UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

HSL-ISK

Neural Attractors and Phonological Grammar

What the sound patterns of language can tell us about the brain

—

Joe Collins

A dissertation for the degree of Philosophiae Doctor – June 2019

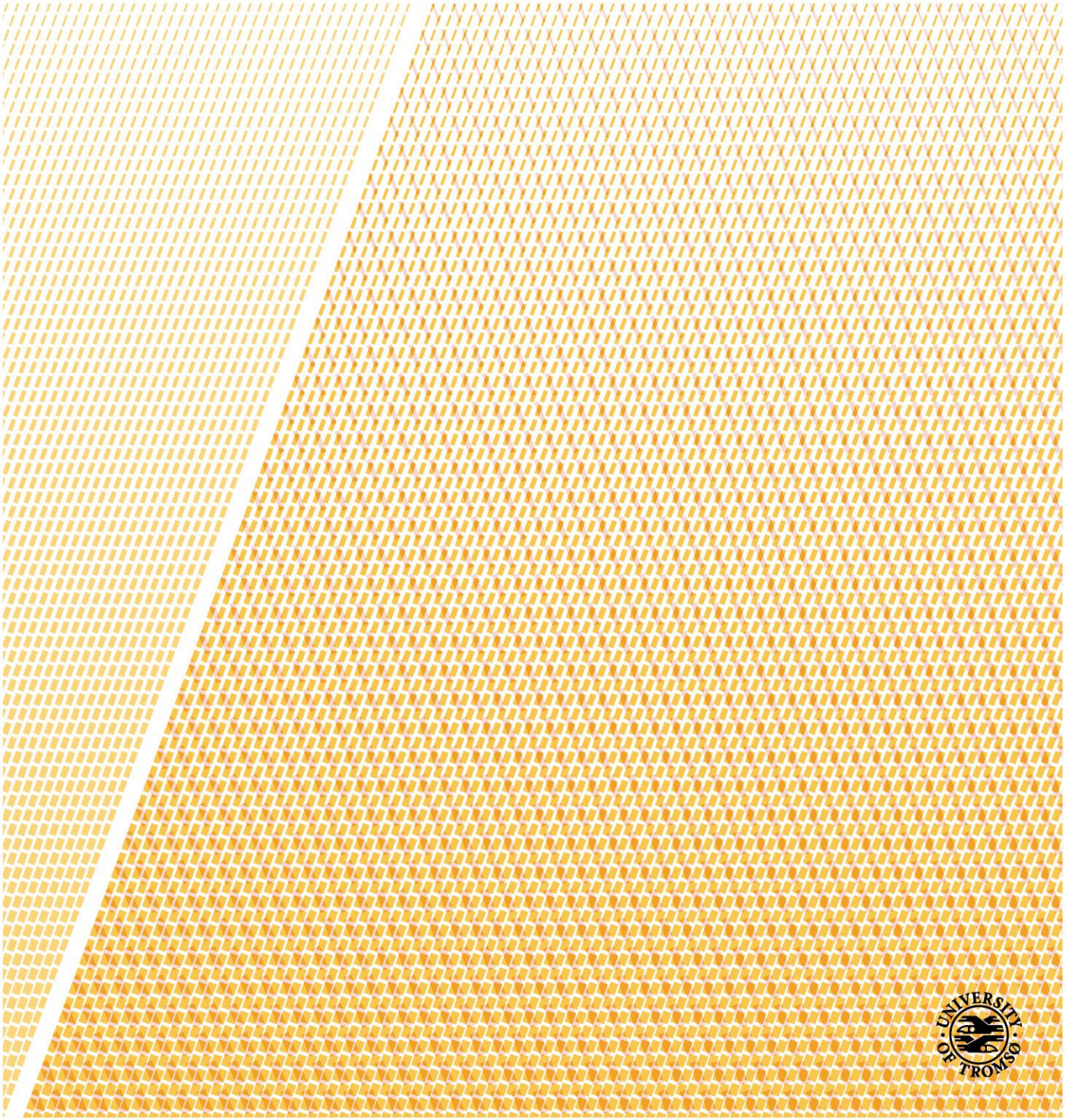

Table of Contents

List of Tables

List of Figures

Foreword

My thanks go to all the people without whom this thesis could not have happened, including:

My supervisor Martin. His intellectual honesty, curiosity, and passion for all things phonological have likely infected me more than he can realise. This thesis isn't quite what either of us imagined when I started, but I haven't given up on calling myself a phonologist just yet.

Everyone at CASTL in Tromsø. I couldn't have hoped for a better group of genuinely interested and interesting colleagues, especially the FISH group under the curatorship of Gillian and Peter. There aren't many linguistics groups that would permit and encourage a thesis like this, and I was lucky enough to end up with the best of them.

My co-supervisor Alessandro for helping me understand the importance of the quantitative , as well as everyone at the LIMBO research group.

My wife for following me all the way to the frozen north, then leading us all back down again.

My parents for all their support and help.

Svigerfamilien for barnepass og kattepass (og ikke minst konepass).

Jacques Koreman at NTNU for kindly lending us his lab.

The countless people who have challenged me with comments and questions at conferences and workshops.

And to everyone else I have forgotten. The list of people to whom I owe my thanks is long enough to fill several chapters.

Kieran: Giggleloop.

1 Introductory Chapter

This volume collects three articles which constitute the bulk of my PhD research. The overarching theme of the volume is the role of attractors - a concept from dynamical systems theory – in the neural realization of phonological grammar.

The motivation for this line of inquiry begins with the claim that the study of language should provide some insight into the workings of the human mind/brain. Indeed this is one of few mantras shared by linguists of the seemingly irreconcilable "Generative" 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.).

1.1.1 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 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 introductory chapter. 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. 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 (e.g.) 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

---

[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.

[2] Conceptually at least, this analogy is not a novel idea in phonology. The same basic assumptions underlie Smolensky's Integrated Connectionist/Symbolic architecture and, by extension, Harmony theory and Optimality Theory (Prince and Smolensky 1997).

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.

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

1.1.2 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. This is where attractor dynamics are critical, because they 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.

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.

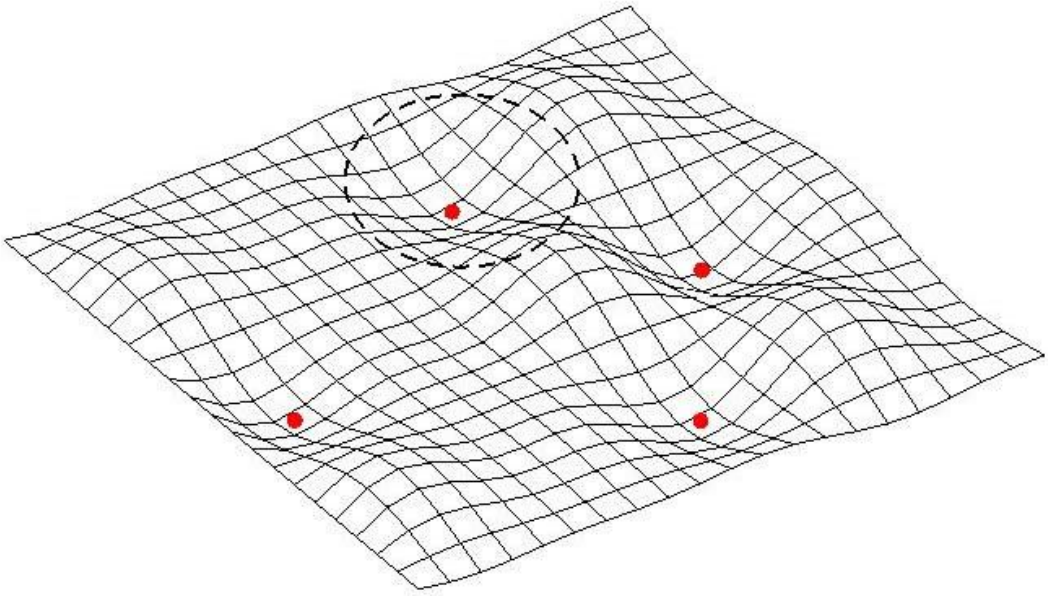

Figure 1: Conception of a network state-space. The z-axis corresponds to the free energy of the network. The red dots are attractors. http://www.scholarpedia.org/article/Attractor_network

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.

1.1.3 Overview of Introductory Chapter

The rest of this introductory chapter is split into two parts: first, a brief summary of each of the three articles in this volume; and secondly, a collection of smaller comments and technical discussions which are of a more general and speculative nature than the articles themselves. These are intended to provide some theoretical background for the articles, as well identifying certain deeper issues for further discussion.

1.2 Summary of the Articles

1.2.1 The Phonological Latching Network

The first paper could be considered the primary contribution of this volume, and it represents by far the largest time commitment of the three articles. It contains an analysis of a model dubbed the Phonological Latching Network (PLN), which is an extension of earlier Potts latching networks. The key claim is that the model appears to reproduce certain quintessentially phonological phenomena, despite not having any of these phonological behaviours programmed or taught into the model. Rather, they 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 natural language patterns can be taken as evidence of the model's plausibility; and secondly, it provides a potential explanation for why these patters appear to frequently in natural language grammars.

The PLN consists of a number of so-called "Potts" units, intended as effective models for small patches of cortex, which are linked via symmetrical, synapse-like connections of varying efficacy. The model belongs to a broader class of neural networks called attractor networks, which are noteworthy for their ability to store quasi-discrete memories as stable, distributed patterns of activity. The PLN is also capable of spontaneously producing strings of discrete elements as it "latches" between the memories stored in the network. The latching behavior is not prescribed by the experimenter, but rather emerges naturally under very specific configurations, due to the fatigue of active units in the network. Previous numerical analyses of latching behavior have shown that the probability of a latch between any two memories depends on the similarity of those memories' representations (broadly: how many units their representations share; see paper for details). In linguistic terms, this notion of similarity can be thought of as shared features. Therefore, latching behavior is one of few explicit hypotheses for how an analogue system, such as the brain, can produce more complex structures of discrete elements, of the sort posited by linguists.

The PLN represents an inventory of phones as distributed patterns of activity, which are split across "motor" and "auditory" subnetworks. Each phone is created algorithmically by superimposing the representations for a given number of phonological features, each of which is defined by a lowly correlated noise pattern. The representations for the phones are then encoded as synaptic efficacies in the network, using a Hebb-rule. Electrophysiological data on

the encoding of speech information in the Superior Temporal Gyrus and premotor areas shows a spacial asymmetry in encoding of place and manner features. Therefore, in the PLN, the features are weighted such that place features are more active in the "motor" sub-network, while manner features are more active in the "auditory" network. For the sake of simplicity, laryngeal features are excluded from the PLN. This is partly because laryngeal processes can often be treated as orthogonal to place and manner, but also because the current electrophysiological data give no clear insight into how laryngeal features should be incorporated into the model.

As the network latches, it produces phonological words of varying length (e.g. Figure 2). By repeating the simulation with fixed variables, but randomly determined initial states, the PLN produces a corpus of data which can be taken to represent a single grammar. Each grammar can then be described using similar tools to those used to describe natural grammar. For the purpose of this study, each transition (or latch) produced by the PLN was characterized using phonological criteria (e.g. "do these two adjacent segments share a place feature?" etc.). These characterizations are then tallied, and then compared to chance level, i.e., a grammar in which the probability that any given segment will occur is equal for all segments, which in turn can be used to calculate the chance occurrence of given phonological feature. The extent to which the PLN grammars diverge from chance level can be taken as an indication of which properties (if any) emerge naturally from the implementation of phones (as defined by phonological features) in a latching network.

The latching network was found to exhibit three types of "phonology-like" behavior. Firstly, the latching strings tend to obey the Sonority Sequencing Principle, which in turn leads to more typologically common syllables (e.g. CV, CVC, etc.). Secondly, the network is near-incapable of immediately repeating a segment, which in turn means that the network obeys the Obligatory Contour Principle (at least at the surface/segmental level – generalization to underlying and/or suprasegmental OCP remains a topic for future investigation). Thirdly, when compared to chance levels, adjacent segments exhibit a preference for place agreement.

These results are striking insofar as the apparent naturalness of the strings produced by the PLN do not depend on stipulating any of these properties *a priori*. Rather they emerge spontaneously from the combination of a neurologically motivated model, with phonologically motivated representations. For this reason, the PLN presents not only a plausible hypothesis for *why* certain properties form a part of the phonological faculty, but also a first step towards understanding their neurological implementation in greater detail. More generally, the model

demonstrates the application of dynamical systems modelling as a way of relating formal linguistics to specific mechanisms for neural computation.

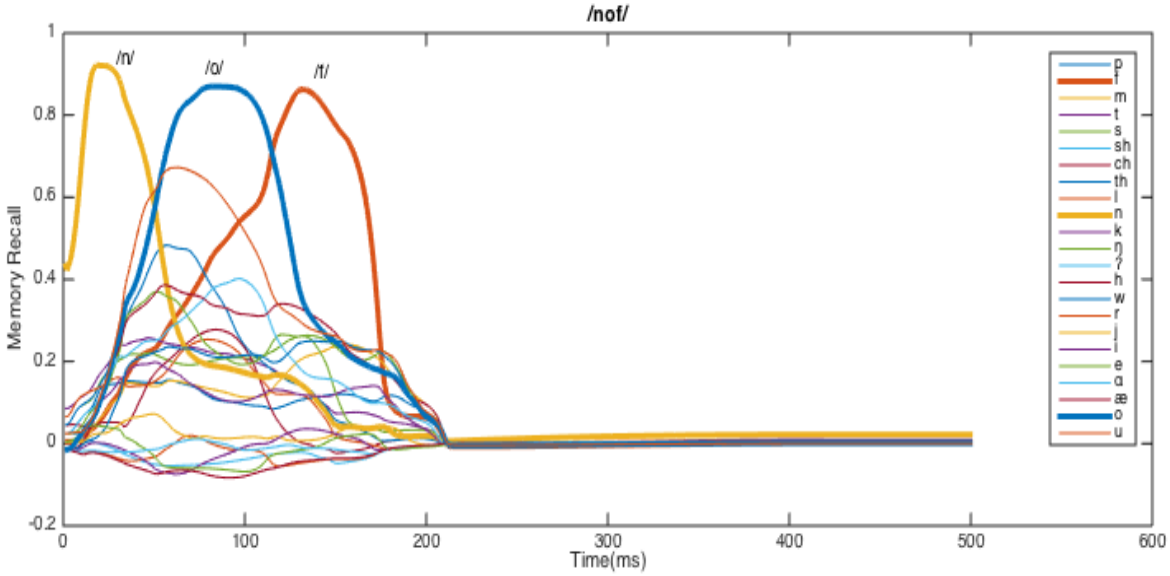

Figure 2: Example of a latching string. The PLN produces /nof/.

1.2.2 Digital Grammar and Analogue Brains

The second paper also features an attractor neural network, albeit a much simpler type than the PLN. The focus of this paper itself is far more conceptual in nature. The contribution is not so much a particular result, but rather an attempt to understand how formal theories of grammar should be understood in relation to "neural" models of cognition. The primary focus of the paper is the apparent incommensurability of digital formalisms with the view of the brain as an essentially analogue machine. Of course, this is not a new topic and many different stances on this issue can be gleaned from the philosophy of mind literature. Rather the rehashing the philosophy however, this paper applies an information theoretic method, *Effective Information* (EI), to an explicit "toy" phonological grammar, and an attractor neural network realization of that same grammar. EI is defined as the mutual information between the interventions on a system, and the effects of those interventions. In this way, EI provides a measure of the causal information conveyed by a scientific model.

The attractor network demonstrates the emergence of discrete categories from an underlyingly gradient system. But it can also be proven that the formal phonological analysis has a higher *Effective Information* (EI) than the neural attractor model. I argue that this shows that discrete formalisms compatible with a gradient view of the brain, but also that they are *causally*

emergent (Hoel 2017), and therefore necessary if we wish to have a complete explanation of natural grammar.

The model itself focuses on the phenomenon of incomplete devoicing, which has been argued to be an example of phonetic gradience that discrete phonological models cannot explain (c.f. van Oostendorp 2008). Therefore, the toy phonological grammar consists of 6 possible phones – 3 places of articulation ([LABIAL], [CORONAL], [DORSAL]), each with a voiced and voiceless variant – and the capacity to distinguish coda and non-coda positions, as well as simple rule which devoices any voiced phone in a coda position. For the attractor network, the 6 phones are encoded as attractor states in the network, while information about syllable structure is supplied to the network as a simple inhibitory signal, which is used to signal a coda-position. Analysis of the network behavior shows that, when the network is told to retrieve a voiced phone in the presence of the inhibitory coda signal, the network spontaneously retrieves the voiceless counterpart. In this way, the model is implementing the devoicing rule of the formal model.

