biolinguistics

ARTICLES

Biolinguistics 14.SI: 102–129, 2020
www.biolinguistics.eu

The Phonological Latching Network

Joe Stephen Bratsvedal Collins[1,2,3,*]

[1] The Center for Advanced Study in Theoretical Linguistics (CASTL), UiT – The Arctic University of Norway, Tromsø, Norway
[2] Cognitive Neuroscience Group, International School for Advanced Studies (SISSA), Trieste, Italy
[3] Department of Language and Literature, Faculty of Humanities, Norwegian University of Science and Technology, Trondheim, Norway

[*] Correspondence: joe.collins@ntnu.no

JSBC: https://orcid.org/0000-0001-7335-5134

This paper gives an analysis of an attractor neural network model dubbed the Phonological Latching Network. The model appears to reproduce certain quintessentially phonological phenomena, despite not having any of these phonological behaviours programmed or taught to the model. Rather, assimilation, segmental-OCP, and sonority sequencing appear to emerge spontaneously from the combination of a few basic brain-like ingredients with a phonology-like feature system. The significance of this can be interpreted from two angles: firstly, the fact that the model spontaneously produces attested natural language patterns can be taken as evidence of the model's neural and psychological plausibility; and secondly, it provides a potential explanation for why these patters appear to frequently in natural language grammars. Namely, they are a consequence of latching dynamics in the brain.

Keywords: phonology; neuroscience; neural networks; attractors; Potts model

1. Introduction

In 1887, Albert Fournie claimed that "[s]peech is the only window through which the physiologist can view the cerebral life" (translation from Lashley 1951). There is nothing novel then, in the claim that the study of language should provide some insight into the workings of the human mind/brain. Indeed, even today, this is one of few mantras shared by linguists of the seemingly irreconcilable "Gener-

Editors: Lluís Barceló-Coblijn, Universitat de les Illes Balears, Spain
Evelina Leivada, Universitat Rovira i Virgili, Spain

Received: 27 December 2019
Accepted: 19 November 2020
Published: 25 March 2021

ISSN 1450–3417

ative" and "Cognitive" schools (e.g., Chomsky 2002; Lakoff 1988). Given this apparent consensus then, it is perhaps surprising that no breakthrough in our understanding of the brain can yet be attributed to some insight from the study of language.

An analysis and critique of this state of affairs is given by Poeppel & Embick (2005), who identify (amongst other things) that we currently have no way of relating the ontologies of linguistics and neuroscience. This *Ontological Incommensurability Problem* (OIP) can be resolved, they argue, by the use of a *Linking Hypothesis*, which spells out linguistic computations at the relevant level of algorithmic abstraction, such that the neuroscientist need only find the exact implementations of those algorithms in the brain. If such a hypothesis were sufficiently complete then it could, in principle, predict the kinds of neural configurations required for natural language processing, using linguistic theories as their starting point. In this way, we could finally realize the long sought-after goal of cashing in theories of language for understanding of the human brain. Simultaneously, a *Linking Hypothesis* also has the potential to unearth lower-level explanations for linguistic phenomena, for example where those explanations might depend on purely neurobiological notions (e.g., neuronal morphology, synaptic density, metabolic efficiency, etc.).

1.1. Emergence as a Linking Hypothesis

The specific approach to the OIP advocated by Poeppel & Embick treats the neurobiological level of analysis as something akin to a decomposition of a linguistic theory. That is, a linguistic theory can be reduced to individual processes (e.g., concatenation, linearization, etc.), and the problem of how to realise each process can be attacked individually. And, while this approach is certainly a logical possibility for resolving the OIP, it rests on assumptions which treat the brain as being fundamentally like a digital, programmable computer. Implicitly, it has borrowed from computer science the idea that the different levels of abstraction for which we might describe a cognitive function, are related to one another through a strict compositional semantics. That is, any property at one level of abstraction can be neatly decomposed to some combination of properties at a lower level of abstraction (e.g., Block 1995).

A full rebuttal of these assumptions is well beyond the scope of this article. It is sufficient to note that this view is by no means the only starting point for constructing a *Linking Hypothesis*. The alternate approach offered here draws inspiration from the natural sciences, where the apparent incommensurability between different levels of abstraction is frequently resolved by treating the higher levels as *epistemologically emergent*[1] from lower ones (e.g., Anderson 1972; Luisi 2002). According to this approach, the goal is not to decompose a macro-level ontology to see how each component is "implemented" at the micro-level.

---

[1] Alternatively: *weakly emergent* (Bedau 1997). Also note that this notion of *emergence* is strictly orthogonal to the notion of *ontogenetic emergence* employed in the study of language acquisition. Whether linguistic ontology is *epistemologically/weakly emergent* does not predict whether it is learned/innate/none of the above.

Rather, the goal is to see what kinds of configurations at the micro-level give rise to a complex system whose behaviour is captured by the macro-level theory.

Therefore, to claim that linguistics is *emergent* from neuroscience entails that linguistic properties do not separately decompose to neuroscientific properties, contra the way that the functions of a high-level computer language reduce to combinations of primitive operations. Instead, the relationship between linguistics and neuroscience would be analogous to, for example, the molecular theory of gasses.[2] Under this view, linguistic properties would be analogous to macro-level concepts like *temperature* or *pressure*, while neuroscientific properties are analogous to molecular explanations of these phenomena. The most relevant aspect of this analogy is that the properties present at each level of abstraction are quite different. So different, in fact, that the different levels of abstraction can seem metaphysically inconsistent. For example, while a notion such as *pressure* can be reduced to the average behaviour of all molecules in a system, no single molecule can be said to possess, explain, or cause *pressure* in any meaningful sense. *Pressure* is simply a concept which exists at the macro-level, but not at the micro-level. Nor can *pressure* and *temperature* be decomposed separately (e.g., there are not two types of molecule which cause *pressure* and *temperature* independently), rather, the properties of the macro-level appear to *emerge*, fully formed, once the micro-level analysis becomes sufficiently complex. In more general terms, there is some point in our analysis at which the collection of molecules ceases to be, and is a replaced by something radically different: a gas (see, e.g., Truesdell & Muncaster 1980).

Applying this analogy, if we allow that the relationship between the brain and phonology is one of *emergence*, rather than a strict compositional semantics, then a *Linking Hypothesis* should take the form of a complex dynamical system, and demonstrate the emergence of phonology-like properties from some specific combination of brain-like elements.

2. Introducing Attractors

The preceding argument leaves us with a well-defined problem: What kind of dynamical system could possibly give us something like a phonological grammar? The first obstacle to answering this question is that, while formal grammars are defined over a set of discrete symbols, dynamical systems (such as the brain) are typically understood as being fundamentally continuous[3]. This is where attractor dynamics are critical, because attractors allow us a way of realizing discrete behavior in an otherwise continuous system. Moreover, they are easily realizable in neural networks, making them a plausible candidate for a neural mechanism capable of underlying the discrete behaviour observable in phonological grammars.

---

2 Conceptually at least, this analogy is not a novel idea in phonology. For example, it appears in Prince & Smolensky (1997) as a proposal for interpreting Optimality Theory.

3 In one sense, this situation is precisely the inverse of the kinetic theory of gasses, which seeks to replace many discrete particles with a continuous field (Truesdell & Muncaster 1980).

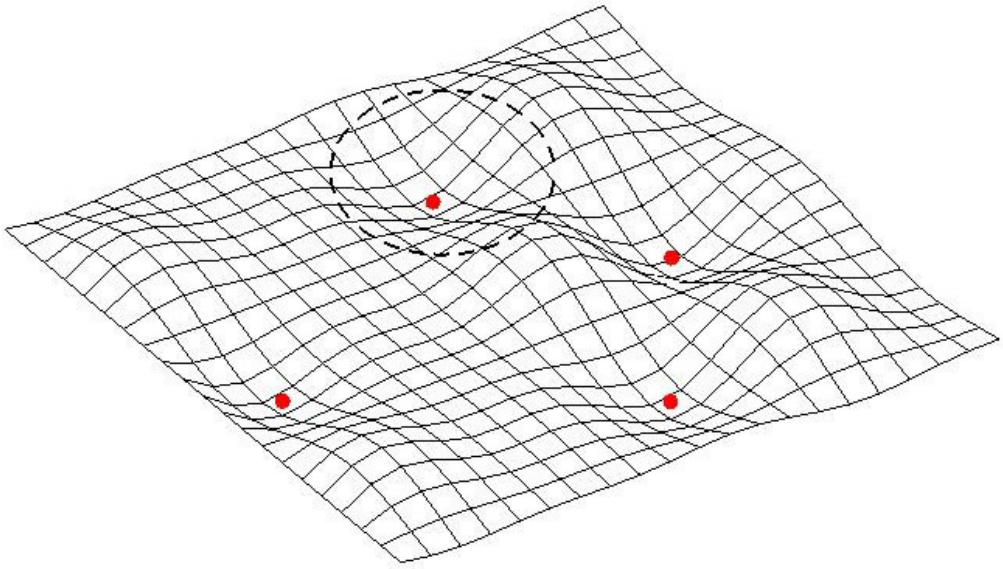

Figure 1: Conception of a network state-space. The z-axis corresponds to the free energy of the network. The red dots are attractors. (Illustration curtesy of Chris Eliasmith, Scholarpedia; source DOI: 10.4249/scholarpedia.1380; license CC BY-NC-SA 3.0.)

