conceptualisation, representing language as a probabilistic mapping between signals and meanings.

The convergence to a common language is reproduced in Section 4, and we observe a 'new' result: a preference of the system for *optimal* communication systems, which maximise intelligibility and minimise ambiguity. This preference is probed mathematically in Section 5, in which we develop a mean-field prescription of the model, and subsequently perform linear stability analysis to understand how consensus about language can emerge. In particular, we highlight the initial convergence of the model to a first mapping, and its strong dependence on the system size.

In Section 6 we investigate the plausibility of language change in the model by implementing a mechanism of innovation in the system. This constitutes a separate, more exploratory part of the project, and as such the results of each section are discussed independently. Finally, Section 7 concludes on the uses of the model in the field of language formation and evolution.

2 Background

2.1 What is language?

Human language can be understood as a system of communication that expresses meanings using words, grammar and syntax. But it is not fixed: rather, it is a dynamic phenomenon shaped by both individual and collective usage, subject to change over time due to social, cultural, and historical factors [2, 3].

In the early 20th century, the linguist Ferdinand de Saussure was among the first to describe language as an association, or *mapping*, between signals and meanings [4]. Signals refer to any form of expression that conveys a message to another person, such as spoken words, written text, or nonverbal gestures. Meanings, on the other hand, represent the ideas, concepts, or emotions that signals convey. In other words, meanings are the content of a signal, the interpretation of which can vary from speaker to speaker. Interestingly, de Saussure argued that the relation between a signal and its meaning was fundamentally arbitrary, with no inherent connection between the two [5].

2.2 Statistical physics of language

Complex systems physics is traditionally concerned with the interaction of many particles, and the resulting collective behaviour that results from it. However, its applicability extends far beyond physical systems: in recent years, interdisciplinary research in biology, ecology, economics and sociology [6] has made use of notions from statistical physics to probe a wide range of systems.

Instead of particles, the systems generally comprise agents—interacting components following simple rules, with an individual 'worldview'—whose contributions result in dynamic, self-organising behaviour that instigates macroscopic trends in the population. This applies well to human language, understood in this sense as a *complex adaptive system* (CAS) [7, 8] of speakers who collectively attempt to develop a shared communication system. Furthermore, speakers use many different signals and meanings in natural language, also corresponding to interacting 'particles', subject to fundamental rules and dynamics.

Additionally, language exhibits physics-like behaviours such as phase transitions (looking ahead, see Figure 2). More conceptually, the universality of dynamical systems in statistical mechanics [9] finds an analogue in linguistic *universals*—properties or patterns shared by all languages, such as grammatical structures (noun, verb), or the use of vowels and consonants [10].

This agent-based approach has proven useful in many areas of language research, such as first and second language acquisition [11], psycholinguistics [12], language evolution and computational language modelling [7]. In particular, it prompted the formulation of numerous mathematical models of language emergence and evolution (broadly termed *language dynamics*), which are studied using tools from statistical physics to investigate the fundamental mechanisms and underlying universalities of language [13]. Generally, such models present a much-simplified description of language, in which the fundamental assumptions about the system are clearly defined, and thus mathematically tractable. In this sense, they can help determine what mechanisms are necessary or sufficient for language to emerge.

Many models of language dynamics utilise a 'Saussurean' representation of language, where each speaker has a mental conception of mappings between the signals and meanings of the language. This describes a speaker's existing knowledge of a language, which can be static or dynamic, depending on the model considered. The latter case describes the perhaps more realistic scenario where speakers continuously alter their conception of language. Indeed, language users have a fundamental social drive [14] by which they seek to understand—and be understood by—other speakers and as such, they continually adapt their language to their interlocutors.

In the language dynamics literature, two approaches to modelling language can be identified: the *sociobiological* viewpoint, rooted in evolutionary theory, positing that successful communicators enjoy a selective advantage (a 'fitness') and are thus more likely to reproduce than worse communicators [15, 16]; and the *sociocultural* one, which sees language as a self-organising social and cultural system shaped by its users, who continuously seek optimal communicative strategies [1]. The consideration of language as a CAS, defined above, generally aligns more with the latter (placing a special importance on the influence of social interactions), although elements of it are present in both approaches [1]. One can establish a parallel with the long-lasting *nature versus nurture* debate, infamous in the cognitive sciences, which opposes innate and learned cognitive abilities. In the case of language, this might translate to the question of how much of language is shaped by biological (innate) or environmental (learned) factors.

2.3 Modelling language formation and evolution

In this section, we review some notable models of language formation and evolution, and briefly outline the studies' results and assumptions. We encourage the reader to refer to the original papers to understand the full intricacies of the models presented here (for a review on the topic, see Loreto et al. [13]).

Relating to the sociobiological approach, several models have applied principles of evolutionary theory to language, and are thus broadly termed *evolutionary language games* (ELG). 'Language-games' were originally developed as a philosophical concept by Wittgenstein to describe the fundamental rules and conventions governing linguistic interactions [17]. In his conception, every conversation is an instance of a language-game, in which we give words their meaning.

ELG studies generally focus on the development of evolutionary stable strategies as a mechanism of language emergence in a population of agents [18]. The strategies are devised by maximising *payoff*, a notion from game theory [19] which in the case of ELG can be associated with an individual's 'fitness': speakers with a higher payoff have a higher survival chance. In such models, one timestep represents one generation of speakers, who produce offspring in the following one. Speakers are each assigned a Saussurean association of signals and meanings, which they are able to transfer to the next generation.

In Nowak et al. [20] for instance, language is transmitted by enforcing a selection mechanism by which speakers with the highest payoff (the fittest communicators) produce more offspring, thereby favouring more communicative languages. The study observes the convergence of a population towards a shared vocabulary, thus showing that language emergence can be understood as an evolutionary process. More recent ELG studies have made efforts in investigating the interaction topology (the network defining interactions between agents) using tools from graph theory [21, 22]. Some studies also made use of concepts of statistical mechanics to approach evolutionary language games, for instance in Kosmidis et al. [23].

Turning to the sociocultural approach, a significant contribution came from Steels [24, 25], who introduced the concept of the Naming Game (NG), in which language is understood as a self-organising system (CAS) where new words can be acquired or invented. In its simplest form, the NG is a computational model in which agents interact in successive conversations, randomly selecting a name (a signal) for an object. If the name is already used, they adjust their vocabulary by choosing a new name; over time, the agents converge on a shared vocabulary through repeated interactions.

Numerous extensions to the Naming Game were considered in subsequent studies. In particular, one study by Ke et al. [26] developed a model in which agents imitate each other to converge on a shared language. The authors make use of a probabilistic mapping between signals and meanings, an approach that is also followed in the model presented in this report. Interestingly, the authors integrate mechanisms of self-organisation and cultural transmission across generations, inspired from Nowak et al. [20], to show an improved convergence to a common language. As such, they describe language formation as the combination of both biological and cultural factors.

Another notable effort in the field came from Zuidema et al. [27], who developed a model inspired from Nowak et al. [20] that focused on the compositionality of language—the creation of complex meanings from simpler elements, such as words or phrases [28]. The model utilises a deterministic mapping of signals and meanings and introduces a topology (or structure) in the signal and meaning spaces; in other words, signals and meanings can be differentiated. In their interactions, speakers attempt to maximise their communicative success (payoff) similarly to ELG strategies. As a result, the emergence of a common mapping across speakers is observed, which enforces what the authors call topology preservation: similar meanings are associated to similar signals. The introduction of topology among signals and meanings developed in this study is considered in our model as a further implementation (see Section 6.4).

2.4 The Utterance Selection Model

The Utterance Selection Model (USM) is a mathematical model of language change developed by Baxter et al. [29], grounded in Croft's *Utterance Selection Theory* [30]. In broad terms, the theory suggests that language users choose their words and phrases based on the social context and the communicative goal of an interaction. In this sense, a speaker's knowledge of the world, as well as their intended meaning, will determine the utterances (or signals) they select in conversations.

The USM considers what the authors refer to as 'linguistic variables', which can represent various linguistic 'units' of language such as a vowel sounds, words, or the grammatical orderings of a sentence. The model focuses on one linguistic variable, to which different variant forms are associated—for instance, the various words used to describe an object. These variants are then reproduced by a community of speakers in their utterances.

The relation between a speaker and their utterances is called their *grammar*, which contains the entirety of a speaker's knowledge about the language. In the model, a speaker's grammar consists of all probabilities that the speaker will choose a specific variant to express the linguistic variable of interest. These probabilities thus indicate the relative frequencies of use of each variant.

The authors are interested in the competition between variant forms, and specifically the propagation of innovative forms in the language which eventually replace the existing convention. The authors posit that this mechanism is social in nature [30], with speakers choosing variants associated with certain social groups—a preference therefore acting as a selection mechanism for those variants.

The interaction between speakers is modelled as a Markovian stochastic process, in which speakers interact in successive interactions. One such interaction consists of two speakers exchanging a set of *tokens* (instances of a variant) in their utterances, and updating their grammar based on the outcome of the interaction. The exchange can be biased or unbiased, the former allowing for innovation in the language, by forming new variants. The speaker modifies their grammar according to the tokens produced and the success of the interaction—this is specified with weights, which quantify how much change is produced in the grammar as a result of the interaction.

A distinctive feature of the USM is that it separates the dynamics of the utterances from that of the speakers—if the latter undergoes change at all. The authors draw a parallel between the USM and evolutionary theory, arguing that ‘natural selection’ acts on the utterances, rather than the speech community as is implemented in ELG. Indeed, studies such as Nowak et al. [20] postulate that language evolves via a selection mechanism specifying that the speakers using more coherent grammars produce more offspring; as such, there is a direct relation between the propagation of a grammar and the ‘fitness’ of its associated speaker. In contrast, Baxter et al. [29] defend that the survival of a linguistic variant is independent of the fitness of the individual uttering it (rather, variants have a ‘fitness’ of their own). Furthermore they assume that the propagation of language occurs purely from exposure with social status as a selection process, rather than via the maximisation of its coherence (or payoff function)—as is done in ELG studies. In this sense, the USM rejoins the work of Steels [24], treating language as a CAS.

In the authors’ formulation, language change thus becomes possible without the alteration of the structure of the speech community, in timescales shorter than that of one generation—a phenomenon widely observed in language [31].

3 The multi-meaning USM

Whereas the USM was concerned with the evolution of variants of one linguistic variable in speakers’ utterances—in a simplified sense, the evolution of different signals to express one meaning—the model considered in this project allows signals to be mapped to multiple meanings. As such, the model is formulated as an extension to the USM, henceforth referred to as the ‘multi-meaning USM’. It is grounded in the same theoretical background of utterance selection, by which members of a speech community select signals to express intended meanings according to their own understanding of the language. Similarly to the USM, this mental representation is defined as a speaker’s grammar, and it consists of a probabilistic mapping between signals and meanings.

The model understands language as a CAS, in which speakers interact in successive stochastic conversations, and where the dynamics of their signal-meaning associations is separate to that of the speakers. Each conversation sees speakers adapt their own grammar to 'imitate' their interlocutor—i.e. they want to produce meaningful utterances that they communicate to other speakers, to ensure mutual understanding [32]. This mechanism resembles the one used in Ke et al. [26], which presents a similar update algorithm. An important difference comes in the formulation of feedback, which is discussed in Section 3.2.

3.1 Conceptual description

We define the model via its core parameters: the size of the speech community N, the number of signals S, and the number of meanings M. In its simplest formulation, we denote the speakers $i = 1, \ldots, N$, the signals $s = 1, \ldots S$ and the meanings $m = 1, \ldots M$, and make no other distinction between them.

To each speaker i we associate the grammar ϕ_i, a $S \times M$ matrix spanning all signals and meanings, with individual elements corresponding to the signal-meaning mappings (s, m). These are denoted $\phi_i(s|m)$, and represent the probability that speaker i will choose signal s to express meaning m. The probabilities must constitute a density distribution over the signal space for each meaning, such that

$$\sum_s \phi_i(s|m) = 1 \tag{1}$$

must hold for all i, m. This mechanism ensures that if a speaker increases their belief that a specific signal s describes a given meaning m, then their belief that any other (non-s) signal describes m must decrease, hence making them less likely to use them.

Speakers can also become listeners in the model, in which case they perceive a signal s' and infer a meaning m' with probability $\psi_j(m'|s')$, which can be retrieved from ϕ by 'inverting' the probability distribution, such that

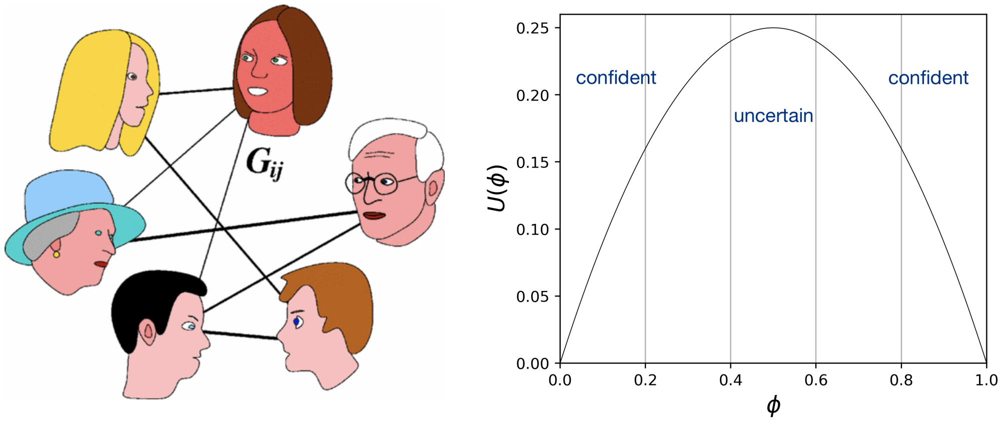
$$\psi_j(m'|s') = \frac{\phi_j(s'|m')}{\sum_{\hat{m}} \phi_j(s'|\hat{m})} \ . \tag{2}$$

This probability dictates how likely a listener is to deduce meaning m' upon hearing signal s' and is normalised along the meaning axis.

The model evolves over time via stochastic *conversations* between two members of the speech community, one counted as the 'speaker' and the other as the 'listener'. Each interaction sees the speaker selecting an intended meaning and producing an associated signal to express to the listener, based on their understanding of the language. The speaker then updates their own grammar based on the success of the interaction. The stochastic formulation of this process is motivated by the inherent uncertainty of conversations in natural language.