I_D *at time=t*	*t+1*	E_D
$<do(b\#)>=\frac{1}{12}$	*[p]#*	*<b#>=0*
$<do(d\#)>=\frac{1}{12}$	*[t]#*	*<d#>=0*
$<do(g\#)>=\frac{1}{12}$	*[k]#*	*<g#>=0*
$<do(p\#)>=\frac{1}{12}$	*[p]#*	$<p\#>=\frac{2}{12}$
$<do(t\#)>=\frac{1}{12}$	*[t]#*	$<t\#>=\frac{2}{12}$
$<do(k\#)>=\frac{1}{12}$	*[k]#*	$<k\#>=\frac{2}{12}$
$<do(b)>=\frac{1}{12}$	*[b]*	$<b>=\frac{1}{12}$
$<do(d)>=\frac{1}{12}$	*[d]*	$<d>=\frac{1}{12}$
$<do(g)>=\frac{1}{12}$	*[g]*	$<g>=\frac{1}{12}$
$<do(p)>=\frac{1}{12}$	*[p]*	$<p>=\frac{1}{12}$
$<do(t)>=\frac{1}{12}$	*[t]*	$<t>=\frac{1}{12}$
$<do(k)>=\frac{1}{12}$	*[k]*	$<k>=\frac{1}{12}$

Interestingly, however, the voiceless outputs which are derived from a voiced input can vary fractionally from those voiceless outputs which are underlyingly voiceless. This small variation is could be easily interpretable as a small, but consistent difference in the voicing of the phone during realization. In this way, this simple model is a proof of concept for how a discrete phonological system, when implemented in an underlyingly continuous system, can exhibit the sorts of gradience observed in phenomena such as incomplete devoicing.

In order to compare the EI of the formal and attractor model we must understand both as kind of dynamics over a state space. The toy grammar can be understood as a system having $n=12$ possible states S={[b]#, [d]#, [g]#, [b], [d], [g], [p]#, [t]#, [k]#, [p], [t], [k]}. The dynamics of the system can be understood as an intervention over each state s_i, at time=t, and a resulting effect at time=$t+1$. With

the formal system defined, we can then determine two probability distributions, *Intervention Distribution (I_D)* and *Effect Distribution (E_D)*, which can then be used to calculate the *effectiveness* of the system. This is slightly simpler than calculating the EI directly, but it stills allows to determine the relative EI of the formal and attractor models. The I_D is considered in the maximum entropy case, where $I_D(i)=n^{-1}$. and the E_D is calculated by observing the effects of the interventions at time=$t+1$ (see table above). These values can then be used to determine the *degeneracy* of the system:

$$degeneracy = \frac{D_{KL}(E_D|I_D)}{log_2(n)} = log_n(2)\sum_i E_D(i) log_2 \frac{E_D(i)}{I_D(i)}$$

This will then allow us to calculate the $effectiveness = [determinism] - degeneracy$. Since out toy grammar is strictly deterministic, the *determinism* is equal to 1. Crunching the numbers gives our toy grammar $eff = \sim 0.93$.

We then repeat this process to determine the *effectiveness* for the attractor model. This is slightly more complicated because the state space is both continuous and intractably large. However, by using a simple approximation method (see paper), we can determine that $eff = \sim 0.174$ for the attractor model.

These two values can be used to determine the relative EI, because it can be proven that a system is only *causally emergent* when the gain in information from increased EI outweighs the loss in information from the smaller state space at the coarser, or more "abstract" level of analysis. Given that the size of the state space is known for the both the toy formal model and the attractor network, it is easy to prove that the formal model must have a higher EI than the attractor network (see paper).

Therefore, even when our discrete phonological representations are taken as emergent phenomena from an underlyingly gradient system, such as an attractor network, it is in fact the phonological model which has the highest *EI*, rather than the neurological model. Thus, the formal analysis of the grammar carries more information about the underlying *causal structure* of the system. This is argued to be the utility of formal linguistics within cognitive science more broadly.

1.2.3 On the Language Specificity of Vowel Maps

The third article focuses on attractor dynamics in the domain of speech perception. Specifically, the way a continuous acoustic space, such as the vowel space, can be perceived by speakers as

being composed of quasi-discrete objects, i.e. the vowel inventory. The paper gives the results from three different vowel perception experiments, carried out with the help of collaborators in several different countries. By comparing the results from participants with different L1s, we can see the way the perception of the vowel space depends on the participants native vowel inventories. Finally, a visualization method, developed by collaborator Zeynep Kaya allows us to generate a deformed map of the vowel space for each language tested.

For our first experiment we tested speakers of Italian, Turkish, Spanish and Scottish English on their ability to discriminate ambiguous pairs of vowels. The experiment is designed around a confusability paradigm, whereby participants are played pairs of CV-syllables and asked to press a key if they believe the two vowels to be the same. The stimuli were generated first by recording a phonetically trained speaker, then using a morphing algorithm to generate new CV-syllables with intermediate vowel qualities. This way, we could produce groups of four CV-syllables whose vowel qualities are approximately evenly distributed along a small continuum within the vowel space. The perception results show definite, albeit small, differences between the language groups.

The second experiment tested speakers of Italian, Norwegian and Turkish. For this experiment we extended the paradigm of the first experiment by generating new, intermediate stimuli. This allowed us to test participants perception over approximately the whole vowel space. In this case the result present a much clearer picture of the differences between the language groups. Moreover, we were able to use participants responses to generate deformed "maps" of the vowel space for each language. While this visualization method does result in some information loss, it nonetheless captures some important differences in vowel perception between the language groups.

Finally, we conducted a variation of the second experiment using only (late-)bilingual Norwegian speakers of English. The paradigm remains the same as before, with the addition of language priming sessions for the participants. These were interspersed during the vowel discrimination test, in the form of aural short stories in either English or Norwegian. The results do not show any evidence that the priming affected participants vowel perception. This supports the hypothesis that L2 learners merge the vowels of the new language onto their existing "vowel map", rather than developing a new map. These results also present an explanation for why the Norwegians exhibited better discrimination over English-like (but non-Norwegian) vowels in

the second experiment: their higher exposure to English compared to the other groups has left them with a vowel map which merges both English and Norwegian vowels.

The subdivision of labour among the three co-authors is approximately as follows:

Zeynep Kaya: Experimental design, coding experiment program, Turkish/Italian data collection, applying morphing algorithm.

Joe Collins: Producing stimuli, Norwegian data collection, coding statistical analyses, writing up and analysis from a phonological perspective.

Alessandro Treves: Supervision over all aspects, especially during experimental design and writing phases.

With additional data collection by Simona Perrona.

1.3 Background, Tangents and Outstanding Issues

This final portion of the introductory chapter collects a number of smaller technical discussions which relate to issues surrounding the articles, but which I have chosen to edit out of the articles themselves. How tangential these topics seem will depend largely on the reader's own technical background. However, they are included here in the hope that they may provide some context for various (potentially contentious) assumptions which motivated the research in this volume.

1.3.1 Linguistics and Neural Networks

This volume is far from the first attempt to fuse insights from Artificial Neural Networks (ANNs) with formal linguistic theory, as the subject has been broached many times before (see Alderete & Tupper 2018). Indeed, the entire formalism of Optimality Theory was largely an attempt to resolve the tensions between the assumptions of ANNs and the symbolic models of formal linguistic (Prince & Smolensky 1997). Nonetheless, there is still an implicit assumption among some that generative models and ANNs are fundamentally competitors (c.f. Pater 2019). The roots of this belief arguably stem from a perception that ANNs and generative grammars belong to different schools in the "theory of mind" debate. ANNs are often thought to be synonymous with "connectionism", while generative grammars are regarded as a form of (classical) computationalism.

There are a variety of reasons to think that this dichotomy is both unhelpful and misleading. Firstly, it is something of an oversimplification to equate all ANNs with connectionism. The

models in this volume are not really connectionist models *per se*, the reasons for which I discuss in section 1.3.2. However, even if we restrict the discussion to connectionist ANNs, the distinction between connectionism and classical computationalism is considerably murkier than some might suppose. Consider, for example, that any sufficiently general definition of "neural network" will end up including digital computer architectures by extension. This is true if only because, for a great many ANNs, the individual units are capable of functioning as Boolean operators. In the case where all the units of an ANN are Boolean operators, the ANN is not merely *simulating* a digital computer, it *is* a digital computer under any reasonable definition. The implication then, is that digital computers are actually a very specific subset of neural network architectures (see Piccinini 2015:ch13 for a more complete account of this argument).

This conclusion might strike us as radical, but in reality it is trivial and fairly uninteresting. A digital computer constructed using modern machine-learning methods would be both deeply implausible as a neural model and fairly useless for machine learning (doubly so given than modern ANNs are usually simulated using digital computers). Thus, in the modern context, the distinction between ANNs and digital computers is more a question of appropriate application, rather than any well-defined difference in the architectures themselves.

But if this is true, why are ANNs and computationalism so often regarded as competitors? The answer I will advance here, is largely sociological. Historically, there does not seem to be much evidence for a strong divide between ANNs and classical computationalism until the connectionist wave of the 1980s, which brought with it a set of long-enduring arguments about the relative merits and failings of ANNs and digital computers. These arguments also spilled over into the realm of linguistics, and were to some extent mirrored by the cognitivist/generativist split at the same time.

With that in mind, what follows then is a terse and (at times) speculative history of ANNs, as it pertains to the connectionist/computationalist divide. I argue that this provides some much needed context and demonstrates just how recent and arbitrary this divide really is.

1.3.1.1 A Terse History of ANNs I: The early days

The early days of computing saw pioneers pursue a multitude of hypothetical computing machines (see e.g. von Neumann 1951). During this era, work on neural networks and classical computers emerged not only at the same time, but largely by the work of the same people. As early as 1948, the father of computing himself, Alan Turing, submitted a technical report on so-called "unorganized machines", which were intended as simplified model of the nervous

system. To the modern eye, these machines are unambiguously a form of neural network, and are arguably no less prescient than Turing's more widely-lauded work on symbolic architectures (Copeland & Proudfoot 1996).

Of course, we've no way of knowing how Turing's version of the neural network would have progressed had his life not been cut short. However, similar ideas would be pursued by others. These include John von Neumann, who is perhaps most famous for creating the architecture for program-loading digital computers which became the standard for all computers as we know them today. Despite this herculean contribution to digital computer design, von Neumann was also deeply concerned with the problems of probabilistic computation in distributed architectures (von Neumann 1956). Indeed, von Neumann expressed a concern that would become critical for the connectionists of the 1980s:

> *"[N]atural organisms are constructed to make errors as inconspicuous, as harmless, as possible. Artificial automata are designed to make errors as conspicuous, as disastrous, as possible. Natural organisms are sufficiently well conceived to be able to operate even when malfunctions have set in. They can operate in spite of malfunctions, and their subsequent tendency is to remove these malfunctions*." (1951:432)

This type of observation would later become a cornerstone argument levied as evidence of the biological plausibility of connectionist models. Namely, that connectionist models exhibit "graceful degradation" (e.g. Rumelhart 1998). However, while von Neumann pre-empted some of the limitations of purely symbolic/logical methods, he did not appear to advocate abandoning them so much as extending them to a "*general and logical theory of automata*" (1951:430). There is some irony then, that recent attempts to delineate neural networks from classical computers fall back on the terms "Turning machine" and "von Neumann architecture" (Fodor & Pylyshyn 1988; Gallistel & King 2009), given that their namesakes were pioneers of both fields, and apparently perceived no great conflict between the two areas of research.

Still, Turing and von Neumann's early work on ANNs is arguably more of a historical curiosity, insofar as it appears to have had limited impact on later ANN developments[3]. Indeed, Turing's proto-connectionist proposal seems to have been something of a secret until the mid-1990s

[3] At least in comparison to the impact their work on digital computers had.

(Copeland & Proudfoot 1996). The same could not be said, however, of Warren McCulloch and Walter Pitts seminal 1943 paper, "A logical calculus of the ideas immanent in nervous activity", which is widely regarded as a foundational paper for neural network research. It is frequently cited, in large part, because it contains a tractable mathematical approximation of single neurons. This model would, in time, be generalized by others (e.g. Rosenblatt 1958) and allow for the creation of ANN simulations of the sort we would recognise today. Interestingly however, McCulloch & Pitts themselves appeared to have a slightly different focus from modern connectionist research. What it is perhaps most striking about McCulloch and Pitts (1943) from a modern perspective, is that they are explicitly concerned with Turing's notion of computability. They themselves regarded their conclusions as being…

> "*of interest as affording a psychological justification of the Turing definition of computability and its equivalents, Church's A-definability and Kleene's primitive recursiveness: if any number can be computed by an organism, it is computable by these definitions, and conversely.*" (1943:113)

Far from attempting to instigate an alternative to Turing's work, McCulloch and Pitts were trying to demonstrate its relevance for the study of cognition. Moreover, as noted by Piccinini (2004), a formalism introduced by McCulloch and Pitts was an important step towards the concept of finite automata – a fundamental concept in computer science - suggesting that McCulloch and Pitts' contribution may be as significant for classical computation as for neural networks.

Of course, the work of McCulloch and Pitts would ultimately pave the way for many others interested in artificially imitating the architecture of the brain. This includes not only Turing and von Neumann (Piccinini 2004), but also theoretically important work by Stephen Kleene, who wrote:

> *"Finally, we repeat that we are investigating McCulloch-Pitts nerve nets* [sic] *only partly for their own sake as providing a simplified model of nervous activity, but also as an illustration of the general theory of automata, including robots, computing machines and the like."* Kleene (1956[1951]:3)

This quote encapsulates the divide between the modern view of neural networks and that of the 1940s and 50s. The notion that studying ANNs could provide insights into a general theory of computation sounds quite radical to the modern ear (c.f. Piccinini 2015:ch13). Within cognitive

science at least, it is perhaps more common to interpret ANN research as an attempt to undermine the classical computational theory of mind (see Marcus 1998). However, there is little evidence that the early pioneers of both ANNs and computation generally perceived any such antagonism. Rather, there seems to have been a sense that all types of automata and computing machines belonged to some larger, common class of systems.

1.3.1.2 A Terse History of ANNs II: The birth (and death) of connectionism

The earliest pioneers of ANNs showed little sense that these models were at odds with the programmable machines that would precede modern digital computers. So, when does this divide begin to emerge? Perhaps the first serious attempt to delineate neural models from purely logical or symbolic architectures comes from Frank Rosenblatt, who could rightly be called the father of connectionism (not least because Rosenblatt seems to have inadvertently given the term its modern meaning in his 1958 paper). However, as we shall see, there are certain key aspects in which even Rosenblatt's views do not fully approach the modern discord between connectionism and computationalism.

Rosenblatt's own model, the Perceptron, differed from earlier ANNs in that the connections between units had an efficacy (or weight) which was represented by a continuous variable. By using a learning algorithm to determine the weights between units, the Perceptron could be taught to classify input data into different categories. For these reasons, the perceptron is often regarded as the first connectionist network.

Rosenblatt himself was clear about presenting the perceptron as a departure from the types of models that preceded it. When discussing the (then) start-of-the-art, he writes:

> "*During the last few decades, the development of symbolic logic, digital computers, and switching theory has impressed many theorists with the functional similarity between a neuron and the simple on-off units of which computers are constructed, and has provided the analytical methods necessary for representing highly complex logical functions in terms of such elements. The result has been a profusion of brain models which amount simply to logical contrivances for performing particular algorithms [...]"*(1958:387)

Rosenblatt is discussing earlier ANNs (e.g. McCulloch & Pitts 1943), however his description is clearly applicable to what we would now call classical or symbolic computation, i.e. the type of computation that would become ubiquitous after the rise of the silicon microchip. Inspired

partly by von Neumann, Rosenblatt argues that these systems are too fragile and idealized to capture the randomness and imperfection of real biological systems. This is what leads him to, in his own words, "*formulate the current model in terms of probability theory rather than symbolic logic.*"(1958:388)

Where Rosenblatt draws the line, between probabilistic and logical systems, seems much closer to the modern distinction between classical and connectionist architectures. Rosenblatt also explicitly relates the perceptron to the empiricist philosophical tradition (1958:386), which would later become a sticking point for nativist critics of connectionism (e.g. Fodor 1975).

It seems then, that Rosenblatt deserves at least some of the credit (blame?) for driving a wedge between ANNs and classical computationalism. Despite this, it is worth noting that Rosenblatt's exposition of the Perceptron is concerned almost entirely with the architecture of the physical brain, and not necessarily cognition or the mind more generally. This is relevant because the great clash between connectionists and classicists in the 1980s focused heavily on the plausibility of connectionism as a *cognitive* architecture (e.g. Fodor & Pylyshyn 1988), whereas it is not clear that Rosenblatt (1958) had an explicit stance on this point. The final sentence of Rosenblatt (1958) is also somewhat revealing for the present discussion:

> *"By the study of systems such as the perceptron, it is hoped that those fundamental laws of organization which are common to all information handling systems, machines and men included, may eventually be understood."*(1958:407).