Like other artificial neural networks (ANNs), attractor networks consist of a number of simple units, which are interconnected with varying degrees of efficacy. Unlike other ANNs, attractor networks are characterized by symmetrical connections between units, which cause the network activity to settle on one of a number of asymptotically stable network states (i.e. attractor states). These stable states can be formally defined as local minima in an energy function and the behaviour of the network can be understood as analogous to the second law of thermodynamics: the entropy of the system increases over time, as the free energy decreases. This is sometimes visualised as a landscape of peaks and valleys (*Figure 1*), with the network always rolling down into the nearest valley.

The dynamics of attractor ANNs were popularized by Hopfield (1982), who noted that, if the attractor states are taken to represent pieces of information, then the network functions as a content addressable memory system.

Crucially for linguists, these attractor-memories are effectively discrete pieces of information. This is even true in cases where the individual units of the network are functionally gradient (Hopfield 1984). Thus, attractor dynamics are arguably our best candidate for explaining how a grammar over discrete elements could emerge in a seemingly analogue system like the human brain.

The model under examination here, the Phonological Latching Network (PLN), represents an attempted first step towards such a model. In its nascent form, it is necessarily an incomplete model of phonological grammar. It has no notion of lexical items, suprasegmental phenomena, or even a distinction between underlying and surface forms. Nonetheless, it does demonstrate how quintessentially phonological phenomena, such as assimilation, the Sonority Sequencing Principle (e.g., Clements 1990), and the Obligatory Contour Principle (e.g.,

McCarthy 1986), can emerge spontaneously from a relatively simple form of neural coding and memory retrieval.

3. Background and Outline of the Model

The PLN is a type of attractor network, similar to the Hopfield network (Hopfield 1982). This means that it stores memories as asymptotically stable states, which the network "self-organises" towards. However, most Hopfield-like ANNs have relatively simple dynamic properties: once switched on, the network will begin rearranging itself into the closest attractor state, where it will remain until the simulation is switched off. This limited degree of complexity has proven sufficient for investigating certain aspects of perception (e.g., Nasrabadi & Choo 1992) and memory capacity (e.g., Tsodyks & Feigelman 1988). However, it is clearly inadequate for modelling natural language grammar, which requires (minimally) the ability to define relationships between discrete memories.

Latching networks can be understood as an attempt to introduce between-memory dynamics into an attractor network. Fundamentally, latching networks behave like a Hopfield network, with the additional property that once an attractor state has been reached; the network begins to "latch" into a different attractor basin. Thus, the network can produce strings of discrete elements, which exhibit a kind of inherent grammar.

The latching dynamics themselves emerge from the introduction of a "fatigue" function (i.e. adaptation or inhibition) to active units, which means that attractor states become increasingly unstable once reached. This is what causes the network to latch into a different, nearby attractor, and ultimately places restrictions on what kinds of strings the network can produce.

3.1. The Potts Unit

The notion of fatigue in a latching network requires that individual units have an inactive state, that is a state which they tend to after periods of activity. The classical binary-unit Hopfield network lacks this property, since its units are either in an excitatory or inhibitory state.

The solution explored here is replace the binary-state Hopfield units with multi-state (or "Potts") units, which have previously been studied as models of associative memory (Treves 2005; Russo & Treves 2012; Pirmoradian & Treves 2012; 2014; Song, Yao & Treves 2014; Kang et al. 2017; Naim et al. 2017). As in the case of the Hopfield network, single unit dynamics can be modelled using a rule based on heat bath dynamics (Treves 2005; Kanter 1988). These dynamics can be conceptualized as something akin to a compass needle being pulled in different directions by the various inputs received from other units in the network. The number of different directions in which the needle can be pulled is determined by the parameter S, which is typically in the order of 5 to 9, with one extra direction for the inactive state. Therefore, the state of a given Potts unit i is a probability vector of S+1 components, denoted below by σ_i^k for the active states, and σ_i^0 for the null-state.

At time *t*, the value for each active state *k* of any given unit *i* is given by the equation:

$$\sigma_i^k(t) = \frac{exp\left[\beta r_i^k(t)\right]}{\sum_{l=1}^{S} exp\left[\beta r_i^l(t)\right] + exp\left[\beta(\theta_i^0(t)+U)\right]} \quad (1)$$

Where r is dynamic input variable, β is the global noise parameter, and U is a global parameter determining input to the inactive state. The time dependent thresholds for each state of each unit are given by the vector θ_i, which also has *S+1* components denoted by θ_i^k for the active-state thresholds, and θ_i^0 for the null-state threshold.

Complimenting *Equation 1*, the value for the inactive state at time *t* is given by:

$$\sigma_i^0(t) = \frac{exp\left[\beta(\theta_i^{0\,k}{}_i(t)+U)\right]}{\sum_{l=1}^{S} exp\left[\beta r_i^l(t)\right] + exp\left[\beta(\theta_i^0(t)+U)\right]} \quad (2)$$

Calculating the values for σ_i at time *t* requires first determining both the values for the dynamic thresholds θ_i and the input variables r_i, which are linked through a system of differential equations (*Equations 3*, *4*, and *5*).

Firstly, the dynamic thresholds for the active-states are calculated from the current state of σ_i:

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t) \quad (3)$$

As the level of activation of a given state, *k*, in σ_i increases, so too will the corresponding threshold in θ_i, modulated by the coefficient τ_2, which is a global parameter controlling the rate of active-state fatigue (or adaptation).

The dynamic threshold for the null-state is given by:

$$\tau_3 \frac{d\theta_i^0(t)}{dt} = \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^0(t) \quad (4)$$

Therefore, θ_i^0 increases relative to the sum of all active-states in σ_i, modulated by the global parameter τ_3.

Note that θ_i^0 and θ_i^k (and their respective parameters τ_3 and τ_2) are intended to model two different forms of fatigue over two different timescales. While τ_2 is typically assumed to represent the rate of short-term depression in synapses, τ_3 is assumed to represent the rate of slow inhibition within a cortical patch.

Finally, once the dynamic thresholds for unit *i* at time *t* are known, the input variables r_i^k, can be calculated with respect to the local field h_i^k:

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t) \quad (5)$$

The local field for each state at time *t* is defined as the summed influence of presynaptic units, added to a local feedback term with the coefficient *w*:

$$h_i^k(t) = \sum_{i \neq j}^{N} \sum_{l=1}^{S} J_{ij}^{kl} \sigma_j^l(t) + w\left(\sigma_i^k(t) - \frac{1}{S}\sum_{l=1}^{S} \sigma_i^l(t)\right) \quad (6)$$

Where J_{ij}^{kl} denotes the connection strength between state *k* of unit *i* and state *l* of unit *j* (see Section 3.3.1 below for explanation of how connections strengths are determined).

Under the standard interpretation, each Potts unit is an effective model for a smaller attractor network (Naim *et al* 2018). Therefore, the *w*-term is intended to subsume the internal dynamics of each cortical patch. Continuing the compass needle analogy, it can be thought of as giving the compass needle an extra push towards whichever direction it is currently closest too.

3.2. Latching Dynamics

The relationship between fatigue on individual units and the emergence of latching dynamics is relatively transparent: an attractor state simply can't be maintained once the active units start switching off. What is less transparent however, is the rich complexity of the latching dynamics themselves.

In one sense, a latching network obeys the same principle of minimizing free-energy that all attractor networks obey, that is it "rolls into the valley" (*Figure 1*). The additional complexity arises from the fact the free-energy of any given network state is continuously changing as the fatigue rises and declines on individual units. In other words, the attractor landscape itself is constantly shifting. What was "downhill" at one moment in time can become "uphill" the next. The sheer mathematical complexity of these dynamics means that attempting to give a deterministic account of why one attractor latches into another is, although theoretically possible, massively intractable in practice.

For this reason, latching dynamics have more commonly been analysed probabilistically, for example, what determines the probability of a latch between any two attractors? This is still a non-trivial problem, but in general terms we can state that the probability of a latch between any two given attractors in the network depends on the overlap in the representations of those attractors (Russo & Treves 2012; Kang et al. 2017). The notion of "overlap" here has two dimensions: Firstly, how many active units do the two attractor states share? Secondly, how many of those shared units are in the same unit state? The interaction between these two types of overlap is quite complex, owing to the fact that they are governed by slightly different fatigue effects. The fatigue on individual unit states is controlled by the parameter τ_2, while the fatigue on whole units is controlled by the parameter τ_3. In the case where $\tau_2 << \tau_3$, an individual unit state will fatigue long before the unit itself begins to switch off (i.e. enter its inactive state). Thus, the degree of fatigue of an individual unit can bias the target of a latch in several ways: If a given unit is not fatigued, then the network will prefer to latch into an attractor in which that unit is both active and remains in the same unit state. However, if an individual unit state is fatigued, but not the whole unit, then the network might prefer to latch into an attractor in which the unit is active but in a

different state. Finally, if the unit itself is fatigued, then it will begin to enter to switch off and the network will prefer to latch into an attractor in which that unit is inactive.