3.2 Conversation algorithm

The model is implemented in Monte Carlo simulations via a stochastic algorithm that is iterated over many times. The algorithm describes one conversation, which is taken to last a duration δt, and it assumes the following stochastic steps:



(a) The interaction topology G_{ij}

(b) The nonlinear impact function U

Figure 1: **(a)** An illustration of the interaction topology, retrieved from Baxter et al. [29], which describes the complex social relationships between the different speakers of the system, and specifies the frequency of interaction for each pair of speakers. This is generally nontrivial, and in the simplest case where speakers and listeners are drawn randomly it is equal to $\frac{1}{N(N-1)}$. **(b)** The nonlinear impact function $U(\phi)$, which modulates the strength of the feedback perceived by the speaker, depending on the value of a specific mapping. If the speaker is confident in their understanding of this mapping, i.e. if ϕ is close to 0 or 1, they are less 'trustful' of another speaker's feedback. It is highest when $\phi = 0.5$ ('uncertain').

1. A speaker i and a listener j are selected from the speech community (without replacement) with a probability G_{ij}. This defines the *interaction topology* of the system, describing the complex social networks existing between the members of the speech community and illustrated on Figure 1a. In the simplest formulation of the model, the topology is uniform across the community and thus equal to $\frac{1}{N(N-1)}$.

2. Speaker i chooses a meaning m^*—the *intended* meaning—to convey to the listener with a probability $\rho_i(m^*)$. The associated probability distribution, subject to $\sum_m \rho_i(m) = 1$, describes the relative frequencies of meanings in conversation. In this sense a structure, or topology, can be integrated in the meaning space. In the simplest case, all meanings are equally likely and m^* is selected with uniform probability $\rho = \frac{1}{M}$.

3. The speaker chooses a signal s^* to present to the listener, that he believes best describes the intended meaning m^*. This is selected with probability $\phi_i(s^*|m^*)$, i.e. by referring to their own mental grammar. Once again this is a stochastic process, such that speakers can attempt to use words not as well aligned to m^*, albeit with a lower probability.

4. The listener thus perceives the signal s^*, and associates it to the *inferred* meaning $m^\dagger$ with probability $\psi_j(m^\dagger|s^*)$—that is, by referring to their own conception of the language. Again it is possible for the listener to infer less likely meanings.

5. The listener then provides feedback λ to the speaker, to inform them of how successful the interaction was—this assumes a similar role to the weights in the USM (cf. Section 2.4). In the simplest case, $\lambda = \mu$ when $m^* = m^\dagger$ and $-\mu$ otherwise, where μ is a small positive number. This approach might correspond to the situation where,

after uttering signal s^*, the speaker and the listener both point at the intended and inferred object. The feedback modulates the amount of change per conversation, and it is generally dependent on the meanings intended and inferred, and the identity of the speaker and the listener. Hence we write the *feedback function* explicitly as $\lambda_{ij}(m^*, m^\dagger)$.

6. Finally, the speaker updates their grammar ϕ_i based on the feedback they received. This is implemented with the *update rule*

$$\phi'_n(s|m) = \frac{\phi_n(s|m) + \lambda_{ij}(m^*, m^\dagger)\ U(\{\phi_i(s^*|m^*)\})\ \delta_{m,m^*}\delta_{s,s^*}\delta_{n,i}}{1 + \lambda_{ij}(m^*, m^\dagger)\ U(\{\phi_i(s^*|m^*)\})\ \delta_{m,m^*}\delta_{n,i}}\ , \tag{3}$$

   where the denominator is a normalisation factor to ensure that (1) remains true after the update, such that $\phi'_i(s|m)$ is the updated probability variable. The δ-functions serve to ensure that only the (s^*, m^*) pairing is updated in speaker i's grammar. Thus the grammar update sees the speaker i increasing or decreasing the probability of the mapping uttered in the conversation by the amount λU, where $U(\{\phi\})$ is the *impact function*, which modulates the strength of the feedback on speaker i's (s^*, m^*) mapping.

   In this model, the impact function takes a nonlinear form such that

$$U(\{\phi\}) = \phi(1 - \phi)\ , \tag{4}$$

   as is shown on Figure 1b. U varies depending on the value of ϕ_i for the specific (s^*, m^*) mapping considered—it is maximal when $\phi_i = \frac{1}{2}$ and tends to zero when ϕ_i nears 1 or 0. Therefore, the nonlinear form of U ensures that the speaker adapts their grammar based on how *confident* they are in their use of the mapping of interest: they will be more susceptible to feedback when they are uncertain about the mapping, and less so the more confident about it they become. In this sense, the impact function incorporates the speaker's *experience* into the update rule. This is the principal difference with the update rule specified by Ke et al. [26], which implemented a constant feedback value.

   The nonlinear form of (4) requires that $|\lambda| < 1$, to ensure that the probability $\phi_i(s|m)$ remains within the range $[0, 1]$. In general, the simulations are run with a feedback value of $\lambda = 0.1$, which is small enough to satisfy the condition, and sufficiently large to ensure appropriate computation times.

An aspect of the dynamics defined above that is perhaps unrealistic is the fact that only the speaker learns from an interaction. The listener here contributes as a 'silent observer' returning feedback to the speaker without integrating the result of the conversation, successful or not, into their own grammar. Its justification will come in Section 5.1, in which we derive a mean-field formulation of the model, allowing a formal mathematical analysis of the model. The 'silent observer' assumption facilitates this formulation, which is not possible if the listener also learns from a conversation.

3.3 Regimes of the model

The model allows for different values of S and M, hence we distinguish between three 'regimes': $M = S$ ('symmetric'), $M > S$ ('homonymous') and $M < S$ ('synonymous').

In the first one, each of the M meanings can be assigned to a different signal. This is the most straightforward formulation of the system, on which most of the analysis is conducted. It is particularly useful in understanding the decision of speakers to map signals to meanings from the update rule.

The second regime corresponds to a system with more meanings than signals, i.e. where at least one signal will be used to express more than one meaning. For instance in English, a tree's 'trunk' or 'bark' might also be that of an elephant for the former, and a dog's for the latter if no context is provided. In this sense, we refer to these configurations as *homonymous*. On a more conceptual level, this might correspond to a view of language where the range of intentions that we wish to communicate (meanings) is much greater than the words (signals) that we are able to produce.

The third regime, where more signals than meanings exist, is also applicable to language in the context of synonyms. For instance, one might use the words 'settee', 'sofa' or 'couch' interchangeably to describe the same object. The formulation of the model, as described in the sections above, stipulates that to every meaning corresponds exactly one signal (this unfolds from the probabilistic mapping of signals to a meaning in (1)). For this reason, simulations in this regime generally result in the discarding of redundant (synonymous) signals by the speech community, to obtain a system behaving similarly to the $M = S$ regime described above.

3.4 Blank slate grammars

The first part of the project was concerned with the emergence of consensus on language use in a community of speakers with no initial, pre-established communication system. This requires an initialisation of the system to what we call *blank slate* grammars, with uniform probabilities of assigning a signal to a meaning. From (1) we must therefore have $\phi_i(s|m) = \frac{1}{S}, \quad \forall\ i, s, m$. The simulations and associated analysis, documented in Sections 4 and 5, investigate the evolution of the system from the blank slate grammars to a shared communicative system across the speakers.

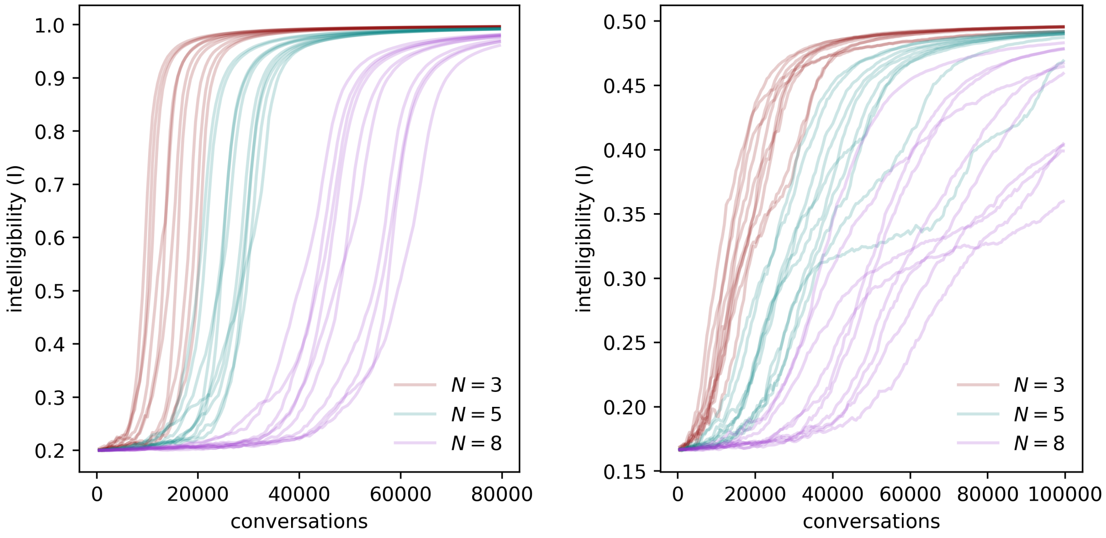
3.5 Language intelligibility

To understand the behaviour of the system and, specifically, how the shared understanding of the language evolves over time, we develop a summary statistic called the *intelligibility*, short-handed as I. This is a measure of how likely it is for a listener to understand a speaker's utterance. The intelligibility of a system is calculated as the probability that any listener will correctly infer another speaker's intended meaning. To compute it, we assume that for the same uttered signal s both the speaker i and the listener j arrive at the same meaning m, and account for the probability distributions involved in one conversation, that is G_{ij}, $\rho_i(m)$, $\phi_i(s|m)$ and $\psi_j(m|s)$. The intelligibility is found by summing over all speakers, signals and meanings in the language, giving

$$I = \sum_i \sum_{j \neq i} G_{ij} \sum_m \rho_i(m) \sum_s \phi_i(s|m) \psi_j(m|s)\ . \tag{5}$$

This is a probability, with low values corresponding to limited understanding and high values to reliable communication among speakers. Its minimal value is $\frac{1}{M}$ for the blank slate grammars across all regimes, where no agreement on communication exists—upon hearing the speaker's utterance, the listener must choose between M equivalent meanings to guess the speaker's intent. As can be seen in Figures 2a and 2b, the intelligibility I is contained in the range $\left[\frac{1}{M}, 1\right]$ for $M \leq S$, and $\left[\frac{1}{M}, \frac{S}{M}\right]$ for $M > S$. The restriction on the maximal intelligibility value in the latter case comes from the unavoidable ambiguities in the final grammars reached by the system (see Figure 3c for illustration).

As shown in Figure 2, intelligibility provides a clear measure of shared understanding in the system over time, with a typical S-shaped curve when evolving from a blank slate grammar to the final shared grammar. It is used in to probe the state of a system, and can aptly represent the different phases that the model goes through, as discussed in Section 5. One general finding is that the model spontaneously maximises the intelligibility of the language across speakers.



(a) Intelligibility for $(S, M) = (5, 5)$

(b) Intelligibility for $(S, M) = (3, 6)$

Figure 2: (a) Evolution of intelligibility I for the simplistic system with $S = M = 5$ and $\lambda = 0.1$ for 10 MC runs for different speech community sizes N. The convergence to an final grammar system with $I = 1$ is slower for larger communities. **(b)** Intelligibility of the homonymous system with $S = 3$, $M = 6$ and $\lambda = 0.1$ for 10 runs per N value. The convergence of the latter occurs over a longer timescale, and with more variability. The maximal intelligibility value for the system is $\frac{S}{M} = \frac{1}{2}$.

4 Understanding language formation

This first part of the project is concerned with the emergence of consensus on language in a community of speakers with no prior means of communication. As such, the system is initialised at the blank slate grammars, in what we call the *simplistic* formulation of the model, where all speakers, signals and meanings are interchangeable. In the simplistic model, speakers have no preconceived preferences or biases for certain languages over others, hence in mathematical terms its functions are defined as follows:

- $\rho_i(m) = \frac{1}{M}$, a uniform meaning topology,

- $G_{ij} = \frac{1}{N(N-1)}$, a uniform interaction topology, and
- $\lambda_{ij}(m^*, m^\dagger) = \mu(2\delta_{m^*,m^\dagger} - 1)$, ensuring $+\mu$ if $m^* = m^\dagger$ and $-\mu$ otherwise.

By disregarding potential asymmetries among the speakers or in the signal and meaning spaces, the simplistic system therefore probes the self-organising behaviour of the model, in which communication is reached only by internal adjustment of the speakers to listeners' feedback. The typical evolution of this system, initialised at the blank slate, first exhibits a phase that we call *hesitation*—this corresponds to a period of uncertainty in the language where every speaker's grammar experiences random fluctuations about the blank slate, with no clear evolution towards consensus.

Hesitation is generally marked by a low and nearly constant intelligibility, fluctuating near its minimal value of $1/M$. Its duration varies greatly from one simulation to the next, underlining the fundamental stochastic nature of the model, and it depends significantly on the system parameters. This is explored in Section 4.3.

Hesitation in the system ends once a majority of speakers enter a *consensus-forming* phase, in which they begin to agree on the same mappings between signals and meanings. This occurs once the system reaches a point of irreversible change (a 'critical point'), making it stable to fluctuations. After this, we generally observe a global differentiation of the signals and meanings across all speakers, taking place on a short timescale. This is akin to a phase transition, from hesitation to consensus, identified by a sharp increase in intelligibility and ultimately attaining a common grammar across all speakers. This observation is in accordance with previous studies (cf. Section 2.3), specifically highlighting the formation of language by a self-organising process of cultural readjustment.

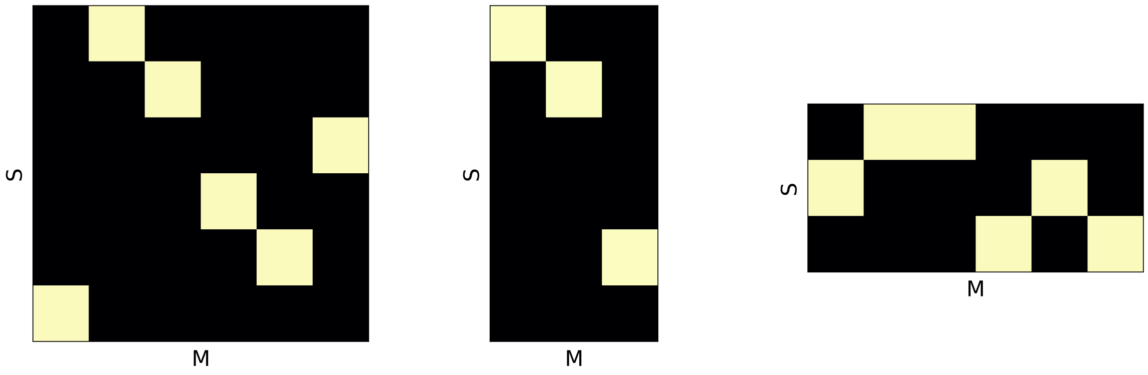
Additionally, we observe an asymptotic convergence towards maximal intelligibility, resulting in the S-shape observed in Figure 2. This feature originates from the impact function U, which diminishes the strength of listeners' feedback the closer the system approaches the final state. The mechanisms responsible for this initial 'irreversible' escape of the hesitation phase are investigated in Section 5.