This seems to echo the earlier comments of McCulloch & Pitts (1943:113) and Kleene (1956 [19561]:3), as well as von Neumann's (1951) speculation of an overarching theory for both analogue and digital automata. Moreover, it strongly suggests that Rosenblatt did not perceive ANNs and digital computers as being two incommensurable classes of machine. Thus, Rosenblatt's criticism of symbolic logic (1958:387), is perhaps better understood as a criticism *of those formalisms* and their limitations. The more modern argument in cognitive science, that the mind/brain is *either* a connectionist network *or* a digital computer, does not seem to have registered as a possibility for Rosenblatt (1958). So, while Rosenblatt may have sown the seeds for the modern connectionist/computationalist divide, there are nonetheless important differences between the aspirations of Rosenblatt (1958) and the later arguments around connectionism that emerged in the 1980s and 90s.

Rosenblatt's Perceptron model sparked a wave of interest into the learning capacities of ANNs, which lasted until the late 1960s. Some have credited Marvin Minsky and Seymour Papert's 1969 book *Perceptrons* with killing interest in Rosenblatt's model and ANNs more generally (cf. Olazaran 1996). According to this argument, Minsky and Papert pointed out a fundamental flaw in the Perceptron (that it couldn't learn XOR relations), which caused almost everyone in AI to lose interest in neural networks and revert to symbolic/logical approaches.

It might be tempting then, to credit Minsky and Papert with firing the first real salvo in the battle between connectionists and computationalists. However, there are several details that question the accuracy of such an account[4]. Firstly, both Minsky and Papert pursued research into neural networks (e.g. Minsky 1954). Furthermore, Minsky's subsequent comments suggest that his focus was on prompting new solutions to ANN problems, rather than attempting to kill interest in ANNs entirely (Web of Stories 2016). Thus, the perception of Minsky and Papert as staunch critics of ANNs and advocates of classical computationalism seems to be a case of retrospective rationalizing.

Secondly, the subsequent method for solving the XOR problem was the same method that led to the reemergence of connectionism in the 1980s (backpropagation training over hidden layers; Rumelhart et al 1985). Thus, the XOR issue was already resolved before the most active debates in cognitive science that pitted connectionism against classical computationalism. So, while the inability to learn XOR might have been a critical factor for ANNs in machine learning, there's no particular reason to think that this was a decisive issue as far as theories of cognition go.

Still, the 1960s and 70s did see the coalescence of an explicitly symbolic research program in AI (Newell & Simon 1963;1976), as well as various criticisms of this approach (e.g. Dreyfus 1972), which set the scene for connectionism's resurgence in the mid-1980s. However, there is some evidence that most researchers in AI during the 1960s and 70s were relatively pragmatic (see Olazaran 1996). And what disagreements did exist appear to have been of a more technical nature than the polemics that characterized later discussions.

1.3.1.3 A Terse History of ANNs III: Rebirth and shots fired

While earlier work had focused on the distinction between systems which were analogue vs digital, deterministic vs probabilistic (etc.), the idea that ANNs and classical computers are

[4] This paragraph draws from an unpublished manuscript by Istvan Berkeley (1997).

fundamentally at odds seems to have solidified in the 1980s. Exactly who deserves responsibility for this is not obvious. However, many connectionists appeared to promote their research as an alternative to the *status quo* which was, for them at least, classical computationalism. As Rumelhart, Hinton & McClelland put it: "*We wish to replace the 'computer metaphor' as a model of the mind with the 'brain metaphor' as a model of the mind.*"(1986:75).

This represents something of a departure from ANNs research in earlier eras. While Rosenblatt (1958) sought to draw a distinction between probabilistic and purely logical models, he nonetheless seemed to view them as all belonging to some broader class of systems, and believed that his Perceptron model could provide insights into brains and computers alike. For Rumelhart *et al*, however, ANNs were a means of replacing outright the computational model of the mind/brain.

Interestingly, this change in perspective between Rosenblatt and Rumelhart *et al* is not obviously related to a development in the ANNs themselves. The primary advancement that distinguished Rumelhart *et al*'s model from the Perceptron was the backpropagation algorithm, which allowed the use of extra layers of units between the input and output layers. While this had a profound impact on the applicability of ANNs, it does not introduce any new conceptual distinctions between ANNs and classical computers. This suggests that the explanation for this shift might not be purely technical but at least somewhat sociological.

Some evidence for this lies in Rumelhart *et al*'s claim that computers are a *metaphor* for a model, rather than a model in-and-of itself. This is a subtle but potentially indicative change compared to earlier eras. Speculatively, we might ascribe this change to a cultural shift in the perception of computers. Indeed, it would have been very odd for McCulloch & Pitts (or even Rosenblatt) to talk about computers as a metaphor for anything, because computers in the modern sense didn't quite exist yet. The only real computers were large and impractical mainframes, which very few people had access to. Consequently, during the early days of ANNs, the people who would have had the most contact with actual computers were also the people who were deeply interested in computation as a field of study. By the time of Rumelhart *et al* however, personal desktop computers were starting to become widespread. GUIs, keyboard interfaces and word processors had taken over, creating a significant number of philosophers and psychologists (etc.) who used computers regularly, without having any deep interest in the mathematics that preoccupied the early pioneers of ANNs and other automata.

Perhaps for this newer generation then, the notion of a "computational theory of mind" would be more easily interpretable as a metaphor with "that machine in my office", rather than an appeal to some overarching theory of automata, information processing, and the like. Moreover, wider exposure to computers would have perhaps provided a widespread intuitive grasp of von Neumann's earlier argument, namely that digital computers exhibit a kind of rigidity which seems deeply at odds with naturally occurring systems[5]. These factors may have led to a subtle reinterpretation of classical computational models, even if not by Rumelhart *et al*, then most likely by the wider community of researchers for whom connectionism suddenly seemed like a new and viable alternative for explaining cognition.

Still, whatever the reason for this shift, the new perspective subsequently percolated into the philosophy department, where the antagonism between the connectionist and computationalist worldview was expounded and reified. Perhaps the most explicit example of this are the philosophers of the so-called eliminativist school (e.g. Churchland 1986). They interpreted connectionist models as a proof-of-concept for rejecting not only classical computationalism, but indeed all "folk psychological" concepts which had underlain much cognitive science up until that point. According to the eliminativists, psychological concepts such as emotions and memories are pre-scientific notions to be replaced by hitherto undiscovered scientific ones, much as Newtonian concepts such as force and gravity came to replace Aristotelian physics. The eliminativists included in their definition of folk psychology many of the foundations of classical computationalism (symbolic representations, etc.), and looked to connectionism instead to provide a new foundation.

The merits and flaws of the eliminativist view will not be discussed here (see Marcus 1998 for criticism) and it is worth noting that far from all connectionists openly endorsed eliminativism. Nonetheless, many at least flirted with the idea that connectionist models could partially replace or outdo classical computation in certain contexts (e.g. Rumelhart & McClelland 1987). And certain aspects of the eliminativist position can arguably be inferred from Rosenblatt (1958). It is perhaps fair to say then, that eliminativism represented the apex of anti-computationalist connectionism, weaker versions of which were favourably regarded by many connectionists.

---

[5] Anyone who has found themselves shouting in exasperation at a nonsensical error message can surely recognize this.

Regardless, it was inevitable the anti-computational rhetoric would generate a response. And while there have been many criticisms of connectionism by computationalists, the most pertinent for linguistics relates to the issue of systematicity, which was perhaps expressed most forcefully by Fodor & Pylyshyn (1988). The term "systematicity" here refers to the lawful relationships between complex representations and their constituents. For example, the representation of a sentence should be lawfully related to the representations for the individual words, as well as the syntactic and semantic bonds between them. For Fodor & Pylyshyn, this is an essential property of cognition, and one which computational models could account for better than connectionist ones.

However, it should be also noted that Fodor & Pylyshyn did not seek to prove that connectionist ANNs were *in principle* incapable of systematicity. Rather they suggested that connectionist models could only recreate the systematicity of computational models, if the connectionist model was a "mere implementation" of the computational model. And therefore, they argued, there is no sense in which the connectionist model can supplant the computational model as a model of cognition. But they were nonetheless quite explicit that connectionist models could still be valid as models of how the brain implements a computational architecture[6]. Thus the core of Fodor & Pylyshyn's argument is concerned with refuting the idea that all cognitive explanations can be *reduced* to an ANN.

Fodor & Pylyshyn's paper also garnered a significant number of responses. Many took issue with their failure to properly distinguish the properties of local and distributed representations (Chalmers 1990), and pointed to cases where distributed representations might seem to solve the problem of systematicity in ways which cannot be dismissed as "mere implementation" (Smolensky 1987; Dawson *et al* 1997). David Chalmers (1990) also argued that Fodor & Pylyshyn faulting connectionists for not solving systematicity is akin to a behaviouralist faulting computationalists for not solving classical conditioning. In other words, Fodor & Pylyshyn's arguments would have been devastating if systematicity were the *only* aspect of cognition in need of a scientific explanation. But as long as systematicity is one of many aspects

[6] In fact, most of the criticism of the neural plausibility of connectionist-style ANNs comes from neuroscience. See section 1.3.2 for more on this point.

in need of an explanation, then Fodor & Pylyshyn's argument might seem like a case of special pleading.

Regardless of the merits (or otherwise) of the arguments, the perspectival shift evident in Rumelhart *et al* (1985), as well as the subsequent expunction among philosophers, created a heated debate which persists to this day. Meanwhile, von Neumann and Rosenblatt's aspiration, a general-mathematical theory for all computing systems, has since become something of a minority position.

1.3.1.4 Linguistics and ANNs: Where we are now

Elements of the connectionist debates naturally spilled over into linguistics. For example, Rumelhart & McClelland's (1987) model of past tense verbs, explicitly sets their connectionist account in opposition to a classical rule-based account. The subsequent response by Pinker & Prince (1988) expresses a similar perspective to Fodor & Pylyshyn (1988), namely that the connectionist model fails to capture the systematicity of regular past tense morphology (see also Pater 2019).

At the same time, the connectionist/computationalist divide among philosophers seemed to map loosely onto to an ever-widening schism within linguistics, namely, the divide between the generativists and cognitivists (see Harris 1995). Chomsky, the father of generative linguistics, described connectionism relatively recently as having "*failed so badly that it was effectively abandoned*" (Chomsky & Guignard 2011)[7]. Meanwhile George Lakoff, a key proponent of the cognitivist school, was clearly positive about connectionism (Lakoff 1988), and, along with coauthor Mark Johnson, would go on to coin the notion of "second generation cognitive science" (Lakoff & Johnson 1999), which can be read as an attempt to unify various, loosely anti-computational approaches to cognitive science. This apparent alliance of anti-generative cognitivists with anti-computational connectionists may well have helped to seal the impression among generativists that ANNs were, in some sense, the enemy.

Despite this, various attempts at integrating ANNs with aspects of linguistics have proceeded undeterred (see e.g. Alderete & Tupper 2018). Harmonic Grammar and Optimality Theory are examples of formalisms which were explicitly designed to integrate aspects of both ANNs and

[7] Somewhat awkwardly, this comment was made the year before the deep convolutional net of Krizhevsky *et al* (2012) won the ImageNet Large-Scale Visual Recognition Challenge.

traditional symbolic grammars (Prince & Smolensky 1997). Meanwhile, advances in machine learning have allowed researchers to probe ANNs for the kinds of computational properties supposed by formal linguistic theories (e.g. Kuncoro *et al* 2017).

So, the animosity generated by the connectionist arguments of the 1980s and 90s was far from fatal. Indeed, a recent special issue of *Language* was dedicated to the topic of neural networks and generative grammars. In this issue, Joe Pater (2019) makes the argument that ANNs can complement generative grammar best if ANNs are treated primarily as theories of learning. The argument differs from the one presented in this volume, where I treat attractor ANNs as a neural realization of formal grammars, without making any strong claims about learning at either the neural or formal levels of abstraction. To a large degree, the distinction between Pater's view and my own can be traced to differing approaches to ANNs in general. While Pater is primarily interested in ANNs which come from the connectionist tradition, the attractor models I examine here come from the tradition of theoretical neuroscience. The distinction between these two is blurry but also potentially relevant. Therefore, I will dedicate the next section of this introductory chapter to this topic.

1.3.2 Connectionism vs Theoretical Neuroscience

The previous section considered the relationship between linguistics and Artificial Neural Networks (ANNs) in terms of the historical divide between connectionist and computationalist theories of mind. However, I've already argued that the attractor networks examined in this volume are a somewhat distinct species from connectionist ANNs. This means that not all arguments that pertain to connectionism are necessarily relevant for attractor networks.

Unlike connectionist models, most attractor networks do not derive from Rosenblatt's Perceptron. Rather, they typically derive from the Hopfield network (Hopfield 1982). The Hopfield network is itself a generalization of models from statistical physics, which were originally posited to study emergent phenomena such as ferromagnets (Hopfield 2007). The key distinction between the Hopfield model and its statistical physics forebears, is that the Hopfield model allows units to (potentially) interact with any other unit via connections of varying efficacy, whereas physical models typical place units on a lattice which only permits interaction between neighbours. Thus, while the physics models are a loose approximation of (e.g.) electrons with individual spins, the Hopfield model is a loose approximation of neurons with individual levels of activity, and synaptic connections of varying efficacy.

Consequently, the general structure of the Hopfield model and its descendants differs from those derived from the perceptron, and each is suited to a somewhat different approach to explaining the mind/brain. Attractor networks are often regarded as belonging to branch of neuroscience varyingly referred to as theoretical or computational neuroscience (e.g. Dayan & Abbott 2001), rather than connectionism in the narrow sense. It should also be noted that the field of theoretical neuroscience is not restricted to ANNs, and broadly subsumes a wide variety of mathematical approaches to brain function, from complex models of single neurons (Brunel *et al* 2014) to holistic models of neural functions (e.g. receptive fields; Jones & Palmer 1987).

Moreover, it should be acknowledged that the divide between connectionism and theoretical neuroscience can be somewhat fuzzy, since ANN models from both fields typically share a number of traits: both typically exploit large numbers of simplified "neurons" connected to one another with varying degrees of efficacy, and both generally assume that cognition emerges through the collective organization of those units. And importantly, there is at least *some* cross-fertilization between the approaches.

Still, while connectionist models are a somewhat familiar concept to many linguists, attractor networks and theoretical neuroscience generally are not. With that in mind, what follows is an approximate guide to delineating connectionist ANNs from those of theoretical neuroscience. Understanding the distinction can help us to understand how different ANNs should be assumed to relate to linguistic theory. Rather than concentrating on purely technical aspects that distinguish the networks, I will concentrate on three areas where the goals or focus of connectionists and computational neuroscientists tend to differ. They are: static vs. dynamic representations, learning vs. intrinsic properties, and biological realism vs. functional application.

Finally, I will briefly compare and contrast a connectionist account of the OCP effect in phonology (Alderete *et al* 2013), with the account given in the first paper of this volume. I will argue that these accounts are not necessarily in competition, but nor is the relationship between the two simple to decipher.

Static vs. Dynamic Representations

The prototypical connectionist ANN has a multi-layer, feed-forward architecture. This means the networks typically contain one layer of units that receive an input, and then pass the signal onto one or more "hidden" layers, before the signal finally arrives at an output layer (e.g.

Rumelhart *et al* 1986). Which mathematical function the network implements can be defined in terms of the difference between the input and output layers. In this regard, the prototypical connectionist network is broadly similar to a classical computer: both take a static input string, do something to that string, and then return a static output string somewhere else.

By contrast, the connections in a Hopfield network are symmetrical, i.e. activation flows in both directions[8], and therefore the network has no input or output layers as such. Consequently, computation in a Hopfield-style network is usually defined as the evolution of the entire network over time. For example, if the experimenter places the network in one state, and the network evolves to a different state, then the network can be said to have computed its ending state from the starting state. For this reason, theoretical neuroscientists tend more to discuss the properties of their models in terms of dynamical systems theory, rather than input/output mappings.

The distinction between static and dynamic representations is particularly relevant for linguistic models in a cognitive/neuroscientific context. This is because formal linguistic models are also, by and large, static models. That is, grammars are assumed to act over entire strings or tree structures. And to the extent that formal linguistic models *do* reference time (e.g. phases, domains, etc.), they are generally not bound by the flow of time in the normal sense. For example, a wh-word can raise from the object position, or vowel harmony can travel backwards from a suffix to the root, even though these descriptions are clearly odds with the order in which they are actually spoken. This is not a flaw of linguistic models *per se*, since they are designed to account more for grammaticality and competence rather than the specific algorithms which underlying processing. And the question of how a linguistic model relates to real time processes in the brain is something of an open question.