The resulting global dynamics produces distinct "latching bands" in the degree of overlap between attractors: for some degrees of overlap, a latch will be highly probable, while for other it will be impossible (Russo & Treves 2012). If we allow ourselves a rhetorical simplification, we could say that the latching obeys a Goldilocks-principle; preferring to latch between memories which are neither too similar nor too dissimilar. In this sense a latching network always has an inherent grammar to it, since encoding multiple attractors in the network will always produce varying degrees of overlap between those attractors. Thus, a given latching network typically cannot produce all possible permutations of the memories represented by its attractors, but only a subset.

Finally, although the description of latching dynamics given so far only considers the probability of a latch between any two attractors, it should not be inferred that the network behaves like a finite-state machine. A latching network typically does exhibit long distance effects. This is a consequence of two facts: Firstly, the recovery time of a fatigued unit will typically be longer than a single latch. Thus, even if a given unit is inactive in the current attractor, it may still be fatigued from some earlier activation, and thus be less inclined to switch on again for the next latch. Secondly, in practice the retrieval of a memory is not actually understood as reaching one specific attractor state, but rather as passing through that state's basin of attraction. This means that there are very many network states that would all be interpreted as a retrieval of the same memory, and each of these network states can behave differently in terms of where they would prefer to latch next.

When viewed from the macro-level then, the behaviour of the network might seem quite opaque: a single memory (or attractor basin) can produce a latch to one of many different targets, for reasons which are only apparent when viewed from the micro-level. This typically precludes reducing the global behaviour of the network to that of a deterministic automaton.[4]

Despite this, it is nonetheless possible to uncover distinct tendencies or biases in the strings produced by latching, when using probabilistic methods. As we shall see, the Goldilocks behaviour of the network can be seen to give rise to common phonological processes such as place assimilation and the Obligatory Contour Principle (OCP), while the slower cycles of fatigue can reproduce a kind of Sonority Sequencing Principle (SSP).

3.3. Constructing a Neurologically Plausible Model

Unlike many ANNs, the Potts units of the latching network do not strive to model individual synapses, firing rates or action potentials. Rather they can be thought of as an effective, or "grey box", model, where certain details are subsumed into a system of differential equations. For this reason, a Potts model is as much a

[4] This does not entail that *no* configuration of a latching network can reproduce *some* level of complexity on the Chomsky hierarchy; this ultimately remains to be seen.

theoretical model of specific system dynamics, as it is a model of neurological reality. Indeed, while many aspects of the latching model are intended to capture known facts about neural function, the exact neural implementation of a Potts unit is somewhat open to interpretation.[5] Under the standard view, each Potts unit is an effective model for small patches of cortex. The active states of each unit represent different local attractors in each patch, while the self-reinforcement term represents the internal attractor dynamics of the patch. Then the behaviour of the network as a whole is taken to model global dynamics between relatively distant areas of the cortex (Naim et al. 2018). This standard view of a Potts network seems well suited to modelling language, which is known to be a widely distributed cognitive faculty (see, e.g., Hickok & Poeppel 2007).

The PLN is intended to model the representation of phonological information in the cortex. While a great deal is still unknown on this topic, recent ECoG studies have uncovered a striking degree of congruence between phonological representations in the cortex and the abstract, discrete features employed by linguists to explain the behaviour of phonological grammars (Bouchard et al. 2013; Mesgarani et al. 2014). Specifically, these studies uncovered the existence of small patches of cortex which are highly sensitive to specific phonological features. Moreover, they hint at a spatial asymmetry between manner and place features, with manner features being distinguished more strongly in the superior temporal gyrus (STG), and place features being distinguished more strongly in the ventral sensorimotor cortex (vSMC). Similarly, both experimental results and theoretical modelling have suggested that phase coupling between these areas may form a critical component of the phonological capacity (Assaneo & Poeppel 2018).

These findings suggest three relevant criteria for the structure of the PLN: Firstly, the network should be split into two sub-networks: an auditory sub-network for manner features, and a motor sub-network for place features, and that production should arise from synchronous activity between these areas. Secondly, phonological similarity between phones should be captured in terms of shared units in the network (i.e. shared patches of cortex), such that the Goldilocks-principle is acting over phonological properties. Finally, the congruity between the ECoG studies and phonological theory suggests that the representations themselves could be constructed using abstract phonological features as a guide.

3.4. Building Phones

Unlike neural networks typically employed in machine learning and connectionist frameworks, the PLN is not subject to any form of supervised learning (cf. Alderete & Tupper 2018). Rather, the patterns of activity which represent memories are generated algorithmically by the experimenter, and then encoded in the connections between units using a simple Hebb-like rule.

---

[5] In Marrian terms (e.g., Marr & Poggio 1973), if the linguistic model is the *computational* level, then the latching network is the *algorithmic* level, while the *implementational* level would be occupied by some exact neural model of the Potts units.

Because the memories in the PLN are intended to represent phones, the algorithm for memory generation in the PLN works from a given phoneme inventory, which is formally defined in terms of a relevant set of phonological features (see Appendix). Broadly, each of the features is defined as a random pattern of activity. These patterns can then be combined into phones, following the definitions in the phoneme inventory. The process for combining features is a competitive one, whereby the individual features are used as competing "suggestions" for the final phone. Contradictions between suggestions are resolved by weighting individual features, such that only the strongest suggestions for each unit will contribute to the phone representation.

The same features are used in both the auditory and manner sub-networks, and the asymmetry is achieved by reversing the weighting of those features. So, the auditory network representations are generated with heavily weighted manner features and weakly weighted place features, and vice-versa for the motor sub-network.

The phone inventory is loosely derived from English phonology, with the important caveat that there are no minimal pairs based on voicing distinction. The large number of features means that phones are redundantly over-specified, as otherwise the algorithm tended to produce phones with excessive overlap. Slowly adding redundant features to the inventory was a way of overcoming this problem. However, it should be noted that some information is lost during phone creation, so not all the features should be regarded as playing a role in the behaviour of the system (by extension, the PLN should not be interpreted as for or against any particular theory of phonological features).

The process for generating representations in the PLN will now be described in detail. First, each phone μ is formally defined as a set of M features:

$$\mu := \left\{\varphi_1^{\mu}, \varphi_2^{\mu}, \dots, \varphi_M^{\mu}\right\} \tag{7}$$

The notation φ^{μ} indicates that feature φ is a member of phone μ.

The features defining a given phone are, in principle, unordered. However, the process for generating phones requires two different orderings of the features in μ (one for each sub-network).

A sub-network is defined as a pool of units and is denoted by Q, which in the PLN can take the value *mot* or *aud*. Any given unit in the network, i, is assigned membership to one, and only one, of the pools. The two pools contain the same number of units: $N/2$.

The auditory and motor components of each phone are defined as ordered tuples of all elements in μ:

$$\mu^{Q} := \varphi_1^{\mu^{Q}}, \dots, \varphi_m^{\mu^{Q}}, \dots, \varphi_M^{\mu^{Q}} \tag{8}$$

The order is always derived from the inventory given in the Appendix. Also note that μ^{aud} and μ^{mot} always contain the same elements, but in the reverse order, that is the relationship always holds that $\varphi_m^{\mu^{mot}} = \varphi_{M-m+1}^{\mu^{aud}}$.

The function *W* assigns a weight to each feature, with respect to its position in μ^Q, such that:

$$W\left(\varphi^{\mu^Q}\right) = e^{\frac{q(m-1)}{M-1}} \tag{9}$$

Where *m* is the index of feature φ in μ^Q, *M* is the total number of features in μ^Q, and q is a global parameter used to control the cumulative influence of lowly weighted features (the smaller the value of q, the greater the influence of the lower weighted features).

The result of the function *W* is that the weightings of the features in μ^Q fall along an exponential scale between *1* (when *m=1*) and e^q (when *m=M*).

The weightings from *W* are used to determine the actual representations for a phone.

First, the representation for phone μ in pool *Q* is denoted as ξ^{μ^Q}, which is defined as a tuple whose components represent the units in pool *Q*, and can take a value from 0 to S.

$$\xi^{\mu^Q} := \xi_1^{\mu^Q}, \dots, \xi_i^{\mu^Q}, \dots, \xi_{\frac{N}{2}}^{\mu^Q} \tag{10}$$

The final representation for a given phone will simply be the concatenation of the two pools: $\xi^{\mu} := \left(\xi^{\mu^{mot}}, \xi^{\mu^{aud}}\right)$.