4.1 Optimal grammars

A remarkable aspect of these common mappings is their spontaneous maximisation of communication (intelligibility), as was seen in Figure 2. The resulting common grammars all assign a different signal to each meaning, and in the $(M > S)$ regime the signals are evenly shared by multiple meanings—minimising ambiguity. As such, we refer to these systems as *optimal grammars*, examples of which are shown in Figure 3.

In the case of a symmetric grammar system $(M = S)$, optimal grammars have each meaning assigned to a different signal, as shown in Figure 3a. In the $M < S$ regime, as in Figure 3b, the same states are observed with the additional $(S - M)$ redundant signals being discarded from the language. Finally, in the case of $M > S$, where homonymy cannot be avoided, we note a preference of the system for minimising the number of redundant meanings used for any given signal, thereby reducing ambiguity.

Although previous studies [18, 26] have already noted that shared mappings could maximise communication, less attention has been given to the structure of optimal grammars,



(a) $M = S$ **(b)** $M < S$ **(c)** $M > S$

Figure 3: Optimal grammars shown for the three regimes of the system. The first one is for a $(S, M) = (6, 6)$ configuration, with each meaning being assigned to a different signal, thereby avoiding redundancy and ambiguity, and optimising communication. The second grammar corresponds to a $(6, 3)$ configuration, with three redundant signals being discarded by the speakers. The third one is a homonymous $(3, 6)$ optimal grammar, in which every signal describes two meanings instead of one: to express all meanings then, ambiguity (via homonymy) is unavoidable. We observe that these grammars tend to minimise the number of meanings associated with each signal. In the simplistic model, all signals and meanings are interchangeable; their mappings are therefore arbitrary, and vary in every MC run.

and the mechanisms leading to their convergence. This implicit 'preference' for optimal grammars in the simplistic system motivates much of the subsequent analysis of the model. This is developed in Section 5, which aims to understand the origin of this preference from a mathematical standpoint.

4.2 Grammar multiplicity

To probe and quantify this result, we investigate the distribution of signal-meaning mappings in these systems, with and without a preference for optimal configurations.

To do so, we calculate the expected probabilities of configurations from their multiplicity—that is, the number of different ways that an arrangement can be obtained from permutations of the signals and meanings. We concentrate on the $M > S$ regime in particular as its optimal configurations must also minimise ambiguity. We thus compare the distribution of configurations observed in the simulations to that expected from the multiplicity calculations.

With this in mind, we consider the number n_k of blocks featuring k redundant meanings with $0 \leq k \leq M$, so that a final grammar can be fully defined by the arrangement $(\{n_k\}_k) = (n_1, n_2, \dots n_M)$, see Figure 4 for illustration. For reference, the homonymous grammar shown in Figure 3c has the arrangement $(n_1 = 0, n_2 = 3)$—also described by '222' when considering the meaning redundancy k.

Additionally, we infer that $\sum_k n_k = S$ and $\sum_k k n_k = M$. In the case of no preference for optimal communication, a speaker will have the choice to associate each of the M meanings with any of the S signals, and this results in a total number of possible arrangements equal to S^M.

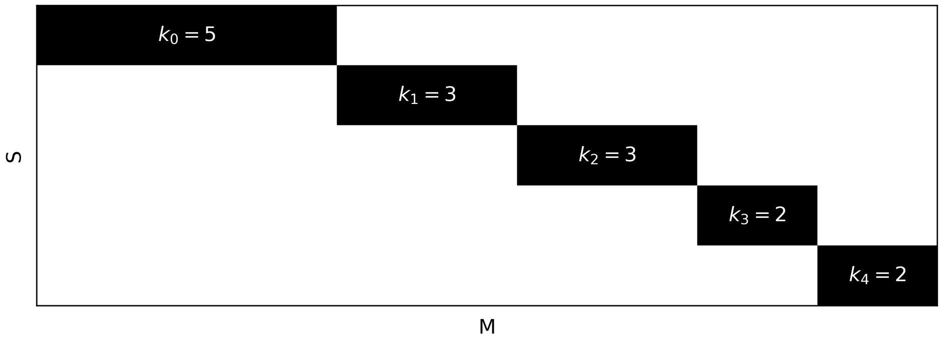


Figure 4: Illustration of multiplicity for a specific homonymous grammar ($M > S$). The k values give the number of redundant meanings for each signal, i.e. the block length. n_k is the number of blocks that have the same k; in this example we find $n_5 = 1$, $n_3 = 2$ and $n_2 = 2$. Looking ahead to Figure 5, this system might also be referenced as '53322'—noting that this denomination represents all systems produced when the signals and meanings are permuted, not just the one shown here.

Finally, we must calculate the multiplicity of the arrangement ($\{n_k\}_k$) to understand how likely the system is to reach it. This is done by considering the number of ways that meanings and signals can be permuted to obtain a given arrangement: since meanings are all equivalent in the simplistic model, the arrangements are not concerned with which the specific meanings in each block, or which order they take—and the same can be said of the signals.

This yields the multiplicity

$$\mathcal{M}(\{n_k\}_k) = \frac{M!}{(1!)^{n_1}(2!)^{n_2}\dots(M!)^{n_M}} \times \frac{S!}{(n_1!)(n_2!)\dots(n_M!)} \tag{6}$$

of the arrangement $(n_1, n_2, \dots n_M)$. The first fraction calculates the number of ways that the meanings can be permuted to obtain a different arrangement of blocks, and the second does the same for permuting signals. The probability of obtaining this arrangement, or grammar, if the system shows no preference for optimal communication, is therefore

$$P(\{n_k\}_k) = \frac{\mathcal{M}(\{n_k\}_k)}{S^M}. \tag{7}$$

In Figure 5, the distribution of configurations reached by the simulations of the $(N, S, M) = (2, 3, 6)$ system is compared to that produced from the multiplicity calculations, which assume that specific grammar configurations are reached by chance. It can clearly be seen that the model presents a bias towards optimal grammars ('222'), with a minor fraction of the languages reaching a sub-optimal ('123') configuration. This result highlights the fundamentally stochastic component of convergence, which can orient the system towards less efficient—more ambiguous—grammars in the $M > S$ regime. In $M \leq S$ systems, the model systematically converges to optimal configurations, in all cases.

4.3 Hesitation in the three regimes

As described above, the hesitation phase marks the time necessary for the speakers to reach a first form of irreversible consensus across the language, and it varies significantly

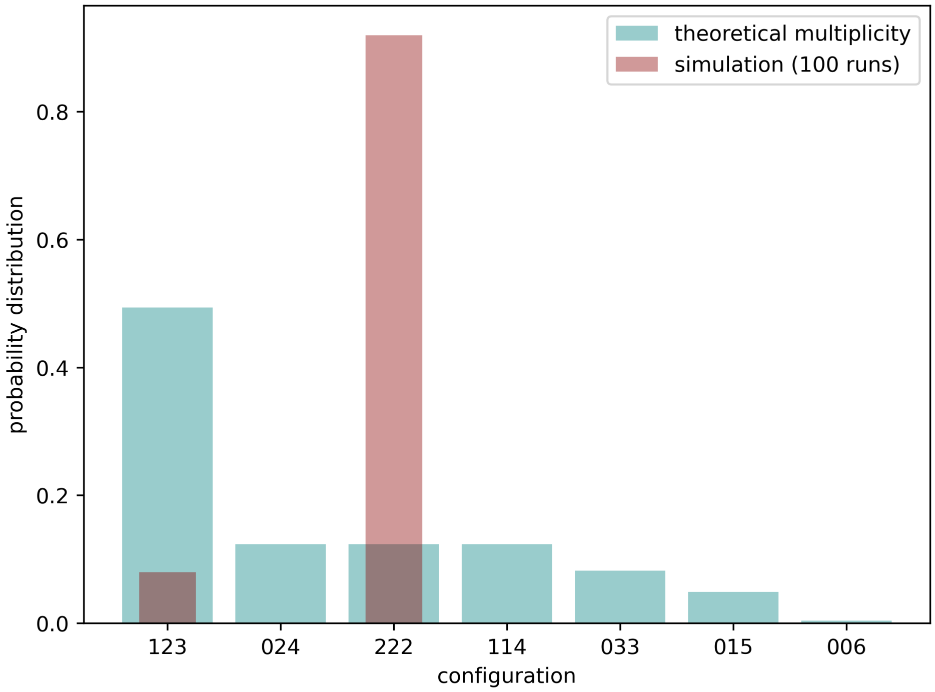
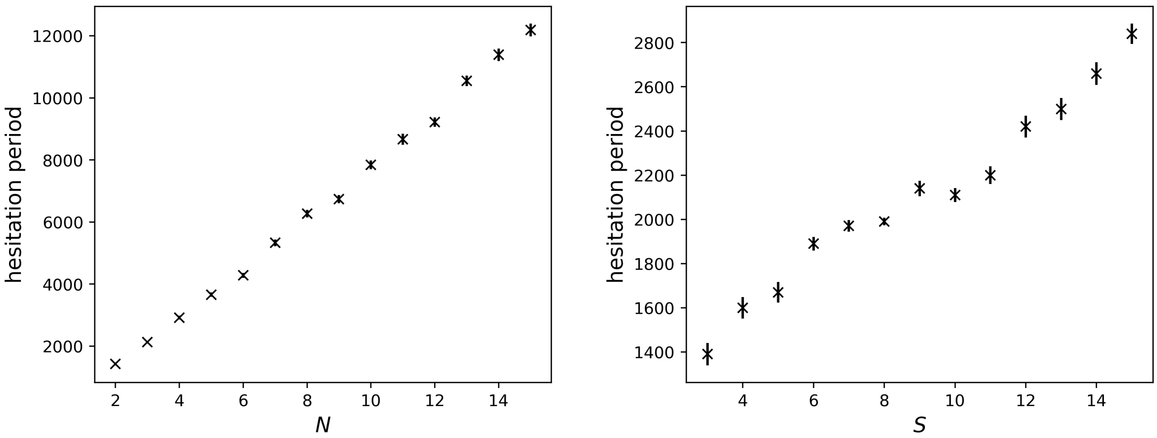


Figure 5: The distribution of configurations for the $(N, S, M) = (2, 3, 6)$ system, obtained from the results of 100 Monte Carlo simulations (red) and from the multiplicity calculations (blue). Clearly, the model does not follow the expected distribution of ‘chance’ configurations, showing instead a bias towards the most communicative arrangements, quoted here as ‘222’ and ‘123’. The configurations are denoted by the different number of meanings associated with each signal, i.e. the meaning redundancy k, regardless of their order, for clarity. The ‘222’ arrangement is that of the homonymous optimal grammar, as shown in Figure 3b. The small probability of the system converging towards the ‘123’ configuration, which is slightly less efficient (more ambiguous), highlights the role of stochastic fluctuations in the system, as well as the relative stability of near-optimal grammars in the $M > S$ regime.

from one simulation to the next. In this section, we explore its dependence on the system parameters N, S, and M for the synonymous $(M < S)$, homonymous $(M > S)$ and symmetric $(M = S)$ regimes. We note that since the consensus phase is relatively short in comparison, the duration of the hesitation period is also a good measure of the time taken by the system to reach a common mapping.

It is measured with a threshold intelligibility value, taken to be $\frac{1.3}{M}$ (130% of the blank slate value), which effectively defines consensus across all three regimes without mistaking it for random fluctuations. The typical hesitation time is estimated by averaging over several MC runs.

First, the dependence on the speech community size N is investigated for the $(S, M) = (3, 3)$ system, with each data point averaging 100 runs. Figure 6a shows a near-linear increase of the hesitation time for increasing N, which makes intuitive sense: with more speakers interacting, we expect an initial consensus to take longer to reach. Figure 6b shows the dependence of hesitation on S in the synonymous regime, with $N = 2$, $M = 3$ and $S \geq M$. We note a relatively small increase in hesitation, which barely doubles when S is increased five-fold. We conclude that the speakers discard redundant signals on relatively short timescales in the simplistic model.

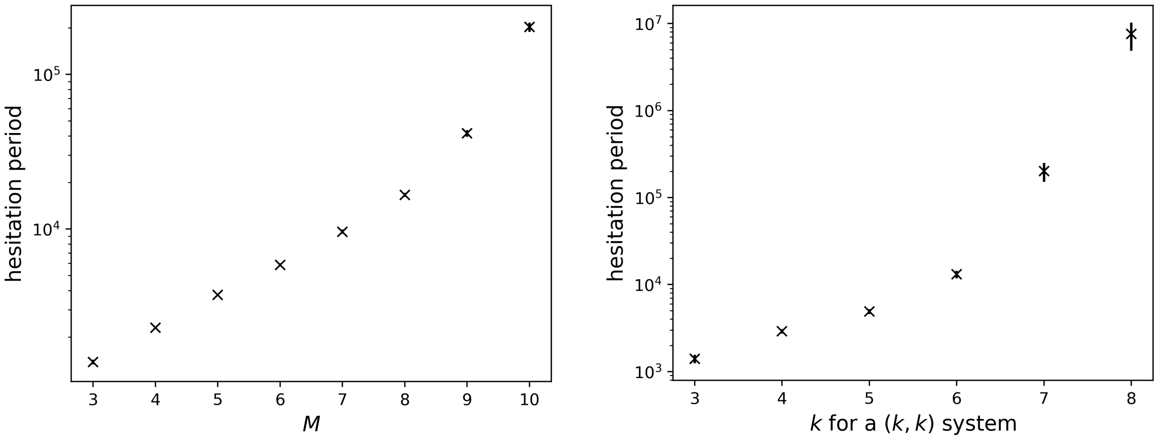


(a) Dependence of hesitation on N

(b) Dependence of hesitation on S

Figure 6: Length of hesitation phase against N and S. Each data point averages 100 runs, assuming a Gaussian error. Hesitation increases nearly linearly with N, implying that larger communities take more time to reach a consensus. The dependence on S for $N = 2$, $M = 3$ is also near-linear, with a relatively small increase for larger S. Hence the 'removal' of redundant signals in synonymous systems is a rapid process.

Figure 7a shows the relation between hesitation and the number of meanings M in the homonymous regime, with $N = 2$ and $S = 3$. We observe a much longer time to consensus, with a strong dependence on increasing M. We can understand this result from the added ambiguity brought by homonymy, which hinders the mutual understanding of speakers in conversations. The additional increase noted for $M \geq 9$, where every signal is assigned 3 meanings instead of 2 (for at least one of them), further confirms this hypothesis.