Still, for the linguist, this presents an interesting distinction between connectionist models and those of theoretical neuroscience. Connectionist models share with linguistic models the assumption of static representations, thus in this regard, drawing comparisons between connectionist models and formal linguistic model might seem easier[9]. On the other hand, any

[8] It is precisely this fact which allows a Hopfield network to implement attractors, since symmetrical connections permit equilibria.

[9] See, for example, Smolensky's tensor product vectors which equate static ANN states with tree-structures (Smolensky 1987).

model of brain function will ultimately have to reduce to some kind of real-time dynamics (e.g. Edelmann 2017), so the problem of how to recast static connectionist models as a real-time process has been pushed back (presumably to some other level of analysis) rather than solved. Conversely, relating static linguistic models to the models of theoretical neuroscience forces us to take the leap from static to dynamic at the same time as the leap from symbolic to "sub-symbolic" (Smolensky 1987).

Learning vs. Intrinsic Properties

A key component of Rosenblatt's Perceptron was the method he proposed for training the model. Rather than be programmed or pre-wired like most computers, the Perceptron could be "conditioned" to categorize data by exposing the network to different inputs and then strengthening or weakening connections to associate each input with the desired output. Subsequent waves of connectionist research were driven in large part by modifications to Rosenblatt's original method, which allowed for ever bigger networks and improved performance. Connectionism then, places a strong emphasis on "supervised" learning, i.e., learning where the algorithm already knows a dataset of "correct" answers (c.f. Bartunov *et al* 2018). The measure of a connectionist model is, to a large extent, what it can or cannot be taught using these methods. Thus, the properties of the network architecture and the learning algorithm are inextricably bound together, since any changes to the network architecture frequently necessitate a change to the learning algorithm and vice-versa.

By contrast, the Hopfield network and its descendants are typically trained using "unsupervised" or Hebbian learning. This method is far more limited for teaching the network (e.g.) to categorize data, but it has the key advantage of turning the memories into a controlled variable. In practice then, this method does not usually involve training datasets, but rather the experimenter deciding what the memories should look like and then solving a simple equation to determine the individual connection weights. Thus, the experimenter is free to examine the intrinsic properties of the network, without worrying about whether the results are indicative of the network architecture or the training algorithm. For example, theoretical neuroscientists have studied the relationship between the storage capacity of a network and various neural parameters, such as the sparsity of connections or the average firing rate of the units (Treves & Rolls 1991). Thus, for theoretical neuroscientists, the ANN can be an *object of study* in-and-of itself, with the deeper assumption that, if the model has some basis in reality, then properties of the model will reveal properties of real nervous systems.

Functional Application vs. Biological Realism

Both connectionism and theoretical neuroscience engage in a serious amount of abstraction from biological reality, and neither can claim an accurate, fine-grained description of a complete nervous system. Nonetheless, as already noted, theoretical neuroscientists often make the assumption that studying ANNs will reveal properties of the brain, an assumption which only holds if the ANNs are grounded in some kind of neural realism. Thus, it is probably fair to say that theoretical neuroscience places a much higher value on biological realism than does connectionism.

This might seem strange, given that one of the early arguments raised in support of connectionism was its supposed adherence to observations about the physical structure of the brain (Rumelhart *et al* 1986). However, this case was always *somewhat* overstated, not least because connectionist training algorithms (mentioned above), are already a fairly significant departure from reality. So the fact that biological realism was slowly deprioritized by many connectionists is perhaps not as surprising as it might seem. In general, the divergence of connectionism from its early biological commitments can be understood as a consequence of two facts: Firstly, the more was learned about real neurons, the less realistic the connectionist models of neural activity seemed (e.g. Connors & Gutnick 1990). Secondly, biologically implausible mechanisms often seem to produce better results in applied domains such as machine learning. To give one example, many convolutional ANNs depend on "weight sharing", whereby connection weights from one portion of the network need to be copied perfectly and entirely to a different portion of the network. This is, of course, a deeply implausible neural mechanism (c.f. Bartunov *et al* 2018), but one which has nonetheless produced impressive results (*ibid*).

Perhaps in response to this, various connectionists have embraced the implausible aspects of their models and argued that this is proven approach for making progress in complex scientific domains. Some have drawn an explicit comparison with the history of aerodynamics, for example:

> *"Birds provided an existence proof of how an object could fly through the air under its own power. However, as the principles of aerodynamics began to be understood, researchers studied artificial man-made systems of flight."* (Schneider 1987:73)

Variants of this analogy appear in the literature, but the general gist is as follows: The earliest attempts at human powered flight involved imitating birds, i.e. flapping wings to generate lift. These attempts all failed (often suddenly and violently). Progress was only made once people abandoned wing-flapping and instead pursued fixed wing aircraft with a propeller attached. In other words, people stopped imitating natural systems and instead attempted to engineer an artificial system from the ground up. It was only *after* artificial flying machines had been built, that scientists began to understand aerodynamic principles (e.g. aerofoils), and then apply those principles back onto natural systems (e.g. birds' wings). By analogy then, the goal of connectionist models is not to imitate natural nervous systems exactly, but rather to engineer artificial systems that perform approximately the same task. The hope, then, is that this will reveal principles which can be applied back to real nervous systems.

The general approach taken by theoretical neuroscientists, however, is slightly different. The limitations of unsupervised learning mean that such networks cannot outcompete a connectionist network in (e.g.) an image classification task. Instead, the measure of a model within theoretical neuroscience is more akin to the measure of other scientific theories (parsimony, empirical coverage, etc.). Thus, models within theoretical neuroscience are often abstracted for the sake of tractability or analytical solvability, but rarely for the sake of chasing improved performance on some task. For example, the PLN in the first paper of this volume uses "Potts" units, which subsume whole patches of cortex into a system of differential equations. The motivation for doing so is not that this system is neutrally exact, but rather that mathematical methods for solving systems of Potts systems exist in the physics literature, which can then be applied to neural models allowing for quantitative analyses of the system's properties (e.g. Naim *et al* 2017).

Of the three dichotomies (static vs dynamic representations, learning vs intrinsic properties, functional application vs biological realism), it is this last one which arguably presents the biggest divide between the connectionist and theoretical neuroscience approaches. Particularly within machine learning, connectionists have generally demonstrated a willingness to integrate elements of dynamics (Elman 1990; Hochreiter & Schmidhuber 1997) or unsupervised learning (e.g. Bengio *et al* 2007) into their models, but only provided it improves the performance of the model in some task. Indeed, many of the most sophisticated deep learning networks are hybrid models which can incorporate any number small tweaks that seem to further blur the static vs dynamic or learning vs intrinsic divides. However, these tweaks are invariably in the service of functional application, and often at the cost of biological realism.

1.3.2.1 Comparison of a connectionist model and the PLN

As already noted, these three distinctions are only an approximate guide to the fuzzy boundary between connectionism and theoretical neuroscience, and exceptions can certainly be found in the literature. Nonetheless, it is perhaps insightful to demonstrate all three with a more specific comparison between a connectionist and theoretical neuroscience model. To that end we can compare a connectionist account of the Obligatory Contour Principle (Alderete *et al* 2013) with the Phonological Latching Network (PLN) account given in the first paper of this volume.

The model of Alderete *et al* can be regarded as a fairly typical feedforward connectionist network. It takes as its input an activation vector representing a triliteral Arabic root (i.e. three consonants), and returns as its output a single numerical value from -1 to 1, representing a grammaticality assessment of the input root.

The training dataset for the model is a combination of attested Arabic roots, and unattested-but-possible roots generated using a second, simpler network. The second network relies on random noise to generate the unattested roots, which provide "negative evidence" during the training phase. The model is trained using a backpropagation algorithm, which adjusts the weights in the network such that attested roots should (ideally) produce the output "1", while unattested roots should provide the output "-1".

While the results of the model are too extensive too list in detail here, it is enough to note that the model does indeed to appear to approximate an OCP-effect during the testing phase: Novel forms which violate the OCP were rated considerably lower than forms which did not.

So, does the model of Alderete *et al* conform to the connectionist side of the three distinctions given above? On the issue of static vs dynamic representations, the PLN and the connectionist model of Alderete *et al* are fairly typical examples of their respective fields. The connectionist model uses purely static representations, i.e. it receives and processes an entire triliteral root at the same time, and then returns an assessment for the grammaticality of the entire root. Conversely, the PLN does not quite process inputs in the same sense, however, it latches between phones one at a time, and the OCP-effect is explained as the model's reluctance to latch between phones which are too similar.

On the issue of learning vs intrinsic properties, the two models again capture the proposed distinction well. In the connectionist model the key issue, as presented by the authors, is whether or not an Arabic-OCP effect can be induced from linguistic data. Thus, it is framed as

a contribution to the nature vs. nurture debate. It does not, however, make strong claims as to why the OCP is a fact of language in the first place. Conversely, the PLN model has no aspirations towards a theory of learnability. Indeed, there is currently no method for training a latching Potts model on real language data (e.g. Arabic roots). Rather, the PLN starts with a set of "phonology-like" representations, with no particular commitment to where those representations come from. However, the PLN attempts to provide an answer to why the OCP exists in the first place: the neural representations themselves cannot be "reactivated" once they are fatigued, which is not an arbitrary restraint on the system, but follows as a consequence of the mechanism which makes string production possible: latching dynamics.

Finally, the issue of biological realism vs functional application is perhaps less clear cut in the case of these two models. The model of Alderete *et al* does not make any specific reference to neurological facts that might inspire its design, since it is not clear that the model is even intended to model a particular cognitive task. Indeed, the basic function of the model is as a classifier for grammaticality – its output is a unit whose sole job is to specify which forms are ungrammatical, and it's debatable whether such a unit would have an analogue in an actual nervous system. However, the authors are quick to draw parallels between the biologically implausible aspects of their model and similar aspects in other, non-connectionist theories of learning (e.g. the non-attested forms used during training). This suggests that the model is perhaps best understood as a model of learning first, and only indirectly a model of neural function. Certainly, this interpretation is congruous with the "connectionism-as-learning" view advanced by Pater (2019), mentioned in section 1.3.1.4.

By contrast the architecture of the PLN is designed explicitly to capture both broad neuroanatomical facts – the separation of motor and auditory areas – as well as certain properties of inhibitory interneurons and resource depletion in synapses. However, the Potts units which constitute the model are heavily abstracted from neural reality. And as such it would be a mistake to regard the PLN as an exact model of neural function. It is rather, designed to occupy an intermediate level of abstraction between the neural and linguistic levels.

It seems then, that the connectionist model of Alderete *et al* (2013) and the PLN have only a partially overlapping domain of explanation, despite ostensibly being models of the same phenomenon (OCP-effects). From one perspective, the models could be seen as contradictory: the PLN claims the OCP is a property of latching dynamics, while the connectionist model claims that the OCP is learned from exposure to linguistic data. However, this interpretation is

likely premature. The OCP effect in the PLN depends on representations of phones having certain properties, but the model has no theory of how those representations come to be. Conversely, the connectionist model is not, in the words of the authors, a "*blank slate [...] that [is] completely free of bias and a priori assumptions*." (2013:58). Perhaps then, we could view these models as complementary, rather than in competition. One could speculate a synthesis of the PLN's dynamic computational insight with the connectionist model's learning, might provide us with a more complete account of phenomena such as the OCP.

Despite this, there is no obviously simple way in which this synthesis could be accomplished. The basic architectures of the two models are not easily commensurable, not least because they sit on opposite sides of the static vs dynamic divide. Moreover, the learning model used by Alderete *et al* is not applicable to the PLN without serious modification, if indeed it is possible at all. So, while it seems that the two models may hint at some deeper continuity, for now such an insight is frustratingly beyond our grasp.

1.3.3 Linguistics and Attractor Dynamics

The models examined in this volume are examples of attractor neural networks. That is, networks of units which self-organise toward stable states (attractors) representing stored memories.

However, attractors are not solely a property of certain neural networks. Rather, "attractor dynamics" can apply to a much broader class of dynamical systems which tend towards a small subset of configurations over time. In practice, their application ranges from ferromagnetism (Ising 1925) to ant cooperation (Feinerman et al. 2018) and much more besides. One classic example of an attractor system in physics is that of a vertical mass-spring system. That is, if one imagines a simple spring with a weight hung from one end, by pulling and then releasing the weight the spring will contract and extend such that the weight "bounces" up and down. Provided the system has some damping (as any real system will via friction, etc.) then the weight will eventually come to a standstill. Crucially, how low the weight hangs once it stops is entirely dependent on the mass of the weight and the rate of the spring – it does not matter how hard one pulls the weight or how many times its bounces up and down, it will always come to rest at the same point (until we change the mass of the weight or the properties of the spring). The point at which the weight comes to rest is an example of a "point attractor", and is conceptually no different from the memories in an attractor neural network.

The study of attractor systems is sufficiently well developed that they can also be studied in purely mathematical terms. That is, in much the same way one can discuss the properties of (e.g.) triangles without needing to discuss any particular triangular object, so too can one discuss attractor systems without needing to refer to any particular physical system. Compared to attractor neural networks, these "holistic" attractor systems are typically quite simple, making them more amenable to analytical study[10]. Similarly, they are also more transparent than attractor neural networks, making them well-suited to conceptual or qualitative arguments about cognition.

Because of this, holistic attractor models have also been applied to the problems of cognition without reference to neural networks, or indeed any explicit neural structure. The late 80s and early 90s saw a rise in the interest of dynamical systems theory to cognitive science (e.g. Haken & Stadler 1990; Serra & Zanarini 1990;). Arguably this reached a zenith with Tim van Gelder's 1995 paper "What Might Cognition Be, If Not Computation?", in which dynamical systems theory is posited as direct competitor to the computational theory of mind. According to this view, the brain is not a computer at all, but rather a type of *control system*, whose explanation can be found in the mathematics of differential equations. Various subsequent authors have argued that van Gelder was probably overstating his case (e.g. Eliasmith 1997), and that dynamical and computational approaches could be unified (e.g. Crutchfield 1998; Dale & Spivey 2005). Still, versions of this anti-computational dynamicism have persisted and even permeated into phonology (e.g. Port & Leary 2005). Nonetheless, this wave of interest in dynamic approaches to cognition brought with it a number of holistic attractor models, some of which would subsequently be applied to the study of language.

One early important example for phonologists is a study of categorical perception in speech by Tuller *et al* (1994). They conducted perception experiments using stimuli which could be ambiguously interpreted as either the words *say* or *stay*. The ambiguity was introduced by a silent pause (up to 76ms) between the /s/ and vowel portion of the utterance, as well as varying the vowel between either a simple synthesized /ei/ (as in *say*), or the same diphthong extracted from a synthesized utterance of *stay*, which contained information of the preceding stop in the formant structure. The experiments studied the effects of the length of the silence and the quality of the vowel on participants perception of the stimulus. Participants performed the

[10] At least within cognitive science.

perception task with both vowel types, and both with stimuli presented in a random order, as well as in a "sequential" order, where the length of the silence was progressively increased or decreased in 4ms increments.

The results from the random order trials were consistent with other research on categorical perception. As the length of the silence increased, participants were more likely to perceive the stimulus as *stay* rather than *say*. Furthermore, the relationship between silence length and perception was approximately sigmoidal, rather than linear, which is characteristic of categorical perception, as participants switch between perception of *say* and *stay*. The effect of the vowel was also fairly predictable; the stay-derived diphthong biased participants towards a *stay* perception at a shorter pause length compared to the *say*-derived diphthong.

However, the results from the sequential trials revealed a slightly more complicated picture. The critical silence length at which participants switch from *say* to *stay* (or vice-versa) was dependent on whether the lengths were presented in an increasing or decreasing order. Moreover, the exact nature of the dependency was not consistent across trials or across participants. Tuller *et al* interpret these results as being indicative of a complex non-linear system that drives categorical perception, rather than a simple, immovable boundary between categories.

Tuller *et al* present a holistic attractor model as a qualitative model of the phenomenon. The model is a simple function in two dimensions, which allows for two adjacent attractor basins, i.e. the function traces a "W" shape when graphed. Each of these basins is understood to represent the words *say* and *stay* respectively. And, as with all attractor systems, the behaviour of the system is understood as seeking local minima, i.e. it "rolls downhill" to one of the bottoms of the "W" shape. This means that it always retrieves either *say* or *stay* and not some ambiguous point in between. However, by manipulating a single parameter in the system, Tuller *et al* demonstrate that the shape of the function morphs such that one of the basins raises as the other one lowers, i.e. the "W" morphs slowly into a wonky "U" shape. In that case, one of the words is no longer a local minimum (because it has risen up), and attempting to retrieve that word would simply cause the system to roll down into the other, deeper basin, thereby retrieving the other word instead. In this way, this single parameter in Tuller *et al*'s model can approximate the changes in participants responses, whereby either *say* or *stay* becomes the preferred perception of a given stimulus, depending on context.