Generating the representations for phones depends on the representations for individual features. Each of the features in the phoneme inventory is defined as a hypothetical network state within each sub-network which, following Pirmoradian & Treves (2012; 2014), are generated using sparse[6] patterns of noise. The random noise pattern representing a feature φ is indicated as ξ^{φ}, where, again, each element takes a value between 0 and *S*.

$$\xi^{\varphi} := \xi_1^{\varphi}, \dots, \xi_i^{\varphi}, \dots, \xi_{\frac{N}{2}}^{\varphi} \tag{11}$$

Crucially, the patterns for features are uncorrelated with one another, i.e. they should be approximately equally dissimilar.

Additionally, the sparsity of these patterns is enforced by the parameter a_{feat}, which represents the probability that the value of any component ξ_i^{φ} is non-zero. In practice, the value of a_{feat} is typically lower than the value of *a*, the sparsity of the phones. This ensures that no phone can be dominated by a single feature.

Note that any given feature pattern, ξ^{φ}, is constant for all phones and all pools. Features vary only in terms of their membership in μ and weighting in μ^Q. Also note that each feature pattern is only defined over half the total units of the network. This is because, in principle, each feature appears in both the auditory and motor sub-networks.

---

6 That is, only a small subset of units are active.

As well as the patterns representing phonological features, each phone also has a corresponding "noise" feature, $\mathcal{N}$, which is introduced as a means of preventing excessive overlap between phones. The noise feature is similarly defined:

$$\xi^{\mathcal{N}} \coloneqq \xi_1^{\mathcal{N}}, \dots, \xi_i^{\mathcal{N}}, \dots, \xi_{N/2}^{\mathcal{N}} \tag{12}$$

Having defined and generated all the relevant feature representations, the final value of any unit in ξ^{μ^Q} is set to the value of k (between 0 and S) which carries the highest weight, from W, which is summed over all features in phone μ.

$$\xi_i^{\mu^Q} = \arg\max_{1 \le k \le S} \sum_{\varphi \in \mu} \delta_{\xi_i^{\varphi} k} \; W\left(\varphi^{\mu^Q}\right) + pe^q \delta_{\xi_i^{\mathcal{N}} k} \tag{13}$$

The Kronecker delta is a function which equals 1 when its arguments are the same, but 0 otherwise. The last term in (13) represents the influence of each phone's unique noise feature, $\mathcal{N}$, where p is a global parameter used to control the influence of all noise features. Note that if $p=1$, then the weight of the noise feature will be equal to the weight of the strongest feature in μ^Q. High values of p (greater than 1), were found to be useful for maintaining an optimum degree of overlap between representations.

Additionally, the sparsity of the representations is maintained by assigning a value of 0 to those units whose weighted suggestion falls below some threshold. The value of this threshold depends on the sparsity parameter a, such that only the $aN/2$ strongest suggestions in ξ^{μ^Q} are non-zero.

Having generated the representations for each phone, the patterns are encoded in the weight matrix as attractors using a Hebb-rule. Each phone μ suggests a connection strength J between state k of unit i and state l of unit j, which is given by the rule in:

$$J_{ij}^{kl}(\mu) = (\delta_{\xi_i^{\mu} k} - \tfrac{a}{S})(\delta_{\xi_j^{\mu} l} - \tfrac{a}{S})(1 - \delta_{k\,0})(1 - \delta_{l\,0}) \tag{14}$$

Here, as before, the Kronecker delta's output is 1 when the two arguments are equal and is 0 otherwise. Therefore, in a pattern, ξ^{μ}, if unit i is in state k and unit j is in state l, where $k=l$, then the connection will be positive (excitatory), else the connection will be negative (inhibitory). The last two factors ensure there are no connections to/from units in the null state (if k or l equal 0).

The final value for each connection is determined by summing over all memories in the network, and multiplying by a normalization factor:

$$J_{ij}^{kl} = \frac{c_{ij}}{Ca(1-\frac{a}{S})} \sum_{\forall \mu} J_{ij}^{kl}(\mu) \tag{15}$$

Where c_{ij} is set to 1 when i and j share a connection and is 0 otherwise. This value is normalized by C, the average number of connections per unit, and a, the sparsity parameter.

The probability that they share a connection is defined by the variable c_{int} if i and j are both in the same sub-network, or c_{ext} if they are not:

$$For\ Q \neq R, \qquad c_{ij}^{QR} = \begin{cases} 1 & with\ probability\ c_{ext} \\ 0 & with\ probability\ (1 - c_{ext}) \end{cases}$$

$$For\ Q = R, \qquad c_{ij}^{QR} = \begin{cases} 1 & with\ probability\ c_{int} \\ 0 & with\ probability\ (1 - c_{int}) \end{cases}$$

This process is intended to ensure that the similarity between the representations of phones in the PLN correlates strongly with their phonological similarity, as is given by the feature definitions in the phoneme inventory. We can see evidence of the non-random structure of the PLN memories, shown in *Figure 2*. Here we can see that, in general, the more units two memories in the PLN share, the more likely it is that those shared units are in the same Potts state. This implies that overlap between representations is a consequence of shared features which suggest specific Potts states for individual units.

4. Analysis of PLN Behaviour

Because the process of generating features depends heavily on randomization, it is possible to generate multiple weight matrices for the same phoneme inventory, which have different latching properties (i.e. they produce different grammars).

Using the same phoneme inventory and network hyperparameters (see Appendix), the latching strings from 125 trials, representing 8 different grammars,

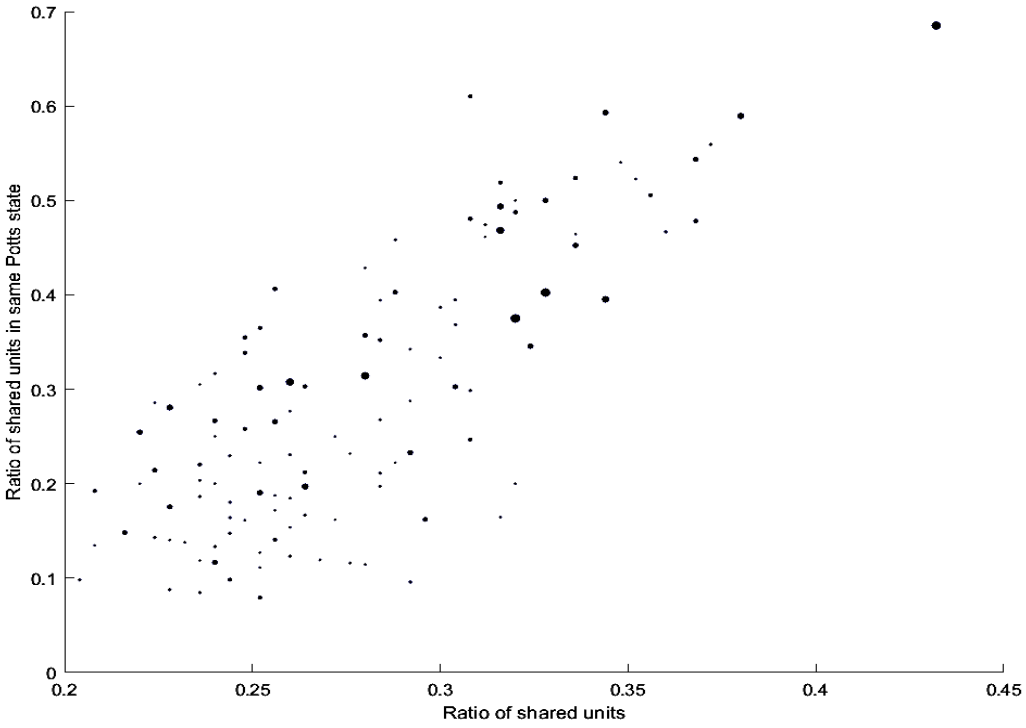

Figure 2: Overlap of memories produced by feature super-position. The size of each circle indicates the total number of attested transitions between the two memories during the simulations.

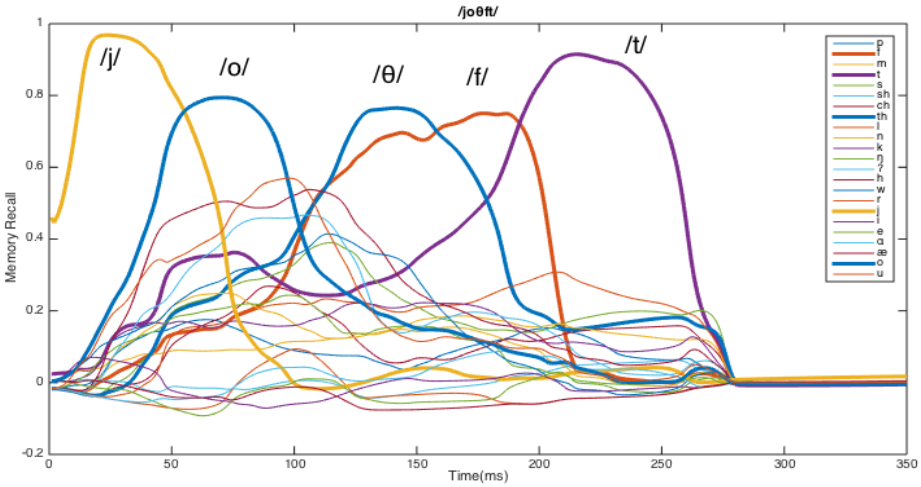

Figure 3: Example of a latching string.

were collected into a corpus containing a total of 464 individual phoneme transitions. This was found to be large enough to allow statistical generalization, but small enough that all latching transitions could be manually checked for network pathologies (failed retrievals, mixed states, etc). Only strings which exhibited no obvious pathologies were included in the corpus. All strings were between 2 and 8 segments long, with an average length of 4.7 segments. Strings were generated by placing the network into a state which matched a 50% memory retrieval and allowing it to run for 400 time steps.