(a) Dependence of hesitation on M

(b) Hesitation for the symmetric system

Figure 7: Dependence of the hesitation phase on M (in the homonymous regime) and on the symmetric system size, denoted as k (the latter only averages over 10 MC runs due to the large timescales encountered). Increasing M in the homonymous regime significantly increases hesitation, due to the increasing ambiguity in conversations. When both S and M are increased simultaneously in the symmetric system, the hesitation time is much larger, reaching 'unrealistic' timescales for relatively small values of k.

Finally, we investigate the hesitation phase for the symmetric system in Figure 7b. A very large increase in hesitation time is observed for systems ranging from $(3,3)$ to $(8,8)$—the latter taking close to 10 million time steps (or conversations) to reach consensus. This underlines an important limitation of the simplistic model, which quickly requires an unrealistic amounts of conversations to converge on a common language when $S \gg 1$, $M \gg 1$.

These results echo those of Ke et al. [26], who also observe a convergence towards shared grammars only for small population sizes and limited vocabulary dimensions, albeit for a slightly different model formulation. The authors' findings lead them to suggest that self-organisation is only one part of the whole story, since natural language evidently relies on a much larger set of signals and meanings. They go on to conclude that language acquisition must also incorporate a mechanism of cross-generational transmission, which speeds up the convergence.

Although the multi-meaning USM does not incorporate such a mechanism, it would be interesting to look for alternative processes that might result in accelerated convergence for greater systems. There are multiple possibilities to consider, for instance with the different topologies that can be introduced in the speech community (G_{ij} and λ_{ij}), and the meaning space ($\rho_i(m)$). One could think of introducing a preference among speakers to first converge on one particular meaning (more relevant to conversation), or inversely, to have social valuation influence interactions in the speech community. This is briefly mentioned in the context of language change in Section 6, however a full consideration of these effects on language formation is left as a future implementation to the model.

5 Stability analysis of the simplistic model

This section aims to probe the convergence of the model to optimal grammars with more rigorous mathematical analyses. To do so, we first derive a mean-field prescription of the model (see Section 5.1), yielding *deterministic* 'equations of motion' (ODEs) by taking a continuous-time limit. These neglect the stochastic contributions of the dynamics, and can therefore be numerically integrated. Subsequently, the concept of linear stability analysis is introduced and defined starting from the deterministic model. Its application to different systems and regimes yields important insights into the final optimal configurations favoured by the simplistic model in the context of language formation.

5.1 Mean-field equations

In the context of complex dynamical systems characterised by a large number of interacting components (or agents), it is often difficult, if not impossible, to obtain a complete description of the emerging phenomena and behaviours that result from their interactions. A common approach to deal with this issue is to assume a *mean-field* description of the system, where the interactions between agents are approximated as a single effective interaction that depends on the average behaviour of the system [33]. This has the effect of significantly reducing the complexity of these systems, notably for those exhibiting nonlinear dynamics such as our update rule (3).

It is therefore a natural approach for the model, yielding a set of deterministic differential equations describing the macroscopic behaviour of the system. By effectively removing the stochastic component of the model, we are able to inspect its *average* evolution analytically.

Our approach focuses on the multidimensional system variables $\phi_n(s|m)$, which experience an interaction averaged out over the whole system for one conversation. We first compute the change in the system after one conversation $\delta\phi_n(s|m) = \phi'_n(s|m) - \phi_n(s|m)$, which, by expanding the update rule (3) to first order in λ, gives

$$\delta\phi_n(s|m) = \lambda_{ij}(m^*, m^\dagger)\ U([\phi_i(s^*|m^*)])\ \delta_{n,i}\delta_{m,m^*}\ [\delta_{s,s^*} - \phi_i(s|m^*)]\ . \tag{8}$$

To observe the average change for any possible conversation allowed by the model, we must sum over the overall probability distribution from which the random variables i, j, m^*, s^* and $m^\dagger$ are drawn:

$$P(i, j, s^*, m^*, m^\dagger) = G_{ij}\rho_i(m^*)\phi_i(s^*|m^*)\psi_j(m^\dagger|s^*)\ . \tag{9}$$

Averaging both sides of (8) and relabelling dummy indices ($n \leftrightarrow i$, $\hat{n} \leftrightarrow j$, $\hat{s} \leftrightarrow s^*$, $m \leftrightarrow m^*$ and $\hat{m} \leftrightarrow m^\dagger$) for clarity, we obtain

$$\langle\delta\phi_n(s|m)\rangle = \rho_n(m)\sum_{\hat{n},\hat{s},\hat{m}} G_{n\hat{n}}\lambda_{n\hat{n}}(m,\hat{m})[\delta_{s,\hat{s}} - \phi_n(s|m)]\phi_n(\hat{s}|m)U(\phi_n(\hat{s}|m))\psi_{\hat{n}}(\hat{m}|\hat{s})\ . \tag{10}$$

This expression gives the average change of any grammatical variable, $\delta\phi_n(s|m)$, in terms of the known probability distributions of the system, thus removing the stochastic components of the model. We introduce a time dependence in the equation above by assuming that the feedback λ is small, and can be written as $\mu_{n\hat{n}}(m,\hat{m})\delta t$ where $\mu_{n\hat{n}}(m,\hat{m})$ is the rate of change of feedback, and δt represents the duration of one conversation. We then divide (10) by δt to facilitate a continuous-time description of the grammar update by taking $\delta t \to 0$, giving

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\phi_n(s|m)\rangle = \rho_n(m)\sum_{\hat{n},\hat{s},\hat{m}} G_{n\hat{n}}\mu_{n\hat{n}}(m,\hat{m})[\delta_{s,\hat{s}} - \phi_n(s|m)]\phi_n(\hat{s}|m)U(\phi_n(\hat{s}|m))\psi_{\hat{n}}(\hat{m}|\hat{s})\ , \tag{11}$$

shorthanded to

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\phi_n(s|m)\rangle = F_{nsm}(\{\phi\})\ . \tag{12}$$

This constitutes a set of differential equations in terms of the variables $\phi_n(s|m)$, spanning the three dimensions (n, s, m).

These facilitate the identification the *fixed points* of the dynamics, denoted ϕ^*, such that the time derivative of all grammar variables (the left hand side of (12) at those points is zero. In other words, they must thus satisfy $F_{nsm}(\{\phi^*\}) = 0$. In this way, we find that blank slate grammars are fixed points of the dynamics, and so are optimal grammars.

In the case of the simplistic system, as defined at the start of this section, the right-hand side of (12) simplifies to

$$F_{nsm}(\{\phi\}) = \\ \frac{\mu}{N(N-1)M}\sum_{\hat{s}}\phi_n(\hat{s}|m)^2[1 - \phi_n(\hat{s}|m)][\delta_{s,\hat{s}} - \phi_n(s|m)]\sum_{n\neq\hat{n}}\left[\frac{2\phi_{\hat{n}}(\hat{s}|m)}{\sum_{\hat{m}}\phi_{\hat{n}}(\hat{s}|\hat{m})} - 1\right]\ . \tag{13}$$

This expression can be implemented in a numerical integration algorithm, to investigate the behaviour of the simplistic model averaged over the whole system at an arbitrary time t. This technique reduces the computation time dramatically compared to the Monte Carlo simulations, at the expense of removing all stochastic contributions from the model; an example of this is shown in Figure 8. It also proves useful for the development of *linear stability analysis*, which is discussed below.

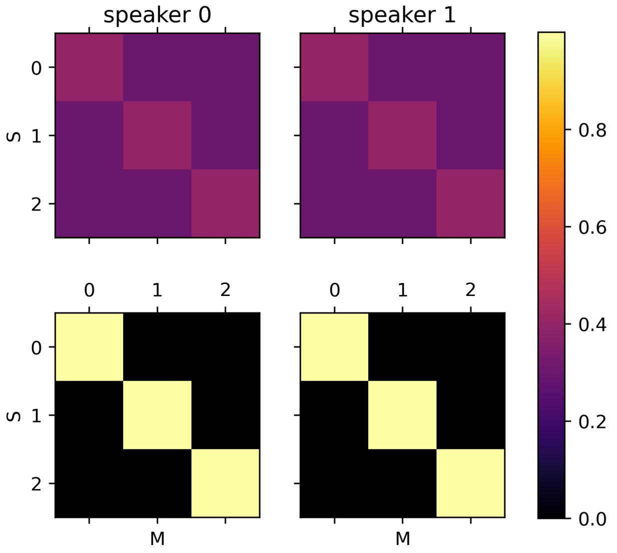


Figure 8: Implementing the deterministic mean-field equations into a numerical integration algorithm. Starting from a 'hazy' $(N, S, M) = (2, 3, 3)$ system with a slight preference for the mappings along the diagonal (top row), the deterministic equations predict a convergence to the corresponding optimal grammar (bottom row). Since the mean-field equations have no stochastic elements, the convergence to a final grammar is always fully defined by the initial state.

5.2 Linear stability analysis

In order to further understand the preference for optimal grammars established by the system, we turn to linear stability analysis, a mathematical method probing the effects of small perturbations about a fixed point of a dynamical system. The technique can determine whether these perturbations will grow or decay with time, therefore yielding insights on the evolution of the system from these points to linear order. Linear stability analysis is regularly used to understand the behaviour of complex physical systems such as fluid flows, chemical reactions, and biological systems [34].

In relation to the model, the effects of fluctuations about the blank slate grammars are investigated to understand how the simplistic model establishes a preference for optimal configurations. To do so, the perturbation $\epsilon_{nsm} = \phi_n(s|m) - \phi_n^*(s|m)$ is defined, representing the evolution of the system from the fixed point $\{\phi^*\}$ of the dynamics. The perturbation must be subject to the constraint

$$\sum_s \epsilon_{nsm} = 0, \tag{14}$$

ensuring that an increase in ϕ at a given m is balanced out by a decrease in all other mappings associated with m. This perturbation is substituted into the deterministic equations (11), which are then linearised—keeping only first order terms in ϵ. This yields a set of linear equations describing the time derivative of the perturbation

$$\frac{\mathrm{d}}{\mathrm{d}t}\epsilon_{nsm} = A_{nsm}^{n's'm'}\epsilon_{n's'm'} \ , \tag{15}$$

in which we have made use of a summation convention where upper indices are contracted with lower indices when their symbols coincide, with an implied sum over n', s' and m'. The coefficient associated with each perturbation component $\epsilon_{n's'm'}$ is given by

$$A_{nsm}^{n's'm'} = \left.\frac{\partial F_{nsm}(\{\phi\})}{\partial \phi_{n'}(s'|m')}\right|_{\phi=\phi^*} \ . \tag{16}$$

Equation (15) can also be written as a matrix equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\epsilon} = A\boldsymbol{\epsilon} \tag{17}$$

where all perturbation terms of the system are combined into a single vector, $\boldsymbol{\epsilon}$, and the linear coefficients form a symmetric matrix A, with a set of eigenvectors $\boldsymbol{u}_{nsm}$ and corresponding eigenvalues w_{nsm}, spanning the entire system.

For a general system of the form (17), linear stability analysis is typically concerned with the real part of the eigenvalues, the sign of which determines the fate of the perturbation over time. If the real part of the eigenvalue is negative, the perturbation decays away and the system returns to its unperturbed state; if it is positive, the perturbation grows exponentially, eventually leading to instability [34].

In the case of the model, we can understand this by expanding the perturbation vector $\boldsymbol{\epsilon}$ in the eigenbasis of A, giving

$$\boldsymbol{\epsilon}(t) = \sum_{n's'm'} a_{n's'm'} \ \exp(w_{n's'm'}t) \ \boldsymbol{u}_{n's'm'} \ , \tag{18}$$

in which we introduce an explicit time dependence for clarity. The coefficients in the expansion are given by $a_{nsm} = \boldsymbol{u}_{nsm} \cdot \boldsymbol{\epsilon}(0)$, where $\boldsymbol{\epsilon}(0)$ is the vector of initial displacements from the fixed point. The implicit $t = 0$ simply comes from the fact that a_{nsm} is time-independent. The exponential term makes the dependence on the sign of the eigenvalue more apparent: a positive w_{nsm} will 'amplify' (favour) a displacement in the direction of $\boldsymbol{u}_{nsm}$, whereas a negative w_{nsm} will 'dampen' (disfavour) it. Thus, by inspecting the eigenvalues of A and their corresponding eigenvectors, we can gain insights on the behaviour of the system before convergence to a final grammar. The eigenvectors with a positive eigenvalue are understood here as the preferred *directions of convergence* of the system near the blank slate grammars, provided that they satisfy the constraint (14).

5.2.1 The simplistic model matrix

In the case of the simplistic model, we turn to (16) with the simplistic $F_{nsm}(\{\phi\})$ in (13). In the summation convention defined previously, we obtain

$$\begin{aligned} A_{nsm}^{n's'm'} = \frac{\mu}{NSM}\Bigg\{ & \left(1-\frac{2}{M}\right)\left(2-\frac{3}{S}\right)\delta_n^{n'}\left[\frac{1}{S}-\delta_s^{s'}\right]\delta_n^{n'} \\ & +\left(1-\frac{2}{M}\right)\left(1-\frac{1}{S}\right)\delta_n^{n'}\delta_s^{s'}\delta_m^{m'} \\ & +\frac{2}{(N-1)M}\left(1-\frac{1}{S}\right)\left[1-\delta_n^{n'}\right]\left[\frac{1}{S}-\delta_s^{s'}\right]\left[\frac{1}{M}-\delta_s^{s'}\right]\Bigg\} , \end{aligned} \tag{19}$$

where we have introduced the Kronecker delta $\delta_\alpha^{\alpha'}$, equal to 1 if $\alpha = \alpha'$ and zero otherwise.

In order to find the eigenvalues of this matrix, we first consider the general operator (representing a general form of the terms in A)

$$K_{nsm}^{n's'm'} = \left(a_N + b_N\delta_n^{n'}\right)\left(a_S + b_S\delta_s^{s'}\right)\left(a_M + b_M\delta_m^{m'}\right) . \tag{20}$$

Different values for a and b allows distinctions to be made between identical and non-identical speakers, signals and meanings, which is consistent with the symmetries of the simplistic model. If all b terms are zero, the operator represents a global interaction; if instead the a term vanishes along a specific axis, say N, the interaction applies only when $n = n'$.