Of course, while the study by Tuller *et al* focuses on categorical perception (which is clearly relevant for phonologists), it is not focused on phonological grammars per se. Arguably the first attempt to apply attractor dynamics to phonological grammar is Gafos & Benus (2006). They discussed the role that attractor dynamics could play in phenomena such as vowel harmony and final devoicing. The latter is particularly relevant since this is also a topic for the second paper in this volume. And, like that paper, Gafos & Benus also advocate attractor dynamics as a way of explaining the gradient effect of incomplete devoicing, i.e. situations where an underlyingly voiceless coda segment appears to have fractionally less voicing than the corresponding devoiced segment.

Gafos & Benus re-employ the same two-dimensional "W" shape system as Tuller et al. However, rather than being a model of categorical perception, the basins represent a categorical distinction within the grammar (in this case [+/-voice]). The devoicing context is equated with the wonky "U"-shape, where the [+voice] basin has been raised, causing the system to roll into the [-voice] basin instead. However, as Gafos & Benus note, in this context the [-voice] basin can shift slightly relative to the underlying voiceless case, which allows for the interpretation of incomplete devoicing.

In many ways then, the model of Gafos & Benus can be regarded as a precursor to the attractor network model in the second paper of this volume. The analysis of incomplete devoicing in terms of shifting attractor basins is very much the same in both models. However, the key difference between the models is that the dynamics investigated by Gafos & Benus are of a holistic type. That is, it is a simple system where the relationship to neural reality is left open to interpretation – one is left to presume that bundles of neurons could implement such a dynamical system, but Gafos & Benus do not advocate any view of exactly how. The second paper in this volume gives a more explicit account of how such dynamics can be realized in the brain, and can therefore be regarded as an extension of the basic premise of Gafos & Benus.

1.3.4 Definitions of Computation

I argue in this volume that attractor networks can be understood as realizations of the formal theories posited by linguists to explain the grammar of natural languages. The second paper in particular, is an exploration of this topic. Nonetheless, there is an underlying issue which the paper only deals with in passing: namely, neural networks can seem to produce the same outputs as a formal grammar, but they do often do so without realizing explicitly the machinery of the formal model. For example, I argue the Phonological Latching Network (PLN) exhibits a kind

of place assimilation between adjacent segments in a string. In a formal theory, assimilation might be accounted for in any number of ways (e.g. rewrite rules, linking nodes, agreement constraints, etc.), but the PLN doesn't have any components that obviously resemble any of these formal mechanisms. So is the PLN really commensurable with these formal accounts, or are they actually competing accounts of how the brain really works?

Opinions on this issue diverge greatly. Some authors would clearly reject the claim that the neural network and the formal are equivalent (e.g. Port & Leary 2005; Gallistel & King 2009). Conversely, there are a great many others who would accept the equivalence without batting an eyelid[11]. This disagreement is interesting insofar as the terminal principles from which these viewpoints derive are rarely stated explicitly. This can result in an *impasse* between different viewpoints, as each side is operating from fundamentally different assumptions, which neither side has properly expressed. And while there might be many different aspects to this disagreement, I argue here that one of the most important (and understated) points of disagreement is differing understandings of the term "computation".

Computation is invoked extensively in neuroscience, linguistics, and cognitive science more generally, however it is rarely defined exactly by anyone other than philosophers. Of course, it shouldn't be controversial that different implicit definitions of this term can greatly affect our reasoning on the topic of cognition. For example, formal linguistic theories are often said to be computational theories, which in turn (depending on definition) might entail that the brain must be a computer, which (depending on definition) might delimit which neural theories we think are commensurable with formal linguistic theories. There is a potentially long chain of deduction here that depends on exactly how we define what is meant by "computation", or a "computational theory". And in practice this creates problems because there is no universally agreed upon definition.

Here I argue that the view of neural and computational models as "in conflict" typically comes from an excessively narrow definition of computation. Under a narrow definition, to claim the brain is a computer is claim that it is very much like the digital, programmable computers that have come to dominate our everyday lives. And to call a linguistic theory "computational" is to claim that it should be implemented in the brain in a manner very similar to a computer

[11] See the various citations on this point in the second paper.

program. However, I will also argue that these narrow definitions introduce more problems than they solve. And that a much broader definition of computation entails easier reconciliation of the neural and the computational, and that this broader definition is in keeping with the way mathematics is usually applied to explicanda in the natural world. However, such a broader definition is not without its own thorn, namely, the specter of *pancomputationalism*.

1.3.4.1 The specter of pancomputationalism

According to the broadest definition, computation is the physical realization of some mathematical operation. While intuitive, this definition has the unintended consequence that *literally everything* can be regarded as a computer. This follows from the arbitrariness of physical representations of mathematical objects. For example, if we take any normal window and specify that "open window = 1" and "closed window = 0", then the act of opening the window is an act of (trivial) computation, i.e., 0→1. With a bit of imagination, we can impute similar computations onto any state of the world. Often these are trivial, like the window example, but sometimes they are less obviously so (e.g. does a planet "compute" its orbit?). This conclusion is referred to as *pancomputationalism*, and whether we accept it can have profound importance for subsequent arguments (certainly, arguing about whether the brain is a computer doesn't make a lot of sense from a *pancomputationalist* perspective).

A full discussion of this topic is far beyond the scope of this introductory chapter (see e.g. Piccinini 2015:ch4). But for the sake of expediency we can loosely distinguish three types of response to *pancomputationalism*:

1. Reject the conclusion - develop a definition of computation which excludes windows, planets, etc. (e.g. Fodor 1981, Piccinini 2015).
2. Accept the conclusion - whether or not something counts as computation is a matter of perspective/utility (e.g. Dennett 1987, Shagrir 2006, Dewhurst 2018).
3. Accept the conclusion - reject entirely "computation" as an explanation for natural systems, including the brain (e.g. Searle 1992).

I'll ignore the third option here and take it as a given that computation is at least *useful* as a way of talking about the mind/brain[12]. Nonetheless, the distinction between the first two options is

[12] See the second paper in this volume for a more rigorous rejection of the third option.

still relevant for our understanding of how a formal linguistic theory relates to (e.g.) a neural network.

Generally, if we accept *pancomputationalism* (option 2) then we should have few qualms equating an attractor network with a formal/computational model, since we've already accepted that everything can be equated with at least *some* computational model. The only outstanding issue is whether or not that computational model tells us anything useful. However, if we reject *pancomputationalism* (option 1) then the issue becomes more complex. Depending on our definition of computation, we might not think that the neural network is actually computing anything. And even if it is, there's no guarantee that it's computing the same thing as the formal model. This is ultimately the starting point for an argument that, even if a neural network and a formal model are *extensionally* equivalent, their *intensional* differences make them into competitors rather than complementary points of view.

My own implicit stance in this volume is option 2, which can be broadly referred to as *perspectival pancomputationalism*. However, given the (often implicit) acceptance of option 1 among many commenters, it is worth explicating why I disagree. Especially given that my on this issue stance underlies much of the work in this volume. While attempts to avoid *pancomputationalism* come in many different forms, for this discussion I'll focus on Piccinini (2015)'s *mechanistic account* of computation. And although I don't suppose that everyone who would pick option 1 (above) would also endorse Piccinini's account, it is arguably one of the most thorough definitions of computation in the literature, and therefore a good starting point for this (condensed) discussion.

1.3.4.2 Piccinini's Mechanistic Account

Piccinini (2015)'s main goal is a principled way of defining what counts as a computer, and thereby avoiding the conclusion of *pancomputationalism*. The crux of Piccinini's solution to *pancomputationalism* is a distinction between computational modelling and computational explanation: Cases like the open/closed window, or the planet's orbit, are examples of a computational model, but not a computational explanation. According to Piccinini, *pancomputationalism* is only true to the extent that every system can be modelled with some computation(s). However, for a system to be explained by computation, a system needs to have certain properties that define it as a computer and distinguish it from other, non-computational systems. Piccinini then proceeds to develop his account of exactly what delineates computational and non-computational systems. This ultimately boils down to the claim that a

system is computational if the system has computing as a *teleological function*, where functional analyses are understood to be a special type of mechanistic analysis.

Since Piccinini's account is far more thorough than I can possibly dedicate space to here, I will focus on two main points of disagreement: Firstly, Piccinini's account has a reductionist flavour insofar as it privileges a single level of explanation, i.e. there is some level at which a system *really is* (or is not) a computer, and that level of explanation has exactly one correct scientific explanation. Although Piccinini does concede that complex systems permit multiple levels of explanation, the relationship between these levels is quite restrained. For Piccinini, micro-level explanations of a computational system can be mechanistic but not themselves computational (e.g. a mechanistic account of the silicon atoms in a transistor), while macro-level abstractions of a computational systems are simply "sketches" which remove certain details from the model, purely for legibility or convenience (e.g. high-level programming languages).

For this reason, it is not obvious to me how Piccinini's definition deals with the notion of emergent computation, as explored in the first two papers in this volume. For example, in the case where a (continuous) neural network is realizing a (discrete) formal grammar, it is not clear which level of abstraction should be privileged as the computational explanation. If we suppose that the formal model is the computer, then the micro-level account of the neural network must be mechanistic but non-computational. However, this is obviously false because both the individual units, and the network itself, are unambiguously performing computations (see Piccinini 2015:ch13). But if we concede that the neural network is itself a computer then the formal account is relegated to a "sketch" of the neural network – a notion which is at odds with the fact of causal emergence in the macro-level of abstraction.

Thus, privileging only one level of abstraction as having computation as its teleological function leaves us with an arbitrary choice. The real problem in this case is not just that both levels of abstraction are computational, it's that each level exhibits a different type of computation (continuous vs discrete), and the explanation of one does not neatly reduce to a proper subset of the other.

This problem can be viewed as a specific case of the limitations of a reductionist program. In many cases, moving to a micro (or more fine-grained) level of abstraction is not simply filling in the gaps of the macro-level account (*pace* Piccinini). It can often mean transitioning to a radically different ontology which can even obscure the scientific problem at hand. Or to put it

another way, there are cases where abstraction or approximation are not merely conveniences or restrictions imposed by limits on computational resources, but are in fact prerequisites for any kind of scientific theory at all.

A pertinent example here would be the faculty of language. That is, it is completely uncontroversial that people speak different languages, and that those languages are fuzzy collections of dialects, which are themselves fuzzy collections of idiolects, etc. There mere notion of a faculty of language is a kind of approximation that cannot be reduced a single micro-level configuration in peoples' heads. Indeed it is quite conceivable that no two human beings have *exactly* the same wirings in their heads. Therefore, simply "zooming in" on a single person's neural wiring does not give us a more detailed theory of the language faculty as a whole – in fact it may well do the opposite.

This situation is not unique to language, or even cognitive science. Many forms of scientific explanation work precisely because they do not distinguish states of the world which are distinct but equivalent in some important regard – not for legibility or convenience, but because this is the crux of the causal explanation. For example, a distant ancestor of attractor neural networks is the Ising model of ferromagnetism (Ising 1925; Hopfield 2007). The Ising model allows for exact prediction of phase transitions, such as the Curie temperature in a ferromagnet[13]. For example, whether or not iron behaves as a magnet depends on the alignment of the spins of the electrons: if they are all aligned then it is a magnet, if they are not aligned then it is simply a regular lump of iron. Of course, for any given number of iron atoms there are only 2 possible configurations where the spins are all aligned (all "up" or all "down"), but a very, very large number of possible unaligned configurations. The key observation of the Ising model is that the unaligned cases are effectively all equivalent as far magnetism goes, and thus the model does not need to distinguish them. The result is that the model can be solved by treating all the interacting electron spins as "averages" to determine when the system transitions from an all-aligned to a not-aligned configuration. Thus, there is no need to ever determine exactly the spins of the not-aligned configurations in a theory of magnetism, because what matters is *that* they are not aligned, not *which* of the many, equivalent not-aligned states the system is in. And it is

[13] i.e. the temperature above which a magnet ceases to be a magnet. The description here also simplifies and ignores various other phenomena (e.g. ferrimagnetisim), as these are not the author's expertise.

important to note that the Ising model is not just an engineer's approximation for sake of computational tractability, but rather has "*given microscopic insight into the many body collective phenomena of phase transitions and [has] developed new areas of mathematics*" (McCoy 2010). To be sure, it is precisely by ignoring individual spins that the model becomes exactly solvable under certain conditions. But it is also the model's solvability that allows it to prove properties of phase transitions. Thus, it is not clear that a version of the Ising model which deterministically modeled each spin could be called a more "detailed" model, since that version of the model would lose the ability to prove anything.

My argument then is that the specific electron spins in the Ising model are like the exact neural wiring in a theory of language: they are, in principle, distinguishable from a God's eye (or Laplacian demon's eye) perspective. That is, a given person's brain or a given lump of iron may have one exact micro-level description at a given moment in time; but that exact description is not only irrelevant for the scientific theory, but in some cases actively unhelpful. Thus, it is misleading to characterize a theory of language or the Ising model as mere "sketches" with conveniently missing details to be filled in later, because it is only by removing irrelevant details that a causal account becomes possible in the first place. This runs counter to Piccinini's rigid distinction between models and explanations, with only the former depending on approximation, and only the latter potentially qualifying a system as *actually* computational. In scientific practice, all explanations are models to some extent, and some explanations presuppose approximation.

Of course, Piccinini is correct when he notes that macro-level (or coarse-grained) theories should constrain micro-level theories (ch.5). And thus, linguistics and neuroscience (for example) could never be truly autonomous enterprises. But exploring the types of models presented in this volume generally entails a more complex interaction (e.g. causal emergence) than the sort allowed by Piccinini's mechanistic account.

This brings us to my second contention with Piccinini's account. Namely that, in practice, it is not clear if the mechanistic account allows for prediction that could not follow from *perspectival pancomputationalism*. This is because, while Piccinini does propose a somewhat loose set of properties that might help identify which *functional mechanisms* can be properly called computational, in practice they are properties which have to be uncovered through the normal scientific method. Piccinini writes:

> "*A system X is a* functional mechanism *just in case it consists of a set of spatiotemporal components, the properties (causal powers) that contribute to the system's teleological functions, and their organization, such that X possesses its capacities because of how X's components and their properties are organized […] To identify the components, properties, and organization of a system, I defer to the relevant community of scientists." (p.119)*

Consequently, Piccinini's mechanistic account can only be distinguished from a perspectival account after the fact. That is, the mechanistic account presumes the existence of a community of researchers who have adopted a pragmatic form of *perspectival pancomputationalism* and are willing to apply computational models to any novel systems they discover. Over time, the community of researchers will converge on a set of conventions whereby certain systems are more profitably described as computational mechanisms, while others are not. At this point, the philosopher is free to declare the computational models as examples of *explanation*, and that the relevant systems have computation as their *teleological function*. My counter-argument is simply that it is not obvious what positing is gained by positing computation as a teleological function after the fact. The scientists' standard heuristic of "*is this mathematical analysis telling me anything?*" seems to lead us to the same outcome.

One could counter that, the scientific process is an iterative one, and therefore the process of discovery may continue even after one has declared a given system to be a computer. However, even then it is not clear what the categorization of a system as a "genuine computer" subsequently enables that the perspectival view does not.

1.3.4.3 Why we can't agree – a speculation

For the reasons discussed so far, nothing in Piccinini's arguments has motivated me to abandon my *perspectival pancomputationalism*, which I regard as a much weaker ontological commitment. Nonetheless, I agree with Piccinini that the general disagreement on the topic of *pancomputationalism* demands some form of explanation. For example, Piccinini begins his chapter on pancomputationalism with the following observation:

> "*I have encountered two gut reactions to pancomputationalism: some philosophers find it obviously false, too silly to be worth refuting; others find it obviously true, too trivial to require a defense. Neither camp sees the need for this chapter. But neither camp seems aware of the other camp.*" (2015:51)

I count myself in the second camp here - pancomputationalism seems obvious and attempts to refute it intuitively strike me as *post-hoc* avoidance of a non-issue. Nonetheless, Piccinini is correct that the sharp bifurcation of opinions on the matter suggests that *something* remains unresolved. My own suggestion here is that the disagreement arises from a failure to distinguish two slightly different things: *computers-as-objects* versus *computation-as-mathematics*.