The strings were assessed for evidence of assimilation, OCP and SSP. The rate at which these phenomena occur was then compared to chance level, that is a grammar in which the probability of a transition between any two phones is the same for all phones in the inventory. The extent to which the PLN grammars deviate from chance level can be taken as evidence of whether these processes are inherent to the PLN.

4.1. Segmental OCP

In its general form, the Obligatory Contour Principle (OCP) requires that there be some minimum degree of difference between adjacent objects. This may or may not be an instance of a more general bias against repetition in language (see, e.g., Walter 2007). In relation to segmental phonology, this can be interpreted in two different ways: firstly, it can mean that the same phone cannot surface twice in a row, or secondly, that adjacent segments cannot be similar with regards to some featural specification (McCarthy 1986).

This first sense of segmental-OCP is a trivial property of the PLN, since the latching dynamics are driven specifically by an active memory becoming unstable. There is simply no way the network could latch out of, and immediately back into, the same memory. The simulations confirmed this, with phone repetitions exhibited in exactly 0% of the recorded transitions.

The PLN also seems to exhibit something closer to the second definition of segmental-OCP. For example, there were no recorded examples of a transition

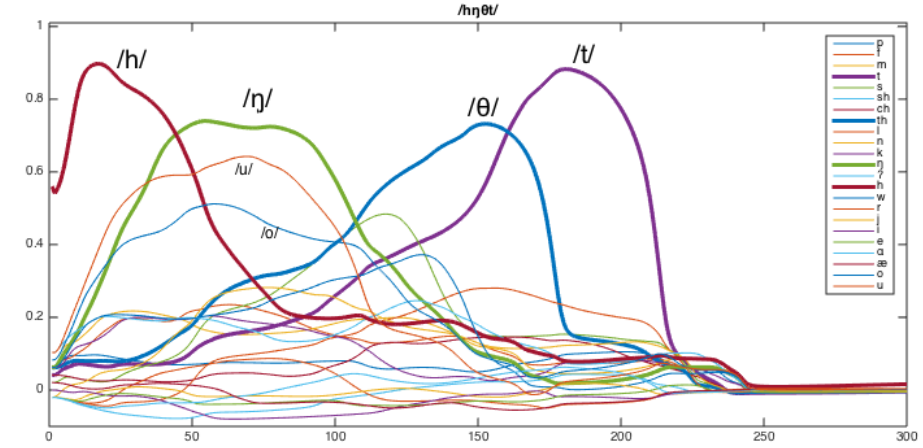

Figure 4: The /θ/ and /t/ phones are similar in both their manner and place of articulation, but are still a possible transition for the PLN.

between /s/ and /ʃ/, suggesting that the network has reproduced something like the OCP-driven epenthesis seen in English plurals and possessives (e.g., bu[ʃ] -> bu[ʃəz] etc.). However, one grammar did spontaneously produce the string [kŋutʃsθu], where the transition from /tʃ/ to /s/ would normally be seen as an OCP violation in the context of English phonology.

A closer examination of the representation overlap of these phones reveals the important difference. Firstly, the total percentage of shared units between /s/ and /ʃ/ in this grammar is much higher (31.2%) than /s/ and /tʃ/ (22.4%). And secondly, of those shared units, a much higher percentage are in the same Potts state when comparing /s/ to /ʃ/ (50%) than /s/ and /tʃ/ (28%). This supports the hypothesis the absence of /s/->/ʃ/transitions in the PLN is an OCP effect, while /s/ and /tʃ/ are dissimilar enough to fall within the "Goldilocks" zone.

4.2. *Assimilation*

Processes in which segments become more similar to their neighbours – in terms of their feature specification – are extremely common cross linguistically (e.g.,

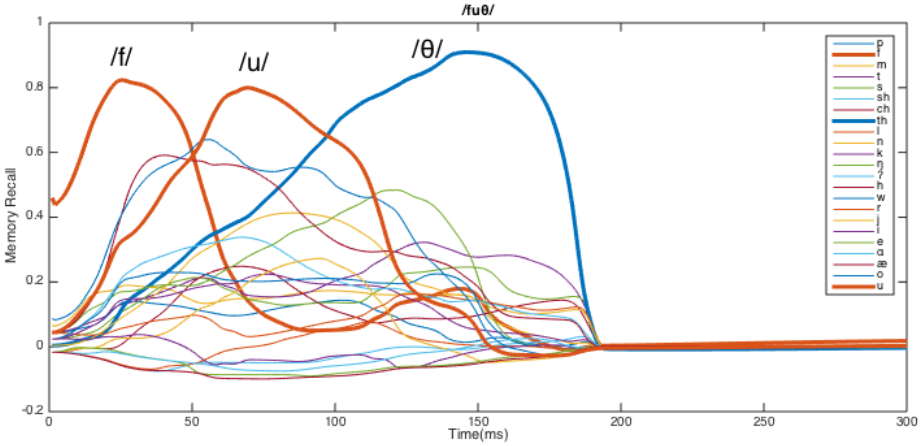

Figure 5: The /f/ and /u/ share the feature [round], so the first transition is interpreted as an instance of place assimilation.

Ohala 1990). With the PLN, a transition was counted as an instance of assimilation if the two phones shared a feature, as defined by the inventory in the Appendix. An example of this is shown in *Figure 5*.

4.2.1. Place

Transitions exhibiting place assimilation were found in 244 (52.6%) transitions, which is slightly above the chance rate (44%). However, the picture becomes more interesting when we break down the assimilation probabilities for each feature. As we can see in *Table 1*, the features HIGH, EXTERIOR, LABIAL, VELAR and ALVEOLAR appear to assimilate at above chance rate, while the others assimilate below chance rate.

These numbers suggest that only some of the features are participating in assimilation. This is arguably a welcome result, since natural phonological grammars typically only exhibit assimilation for one or, at most, a few place features.

However, these numbers alone do not immediately inform us of why some features participate in assimilation, but not others. This picture is further complicated by the fact that not all of these features are independent. In cases where the phones delineated by one feature are a strict subset of the phones delineated by another feature (e.g., all labials are also exterior, etc.), then a naive statistical method doesn't allow us to determine which feature is decisive for causing assimilation.

We can partially circumvent this problem by comparing mutally exlusive pairs of features, for example, HIGH vs LOW, LABIAL vs CORONAL, and ALVEOLAR vs POST-ALVEOLAR. Each phone may have, at most, one of the features from each of these pairs.

Looking at *Table 1*, we can see that within each exclusive pair, it is the feature with the highest weight during phone generation (Section 3.3.1) which appears to assimilate at above chance rate, while the feature with the lower weight assimilates at a below chance weight.

FEATURE	ASSIMILATION %	BASELINE %
HIGH	3.66	1.7
LOW	1.08	4.73
FRONT	3.45	6.8
EXTERIOR	35.34	18.9
LABIAL	15.73	6.8
DORSAL	2.16	6.8
CORONAL	9.27	12.1
VELAR	2.59	0.76
GLOTTAL	0.22	0.76
ANTERIOR	0.65	1.7
ALVEOLAR	6.03	4.73
POST-ALVEOLAR	0.43	0.76

Table 1: Place assimilation probabilities by feature, ordered from strongest weight in motor sub-network (HIGH) to lowest (POST-ALVEOLAR).

FEATURE	ASSIMILATION %	BASELINE %
APPROXIMANT	0.22	2.01
CONTINUANT	76.29	54.63
NASAL	3.01	1.7
SONORANT	61.42	31.94
VOCALIC	17.89	7.05
CONSONANTAL	26.5	37.05

Table 2: Manner assimilation probabilities by feature.

This gives us some indication that the relative weighting of features during phone creation plays a role in determining assimilation in the emergent grammar. Intuitively, this makes sense insofar as features with heavier weights will "suggest" more unit states for the final representation of each phone. Therefore, the heavier the weight of a feature, the more overlap we should expect between any two phones which share that feature, and the greater the probabilty that the network will prefer to latch between them.

4.2.2. Manner

The random baseline for manner assimilation is much higher at 81.1%, owing to the smaller number of manner features, and the larger number of individual phones delimited by each manner feature. The actual rate of manner assimilation within the network is, again, slightly above chance at 89.4%.

Similar to place features, we also see a difference between individual manner features (*Table 2*).