We first concentrate on the operator

$$O_\alpha^{\alpha'} = a + b\delta_\alpha^{\alpha'} , \tag{21}$$

which has the eigenvector

$$u^{(k=1)} = (1\ 1\ 1\cdots)^T \tag{22}$$

with eigenvalue

$$w^{(k=1)} = a|O| + b , \tag{23}$$

where $|O|$ is the dimensionality of O. The remaining eigenvectors are $(|O|-1)$-fold degenerate and can be specified as

$$u^{(k>1)} = (0\cdots 0\ 1\ \ -1\ 0\cdots 0)^T \tag{24}$$

with eigenvalue

$$w^{(k>1)} = b . \tag{25}$$

The eigenvalues of O can be generalised as

$$w^{(k)} = a|O|\delta_{k,1} + b . \tag{26}$$

The eigenvectors of K and their degeneracy are thus specified by the three 'quantum numbers' $k_N = 1, \ldots N$, $k_S = 1, \ldots S$ and $k_M = 1, \ldots M$. We write the generalised eigenvector and its associated eigenvalue as

$$\boldsymbol{u}_{nsm}^{(k_N,k_S,k_M)} = u_n^{(k_N)}u_s^{(k_S)}u_m^{(k_M)} , \tag{27}$$

$$w^{(k_N,k_S,k_M)} = (a_N N\delta_{k_N,1} + b_N)(a_S S\delta_{k_S,1} + b_S)(a_M M\delta_{k_M,1} + b_M)\ . \tag{28}$$

The latter introduces degeneracy in the system depending on the value of k_N, k_S and k_M. As for the operator O, the degeneracy is 1 if $k_\alpha = 1$, and $(\alpha - 1)$ if $k_\alpha > 1$ for N, S and M separately.

Bringing it all together, we can therefore find the eigenvalues and eigenvectors of $A_{nsm}^{n's'm'}$ in (19) by inspecting each of its terms separately. This is done by determining the a and b coefficients along each axis, yielding the eigenvalues $w^{(k_N,k_S,k_M)}$ which depend on the quantum numbers of the system. We find the overall eigenvalue by summing the contributions of each term, considering the different combinations of (k_N, k_S, k_M). This is done below for the simplistic system.

5.2.2 Finding the eigenvalues

We start by considering k_S: noting the factors $\left(\frac{1}{S} - \delta_s^{s'}\right)$ in the first and third term, we infer $a_S = \frac{1}{S}$ and $b_S = -1$. Hence when $k_S = 1$, the contribution to the eigenvalue is $\frac{S}{S} - 1 = 0$, and only the second term contributes, for any values of k_N, k_M (since all of its terms a_N, a_S and a_M are zero).

We thus deduce the first eigenvalue

$$w^{(k_N,k_S=1,k_M)} = \frac{\mu}{NSM}\left(1 - \frac{2}{M}\right)\left(1 - \frac{1}{S}\right) \tag{29}$$

which is NM-fold degenerate and non-negative for $S > 1$ and $M > 1$. Since $k_S = 1$ the corresponding eigenvectors are uniform along the signal axis, thus violating the constraint (14)—this also implies that their coefficient in the expansion (18) is zero. In other words, the 'unphysical' eigenvectors are removed from the expansion and can be disregarded.

For $k_S > 1$, we consider the case where $k_M = 1$, which sees the third term vanishing due to the $\left(\frac{1}{M} - \delta_m^{m'}\right)$ factor, for the same reason as above. The first term has $b_N = 1$, $b_S = -1$ and $b_M = 1$, and its eigenvalue contribution is added to that of the second term, yielding

$$w^{(k_N,k_S>1,k_M=1)} = -\frac{\mu}{NSM}\left(1 - \frac{2}{M}\right)\left(1 - \frac{2}{S}\right) \tag{30}$$

after summing the terms. This is $N(S-1)$-fold degenerate and always negative for $S > 2$ and $M > 2$. These suppressed eigenvectors have a uniform meaning axis, i.e. they associate the same signal to all meanings simultaneously.

In the case $k_S > 1$, $k_M > 1$ we first consider $k_N = 1$. The first two terms yield the same contribution to the eigenvalue, as they are independent of k_N; the third term on the other hand features the factor $\frac{1}{N-1}(1 - \delta_n^{n'})$ which contributes $\frac{1}{N-1}(N-1) = 1$ to the eigenvalue when $k_N = 1$. Thus by summing the terms the eigenvalue is

$$w^{(k_N=1,k_S>1,k_M>1)} = \frac{\mu}{NSM}\left\{\frac{2}{M}\left(1 - \frac{1}{S}\right) - \left(1 - \frac{2}{M}\right)\left(1 - \frac{2}{S}\right)\right\}\ . \tag{31}$$

This is $(S-1)(M-1)$-fold degenerate and can be positive or negative depending on S and M. The corresponding eigenvectors have all speakers differentiating the signals and meanings simultaneously. This is of special interest to our search for the directions of convergence of the system near the blank slate grammars, and will be investigated in detail in the following section.

Finally we are left with the case $k_S > 1$, $k_M > 1$ and $k_N > 1$, which only gathers the b-contributions from each term. Taking special care with the signs of the terms, we find

$$w^{(k_N>1,k_S>1,k_M>1)} = -\frac{\mu}{NSM}\left\{\frac{2}{(N-1)M}\left(1-\frac{1}{S}\right)+\left(1-\frac{2}{M}\right)\left(1-\frac{2}{S}\right)\right\} \tag{32}$$

which is $(N-1)(S-1)(M-1)$-fold degenerate and always negative. This corresponds to different speakers differentiating the signals and meanings in opposite directions, i.e. where speakers contradict each other. As above, because of the negative eigenvalue these modes are suppressed.

In summary, this preliminary inspection of the simplistic system using linear stability analysis yielded the different eigenvalues taken by the model once a perturbation is introduced near the initial blank slate grammars. In turn, by considering their sign and degeneracy, the corresponding eigenvectors can be understood as the possible directions taken by the system away from the equilibrium.

5.2.3 Linear stability results for the simplest system

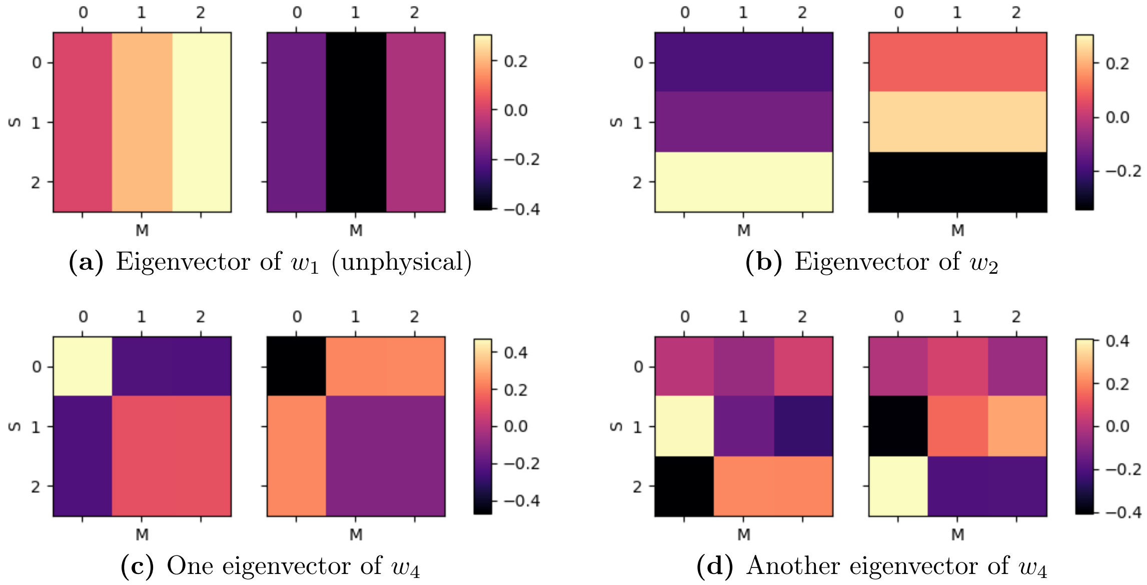
The mathematical formulation of linear stability analysis developed above is investigated for the most basic system configuration, $(N, S, M, \mu) = (2, 3, 3, 0.1)$. This is combined with a standalone computational approach to find the eigenvalues and eigenvectors of the A matrix defined in (19) in order to verify the mathematical derivation of the eigenvalues and their degeneracy, as well as the configuration of the eigenvectors. Subsequently, the implications of these results for different systems and regimes of the model are explored and related to the evolution of the system from the blank slate grammars towards the optimal language systems. Finally, the limitations of the method are discussed.

For the $(2, 3, 3, 0.1)$ system, the eigenvalues in (29–32) are calculated as

- $w_1 = \frac{1}{810}$ (6-fold degenerate) – ‘unphysical’
- $w_2 = -\frac{1}{1620}$ (4-fold degenerate)
- $w_3 = \frac{1}{540}$ (4-fold degenerate)
- $w_3 = -\frac{1}{324}$ (4-fold degenerate)

Computing the A matrix and finding its eigenvalues and eigenvectors numerically returns the same eigenvalues, with identical degeneracies. In fact, the mathematical derivation was found to match the numerical computation of the eigenvalues for every tested system, in all three regimes.

The eigenvectors of the system were also obtained from the matrix, and are shown in Figures 9 and 10. The numerical computation returned a total of 18 eigenvalues and eigenvectors, in agreement with the expected degeneracies. The specific values of each (n, s, m) component of the eigenvectors are only significant in their relative amplitudes, with greater values (positive or negative, since a global sign flip can be performed) being associated with stronger 'preferences' towards a given direction of convergence.



(a) Eigenvector of w_1 (unphysical)

(b) Eigenvector of w_2

(c) One eigenvector of w_4

(d) Another eigenvector of w_4

Figure 9: Eigenvectors discarded **(a)** or suppressed from the expansion **(b–d)** for the $(N, S, M, \mu) = (2, 3, 3, 0.1)$ system. The first one corresponds to the 'unphysical' eigenvalue w_1, with uniform components along the signal axis. The second one relates to w_2, assigning the same signal to every meaning, which is suppressed. The bottom two correspond to w_4, which sees the two speakers differentiating signals and meanings, but in opposite directions: a mapping selected by one is rejected by the other. All of these result in no convergence towards language formation, and are suppressed from the linear stability expansion since $w_2, w_4 < 0$.

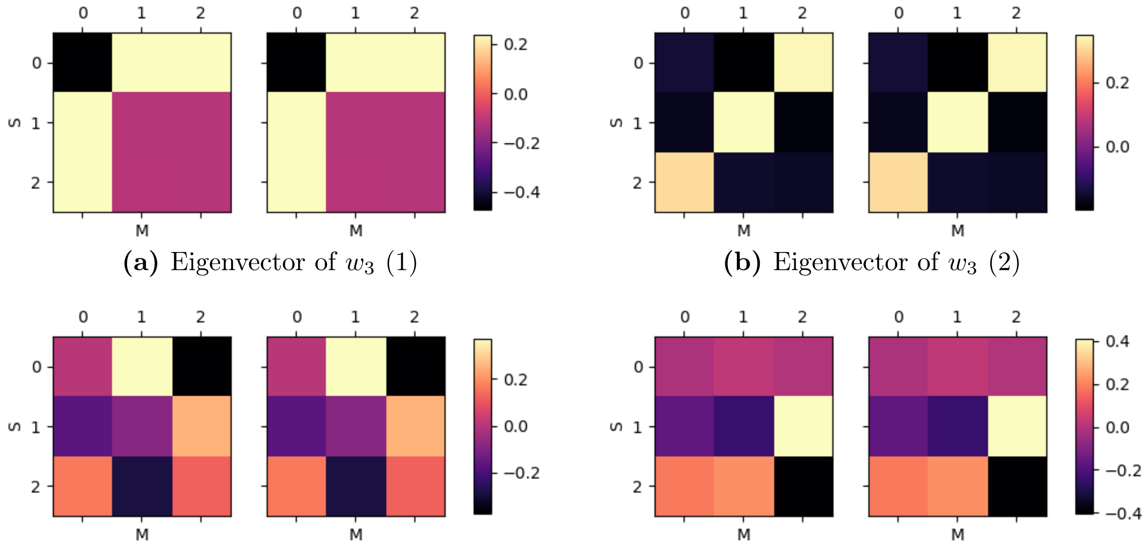
As can be seen in Figure 9, the particular eigenvectors returned by the eigenvalue equation solver display arbitrary preferences for certain signal-meaning mappings, assumed to originate from the numerical algorithm used. Since signals and meanings are interchangeable in the simplistic model, they can be permuted to recover the full range of displacements for a given eigenvalue.

6 eigenvectors were associated to the eigenvalue w_1, each of them having uniform components along the signal axis (violating the constraint). An example of this unphysical case is shown in Figure 9a, where the two speakers choose to associate all signals equally to the same meaning (without necessarily agreeing on the same mappings).

Figure 9b shows one eigenvector associated with w_2, which is negative and 4-fold degenerate: this corresponds to the case where the same meaning is assigned to all signals equally, which is suppressed by the dynamics. The specific example shown in the figure further exhibits disagreement among the speakers, with uniform mappings going in opposite directions.

Figures 9c and 9d are eigenvectors of w_4, which is also negative and 4-fold degenerate. As was inferred from the mathematical derivations, these correspond to directions that differentiate signals and meanings, but with speakers contradicting each other for all mappings.

All eigenvectors in Figure 9 correspond to cases that are either suppressed or discarded from the expansion in (18), i.e. directions that the system does not take to form a communicative language. On the other hand, for the final eigenvalue w_3, which for $S = 3$, $M = 3$ is positive, we find the four eigenvectors shown in Figure 10. These therefore correspond to directions of convergence that the system follows from the blank slate grammars, at least according to linear stability analysis. As mentioned above, the full set of possible displacements are found by permuting signal and meaning axes, as well as performing global sign switches.



(a) Eigenvector of w_3 (1)

(b) Eigenvector of w_3 (2)

(c) Eigenvector of w_3 (3)

(d) Eigenvector of w_3 (4)

Figure 10: All four degenerate eigenvectors corresponding to the positive eigenvalue w_3, in the $(N, S, M, \mu) = (2, 3, 3, 0.1)$ system. The specific configurations shown here each display a different type (or 'mode') of displacement away from the blank slate grammars. For each of those types, a greater set of displacements can be obtained by permuting the axes, since in all signals and meanings (labelled '0', '1' and '2') are equivalent in the simplistic model. The exact value found at each point stems from the eigenvalue equation solver; however one can note that the components of each eigenvector along the signal axis sum to zero, therefore satisfying the constraint (14). Speakers agree in their displacement from equilibrium, thus favouring consensus in their language.