Computers-as-objects is a fuzzy, natural language category, which is defined by similarity to a *prototypical* computer-as-object - for example the laptop on which I am typing this sentence. If one implicitly derives one's definition of computation from this category, then it seems strikingly obvious that planets and windows don't belong in this category, and thus one is forced to conclude that something in the *pancomputational* line of reasoning is wrong, and so begins the search for a more selective definition of computation.

However, if one accepts *pancomputationalism* then one has implicitly accepted that similarity to prototypical computers-as-objects is not the main criterion for whether a system can be described as computational. Instead, one is treating computation as a type of mathematics, and as such it seems reasonable to apply it as freely as one would any other mathematics. For example, consider the claim that a fair coin toss has a 50/50 chance of turning up heads. Surely no-one disputes this, and yet the scientific worldview supposes that we live in a deterministic universe - some Laplacian demon with an understanding of the forces involved could accurately predict outcome of a coin toss 100% of the time. Does this mean it is wrong to apply probability theory to a coin toss? Are statisticians mistaken for thinking that the world is non-deterministic when the reverse is true...? Probably not. A less radical interpretation is that probability is simply a powerful tool for understanding systems with an intractable number of variables and/or interactions[14]. It would be a bizarre kind of sophistry to insist that probability theory entails that coins possess a physics-defying, non-deterministic essence. By analogy then, we might simply say that computational analyses of complex systems are not actually claims about whether the system *is* a computer, just as a probabilistic analysis is not a claim that the system *is* probabilistic.

Of course, this much is compatible with Piccinini's notion of computational modelling:

[14] Once you hit bedrock, all probabilities are epistemic in a Galilean universe.

> *"The computational descriptions play a role fully analogous to the role played by differential equations, diagrams and other modelling tools […] Just as being described by a system of differential equations does not entail being a systems of differential equations in any interesting sense, being described as a computing system in the present sense does not entail being a computing system in any deep sense..."*(2015:64)

This quote is perfectly in line with the probability example given above. However, Piccinini finishes the paragraph:

> *"…So, computational descriptions in the present sense say nothing about whether something literally computes. They are not a basis for computational explanation in computer science or cognitive science."*(2015:64)

And this is where my stance diverges with Piccinini's – I'm simply not convinced that whether something *literally* computes is a useful distinction, in much the same way that probability theory does not need us to distinguish things which are *literally* probabilistic (i.e. nothing).

As a rejoinder, one might contest that computation is special precisely because it is so intimately connected to the physical machines which implement it. However, computation is not the only branch of mathematics to be closely associated with its own canonical, non-Platonic implementation. One other potential example is the relationship between clocks and modular arithmetic (sometimes called clock arithmetic), which studies number systems which "wrap around" at some particular value. An analogue clock is the classic example of a modulo 12 system, i.e., 12 o'clock plus 14 hours equals 2 o'clock (not 26 o'clock). But other systems are useful in other contexts, such as modulo prime systems for encryption, or modulo 2π for anything involving circles or sinusoidal functions.

So, we could say that a digital computer is a canonical example of computation in the same sense that a clock is a canonical example of modular arithmetic. But of course, while it is true that some implicit understanding of modular arithmetic is a prerequisite for understanding clocks, there are also a great many cases where modular arithmetic can be applied to systems which are not clocks in any interesting sense (oscillations, cryptography, etc.). Indeed, at some banal level, everything can be given a trivial treatise in clock arithmetic, as is also true for probabilities and computations. However, to my knowledge, no one treats "*panclockism"* as a

serious problem to be avoided. That is, no one worries whether the application of clock arithmetic to a system is an implicit claim that the system is a type of clock.

My proposal then, is that the difference between *panclockism* and *pancomputationalism* is a sociological one. That is, clocks and clock arithmetic are conventionally understood to be separate things in a way that computers and computation are not. Perhaps it is because the invention of sundials precedes modern modular arithmetic by several millennia, whereas computers-as-objects and computation-as-mathematics were effectively invented at the same time by largely the same people. If we allow ourselves a somewhat trite account of history: Turing developed his ideas largely as a response to the ongoing discussion about the foundation of mathematics and the limitations of formal systems (Eberbach *et al* 2004). He invented a hypothetical machine to prove a mathematical point. Of course, once the hypothetical machine existed, it made quite a lot of sense to turn them into actual machines.

The solving of a mathematical problem by first envisioning a type of machine that solves that problem might well be unique to Turing, and this may explain why computers-as-objects and computation-as-mathematics are so intermingled in the popular imagination. However, I do not see that this justifies treating computation differently to other forms of mathematics. In the end, they are all potential tools in our toolbox. And for a system as complex as the mind/brain, we are likely to need every tool at our disposal.

1.3.5 The PLN and Exemplar Theory

Anecdotally, presentations of attractor networks to groups of phonologists have tended to elicit questions about Exemplar Theory (ET). In part this is because, like attractor models, ET also trucks heavily in the concept of emergent representations. However, there are some important differences between ET and the Phonological Latching Network (PLN) which prevent us from assuming a simple equivalence between the two. But first it is worth defining what is meant by ET. One recent definition is given in Frisch (2018):

> *"Exemplar theory is a theory of the representation and processing of categories in which stimuli are processed by comparing them to a set of previous experiences stored in memory"*

Therefore, ET proposes that categories are emergent in the sense that they are not stored but rather computed on the fly from a large number of stored exemplars. This can be contrasted

with “categorical” or “abstractionist” theories, which assume that speakers store only the categories themselves (c.f. Krämer 2012).

For a long time, ET and abstractionist theories of phonology were generally regarded as competitors. That is, it was assumed that phonological representations must be either abstract categories, or richly detailed episodic memories. Arguably ET’s strongest argument is its ability to deal with frequency effects, which are typically not expressible in abstractionist frameworks. This fact (amongst others) led some to conclude that abstractionist theories must be fundamentally incorrect (e.g. Bybee 1999).

However, in recent years this antagonism has given way to a more conciliatory approach. For example, Pierrehumbert (2016) makes a strong case that both episodic and abstract phonological representations are necessary for a complete account of phonological behaviour. Amongst the pieces of evidence given in support of abstract representations, is the observation that type frequency, rather than token frequency, is more predictive of phonological behaviour in adults. And as Pierrehumbert notes: *“Type frequency can only be defined by forming generalizations over an abstract phonological code, rather than directly over the surface realizations.” (p.11)*

This conciliatory approach adopted by Pierrehumbert was pre-empted by others, most relevantly Nguyen *et al* (2009), who cited attractor dynamics as a way of breaking the apparent deadlock between exemplar and abstract models of phonology. In addition to reviewing evidence for both episodic and abstract phonological representations, Nguyen *et al* also argue that point-attractors exhibit the crucial property of chopping up (or quantizing) a continuous space into a set of quasi-discrete categories. To a first approximation at least, this seems like a plausible mechanism for incorporating both fine-detailed episodic memories and abstract categories into the same system.

So, given that the PLN exploits point-attractors as memories, does this mean that the PLN is also a way incorporating both episodic and abstract representations? The answer is “no”, or at least “not quite”. In the PLN there are no exemplars to speak of - the system is moving through a small set of quasi-discrete attractor basins, which are understood to represent phonological categories (i.e. phones). And although the system does exploit the continuity of the state space in order to encode similarities or relationships between those phones, there is nothing we can

point to in the system as a kind of exemplar. In one sense then, the PLN is closer to an abstractionist theory, because what it stores are categories and the relationships between them.

Nevertheless, this observation does not entail that the PLN is necessarily at odds with ET either, because the PLN itself provides no strong prediction as to how the attractors in the system came to be. In isolation, the model is ambivalent as to whether the attractors are genetically specified or the result of repeated exposure to stimuli (or both, or something else entirely). This is relevant in light of recent work by Boboeva *et al* (2018), which allows us a tantalizing speculation as to how the PLN could be reconciled with a dualistic architecture that incorporates both episodic and abstract representations. Specifically, they studied the behaviour of Potts networks under very high memory load, i.e., networks where the number of memories is too great for the network to be able to reliably retrieve all of them. In these contexts, when the network fails to retrieve a cued memory, it tends to instead retrieve a similar memory which shares some (but not all) of the properties of the cued memory. Simplifying the results of Boboeva *et al* somewhat, this means that, under certain conditions, a high-load network causes related memories to cluster together, in the sense that the network loses the ability to distinguish among the memories in the cluster, but can still distinguish the cluster itself from other clusters. Thus, this spontaneous clustering presents a general mechanism for the emergence of categories from episodic memories. However, it should be emphasized that, because of the clustering, the episodic details are largely lost in the high-load network. Thus, it is ultimately the newly emerged categories which would drive subsequent latching behaviour in the network. This entails that if the lost episodic details are to have any role in cognitive behaviour, then they must simultaneously be stored in a lower-load network. To put this in context, it is quite conceivable that the memories in the PLN could be the result of such spontaneous clustering, in which case the PLN itself would represent only the abstract portion of phonological memory. Episodic details would ultimately have to be stored elsewhere (e.g. the lexicon).

This story is quite congruent with the view expressed by Pierrehumbert (2016), that both episodic and abstract representations play a role in phonological behaviour. Moreover, Pierrehumbert's observation that token frequency is critical for pre-lexical infants, while type frequency is critical for adults, seems to fit the general story implied by Boboeva *et al*'s findings, namely that children would first learn large amounts of episodic speech information, before consolidating that information into more abstract phonological categories, such as the categories in the PLN. This general hypothesis could provide an interesting avenue for future research using theoretical models such as the PLN.

1.3.6 References for Introductory Chapter

Alderete, J., & Tupper, P. (2018). Connectionist approaches to generative phonology. *The Routledge Handbook of Phonological Theory*. Routledge.

Alderete, J., Tupper, P., & Frisch, S. A. (2013). Phonological constraint induction in a connectionist network: learning OCP-Place constraints from data. *Language Sciences, 37*, 52–69.

Anderson, P. W., (1972) More is different. *Science.* New Series, Vol. 177, No. 4047. pp. 393-396.

Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G. E., & Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems* (pp. 9368-9378).

Bedau, Mark (1997). "Weak Emergence," Philosophical Perspectives, 11: Mind, Causation, and World, Oxford: Blackwell, pp. 375–399.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in neural information processing systems* (pp. 153-160).

Berkeley, I. S. (1997). A revisionist history of connectionism. *Unpublished manuscript*.

Brunel, N., Hakim, V., & Richardson, M. J. (2014). Single neuron dynamics and computation. *Current Opinion in Neurobiology, 25*, 149–155.

Bybee, J. (1999). Usage-based phonology. *Functionalism and formalism in linguistics,* 1, 211-242.

Chalmers, D. (1990). Why Fodor and Pylyshyn were wrong: The simplest refutation. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, Cambridge, Mass (pp. 340-347).

Chomsky, N. (2002) *On Nature and Language*. Cambridge University Press.

Chomsky, N., & Guignard, J. B. (2011). Beyond Linguistic Wars. An Interview with Noam Chomsky. *Intellectica, 56*(2), 21-27.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. MIT press.

Connors, B. W., & Gutnick, M. J. (1990). Intrinsic firing patterns of diverse neocortical neurons. *Trends in neurosciences*, *13*(3), 99-104.

Copeland, B. J., & Proudfoot, D. (1996). On Alan Turing's anticipation of connectionism. *Synthese, 108*(3), 361-377.

Crutchfield, J. P. (1998). Dynamical embodiments of computation in cognitive processes. *Behavioral and Brain Sciences, 21*(5), 635-635.

Dale, R., & Spivey, M. J. (2005). From apples and oranges to symbolic dynamics: a framework for conciliating notions of cognitive representation. *Journal of Experimental & Theoretical Artificial Intelligence, 17*(4), 317-342.

Dawson, M. R., Medler, D. A., & Berkeley, I. S. (1997). PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology, 10*(1), 25-40.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge, MA, USA: MIT Press.

Dennett, D. C., 1987, *The Intentional Stance*, Cambridge, MA: MIT Press.

Dewhurst, J. (2018) Computing Mechanisms Without Proper Functions. *Minds & Machines 28*: 569.

Dreyfus, H. L. (1972). *What computers can't do*. MIT Press.

Eberbach E., Goldin D., Wegner P. (2004) Turing's Ideas and Models of Computation. In: Teuscher C. (eds) *Alan Turing: Life and Legacy of a Great Thinker*. Springer, Berlin,

Edelman, S. (2017). Language and other complex behaviors: Unifying characteristics, computational models, neural mechanisms. *Language Sciences, 62*, 91–123.

Eliasmith, C. (1997). Computation and dynamical models of mind. *Minds and Machines, 7*(4), 531-541.

Elman, J. L. (1990). Finding structure in time. *Cognitive science, 14*(2), 179-211.

Feinerman, O., Pinkoviezky, I., Gelblum, A., Fonio, E., Gov, N. S. (2018) The physics of cooperative transport in groups of ants. *Nature Physics*: 1745-2481.

Fodor, J. A. (1975). *The language of thought*. Harvard university press.

Fodor, J. A. (1981) The Mind-Body Problem. *Scientific American, 244*: 114–125.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1-2), 3-71.

Frisch, S. A. (2018) Exemplar theories in phonology. In: Hannahs, S. J., & Bosch, A. (Eds.). *The Routledge Handbook of Phonological Theory*. Routledge.

Gafos, A. I., & Benus, S. (2006). Dynamics of phonological cognition. *Cognitive science, 30*(5), 905-943.

Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. John Wiley & Sons.

Haken, H. E., & Stadler, M. E. (1990). Synergetics of cognition: *Proceedings of the International Symposium* at Schlo–S Elmau, Bavaria.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hoel, E. P. (2017). When the Map Is Better Than the Territory. *Entropy* 19:188.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational properties. *Proc. Nat. Acad. Sci*. (*USA*) 79, 2554-2558.

Hopfield, J. J. (1984) Neurons with graded response have collective computational properties like those of two-sate neurons. *Proc. Nat. Acad. Sci. (USA) 8*1, 3088-3092.

Hopfield, J. J. (2007). *Hopfield network.* Scholarpedia, 2(5):1977.

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. *Experimental approaches to phonology. In honor of John Ohala,* 25-40.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol 58*(6): 1233-1258.

Krämer, M. (2012). *Underlying representations*. Cambridge University Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Kleene, Stephen C. (1956)[1951]. Representation of Events in Nerve Nets and Finite Automate. *Automata Studies, Annals of Math. Studies*. Princeton Univ. Press. 34.

Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G., & Smith, N. A. (2017). What Do Recurrent Neural Network Grammars Learn About Syntax?. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 1, Long Papers (pp. 1249-1258).

Lakoff, G (1988). A Suggestion for a Linguistics with Connectionist Foundations, in Touretzky, D ed. *Proceedings of the 1988 Connectionist Summer School*. UC Berkeley

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh*. New york: Basic books.

Luisi, P. L. (2002) *Foundations of Chemistry 4*: 183–200.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive psychology*, *37*(3), 243-282.

McCoy, B. (2010) *Ising model: exact results*. Scholarpedia, 5(7):10313.

Minsky, M. L. (1954). *Theory of neural-analog reinforcement systems and its application to the brain model problem* (PhD Thesis) .Princeton University.

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. MIT press.

Naim,. M., Boboeva, V., Kang., C. J., Treves, A. (2017) Reducing a cortical network to a Potts model yields storage capacity estimates. arXiv:submit/2036185 [q-bio.NC]

von Neumann, J. (1951). *The general and logical theory of automata*. 1951.

von Neumann, J. (1956). Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies*, *34*, 43-98.

Nguyen, Noël & Wauquier, Sophie & Tuller, Betty. (2009). The dynamical approach to speech perception: From fine phonetic detail to abstract phonological categories. *Approaches to phonological complexity*, 5-31.

Newell, A. & Simon, H. A. (1963). GPS: A Program that Simulates Human Thought, in Feigenbaum, E.A.; Feldman, J. (eds.*), Computers and Thought*, New York: McGraw-Hill.

Newell, A. & Simon, H. A. (1976). Computer Science as Empirical Inquiry: Symbols and Search, *Communications of the ACM, 19* (3): 113–126.

Olazaran, M. (1996). A Sociological Study of the Official History of the Perceptrons Controversy. *Social Studies of Science, 26*(3), 611–659.

van Oostendorp, Marc, (2008) Incomplete devoicing in formal phonology. *Lingua 188,* no. 9:1362.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language 95*(1), e41-e74. Linguistic Society of America.

Piccinini, G. (2004). The First computational theory of mind and brain: a close look at McCulloch and Pitts's "logical calculus of ideas immanent in nervous activity". *Synthese*, *141*(2), 175-215.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. OUP.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics 2*(1), 33-52.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73-193.