That the CONTINUANT and NASAL features exhibit assimilation is broadly in keeping with natural phonology, for example, intervocalic spirantization (Kaplan 2010) and vowel nasalization (Krämer 2019). More surprising, perhaps, is the apparent assimilation of the features SONORANT and VOCALIC, which are typically not thought to spread or assimilate (see, e.g., Clements & Hume 1995 where these features appear on the root node). However, this can actually be explained as an effect of the sonority sequencing effect in the network (see Section 4.3), whereby the network tends to slowly oscilate between greater and lesser sonority. Since the features SONORANT and VOCALIC are the main delineators between degrees of sonority, the sonority sequencing will naturally cause phones with these features to cluster together, rather than being even distributed. Thus, the statistical effect need not be regarded as a consequence of spreading or assimilation *per se*, but rather of sonority sequencing.

4.3. Sonority Sequencing Principle

The Sonority Sequencing Principle (SSP) refers to the tendency for sonority to follow a monotonically rising-then-falling pattern across a single syllable. Arguably, this forms the very definition of a syllable: it is a sonority peak (Clements 1990). For this reason, the SSP represents a good measure for the "naturalness" of the strings produced by the PLN. For example, strings which

VOWELS	GLIDES	LIQUIDS	NASALS	OBSTRUENTS
0	1	2	3	4

Table 3: Sonority scale.

neatly transition between consonants and vowels could be regarded as more natural than strings which consist only of stops.

Unlike the other measures, the extent to which the network obeys sonority sequencing is defined in relation to whole syllables, not individual transitions. And since the PLN does not itself process any information relating to syllable structure, the experimenter must parse the strings into syllables manually. This requirement presents the basis for a simple metric for approximating the model's preference for strings which obey SSP. Specifically, each string produced by the PLN is given the best possible parse according to the SSP. The string is then assigned a value from the sonority scale (*Table 3*), according to the *least* sonorant nucleus required when parsing (*Table 4*).

Note that this method ignores syllable plateaus and size of the sonority "jump" between adjacent segments. Some examples of how these scores would be assigned to example strings are given in *Table 4*.

Once every string in the database has been assigned a sonority score, the mean score (across all strings) is compared to a random baseline, whose sonority sequencing score has been computed for strings of length 3, 4, 5, 6, 7.[7] The sonority scores for different string lengths, both from the PLN and the baseline, are given in *Figure 6*.

The SonSeq score for the latching strings is lower than the baseline for all string lengths, suggesting that the PLN tends towards strings which can be parsed by to the SSP.

Naturally, this simple metric inherently ignores various complexities associated with sonority sequencing in natural grammars (minimum/maximum distance, permissible plateaus, onset/codas asymmetries, etc.). However, it does capture the extent to which the PLN wants to oscillate monotonically between vowels and obstruents. This is informative insofar as it presents an unbiased measure of how well the latching strings conform to sonority sequencing, within the confines of a system which has no actual notion of syllable structure.

STRING	SYLLABLE PARSE	LEAST SON. NUC.	SONORITY SCORE
"Σ L O"	ʃlo̲	o	0
"L Σ O"	l̲. ʃo	l	2
"Θ N Æ L P F"	θnæ̲lpf	æ	0
"Θ N Æ L P F M"	θnæl.pfm̲	m	3

Table 4: Example sonority scores.

[7] Note that the SonSeq score worsens (increases) as the strings lengthen by simple virtue of the fact that the longer the string, the greater the probability of encountering a low sonority nucleus.

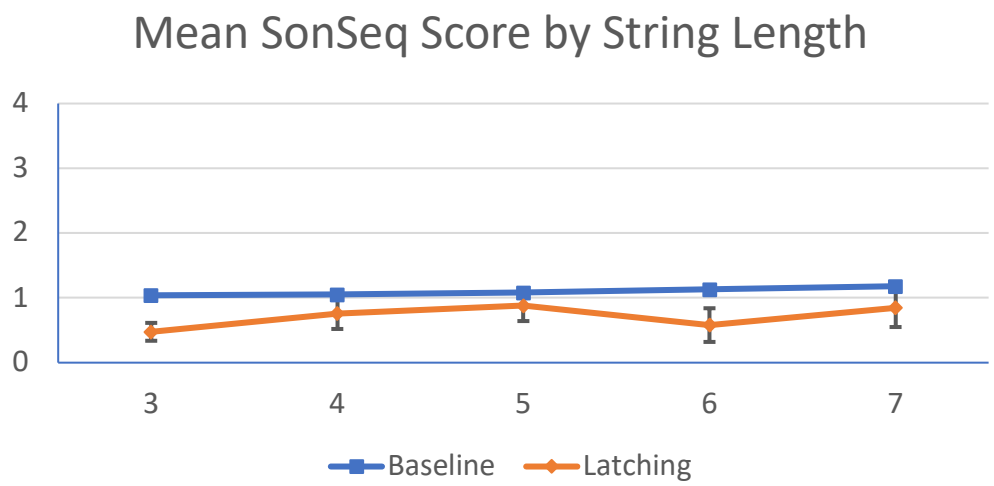

Figure 6: Sonority Sequencing score for latching strings (red) versus random baseline (blue).

4.3.1. *SSP as Oscillation*

Having established the PLN's propensity for oscillating between sonorous and non-sonorous segments, it remains to determine *why* the network exhibits this behaviour. Much like the OCP effect, the SSP effect can also be understood as following from the fatiguing of individual units. In simple terms, because the network representations are intended to reflect phonological properties, we should expect that certain units will be more active when representing sonorous phones than non-sonorous ones (and vice-versa). Thus, if these "sonority units" are fatigued from repeated activation, then we should expect the network to latch into non-sonorous memories for a time, at least until the "sonority units" have recovered from their fatigue. Similarly, the converse will be true for any "non-sonority units" which are most active for non-sonorous phones. Therefore, we should expect the network to slowly oscillate between sonorous and non-sonorous states, driven by the slow fatiguing and recovery of the individual units.

Of course, this oscillation can only persist if sonority is indeed encoded in the network in this way. As already noted in Section 4.2.2, the degrees of sonority within the phone inventory are determined primarily by the features SONORANT and VOCALIC. However, because the process of generating representations relies on randomisation, we need to look at the network representations themselves to see whether or not these features actually play a role in producing the SPP effect. We can get a sense of this by grouping the individual phones in the network into 3 broad sonority categories: vowels, sonorant consonants, and obstruents (which correspond to the features SONORANT+ VOCALIC, just SONORANT, or neither, respectively) and examining the average representation overlap within and across these categories.[8]

The data in *Table 5* show the overlaps across these categories from a single randomly chosen grammar of the PLN. As we might expect, the average overlap

[8] Distinguishing the entire sonority hierarchy requires additionally the features APPROXIMANT and NASAL. However, for legibility we can restrict ourselves to this tripartite distinction.

SONORITY CATEGORIES	AVERAGE % OF SHARED UNITS	% OF THOSE SHARED UNITS IN SAME STATE (ABSOLUTE %)
SON-SON	29.24	30 (8.8%)
OBS-OBS	27.6	31 (8.55%)
V-V	27.6	33 (9.2%)
SON-V	26.05	25 (6.47%)
OBS-SON	25.63	25 (6.38%)
OBS-V	24.83	20 (5.05%)

Table 5: Overlap across sonority categories within a single grammar.

is highest within each category (obstruent, sonorant, vowel), somewhat lower when comparing obstruents to sonorants and sonorants to vowels, and lowest when comparing obstruents to vowels. The divide is even sharper when we examine the ratio of those shared units which are in the same Potts state, where we also see a much higher ratio of shared unit states within categories, when comparing across categories (*Table 5*).

This pattern, taken with the high rate of SONORANT and VOCALIC assimilation (Section 4.2.2), supports the oscillation explanation outlined above. To understand why, recall that the network has two types of fatigue, one which applies to individual Potts states, and one which applies to whole units. The tension between these two types of fatigue are critical for determining the behaviour of the latching network. Specifically, latching is driven by memory overlap in the case where unit fatigue is slower than individual Potts state fatigue (Kang et al. 2017), which is the case in the PLN. This is because latching occurs when an attractor becomes unstable due to fatigue, and since unit states fatigue faster than whole units, then latching will be driven the competing drives to maintain active units but to deactivate fatigued unit states. The consequence in this case will be a latch between memories which share the most units, but only if those units differ enough in their individual states.

5. Discussion

The analysis of the latching corpus presented here suggests that the PLN exhibits a degree of place assimilation and sonority sequencing, with a near-absolute kind of segmental OCP, or anti-adjacent-repetition of phones.

In terms of understanding why the network exhibits certain behaviours, arguably the most straightforward of the three is the segmental OCP. The "Goldilocks" behaviour of the PLN—preferring latching targets which are sufficiently dissimilar but not too dissimilar—will naturally prohibit latching out of and back into the same phone. Of course, depending on the specific overlaps of the memories in the network, this OCP effect can also to extend to phones which are similar though not identical. Thus, as seen in Section 4.1, it is perfectly possible to create an English-like grammar where /s/ and /ʃ/ are separate phones, but where transitioning from one to the other is strictly impossible, by virtue of the high degree of overlap in their representations.