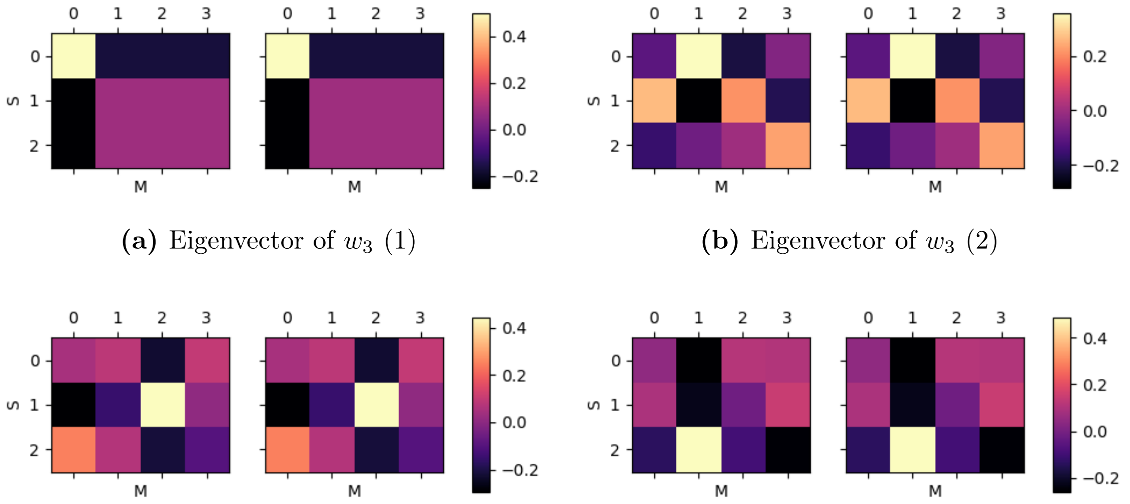
The first eigenvector in Figure 10a differentiates one mapping in particular, exhibiting a maximal (negative—although the sign is irrelevant) value in the top-left corner. Interestingly, the remaining mappings chosen by the speakers exhibit a uniform displacement, rather than a complete differentiation between signals and meanings as is observed in the other eigenvectors. This hints at a specific mechanism by which one mapping is chosen, with a simultaneous displacement *away* from the mappings sitting in the same row and column (signal and meaning). This could be associated with the convergence to a first mapping observed in the simulations, which is further investigated in the following section.

The second eigenvector, shown in Figure 10b, displays a clear pattern of convergence towards an optimal grammar for the $(S, M) = (3, 3)$ system. We note that all mappings are not assigned exactly the same value: some converging mappings will be more favoured by the system than others.

The remaining two eigenvectors in Figures 10c and 10d present a slightly less obvious configuration, although for both of them we note the 'designation' of a pair of mappings with maximal values of opposite signs (top row, second and third columns for the first eigenvector (10c); third column, second and third rows for the second (10d)). The differentiation occurs along the rows (meaning axis) for the first one, and the columns (signal axis) for the second, possibly hinting at two separate mechanisms to escape the blank slate. Although less evident from the colours, the eigenvector in Figure 10c has all maximal positive values arranged in an optimal way, and the same is true for its negative values—again, a global sign shift can be applied to invert the vector. Additionally, we note that these directions of convergence favour certain mappings over others, and that other mappings are simultaneously disfavoured.

In all four cases, there seems to be an implied 'choice' for one or more mappings, with an ensuing differentiation of the signals and meanings, ensuring that the constraint (1) is respected and signal redundancy is avoided.

5.2.4 Linear stability analysis in the non-symmetric regimes



(a) Eigenvector of w_3 (1)

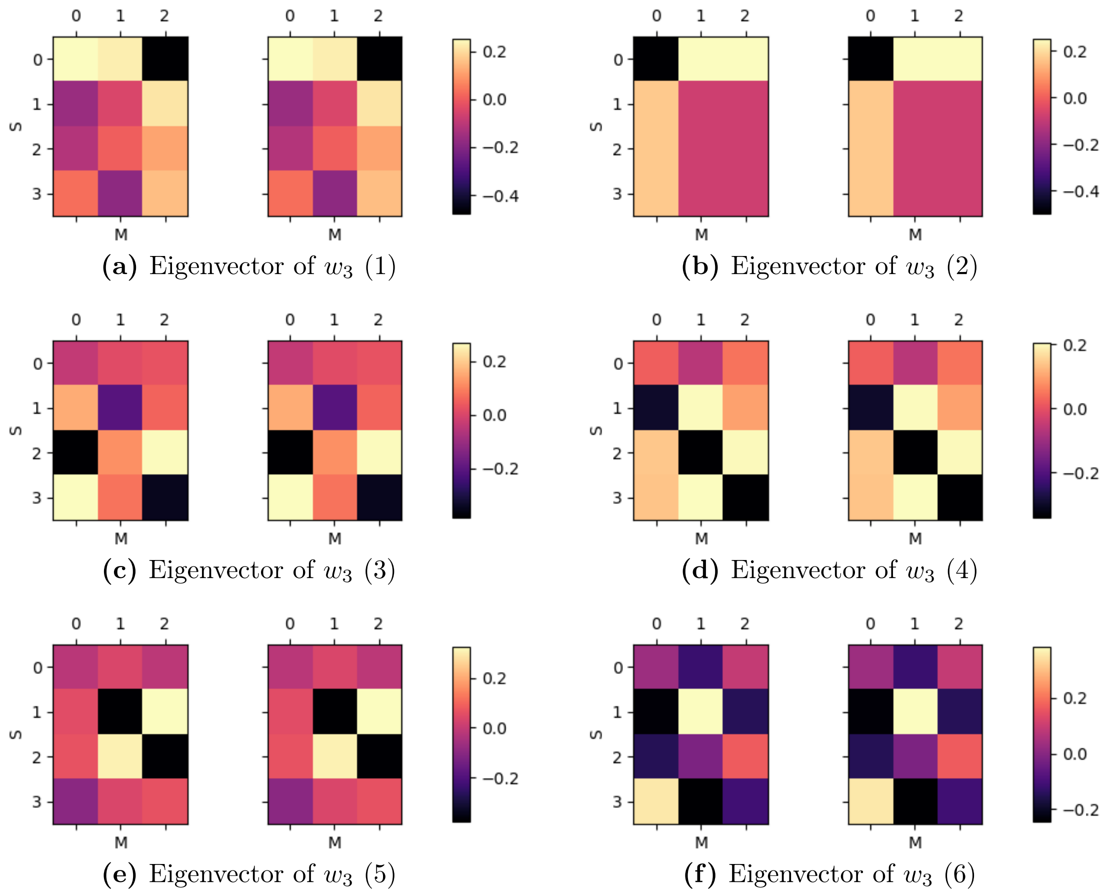
(b) Eigenvector of w_3 (2)

(c) Eigenvector of w_3 (3)

(d) Eigenvector of w_3 (4)

Figure 11: Positive-eigenvalue eigenvectors satisfying the constraint for the $M > S$ regime, with $(N, S, M) = (2, 3, 4)$. Similar configurations are observed, with the convergence to one mapping in particular and the displacement away from the row and column of that mapping. Eigenvector 2 also clearly displays homonymy.

We briefly review the results of linear stability analysis for the $M < S$ and $M > S$ regimes by investigating the $(N, S, M) = (2, 3, 4)$ and $(N, S, M) = (2, 4, 3)$ systems. Both of them have 24 total eigenvalues, with the positive w_3 being 4-fold and 6-fold degenerate respectively (as was expected from (31)). Similar negative-eigenvalue and 'unphysical' positive-eigenvalue configurations to the ones shown in Figure 9 are obtained. The



(a) Eigenvector of w_3 (1)

(b) Eigenvector of w_3 (2)

(c) Eigenvector of w_3 (3)

(d) Eigenvector of w_3 (4)

(e) Eigenvector of w_3 (5)

(f) Eigenvector of w_3 (6)

Figure 12: Positive-eigenvalue eigenvectors for the $M < S$ regime, with $(N, S, M) = (2, 4, 3)$. The 'one-mapping' convergence is also observed, with more elaborate configurations showing displacements towards two or three chosen mappings. Eigenvectors 3 and 4 seem to display the elimination of one redundant signal, useless in conversation.

positive-eigenvalue eigenvectors, thus amplified in the eigenbasis expansion, are shown in Figures 11 and 12 for the two systems.

Starting with the $M > S$ case, we note similarities with the symmetric regime: the first eigenvector (Figure 11a) shows the same 'one-mapping' convergence, simultaneously and uniformly disfavouring the mappings in the same row or column. The bottom two eigenvectors exhibit a similar distribution of mappings, with one clear choice and a nonuniform displacement towards mappings from different rows and columns. The eigenvector in Figure 11b displays a more interesting configuration in which we observe homonymy—that is, one signal (in this case, '1') is assigned to two meanings. This remains true if a global sign flip is performed.

In the case where $M < S$, in Figure 12, the same similarities are observed with the 'one-mapping' configuration with both uniform and nonuniform displacements away from the row and column of the initial mapping. The other eigenvectors exhibit a clear convergence towards a two- or three-mappings configurations. Figure 12d for instance shows a displacement towards three mappings, and the resulting 'elimination' of a signal ('0' in this case), which would have no use in communication. Figure 12e on the other hand displays a clear differentiation of two mappings, with less certainty on which remaining signal will be chosen or discarded from the language.

Hence, linear stability analysis in the non-symmetric regimes yields similar results to those noted for $(N, S, M) = (2, 3, 3)$. The directions of convergence noted from the eigenvectors of w_3 are in sensible agreement with the behaviour noted in the simulations, in which the system converges to optimal grammars that avoid signal redundancy (and ambiguity). Some of the eigenvectors indicate a preference for one or two mappings in particular, and for $M > S$ we note the presence of homonymy in at least one of them.

5.3 Convergence to a first mapping

We now turn to a possible mechanism (or 'strategy') by which the system escapes the hesitation phase—moving away from the blank slate fixed point—by reaching consensus on a first signal-meaning mapping. This behaviour was observed in a majority of Monte Carlo simulation runs, in which the fluctuations of the hesitation phase eventually reach a critical point where all (or most) speakers show a displacement in the same direction for the same mapping. In general, the system quickly converges to other mappings along different rows and columns once this first mapping is attained.

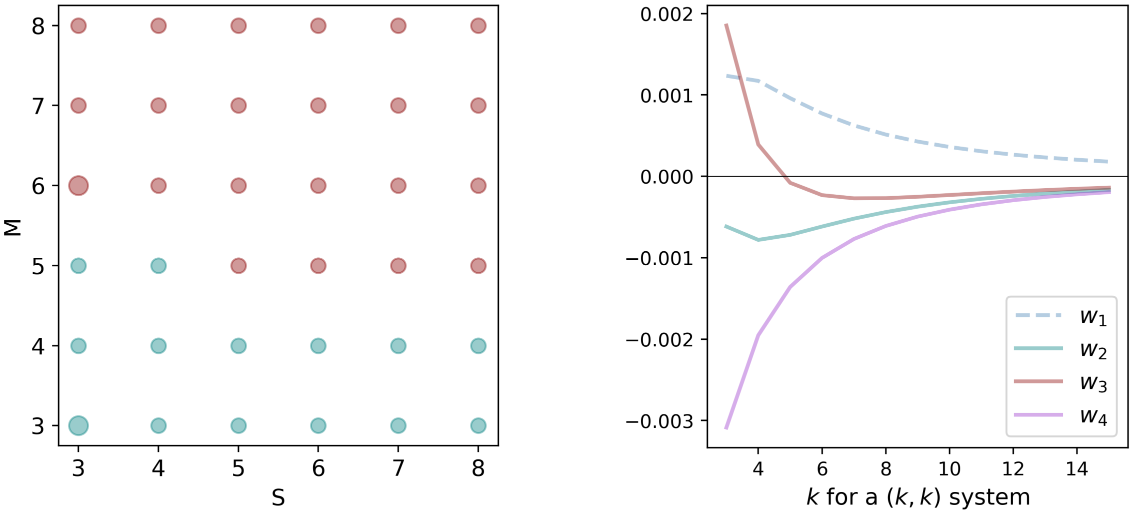
The eigenvectors of w_3 from linear stability analysis for the systems above generally exhibit a stronger concurrence of the speakers for one mapping in particular, with a simultaneous attenuation ('depletion') of the mappings along the same signal or meaning axis. This supports the above hypothesis, and suggests a simultaneous suppression of homonymy and synonymy in the system, which are communicatively unfavourable. As such, the convergence to a first mapping can be tied back with optimal grammars.

Furthermore, once the first mapping is 'chosen' by the system, the speakers are effectively left with a $(S-1, M-1)$ fluctuating subsystem, for which some preferences might already exist from the initial convergence to the first mapping. Interestingly, grammars with a singular, fully-determined mapping and depleted row and column are also fixed points of the dynamics (cf. Section 5.1). As we have seen in Section 4.3, the hesitation duration is strongly dependent on the system size, hence we expect the ensuing convergence in the subsystem to be much shorter. Although not investigated in detail, this effect could perhaps be related to the rapid transition from hesitation to consensus in the system.

5.4 Limitations of linear stability analysis

Linear stability analysis for the $(N, S, M) = (2, 3, 3)$, $(2, 3, 4)$ and $(2, 4, 3)$ systems shown in the sections above suggests that the convergence of the system towards optimal grammars are well explained by the eigenvectors of the eigenvalue w_3 (defined in (31)) discussed above. However, when considering larger systems, this eigenvalue eventually becomes *negative*. Its sign is only dependent on S and M, and we investigate it for different combinations of the parameters in Figure 13a.

The sign of w_3 is positive for 'small' systems, but becomes negative as early as $(S, M) = (5, 5)$. If we take a negative sign to imply no convergence, this is in direct contradiction of the Monte Carlo simulations which exhibit convergence to optimal grammars for all



(a) Sign of w_3

(b) Eigenvalues for the square system

Figure 13: **(a)** The sign of the w_3 eigenvalue for different combinations of S and M. It is positive (blue) for small system size, but becomes negative (red) for greater systems. Although not shown in the Figure, the eigenvalue becomes increasingly negative when S and M are increased, and it appears to remain positive for any S when $M = 3$ or 4. If a negative w_3 implies no convergence to optimal grammars, these results directly contradict Monte Carlo simulations. **(b)** The value of all four eigenvalues for the symmetric system with $N = 2$ and $\mu = 0.1$, with the 'unphysical' one shown with a dashed line. Past $k = 4$, all eigenvalues are negative and converge to the same value. w_3 is the least negative one, perhaps indicating that the system has a small preference for these optimal states, which is increasingly small for greater k.

system combinations in Figure 13a. The degenerate eigenvectors associated with w_3 for these systems display the same relative configurations, where all speakers differentiate signals and meanings in the same direction. We also note from Figure 13b that the value of w_3 remains the least negative out of all eigenvalues satisfying the constraint.

The reason for this result is not quite clear, and needs to be further investigated to determine whether the predictions from linear stability analysis can be trusted for larger systems. Perhaps the nonlinear contributions of the update rule play a non-negligible role in the convergence towards optimal grammars, which might become more important the greater S and M get. Or perhaps the significant increase in hesitation time seen in Section 4.3 has something to do with it, where large systems have increasingly low probability of reaching an 'irreversible' point by chance.

Speculatively, we can also think of the system as a potential landscape in which the blank slate is a local minimum (fixed point) surrounded by potential 'barriers', which the system overcomes by random fluctuations to converge to its final grammar. Under this perspective, perhaps the eigenvalue can be related to the relative height of these barriers, which will be greater (and therefore less likely) for w_2 and w_4 than for w_3. Lastly, it remains possible that the derivations shown above are mistaken in their assumptions, or simply incorrect, although they have been verified multiple times over.