Poeppel,D., Embick, D. (2005). Defining the relationship between linguistics and neuroscience. In A. Cutler ed. *Twenty-first century psycholinguistics: Four cornerstones*, Lawrence Erlbaum.

Port, R.F. Leary, A.P. (2005) Against formal phonology. *Language* 81.

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science*, *275*(5306), 1604-1610.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rumelhart, D. E. (1998). The architecture of mind: A connectionist approach. *Mind readings*, 207-238.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, *1*(45-76), 26.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, D. E. & McClelland, J. L. (1987) Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach.* 3rd edition. Malaysia; Pearson Education Limited,.

Schneider, W. (1987) Connectionism: Is it a paradigm shift for psychology? *Behavior Research Methods, Instruments, & Computers 19*.

Searle, J. R., 1992, *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Serra, R., & Zanarini, G. (1990). *Complex Systems and Cognitive Processes*.

Shagrir, O. (2006). Why we view the brain as a computer. Synthese, 153(3), 393-416.

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, *26*(Supplement), 137-161.

Treves, A., & Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain?. *Network: Computation in Neural Systems*, *2*(4), 371-397.

Tuller, B., Case, P., Ding, M., & Kelso, J. A. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human perception and performance*, *20*(1), 3.

[Web of Stories - Life Stories of Remarkable People] (2016) Marvin Minsky - The problem with perceptrons [Video File]. Retrieved from https://www.youtube.com/watch?v=QW_srPO-LrI

Paper 1

"Our theoretical objective is not dependent on the assumptions fitting exactly. It is a familiar stratagem of science, when faced with a body of data too complex to be mastered as a whole, to select some limited domain of experiences, some simple situations, and to undertake to construct a model to fit these at least approximately.

Having set up such a model, the next step is to seek a thorough understanding of the model itself. It is not to be expected that all features of the model will be equally pertinent to the reality from which the model was extracted. But after understanding the model, one is in a better position to see how to modify or adapt it to fit the limited data better or to fit a wider body of data and when to seek some fundamentally different kind of explanation."

(Kleene 1956 [1951])

2 The Phonological Latching Network

Joe Collins

2.1 Introduction

As noted the introductory chapter, the overarching assumption of this volume is that an attractor neural network can function as a L*inking Hypothesis* (Poeppel & Embick 2005) for linguistics and neuroscience. 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.

2.2 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 and Choo 1992) and memory capacity (e.g. Tsodyks and Feigelman 1988). However, it is clearly inadequate for modelling natural language grammar, which requires (minimally) the ability to define relationships between discrete elements.

Latching networks can be understood as an attempt to introduce between-element dynamics into an attractor network. Fundamentally, they 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.

2.2.1 The Potts Unit

The notion of fatigue in a latching network requires that individual units have an inactive state. Thus, the model differs from the binary-unit Hopfield network in being comprised of multi-state (or "Potts") units. 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:

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

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

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

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 (eqns. 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 2.2.3.1 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 2017). Therefore, the *w*-term is intended to subsume the internal dynamics of each cortical patch. Using 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.

2.2.2 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, i.e. it "rolls into the valley" (**Error! Reference source not found.**). 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, e.g. 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[15].

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).

[15] 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.

2.2.3 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 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[16]. 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. 2017). 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 found 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 and 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

[16] 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.

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.

2.2.3.1 Building Phones

Unlike neural networks typically employed in machine learning and connectionist frameworks, the PLN is not subject to any form of supervised learning. 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.

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: Parameters and phonological inventory). 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 not 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:

$$\mu := \{\varphi_1^{\mu}, \varphi_2^{\mu}, \dots, \varphi_M^{\mu}\} \quad (7)$$

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.

$$\mu^Q := \varphi_1^{\mu^Q}, \dots, \varphi_m^{\mu^Q}, \dots, \varphi_M^{\mu^Q} \quad (8)$$

The order is always derived from the inventory on page . Also note that μ^{aud} and μ^{mot} always contain the same elements, but in the reverse order, i.e., 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:

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

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 a logarithmic 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.

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

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 and Treves (2012), are generated using sparse[17] 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*.

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

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.

[17] i.e. 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:

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

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.

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

The Kronecker delta is a function which equals 1 when its arguments are the same, but 0 otherwise. The last term represents the influence of each phones' 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:

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

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:

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

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 3. 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.

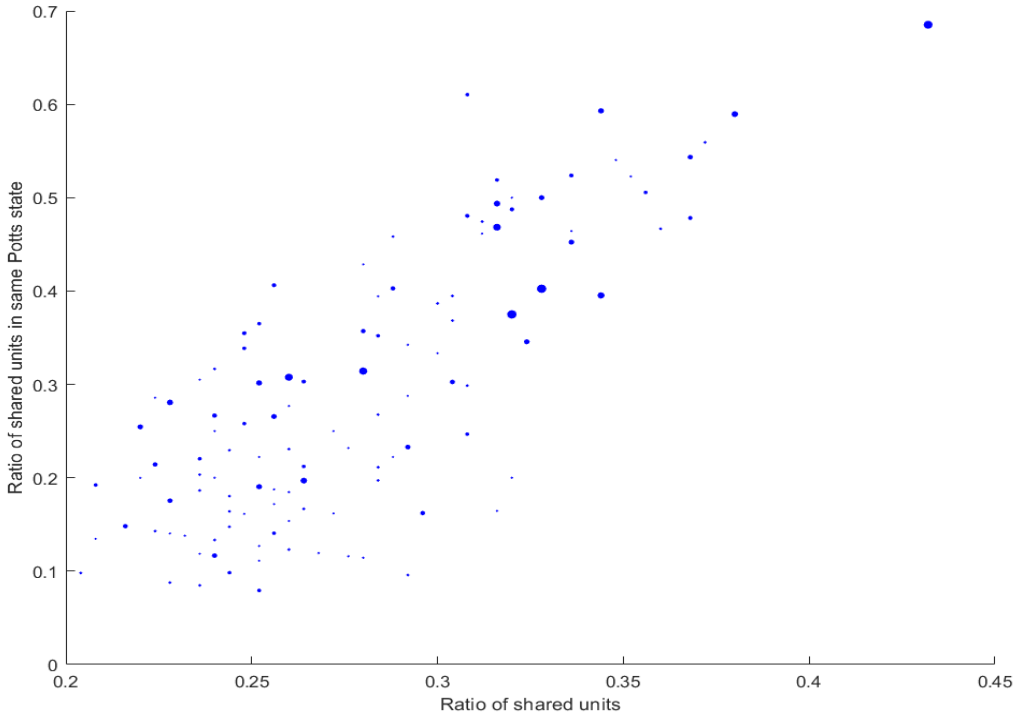

Figure 3:Overlap of memories produced by feature super-position. The size of each circle indicates the total number of attested transitions between the two memories during the simulations.

2.3 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 (see appendix), the latching strings from 125 trials, representing 8 different grammars, 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, i.e. 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.

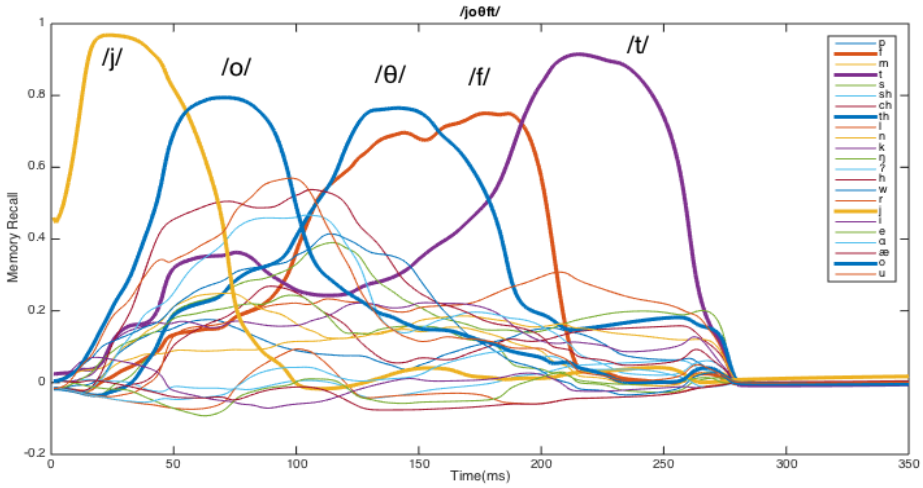

Figure 4: Example of a latching string

2.3.1 Segmental-OCP

In its general form, the Obligatory Contour Principle (OCP) requires that there be some minimum degree of difference between adjacent objects. 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 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[ʃə̲z] etc.). However, one grammar did spontaneously produce the string [kɲut͡ʃsθu], where the transition from /t͡ʃ/ 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 /t͡ʃ/ (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 /t͡ʃ/ (28%). This supports the hypothesis the absence of /s/->/ʃ/transitions in the PLN is an OCP effect, while /s/ and /t͡ʃ/ are dissimilar enough to fall within the "Goldilocks" zone.

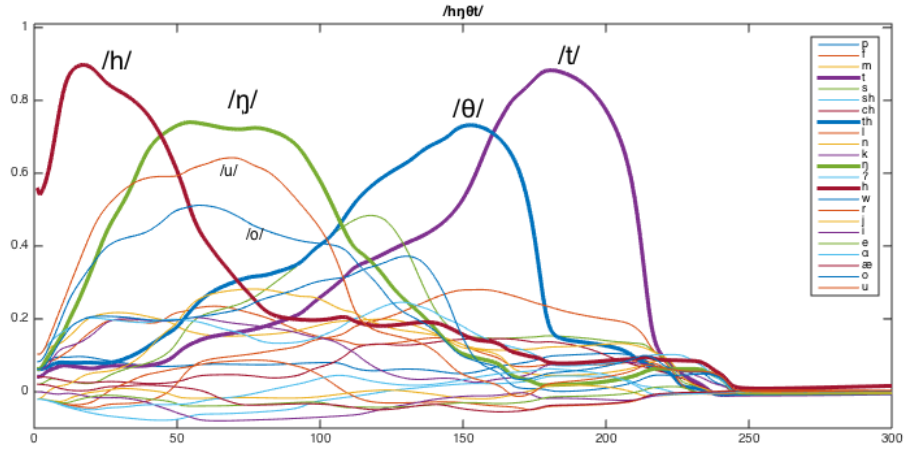

Figure 5: The /θ/ and /t/ phones are similar in both their manner and place of articulation, but are still a possible transition for the PLN.

2.3.2 Assimilation

Processes in which segments become more similar to their neighbours – in terms of their feature specification – are extremely common cross linguistically (cite). 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 6.

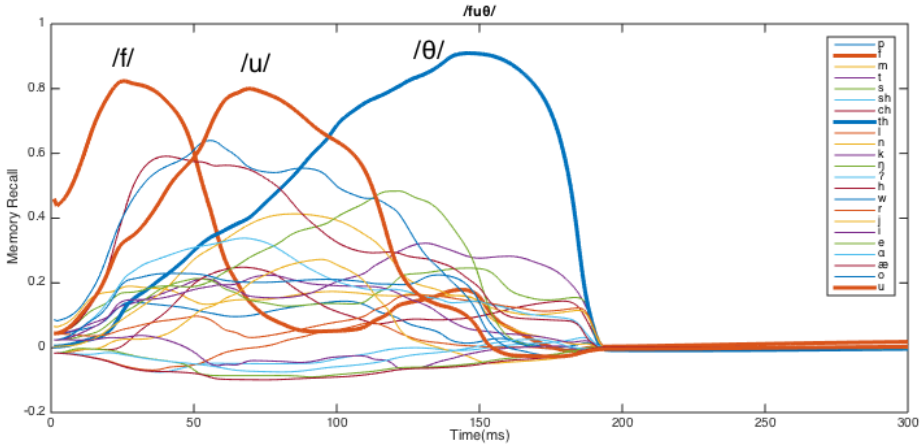

Figure 6: The /f/ and /u/ share the feature [round], so the first transition is interpreted as an instance of place assimilation.

2.3.2.1 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.

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).

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 purly 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, e.g., 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 pair, it is the feature with the highest weight during phone generation (section 2.2.3.1) which appears to assimilate at above chance rate, while the feature with the lower weight assimilates at a below chance weight.

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.

2.3.2.2 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:

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**

That the CONTINUANT and NASAL features exhibit assimilation is broadly in keeping with the phonological literature (e.g. intervocalic spirantization and vowel nasalization). 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 sections 2.3.3), 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 needn't be regarded as a consequence of spreading or assimilation *per se*, but rather of sonority sequencing.

2.3.3 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 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 2), according to the *least* sonorant nucleus required when parsing (Table 3).

Table 2: Sonority scale

Vowels	***Glides***	***Liquids***	***Nasals***	***Obstruents***
0	1	2	3	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 the table below:

Table 3: Example sonority scores

String	*Syllable parse*	*Least Son. Nuc.*	*Sonority Score*
"ʃ l o"	ʃlo̲	o	0
"l ʃ 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

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[18]. The sonority scores for different string lengths, both from the PLN and the baseline, are given in the figure below:

[18] 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.

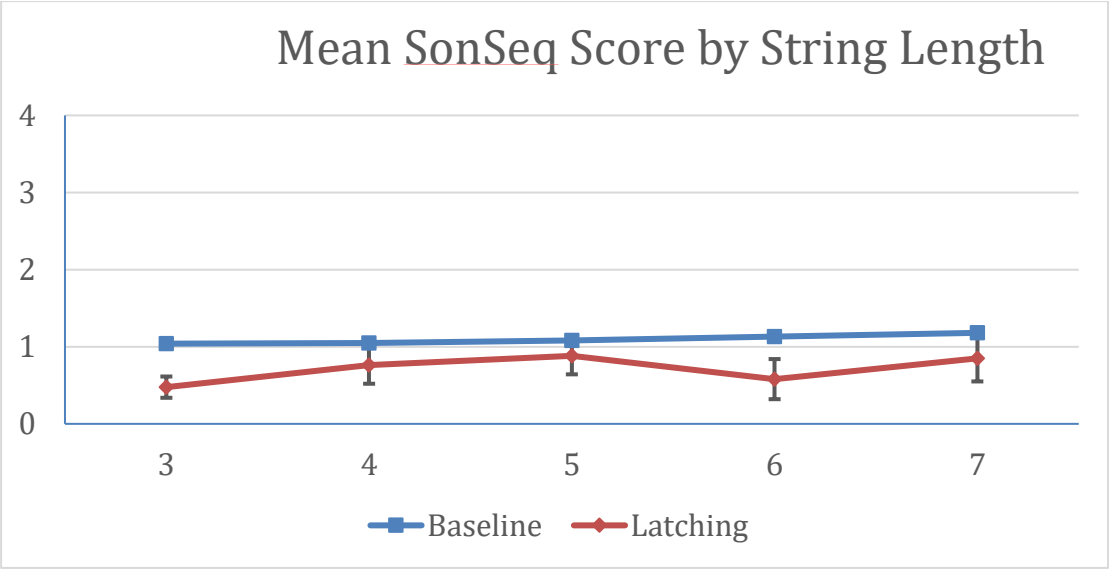

Figure 7: Sonority Sequencing score for latching strings (red) versus random baseline (blue).

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.

2.3.3.1 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 certain "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 2.3.2.2, 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[19].

The data in Table 4 show the overlaps across these categories from a single grammar of the PLN. As we might expect, the average overlap 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:

Sonority categories	*Average shared units %*	*% of 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 4: Overlap across sonority categories within a single grammar.

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

This pattern, taken with the high rate of SONORANT and VOCALIC assimilation (section 2.3.2.2), 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.

2.4 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 2.3.1, 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 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 effects 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 (e.g.) 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 an /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 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[20].

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 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 2.2.3.1, is somewhat volatile, insofar as it frequently produces grammars with obvious pathologies (failing to retrieve phones, mixed-state 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.

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

2.5 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. In general, we should be cautious about drawing direct correlations between the ontologies of neutrally inspired models and formal linguistic theories[21]. However, the PLN could nonetheless inform the discussions and

[21] See the second paper in this volume for more discussion of this point.

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, etc). 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 (Chomsky 2005, etc.). 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 this 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 (Mielke 2008, etc.), 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, neutrally-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 "interdisciplinary 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.

2.6 Bibliography

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

Block, N. (1995) The Mind as the Software of the Brain in E. E. Smith and D. N. Osherson eds. *An Invitation to Cognitive Science vol. 3: Thinking (2nd edition)*. MIT Press.

Boboeva, V., Brasselet, R., & Treves, A. (2018). The capacity for correlated semantic memories in the cortex. *Entropy*, *20*(11), 824.

Bouchard, K. E., Mesgarani, N., Johnson, K., and Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature* 495, 327–332.

Chomsky, N. (2005). Three factors in language design. *Linguistic inquiry*, *36*(1), 1-22.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston and M. E. Beckman (eds.) *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press. 283-333.