Similarly, the PLN's bias towards assimilation can be straightforwardly understood as a result of the "Goldilocks" principle – the network prefers latching targets which are sufficiently different from the current state (OCP), but not *too* different (assimilation). Once again, whether or not a given grammar actually exhibits a given type of assimilation depends on the exact network representations that constitute the phones in the inventory: If two phones share a feature with a higher weight (during phone creation), then more of the overlap between the phones will be determined by that feature, ergo strongly weighted features are more likely to cause assimilation.

Finally, the PLN's apparent preference for oscillating between greater and lesser sonority can also be understood as a cumulative effect of the fatiguing of individual units in the network. However, unlike the OCP and assimilation effects, we need to consider the role of fatigue over a longer timescale.

Nonetheless, because the PLN is, in some sense, an incomplete model of phonological processing, a certain degree of care is required when attempting to draw direct comparisons with concepts taken from phonological theory. With that in mind, it is worth considering some of the limitations of the PLN model, how that affects our interpretation in phonological terms, and what that might mean for future research.

For example, the OCP-like effect exhibited by the PLN does not, by itself, capture the variety of different phonological effects which phonologists might ascribe to the OCP. This is true even if we ignore suprasegmental phenomena (tone, etc.) of which the PLN has no notion. Indeed, even at the segmental level, we might cite the OCP as a motivator for epenthesis, deletion, gemination, metathesis, etc. But whether or not the PLN can exhibit any of these processes is a moot point, since they are defined as the relationship between a surface form and a corresponding underlying form, whereas the PLN has only a single level of representation.

However, this should not be regarded as a fatal flaw in the PLN *per se*, but rather as an indication of how the PLN should be expected to interact with the other components of a complete linguistic system. Speculatively, if the representations in the PLN were interpreted as surface phonological representations, then the underlying representations should correspond to the lexical representations which trigger a given latching string. In this way, input-output mappings in the phonology could be understood as the interaction between the lexical input and the PLN itself.

Again, the PLN does not have a lexical-memory component, so exactly how the activation of a lexical item triggers a latching string is not yet modelled explicitly. But the possibilities here are clearly bounded. For example, the PLN simulations are conducted by "giving" the network a single, incomplete pattern. The exact properties of this initial pattern are what determine the trajectory of the subsequent string. Moreover, it has already been established that small differences in the initial pattern can produce large differences much later in the string—an effect loosely analogous to a butterfly's flapping wing causing a hurricane on the other side of the world. For example, consider these three strings, taken from the same grammar in the PLN corpus:

(1) a. ʔ m u o i ʃ n m
b. ʔ m u o i s n m
c. ʔ m u o ɑ ʃ n m

Each string begins with an incomplete version of the same phone, /ʔ/, and the strings follow the same trajectory for the subsequent 3 latches, before diverging at the 4th and 5th latches, and then returning to the same trajectory for the final two latches. Note that the cause for the differences in each string lies solely in the subtle differences in the initial state for each case, which are invisible when the system is viewed from the macro-level (recall: memory retrieval is understood as passing through an attractor basin, not arriving at an exact point).

This presents an obvious hypothesis that lexical items could trigger a given string simply by sending a short, initial cue to the phonological system. If we suppose that one such cue is sent every time, for example, the syntax/morphology picks a new morpheme, then the cues sent to the phonology would correspond to word/morpheme boundaries, and phonological processes could be understood as the latching network resolving the mismatch between the input from syntax/morphology and its own internal bias for preferred latching targets.

To give an explicit example, suppose we have a network which has latched into an /ʃ/, and then receives a new initial cue in the form of a /z/, as in the case of an English plural like bu/ʃ-z/. If, in the given language, the representation for these two phones are too similar, then directly latching into the /z/ will be impossible. Therefore, the network could react in a number of ways. For example, additional excitation might lengthen the duration of the current retrieved memory (gemination), the network might latch to a similar but sufficiently different memory (dissimilation), it might latch to an intermediate memory before latching to the /s/ (epenthesis), or might fail to latch to the /s/ entirely (deletion). Exactly which strategy the network adopts will depend on the exact nature of the input received from the lexicon. Thus, the phonological grammar for a given language would be localized both within the PLN, and the connections to the lexicon themselves.[9]

Whether or not this model is workable in practice is a topic for future research, since it presupposes a model of lexical storage and retrieval. Currently, there exists no method for exactly "controlling" the strings produced by a latching network. In part, this is because the number of possible initial states for the network is unfathomably large, 8^{200} in the case of the PLN (which is a number 180 digits-long if expressed in regular notation). However, while it is quite conceivable that the majority of those possible initial states do nothing interesting, it need only be true that a tiny subset of them produce unique strings in order for the PLN to be able to produce a vocabulary of lexical items which is comparable in size to that of a typical adult speaker (i.e. in the order of 10s of thousands).

Finally, it should be noted that the method for producing representations, outlined in Section 3.3.1 is somewhat volatile, insofar as it frequently produces grammars with obvious pathologies (failing to retrieve phones, mixed-state

[9] Conceptually, this is strongly analogous to the Optimality Theoretic concepts of markedness (PLN representation) and faithfulness (connection to lexicon).

retrievals, etc.). The solution pursued here was to produce large numbers of grammars and filter out the pathological cases before conducting the analysis. However, in addition to being time-consuming, this method does not allow for a detailed analysis of exactly which variables distinguish the pathological cases from the phonology-like cases. A preferred approach would be the development of a memory-generating algorithm which allows for a more exact control over the variables that differentiate the possible configurations of the network. Such an algorithm has been developed in the context of semantic memories (Boboeva et al. 2018) but has not yet been generalised to a phonology-like case. Of course, semantic memories are fundamentally different to phonological memories insofar as the semantic system is much larger and depends on radically different associations between those memoires. However, it is quite conceivable that the method employed by Boboeva et al might be modified for a smaller phonology-like system. This remains a plausible topic for future research.

6. Conclusion

At the start of this paper I claimed that the PLN can be understood as a *Linking Hypothesis* which bridges the ontological incommensurability between neuroscience and phonological theory. It does not do so by decomposing specific linguistic models into simpler computational mechanisms, but rather by demonstrating how to produce strings which exhibit phonology-like behaviour (assimilation, OCP, SSP), using only a small number of brain-like ingredients (recurrent connections, distributed representations, short-term adaptation), plus a system of memories defined in terms of phonological features. In this way, the components of the linguistic formalism are understood to be emergent from a complex dynamical system.

The relevance of the results from the model can be understood from two perspectives: that of the neuroscientist and that of the linguist. From the neuroscientist's perspective, it is significant that the phonological behaviours exhibited are not explicitly taught to the network, nor are they pre-programmed in any way. Rather, they seem to emerge spontaneously from the specific combination of phonologically-inspired representations and neurally-inspired network dynamics. This fact supports the plausibility of latching dynamics as a real neural mechanism. This type of indirect evidence is crucial because, although latching dynamics have been studied theoretically in a variety of contexts, measuring them directly is likely beyond current neuroscientific techniques. Of course, the PLN still leaves open a number of questions about the underlying neurological reality. Most notable is the specific neural correlate of the Potts units themselves, which are intended to subsume a large amount of potential complexity into a relatively simple and tractable approximation. However, the Potts units are not totally opaque, and the specific parameters of the model implicitly delimit the range of possible underlying biological mechanisms that we can posit. Further research into the PLN is likely to yield clearer predictions in this regard, because as the parameters of the model become more fine-tuned, so too do the neural predictions. Thus, the PLN presents us with an interesting case

where linguistic facts could be used to deduce relatively fine-grained neural properties.

From the linguist's perspective the implications of the PLN are less direct, since we are discussing across two quite different levels of abstraction, that is linguistics and neuroscience. In general, we should be cautious about drawing direct correlations between the ontologies of neutrally inspired models and formal linguistic theories. However, the PLN could nonetheless inform the discussions and assumptions *surrounding* formal linguistic theories, if not the theories themselves. One example of this is the topic of innateness and learnability which, although not necessarily properties captured within a formal theory, are nonetheless topics of thorough debate by linguists (e.g., Odden 2013).

Indeed, under one reading, Chomsky's articulation of Universal Grammar (UG) could lead one to believe that the primary goal of formal linguistics is precisely to disentangle the innate parts of language from the rest (e.g., Chomsky 2005). Of course, it should also be noted that the PLN itself is not a theory of language acquisition. However, if the PLN is remotely plausible then it suggests that the UG/disentangling project is not something that could be properly expressed at the level of a linguistic theory. That is, the components of linguistic theory are themselves an irreducibly complex mixture of genetic and environmental factors.