In any case, the validity of the linear stability results is put into doubt. Efforts in understanding the contribution of nonlinear terms in the expansion are needed to determine whether the sign of w_3 really is a fundamental limitation of the analysis or if it has a sensible explanation, such that the eigenvectors are good descriptors of the evolution of the system from the blank slate. We are inclined to believe the latter, and to see at least some physical significance in the eigenvectors of w_3, because the results for the smaller systems closely match observations from the stochastic simulations. Nevertheless, we must treat the linear stability results with a reasonable amount of scepticism.

5.5 Conclusions on language emergence in the simplistic model

The last two sections considered language formation in a community of speakers using the simplistic model, to inspect the self-organising behaviour of the system when initialised at the blank slate grammars. The hesitation phase undergone by the model was identified, and its dependence on the different system parameters in all three regimes was investigated. We found it to depend most strongly on the system size in the homonymous and symmetric regimes, with a smaller dependence on the number of speakers N or the number of signals S in the synonymous regime. The formation of consensus becomes increasingly unlikely for relatively small systems, underlining the limitations of the simplistic model in describing natural language formation.

Additionally the system displays a preference for optimal grammars which maximise communication (and minimise ambiguity) among the speakers, in all three regimes. This preference was investigated by developing a mean-field description of the dynamics and performing linear stability analysis at the blank slate grammars. In doing so, we identified eigenvectors that the model favours and disfavours, depending on the magnitude and sign of their corresponding eigenvalue. Only one of them, w_3, satisfies the constraint and is positive for 'small' systems; its associated eigenvectors differentiate signals and meanings in the same direction for all speakers. We posit that these eigenvectors may correspond to the preferred directions of convergence taken by the system to reach optimal grammars. However, this eigenvalue becomes negative for systems in which convergence towards optimal grammars is observed, implying that the results from linear stability analysis are at the very least partly incomplete.

Finally, we identify the convergence to a first mapping as an important strategy to escape hesitation and reach consensus, which according to linear stability analysis favours optimal configurations by simultaneously suppressing homonymy and synonymy.

A further step in the stability analysis of the model would be to understand the mean-field description (11) as the gradient of a potential function,

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_n(s|m) = -\frac{\partial}{\partial\phi_n(s|m)}V(\{\phi\}) , \tag{33}$$

in which stable fixed points (e.g. the blank slate grammars) would correspond to local minima. Optimal grammars are also fixed points of the dynamics and might therefore to global minima, reached by the system from the blank slate via a specific path in the multidimensional potential landscape. The difficulty of moving between the two minima (i.e. the length of the hesitation phase), would then depend on the height of the potential barrier (the same as the one mentioned above). This approach, although not considered here by lack of time, would represent a key addition to our understanding of language consensus in the system.

In the dynamics of the simplistic system, optimal grammars are not just fixed points; they are also *absorbing states* (due to $U \to 0$ when $\phi \to 0$ or 1). Therefore, once reached, it is extremely unlikely that the system escapes them, and this is never observed in the simulations. This brings about an obvious question: how then, can language *change*?

6 Language evolution

The second part of this project is concerned with language change, or evolution, and the potential mechanisms that could result in a language *turnover* in the model, where all speakers agree on a new mapping between signals and meanings. First, we turn to the literature to investigate the variation observed in language over time, and subsequently we consider an implementation of language turnover in the model.

6.1 How can language change?

In natural language, change is an everyday process, in which very minute variations are constantly observed in the words used and the meanings they carry. Such an observation is most evident from the wide range of ever-changing trends that are observed on the Internet, in which the meaning of existing words shift (e.g. the new meanings of 'cloud' or 'tweet') and words are invented to describe new concepts and ideas ('greenwashing', 'deepfakes' or 'crypto').

This change takes many forms in natural language, categorised by linguists as syntactic change, in which the structure of language evolves over time (e.g. word order), phonetic change, concerning the sounds and pronunciations of words, or spelling change, referring to their written representation [35, 36].

In this project however, our interest lies primarily in *semantic change*, in which the meanings of words or phrases evolve over time—words can acquire new meanings or lose existing ones. Several forms of semantic change, or drift, have been suggested in the literature [37–39], such as metaphoric extension (the 'foot' of a person can also be that of a mountain), broadening (e.g. the wider significations of 'mouse', 'bug' or 'virus') and narrowing (the word 'girl' originally designated a young person of either gender). Melioration and pejoration of meaning is also observed in English; for instance the word 'nice' originally meant foolish, whereas 'villain' referred to a peasant or farmhand.

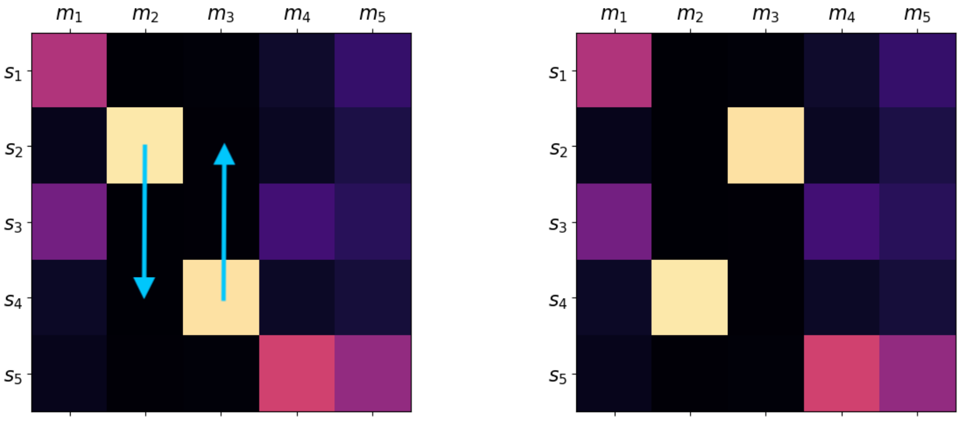
Semantic change is generally driven by social and cultural factors, as language is intricately intertwined with the society that uses it [40]. Technological advancements, for example, can prompt the creation of new words to describe emerging concepts and ideas. Similarly, societal shifts can lead to the evolution of the meanings of words, reflecting changing beliefs, attitudes, and values [31]. Some sociolinguists, such as Labov [41], further argue that language change occurs through the gradual diffusion of new linguistic forms or variants across a speech community. Over time, these variants become widely adopted and eventually establish themselves as the new norm.

6.2 Implementation in the multi-meaning USM

In the present formulation of the model, we look for language change in the form of a *turnover* in the speakers' grammars, where the mapping of signals and their corresponding meanings evolves over time. This thus reproduces the phenomenon of semantic change, albeit in a simplified formulation, where the systems considered feature a small number of speakers, signals and meanings. In this sense, the model is a conceptual description of language evolution, in which speakers might represent subgroups of society; turnover could illustrate the process by which words gain and lose meanings over time. We do not consider here the possibility of *neologisms*, where new signals and meanings are added to the system, although this constitutes an interesting application for future work.

In our approach, we choose to probe linguistic turnover by considering mapping 'swaps' in the system, in which signal-meaning associations are exchanged. This mechanism is implemented by designating a *trendsetter* in the speech community, able to innovate—going against the established convention of the language—by performing swaps in their mappings. We are interested in the possible propagation of these innovations in the community of speakers, corresponding to global semantic change.

The swapping mechanism is illustrated in Figure 14, where the trendsetter permutes two nonidentical *conversational* mappings—that is, the signal-meaning associations (s, m) with the highest probability $\phi_i(s|m)$ along the meaning axis m (in speaker i's grammar). This ensures that the altered grammars remain optimal, thus preventing swaps that reduce communicability.



(a) Before the swap

(b) After the swap

Figure 14: Illustration of innovation in the model, showing a snapshot of the trendsetter's grammar in conversation. The trendsetter swaps two conversational mappings, such that the meanings of the two signals are exchanged. We are interested in whether this innovation can propagate in the rest of the speech community.

The ability of the trendsetter to perform swaps in successive conversations with other speakers is modulated by a characteristic timescale, the *waiting time* T_w, which determines the number of conversations between each swap. Additionally, the designation of a trendsetter in the population implies an asymmetry in the speech community, which is implemented with a social valuation effect in the feedback function. In its simplest form, we choose to augment the feedback returned by the trendsetter to the other speakers by a factor α such that, when in conversation with the trendsetter, other speakers will adjust their grammar with $\pm\alpha\lambda U$ (the trendsetter still adjusts theirs with $\pm\lambda U$). A value too low sees none of the trendsetter's innovations propagate in the population (the majority wins every time), whereas a value too high results in a 'disconnected' system, where the trendsetter changes their grammar on much slower timescales than the other speakers.

Finally, an important effect originating from the impact function U is the relative *stubbornness* of speakers once they reach an optimal grammar, due to the overall feedback strength being attenuated when the mappings $\phi_i(s|m)$ approach zero or one (cf. Figure 1b). Hence the speakers exhibit a resistance to change of sorts, by opposing the trendsetter's innovations. It is also responsible for the *entrenchment* of speakers' grammars, where the speakers 'agree to disagree' with the trendsetter if they are sufficiently confident in their own grammar (see Figure 15c below). This corresponds to a near-constant evolution of the intelligibility, since 'stubborn' grammars change very little.

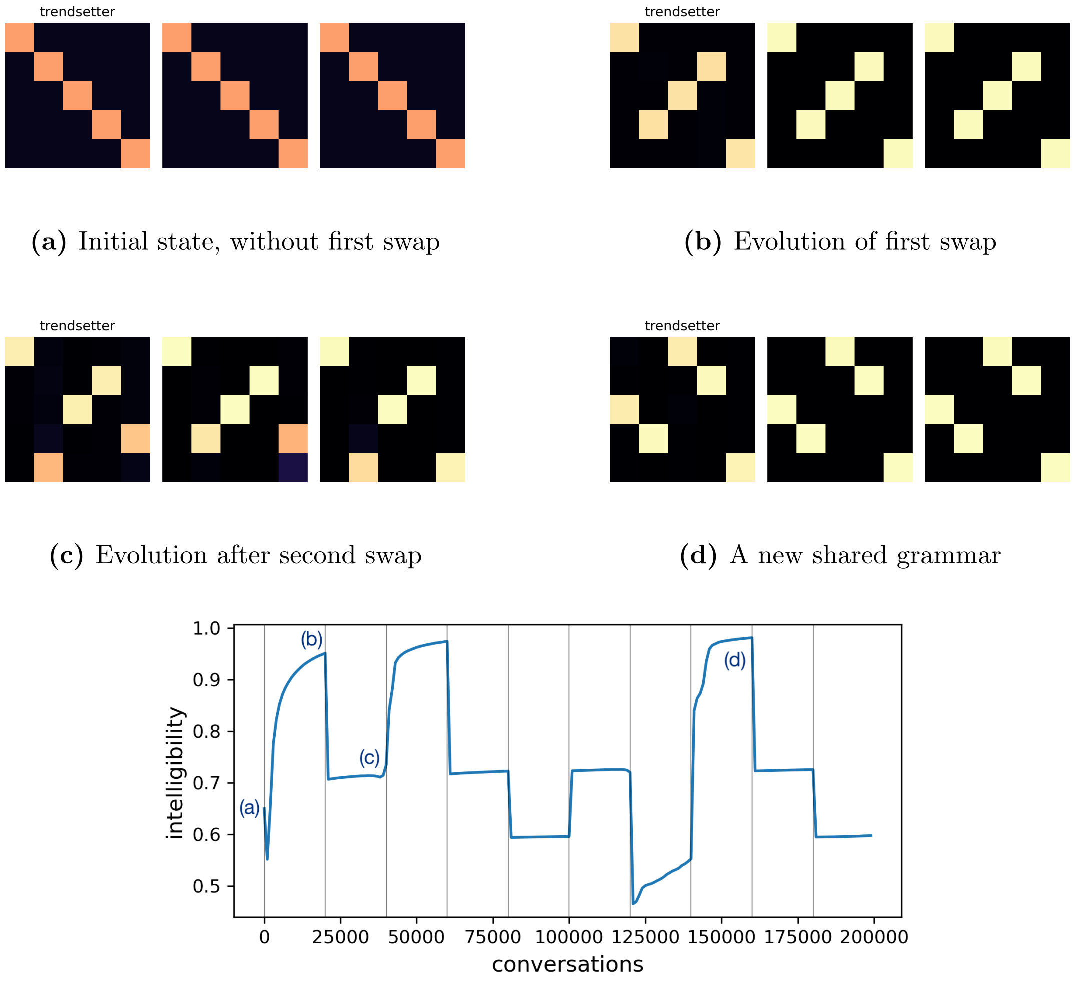
6.3 Trendsetter dynamics

We choose to confine our investigation of this dynamics to the $(N, S, M, \lambda) = (3, 5, 5, 0.01)$ system with one trendsetter, and a feedback augmentation of $\alpha = 10$. Specifically, we probe how the waiting time T_w affects whether language turnover can occur in the system. The value of α was found to win over the majority whilst maintaining similar timescales of evolution for the trendsetter and the other speakers, thus avoiding a 'disconnected' system. In the simulations, the system is initialised at an optimal language with nonmaximal mappings (to avoid entrenchment), to which a swap is immediately performed.

Figure 15 shows the evolution of this system with $T_w = 20000$, for which successful language turnover is observed on three occasions. As this is a variable process, a turnover resulting in efficient communication is not guaranteed—indeed, most of the swaps result in relatively low intelligibility in Figure 15e, showing signs of entrenchment. On some occasions, the trendsetter's innovation propagates among the other two speakers (as in Figures 15b and 15d), and the convergence to a new common mapping is observed. Figure 15c illustrates a case of entrenchment in which homonymy is observed among the two speakers.

In order to probe the impact of different characteristic timescales T_w on the turnover in language, we run multiple simulations and average over the intelligibility to gain insights on the general behaviour of the system, since it varies significantly from one run to the next, as seen in the previous Figure. Doing so, we observe three general 'regimes' (to be distinguished from the symmetric and non-symmetric regimes of the system from the previous sections): low-T_w ('irresolute'), middle-T_w and high-T_w ('stubborn'). These are investigated in Figure 16.

The first regime corresponds to the range of T_w where the 'irresolute' trendsetter generates innovations too frequently for the other speakers to integrate the change, illustrated in the Figure by $T_w = 500$ and 2000. As a result, the system intelligibility drops and the system dissipates in a state of confusion—akin to the hesitation phase of the simplistic model.