For example, if the OCP or SSP are consequences of latching dynamics (as the PLN suggests), then they neither need to be independently learned nor innately specified, since they appear to be largely coextensive with latching dynamics. They could perhaps be equated with Chomsky's *third factor* (Chomsky 2005), however even this categorisation may be too coarse. Because although the OCP and SSP do seem to follow from a purportedly more general mechanism (i.e. latching), it is also true that these behaviours appear to depend on the way the memories themselves are encoded, which seems to be a fact about phonological inventories and the features which define them.

The SSP, for example, is dependent on the particular properties of manner features—namely that they loosely cluster the inventory into two groups along a single dimension: sonorants and obstruents. Given this clustering, latching dynamics seems to naturally produce oscillation between the two clusters. Thus, the SSP is the result of a complex interaction between something specific to phonology (sonority) and something much more general (latching dynamics). Of course, this interaction is not necessarily captured at the level of linguistic formalisms, meaning that the relevant subdivision into innate/learned/third-factor cannot occur at the level of the linguistic theory itself.

This does not necessarily entail that UG is a doomed project, merely that the complex influence of genetic and environmental factors on language acquisition may only be understandable when we integrate insights from linguistic theory into neutrally inspired models such as the PLN (and beyond, into neurobiology, etc.). Thus, properly defining UG may not be a problem that linguists can solve in isolation. This conclusion could render moot long standing discussions about the innateness of (e.g.,) phonological features (e.g., Mielke 2008), since features might not be atomic objects which can be neatly described as either innate or learned.

Of course, this brief discussion of learning is by no means exhaustive. It is intended merely to demonstrate how intermediate, neurally-inspired models such as the PLN can help to bridge the gap between linguistics and neuroscience in a way that permits more nuanced argumentation, rather than causing "inter-disciplinary cross-sterilization" (Poeppel & Embick 2005). The ultimate goal is integration of linguistic and neuroscientific theories into a grander understanding of the mind/brain and, while this goal is certainly a long way off, models such as the PLN do present us with a potential way forward.

Data Availability

The data and code used for analysis in this article are available by contacting the author.

References

Alderete, John, & Tupper, Paul. 2018. Connectionist approaches to generative phonology. In S. J. Hannahs & Anna R. K. Bosch (eds.), *The Routledge Handbook of Phonological Theory*. Routledge, New York, 360–390.

Assaneo, Maria Florencia, & Poeppel, David. 2018. The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advances* 4(2): eaao3842.

Block, Ned. 1995. The mind as the software of the brain. In E. E. Smith & D. N. Osherson (eds.), *An Invitation to Cognitive Science*, vol. 3: *Thinking* (2nd edn.). Cambridge, MA: MIT Press.

Boboeva, Vizhe, Brasselet, Romain, & Treves, Alessandro. 2018. The capacity for correlated semantic memories in the cortex. *Entropy* 20(11), 824.

Bouchard, Kristopher E., Mesgarani, Nima, Johnson, Keith, & Chang, Edward F. 2013. Functional organization of human sensorimotor cortex for speech articulation. *Nature* 495, 327–332.

Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36(1), 1–22.

Clements, George N. 1990. The role of the sonority cycle in core syllabification. In J. Kingston & M. E. Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge, United Kingdom: Cambridge University Press, 283–333.

Hickok, Greg, & Poeppel, David. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience* 8, 393–402.

Hopfield, John J. 1982. Neural networks and physical systems with emergent collective computational properties. *Proceedings of the National Academy of Sciences of the United States of America* 79, 2554–2558.

Hopfield, John J. 1984. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America* 81, 3088–3092.

Kang, Chol J., Naim, Michelangelo, Boboeva, Vizhe, & Treves, Alessandro. 2017. Life on the edge: Latching dynamics in a Potts neural network. *Entropy* 19, 468.

Kanter, Ido. 1988. Potts-glass models of neural networks. *Physical Review A* 37(7), 2739–2742.

Kaplan, Abby. 2010. *Phonology shaped by phonetics: The case of intervocalic lenition.* Santa Cruz, CA: University of California dissertation. doi:10.7282/T30G3J2K

Krämer, Martin. 2019. Is vowel nasalisation phonological in English? A systematic review. *English Language & Linguistics* 23(2), 405–437.

Lakoff, George. 1988. A suggestion for a linguistics with connectionist foundations. In D. Touretzky (ed.), *Proceedings of the 1988 Connectionist Summer School.* Burlington, MA: Morgan Kaufmann.

Lashley, Karl S. 1951. *The Problem of Serial Order in Behavior*. Oxford, United Kindom: Bobbs-Merrill.

Marr, David & Poggio, Tomaso. 1976. From understanding computation to understanding neural circuitry. *Artificial Intelligence Laboratory. A.I. Memo*. MIT. Retrieved from https://dspace.mit.edu/handle/1721.1/5782

McCarthy, John J. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17, 207–263.

Mesgarani, Nima, Cheung, Connie, Johnson, Keith, & Chang, Edward F. 2014. Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010.

Mielke, Jeff. 2008. *The Emergence of Distinctive Features*. Oxford, United Kingdom: Oxford University Press.

Naim, Michelangelo, Boboeva, Vizhe, Kang, Chol J., & Treves, Alessandro. 2018. Reducing a cortical network to a Potts model yields storage capacity estimates. *Journal of Statistical Mechanics: Theory and Experiment* 4, 043304.

Nasrabadi, Nasser M., Choo, C. Y. 1992. Hopfield network for stereo vision correspondence. *IEEE Transactions on Neural Networks* 3, 5–13.

Ohala, John J. 1990. The phonetics and phonology of aspects of assimilation. *Papers in Laboratory Phonology* 1, 258–275.

Odden, David. 2013. Formal phonology. *Nordlyd* 40(1), 249–273.

Pirmoradian, Sahar, & Treves, Alessandro. 2012. A talkative Potts attractor neural network welcomes BLISS words. *BMC Neuroscience* 13(Suppl. 1): P21.

Pirmoradian, Sahar, & Treves, Alessandro. 2014. Encoding words into a Potts attractor network. In Julien Mayor & Pablo Gomez (eds.), *Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop*. Singapore: World Scientific, 29–42.

Poeppel, David & Embick, David. 2005. Defining the relationship between linguistics and neuroscience. In A. Cutler (ed.), *Twenty-First Century Psycholinguistics: Four Cornerstones*. Mahwah, NJ: Lawrence Erlbaum, 103–118.

Prince, Alan & Smolensky, Paul. 1997. Optimality: From neural networks to universal grammar. *Science* 275(5306), 1604–1610.

Russo, Eleonora, & Treves, Alessandro. 2012. Cortical free-association dynamics: Distinct phases of a latching network. *Phys. Rev. E*. 85(5), 051920.

Song, Sanming, Yao, Hongxun, & Treves, Alessandro. 2014. A modular latching chain. *Cogn Neurodyn* 8, 37–46.
Treves, Alessandro. 2005. Frontal latching networks: A possible basis for infinite recursion. *Cognitive Neuropsychology* 22(3–4), 276–291.
Truesdell, Clifford, & Muncaster, Robert G. 1980. *Fundamentals of Maxwell's Kinetic Theory of a Simple Monatomic Gas.* New York: Harcourt Brace Jovanovich.
Tsodyks, Misha V., & Feigelman, Mikhail V. 1988. The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters* 6(2), 101–105.
Walter, Mary A. 2007. Repetition avoidance in human language. Cambridge, MA: MIT doctoral dissertation.

Appendix

This appendix contains the parameters and phonological inventory used in this study.

The results in Section 4 were all obtained from simulations using a constant set of network parameters:

$S = 5$
$N = 200$
$a_{feat} = 0.2$
$a = 0.25$
$p = 1.1$
$q = 0.1$
$\tau_1 = 1.5$
$\tau_2 = 70$
$\tau_3 = 100$
$\beta = 4$
$w = 1.8$
$U = 0.45$
$c_{int} = 0.2$
$c_{ext} = 0.2$

The inventory of phones and their featural specification is given in the table on the next page (*Table Appendix 1*). Note that the ordering of the features in the table reflects the weighting of the features within each sub-network.

Manner ➔	p	f	m	t	s	ʃ	tʃ	θ	l	n	k	ŋ	ʔ	h	w	r	i	i	e	a	æ	o	u
consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+									
vocalic																		+	+	+	+	+	+
sonorant			+						+	+		+			+	+	+	+	+	+	+	+	
nasal			+							+		+											
continuant		+	+		+	+		+	+	+		+		+	+	+		+	+	+	+	+	+
approx.															+	+	+						
del.rel.							+																
lateral									+														
sibilant					+	+																	
dental								+															
apical				+																			
retroflex																+							
post-alv							+																
alveolar				+	+			+	+	+													
anterior		+			+												+						
glottal													+	+									
velar											+	+											
coronal				+	+	+	+	+	+	+						+							
dorsal											+	+			+		+	+				+	+
labial	+	+	+												+							+	+
exterior	+	+	+				+				+	+			+							+	+
front						+	+											+	+		+		
low													+	+		+				+	+		
high																	+	+					+
Place ➔																							

Appendix Table 1: Inventory of phones and their featural specification.