(a) Initial state, without first swap

(b) Evolution of first swap

(c) Evolution after second swap

(d) A new shared grammar

(e) Intelligibility evolution of the system

Figure 15: Language turnover observed for $T_w = 20000$ and $\alpha = 10$. **(a)** The initial state assumes a common mapping with nonmaximal intelligibility, and a first swap is performed upon initialisation. **(b)** The evolution of the system 20000 conversations after the initial swap: the trendsetter's innovation has successfully propagated in the system. **(c)** A second swap is performed to which the system adapts; surprisingly, this generates homonymy, and the system has reached a form of entrenchment (near-constant intelligibility). **(d)** After successive states of entrenchment and low intelligibility, the system converges to a different common mapping to the initial one. In this sense, we observe successful language turnover in **(b)** and **(d)**, and after the swap in **(c)**.

From investigation of the simulations, this is observed for values up to $T_w \sim 3000$. The second regime corresponds to the approximate range $5000 \lesssim T_w \lesssim 50000$ where language turnover (an innovation propagating in the system) can occur—illustrated by $T_w = 20000$ in both Figures 15 and 16. As noted above, successful turnover in this regime remains quite a rare process, whereas momentary entrenchment in less intelligible states is more likely. Beyond $T_w \sim 100000$ (illustrated in Figure 16) the third regime is entered, in which the initial convergence to a common grammar is maintained long enough for the system to become 'stubborn'. As a consequence, the system becomes entrenched when the next swap occurs—none of the speakers' grammars change significantly despite their misalignment. In other words, the trendsetter's innovations cannot 'convince' the other two speakers and no turnover is observed.

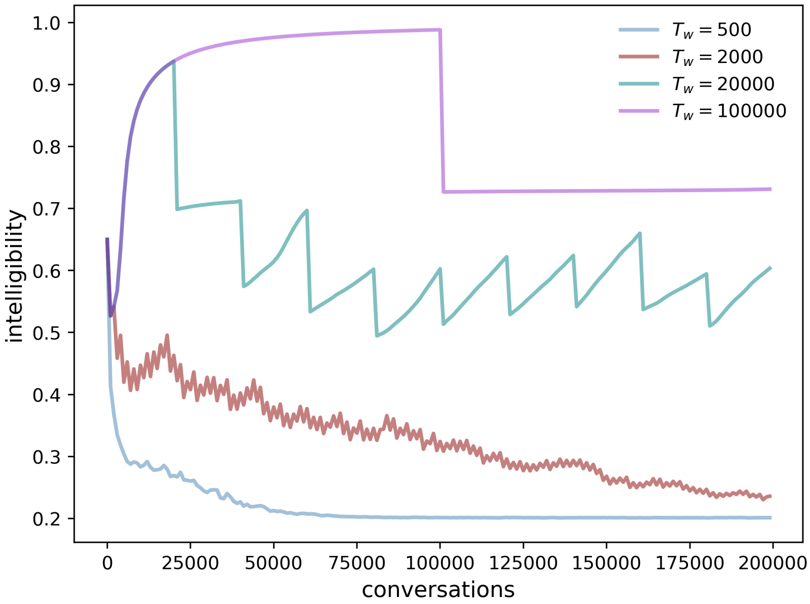


Figure 16: The intelligibility (averaged for 100 runs) for different values of T_w in the $(N, S, M, \lambda) = (3, 5, 5, 0.01)$ system, with $\alpha = 10$. Three 'regimes' are observed, corresponding to different orders of characteristic timescales. Low values (500, 2000) make the system irresolute, and therefore unintelligible. High values (100000) lead to stubborn grammars and subsequently to entrenchment in the following swaps. Middle values (20000) display a combination of momentary entrenchment, and a relatively rare convergence to common mappings—successful language turnover.

Due to time constraints, the phenomenon of language turnover was not probed in other system configurations, for instance with more speakers (and trendsetters), or larger S and M. This is an obvious component to add to the analysis of the trendsetter approach in future work. In addition, the precise dependence of linguistic turnover on α should also be inquired, to determine an 'optimal' value for different system sizes and configurations. Finally, more appropriate metrics in probing language turnover should be developed. As the previous Figures demonstrate, the intelligibility only offers limited insights on the state of the system and its progression towards a turnover.

6.4 Discussion of the approach and further implementations

The section above detailed an implementation of semantic change in the model, by which the innovations of a trendsetter could propagate across the speakers and replace the existing linguistic convention. Its implementation in the $(N, S, M, \lambda) = (3, 5, 5, 0.01)$ system yielded varying results depending on the characteristic timescale of innovation T_w; language turnover was observed, albeit infrequently, between the 'irresolute' and 'stubborn' regimes.

As such, the approach of a trendsetter presents a plausible cultural mechanism to generate language turnover in the system. However, we note that a successful turnover is relatively rare, with many of the trendsetter's innovations not 'convincing' the other two speakers.

In these cases, the intelligibility is greatly reduced, partly because of the small size of the system considered. In this sense, the implementation is not a perfect one, as semantic change is generally a smooth and continuous process in natural language, as opposed to the abrupt innovations of the trendsetter. It also generates much less confusion than that observed in the simulations. Another limitation of the approach comes from the nonlinear formulation of the impact function in the update rule, which facilitates stubbornness, and thus fundamentally restrains the possibility of change once a language is chosen by its speakers.

This prompts the search for alternative mechanisms of linguistic turnover in the system, which could not be explored in detail due to the time constraints of the project. For instance, another 'cultural' approach that was briefly investigated introduced a 'social ladder' in the system, in which speakers were assigned a hierarchical place in the speech community. Such a hierarchy could see the designation of social leaders—akin to trendsetters—and outcasts, whose input in conversations is modulated by the feedback function. Limited results indicated that language turnover could be probed by varying the social ladder over time, thus reorganising the society of speakers and their linguistic conventions with it. Social structure can also be implemented via the interaction topology, which was left untouched in the trendsetter dynamics to limit complexity.

In contrast, semantic change can also be studied by introducing asymmetry (nonuniform topology) in the signal and meaning spaces, rather than among the speakers. Language clearly shows preference for certain words over others; an example of this is the widely-observed *Zipf's law* [42, 43], where a word's frequency of use displays an inverse relation to its rank. As such, the rank of signals and meanings could be set to change over time, reflecting the evolution of society and its linguistic usages. Innovation, and possible turnover, would then originate not from a particular speaker or subgroup of society (the trendsetter), but from general societal shifts; this alternative approach is a promising implementation for future work on the model.

Finally, the possibility of neologisms—adding new signals and meanings to the existing ones—also presents interesting possibilities of language change, which could be tied to the mechanisms of language formation discussed in Sections 4 and 6.

7 Conclusion

In this project, the multi-meaning extension of the Utterance Selection Model was formulated and implemented in Monte Carlo simulations, to explore its uses in understanding language formation and evolution in a community of speakers. The model comprises multiple speakers, each with a probabilistic mapping between signals and meanings, who interact in successive conversations with a nonlinear update rule. In its simplistic formulation, where speakers, signals and meanings are all interchangeable, we observe the convergence to shared mappings across all speakers, a result previously noted in the literature.

This convergence is investigated for the different regimes of the system, and we note its strong dependence on the system size in the symmetric and homonymous regimes. For relatively small systems, it becomes increasingly difficult to escape hesitation, thus suggesting that the self-organising properties of the simplistic model cannot fully account for the formation of language in a speech community. Additional mechanisms must be considered to facilitate this convergence, and this exploration represents an important implementation to the analysis in future work on language formation.

Multiplicity calculations were performed to probe the preference of the simplistic model for optimal grammars, which maximise intelligibility and minimise ambiguity. Subsequently, a mean-field prescription of the model was derived, allowing the application of linear stability analysis to the blank slate grammars. The eigenvalues and associated eigenvectors it returns are interpreted as the possible directions of convergence that the model can take. Only one of the eigenvalues was found to take positive values for small systems while satisfying the constraint; its associated eigenvectors each display a preference for optimal systems in which certain mappings are favoured, with a simultaneous suppression of homonymy and synonymy. Additionally, these states support the hypothesis that the system escapes hesitation by converging to a first mapping.

Puzzlingly, this eigenvalue becomes negative for larger systems, despite their convergence to optimal systems in the simulations. This effect is not well understood, and we postulate that it might originate from the nonlinearities of the system, possibly relating to increasing potential barriers that the system needs to overcome to exit hesitation.

The optimal grammars are absorbing states, thus disfavouring subsequent evolution once they are reached. To investigate language change, in the form of a turnover of the mappings emulating semantic drift, we implement a mechanism of innovation against the established (optimal) conventions via a trendsetter. Our approach incorporates a social valuation effect, enabling the innovations to propagate in the population, and a characteristic timescale, modulating how frequently they are generated.

The possibility of attaining successful turnover is investigated with the integration of one trendsetter in the $(N, S, M, \lambda) = (3, 5, 5, 0.01)$ system. We identify a range of timescales where it is observed, albeit relatively rarely and with low overall intelligibility. Thus the approach yields a plausible cultural mechanism to instigate language change with some obvious limitations, which can be partly attributed to the nonlinear impact function. A deeper analysis of this approach is required to understand its effects on different system sizes and regimes.

Finally, other implementations of language change are briefly discussed, such as the introduction of a changing social hierarchy, or the variation of signal and meaning frequencies over time (through the implementation of Zipf's law for instance). Thanks to its modularity, the multi-meaning USM allows for a broad range of approaches to language change extending beyond those discussed in this report. The exploration of these approaches holds significant promise not only for further research on the model, but for the field of language evolution in general.

References

[1] C. Castellano, S. Fortunato, and V. Loreto, “Statistical physics of social dynamics”, Reviews of Modern Physics **81**, 591 (2009).

[2] K. Campbell-Kibler, “Sociolinguistics and perception”, Language and Linguistics Compass **4**, 377–389 (2010).

[3] M. Tomasello, *Origins of human communication* (MIT Press, 2010).

[4] J. D. Johansen and S. E. Larsen, *Signs in use: an introduction to semiotics* (Psychology Press, 2002).

[5] F. De Saussure, *Course in general linguistics* (Columbia University Press, 2011).

[6] Y. Holovatch, R. Kenna, and S. Thurner, “Complex systems: physics beyond physics”, European Journal of Physics **38**, 023002 (2017).

[7] C. Beckner et al., “Language is a complex adaptive system: position paper”, Language Learning **59**, 1–26 (2009).

[8] L. Steels, “Language as a complex adaptive system”, in Parallel Problem Solving from Nature, edited by M. Schoenauer et al. (2000), pp. 17–26.

[9] L. P. Kadanoff, “Scaling and universality in statistical physics”, Physica A: Statistical Mechanics and its Applications **163**, 1–14 (1990).

[10] R. Mairal and J. Gil, *Linguistic universals* (Cambridge University Press, 2006).

[11] N. C. Ellis and D. Larsen-Freeman, *Language as a complex adaptive system*, Vol. 11 (John Wiley & Sons, 2009).

[12] A. Baicchi, *Construction Learning as a Complex Adaptive System* (Springer, 2015).

[13] V. Loreto et al., “Statistical physics of language dynamics”, Journal of Statistical Mechanics: Theory and Experiment **2011**, P04006 (2011).

[14] W. Croft, “Toward a social cognitive linguistics”, New directions in cognitive linguistics **24**, 395–420 (2009).

[15] S. Pinker and P. Bloom, “Natural language and natural selection”, Behavioral and Brain Sciences **13**, 707–727 (1990).

[16] M. A. Nowak, *Evolutionary dynamics: exploring the equations of life* (Harvard University Press, 2006).

[17] L. Wittgenstein, *Philosophical investigations* (John Wiley & Sons, 2010).

[18] M. A. Nowak and D. C. Krakauer, “The evolution of language”, Proceedings of the National Academy of Sciences **96**, 8028–8033 (1999).

[19] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior* (Princeton University Press, 2007).

[20] M. A. Nowak, J. B. Plotkin, and D. C. Krakauer, “The evolutionary language game”, Journal of Theoretical Biology **200**, 147–162 (1999).

[21] G. Szabó and G. Fath, “Evolutionary games on graphs”, Physics Reports **446**, 97–216 (2007).

[22] H. Ohtsuki et al., “A simple rule for the evolution of cooperation on graphs and social networks”, Nature **441**, 502–505 (2006).

[23] K. Kosmidis, A. Kalampokis, and P. Argyrakis, “Statistical mechanical approach to human language”, Physica A: Statistical Mechanics and its Applications **366**, 495–502 (2006).

[24] L. Steels, “A self-organizing spatial vocabulary”, Artificial Life **2**, 319–332 (1995).

[25] L. Steels, “Emergent Adaptive Lexicons.”, in Maes, P. and R. Brooks (ed.), From Animals To Animats 4: Proceedings of the Simulation of Adaptive Behavior Conference (1996), pp. 562–567.

[26] J. Ke et al., "Self-organization and selection in the emergence of vocabulary", Complexity **7**, 41–54 (2002).

[27] W. H. Zuidema et al., "The major transitions in the evolution of language", PhD thesis (University of Edinburgh, 2005).

[28] J. Fodor, "Language, thought and compositionality", Royal Institute of Philosophy Supplements **48**, 227–242 (2001).

[29] G. J. Baxter et al., "Utterance selection model of language change", Physical Review E **73**, 046118 (2006).

[30] W. Croft, *Explaining language change: an evolutionary approach* (Pearson Education, 2000).

[31] N. Tahmasebi et al., *Computational approaches to semantic change* (BoD–Books on Demand, 2021).

[32] M. H. Christiansen and N. Chater, "Language as shaped by the brain", Behavioral and Brain Sciences **31**, 489–509 (2008).

[33] M. Opper and D. Saad, *Advanced mean field methods: theory and practice* (MIT Press, 2001).

[34] N. P. Bhatia and G. P. Szegö, *Stability theory of dynamical systems* (Springer Science & Business Media, 2002).

[35] J. Bybee, *Language change* (Cambridge University Press, 2015).

[36] A. M. McMahon and M. April, *Understanding language change* (Cambridge university press, 1994).

[37] S. Ullmann et al., *Semantics: an introduction to the science of meaning* (Basil Blackwell Oxford, 1962).

[38] A. Blank and P. Koch, *Historical semantics and cognition*, Vol. 13 (Walter de Gruyter, 2013).

[39] H. H. Hock and B. D. Joseph, *Language history, language change, and language relationship* (De Gruyter Mouton, 2019).

[40] B. W. Fortson IV, "An approach to semantic change", The Handbook of Historical Linguistics, 648–666 (2017).

[41] W. Labov, *Principles of linguistic change, volume 3: cognitive and cultural factors* (John Wiley & Sons, 2011).

[42] J. Kanwal et al., "Zipf's law of abbreviation and the principle of least effort: language users optimise a miniature lexicon for efficient communication", Cognition **165**, 45–52 (2017).

[43] L. A. Adamic and B. A. Huberman, "Zipf's law and the internet.", Glottometrics **3**, 143–150 (2002).

A Appendix

Link to the code used in the project

Follow the GitHub link below to access the Python code developed in this project:

`https://github.com/cazfisch/MPhys-project.git`