Hickok, G., Poeppel, D. (2007) The cortical organization of speech processing. *Nature Reviews Neuroscience* 8, 393-402.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational properties. *Proc. Nat. Acad. Sci*. (*USA*) 79, 2554-2558.

Hopfield, J. J. (1984) Neurons with graded response have collective computational properties like those of two-sate neurons. Proc. Nat. Acad. Sci. (USA) 81, 3088-3092.

Ising, E. (1925) Beitrag zur Theorie des Ferromagnetismus (Contribution to the Theory of Ferromagnetism). *Zeitschrift für Physik*. 31.

Kang,C. J., Naim, M., Boboeva, V., Treves, A. (2017) Life on the edge: Latching Dynamics in a Potts neural network. *Entropy*, 19, p.468.

Kanter, I. (1988). Potts-glass models of neural networks. *Phys. Rev. A* 37, 2739.

Lakoff, G (1988). A Suggestion for a Linguistics with Connectionist Foundations, in Touretzky, D ed. *Proceedings of the 1988 Connectionist Summer School.* UC Berkeley

Marr, D.; Poggio, T. (1976). From Understanding Computation to Understanding Neural Circuitry. *Artificial Intelligence Laboratory. A.I. Memo*. MIT.

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

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*. 343, 1006–1010.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford University Press.

Naim,. M., Boboeva, V., Kang., C. J., Treves, A. (2017) Reducing a cortical network to a Potts model yields storage capacity estimates. *unpublished work* arXiv:submit/2036185 [q-bio.NC]

Nasrabadi, N. M., Choo, C. Y. (1992) Hopfield network for stereo vision correspondence, in *IEEE Transactions on Neural Networks*, vol. 3, no. 1, pp. 5-13.

Odden, D. (2013). Formal phonology. Nordlyd, 40(1), 249-273.

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

Poeppel,D., Embick, D. (2005). Defining the relationship between linguistics and neuroscience. In A. Cutler ed. *Twenty-first century psycholinguistics: Four cornerstones*, Lawrence Erlbaum.

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science, 275*(5306), 1604-1610.

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

Treves, A. (2005) Frontal latching networks: A possible basis for infinite recursion. *Cognitive Neuropsychology*. 22(3-4).

Tsodyks, M. V, Feigelman, M. V. (1988) The Enhanced Storage Capacity in Neural Networks with Low Activity Level. *Europhysics Letters*, vol. 6, no. 2.

2.7 Appendix: Parameters and phonological inventory

The results in section 2.3 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. Note that the ordering of the features in the table reflects the weighting of the features within each sub-network:

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

Paper 2

"What we should do is to pursue all approaches to the brain as best we can, seeing what one can learn from the other. The discoveries of the chemist provided certain guidelines for the revolution in physics, and it could turn out that the discoveries of the cognitive scientists will do the same for the brain sciences. Or, the latter might develop some new approach to properties of language and other aspects of cognition that would suggest new directions for the cognitive sciences. One can have no doctrines about such matters."

(Chomsky 1994)

3 Digital Grammar and Analogue Brains: A Defence of Formal Linguistics

Joe Collins

3.1 Introduction

Various critics have made the claim that the formal linguistics is at odds with the prevailing view of the mind/brain (Edelman 1992; Lakoff 1993; Port & Leary 2005, etc.). This critique often centres around the discrete (or digital) nature of linguistic formalisms, which are assumed to be incompatible with a nervous system which is widely thought to be continuous (or analogue) in character.

The argument against discrete, formal linguistic analysis is articulated perhaps most forcefully in Port and Leary 2005, who write:

> *"Generative phonology seems to have overlooked that, in order for there to be a formal system, there must be something or someone to execute the rules, whether a computer, a linguist or the subconscious brain. Unfortunately there is no evidence whatever that human brains automatically and unconsciously implement any formal system at all." (Port & Leary 2005)*

Port and Leary are arguing from a philosophical stance of *Dynamicism*, which contends that the brain is a form of control system, and that notions like "computation" or "symbols" are fundamentally inapplicable (see Port & van Gelder 1995). This stance can be contrasted with a synthetic approach to theories of the mind/brain, which contends that dynamical and symbolic approaches to cognition can be reconciled with a sufficiently nuanced definition of what each framework contributes to our understanding, e.g.:

> *"One consequence of such a synthetic framework is that it renders entirely moot arguments between symbolic and dynamical approaches to cognition and also debates about discrete versus analog embodiments of computation. In short there simply is no antagonism between a dynamical view and a computational view as long as one is willing to fairly assess where each field is and be willing to extend the notions each brings to the problems of cognition."* (Crutchfield 1998)

This paper adopts the synthetic position advocated by Crutchfield. I claim that, not only are formal linguistic models commensurable with non-discrete systems, such as the brain, but a formal model can in many cases provide a better account of *why* the grammar behaves the way it does.

A key flaw in Port and Leary's argument is their assumption that because formal grammars are discrete, that means that they can only be implemented in a system which is also a "*discrete-time symbol processing device*" (p.937). They then note that this implementation is implausible in the context of a seemingly analogue system like the human brain[22], and thus conclude that the formal grammar must be fundamentally incorrect. However, this reasoning is ultimately an argument against one particular implementation of a formal grammar, and Port and Leary do not give any convincing argument as to why formal linguistic models presuppose such an implementation. Indeed, the performance/competence distinction adopted by many formal linguists (and maligned by Port and Leary (p.932)) deliberately renders linguistic models agnostic with regards to their implementation in human wetware.

As a rebuttal to Port and Leary's argument, we need only consider an alternate approach to implementing a formal grammar - one which does not suppose any discrete or otherwise biologically implausible elements. This is the topic of Section 3.2, which will demonstrate in detail how a discrete, formal phonological model of grammar can be equated to a neural-attractor model which is continuous. This provides an explicit example of how a complex neural system can appear to be both continuous or discrete, depending one's level of analysis. Additionally, this model is designed to capture the phenomenon of incomplete devoicing – cited by Port and Leary as an argument against discrete phonology – which also allows us some insight into relationship between the rigid characterization of formal phonological grammars and the gradient reality of phonetic details.

[22] For example, although action potentials are "digital" in the sense of being an all-or-nothing event, neurons cannot transmit a binary code because (unlike the transistors of a CPU) their firing rates are not synchronized. Moreover, the calculus of post-synaptic potentials is unambiguously continuous in nature, depending not only on the number, efficacy, and distribution of synapses, but also on the morphology of the membrane itself.

Section 3.3 builds on the model in section 3.2, and argues for the significance, and even necessity, of the formal/linguistic level of analysis given a continuous attractor model. That is, the formal model is not rendered obsolete by its implementation in a continuous model of the brain. This argument will depend on the application of *Effective Information (EI)* (Tononi & Sporns 2003; Hoel *et al* 2013) to the model from section 3.2. *EI* is an information theoretic measure which attempts to quantify the extent to which a model informs us about the underlying causal structure of a given system. Crucially, it can be shown that the formal/linguistic level of analysis has a higher *EI* than the neural model. In Hoel's terminology, this is an example of *Causal Emergence*, and it represents a profound and robust cornerstone for understanding the role of formal linguistic models within a broader, cross-disciplinary understanding of cognitive function. Specifically: the formal analysis elucidates the causal structure of the grammar.

Finally, it should be noted that there are multiple prongs to Port and Leary's paper which will not be dealt with in detail here. These include the issues of universal phonetics and discrete vs continuous time. The former is arguably tangential given that it is not a position endorsed by all, or even most phonologists in the 21st century. While the latter appears to follow entirely from Port and Leary's own debatable interpretation of phonological theory[23].

3.2 Macro vs. Micro

A cornerstone of the counter-argument presented here is a rejection of naïve reductionism, in favour of the idea that complex systems can be analysed at different levels of abstraction, and that the phenomenology of these different levels can seem radically different.

Certainly, this is not a novel idea. The importance of understanding the mind/brain at multiple levels of abstraction has been stressed by various commenters, e.g. Marr's tri-level analysis (Marr 1982) and Smolensky's integrated connectionist/symbolicist architecture (Smolensky & Legendre 2006) and many others besides (Cruchfield 1998; Dale & Spivey 2005; Edelman 2008). Moreover, the various conceptual and philosophical underpinnings of emergent properties in scientific theories are so thoroughly dissected by philosophers that they need not be rehashed here (see O'Connor & Wong 2015). For our purposes we need only acknowledge

[23] Certainly, the claim that phonological theories depend on "*serial, discrete time*"(p932.), seems deeply at odds with the widespread acceptance of (e.g.) regressive rules and hierarchical prosodic domains, which are clearly not bound by the laws of serial time in any conventional sense.

that complex things may possess properties that their individual components do not (and vice-versa).

With this in mind, this paper will consider a case where a phonological grammar is a macro-level abstraction of a substrate which is non-symbolic at the micro-level. That is, the components of the system are clearly continuous when viewed up close, but the behaviour of the system as a whole appears to be discrete, thereby demonstrating how a digital grammar can be implemented in an analogue system.

The first question to be answered is: what form might such a micro-level take? There are, of course, many potential answers to this question. Here however, we will consider the case where the micro-level is an attractor neural network, which has been specifically designed to implement a simplified (or "toy") phonological grammar.

3.2.1 Attractor Model

Attractor networks compose a class of complex dynamical systems which have been posited to explain certain computational properties of nervous systems (Hopfield 1982, Amit 1989, Conklin & Eliasmith 2005, Kropff & Treves 2008).

To understand how a dynamical system can have computational properties, we need to think of the system as being a kind of behaviour over a *state space*, i.e. the set of all possible configurations of the systems, expressed such that each configuration is a unique point in the space. If the individual states of a system are understood as representing information (e.g. numbers), then any movement from one point in the space to another can be understood as implementing a function (e.g. arithmetic). Therefore, computation is ultimately a form of system dynamics.

Attractor networks are a class of dynamical system, named after the so-called attractor states which characterize the network dynamics. In simple terms, attractors are states which the network will always tend to over time. This is often visualized as if the state space were a landscape of peaks and valleys, and attractor states are the lowest points of the valleys which the network always rolls down into. Consequently, attractor dynamics can be treated as a model of memory, since the attractor states can always be retrieved by the system.

Moreover, one relevant property of these memories is that they are effectively discrete (Hopfield 1982), since the basins of attraction are non-overlapping. In effect, they quantize or

"chop up" the continuous space into distinct regions. For this reason, attractor models are a means of crossing the continuous/discrete divide which is taken as a fundamental sticking point for criticisms of formal linguistics (e.g. Port and Leary 2005).

What follows then is an attractor model where the individual attractors are taken to represent discrete phonological entities (in our case, phones), and the dynamics of the system are such that certain attractors are only stable under certain contexts. That is, there is some context in which attempting to retrieve one phone will result in the retrieval of a different phone – a simple form of context dependent allophony (see also Gafos & Benus 2006).

3.2.2 Incomplete Devoicing

Clearly, attempting to implement an entire phonological system in an attractor network would exceed not only the space restrictions of a single paper, but most likely our current understanding of both neuroscience and linguistics. A more fruitful goal for our purposes then, is to restrict our focus to a single phonological pattern: that of final devoicing, where a voiced obstruent loses its voicing in a syllable coda or word-final position. This is a pattern which is both widely attested (in familiar languages such as German, Russian, Dutch, etc.) and is arguably a prototypical example of a phonological process, which finds a home on many Phonology 101 syllabi. Consider the following examples from Dutch:

1. Dutch (from Van Oostendorp 2008):

a) *kwaa*[t] 'angry (PRED.)' - *kwa*[də] 'angry (ATT)'

b) *laa*[t] 'late (PRED.)' - *la*[tə] 'late (ATT)'

In the case of 1a, the final [t] appears as a [d] when a vowel is suffixed to the stem, while in 1b, the final [t] remains as a [t], even in the presence of the vowel. The standard analysis of these facts is that the [t] in 1a is derived from an underlying /d/, which then devoices in a word final position, while the final [t] in 1b is derived from an underlying /t/, and therefore appears as such regardless of whether it is in word final position.

In fact, this analysis generalises to all obstruents in Dutch – the voiced variants never appear in a final position. Moreover, similar patterns appear in many other languages. Thus, the pattern is not an arbitrary fact about /t/ and /d/ in Dutch, but rather one aspect of a deeper insight into

the relationship between laryngeal specification and vowels, syllables and prosodic structure more generally.

Yet a number of studies have shown that, contrary to the classical phonological analysis of the phenomenon, the contrast between devoiced and underlyingly voiceless segments is not fully neutralized. That is, there exists a small difference in voicing between those segments which are underlying voiceless and those which are derived via devoicing. The difference is not salient enough to be consistently perceivable by speakers, but there is nonetheless evidence that speakers are able to "guess" whether they are hearing the voiceless or devoiced segment at an above-chance rate (see Roettger *et al.* 2014).

Under the classical, modular view of the phonetics-phonology interface, incomplete neutralization poses a problem since it seems to suggest that the phonetics has access not only to the surface phonological form, but also to the underlying form. Clearly, this breaks the entire premise of the classical-modular system, which treats information encapsulation and shallow inputs as critical properties (Fodor 1983).

Various solutions to account this fact have been proposed within the phonological literature (e.g. Van Oosteendorp 2008; Iosad 2012), however the validity of these approaches will not be discussed here. Instead, this article merely argues that the existence of gradience in phonological categories does not undermine the enterprise of formal phonology. Rather than being interpreted as evidence that discrete categories don't exist, the gradience is better understood as "wiggle room" which emerges from implementing discrete categories in a continuous system.

This interpretation stands in contrast to, e.g., Port & Leary (2005), who argue that this gradience constitutes evidence that the categories posited by phonologists have no neurological or psychological reality. That is to say, there are no "symbols" in the brain.

3.2.3 Constructing a Model

To demonstrate how gradient final devoicing can be accomplished in an attractor network, we can construct a minimal "toy" grammar, consisting of 6 phones (/p/,/t/,/k/,/b/,/d/,/g/), which distinguish 3 places of articulation ([LAB],[COR],[DOR]) and a 2-way voicing distinction, as well as a distinction between final and non-final positions.

For a formal analysis of this toy grammar, we would need a rule to trigger the devoicing. Using traditional phonological notation, we could write:

2. $\begin{bmatrix} voice \\ \alpha place \end{bmatrix} \rightarrow [\alpha place] / \_\#$

This ensures that /b/,/d/ and /g/ will be realized as [p],[t] and [k], respectively, in a final context.

In dynamical systems' terms, this grammar can be thought of as a landscape of 6 attractors, where each attractor basin corresponds to a phone. Additionally, the system should be sensitive to information about final/non-final position, such that a final position causes the voiced phone attractors to become unstable, forcing the network instead into the basin of a voiceless counterpart.

Our system for realising this grammar will consist of 40 individual units, loosely analogous to neurons, connected to one another via symmetrical "synapses" of varying efficacies. The size of the network is approximately the smallest still large enough to allow successful retrieval of all 6 memories in the system.

Each unit, *i*, in the network obeys a simple rule, whereby its state at any given moment is a continuous value, σ, between 0 and 1, representing "inactive" and "fully active" respectively. The value of σ^i is given by equation 1:

$$\sigma^i = \theta\left(1 - e^{-h^i T}\right) \quad (1)$$

The summed, weight input from the other units to unit i is denoted by h^i. The symbol θ represents a threshold function whose output is 1 if h^i falls within some pre-defined values[24] and 0 otherwise. And *T* is a gain parameter used to regulate the activity level of the network.

The value of h^i is given by equation 2, where w_{ij} represents the efficacy of the connection between any two units *i* and *j*, and *n* is the total number of units in the network.

[24] These values are set to accomplish two things: Firstly, to ensure that very small inputs cannot cause a unit to activate. And secondly, to ensure that large inhibitory inputs cannot cause a unit to go into a negative state (i.e. $\sigma^i < 0$).

$$h^i = \sum_{j \neq i}^{n} w_{ij}\sigma^j \tag{2}$$

Thus, the exact value a given unit takes is determined by the weighted sum of the inputs from the other units, modulated by a squashing function. Consequently, all the units in the network are engaged in a feedback loop where each unit is adjusting its activity in accordance with every other unit. If the network arrives at a state where the activity levels of every unit are optimally in balance with each other, then the network will simply remain in that state indefinitely. These states are the attractors, and which of the network states constitute attractors is determined entirely by the configuration of synaptic efficacies between the units. Therefore, the process of "teaching" the network a set of memories means adjusting the efficacies until the states representing each memory are attractors. This is accomplished with a simple Hebb-like rule, given in equation 3:

$$w_{ij} = \frac{1}{na(1-a)} \sum_{\mu}^{m} \sum_{i \neq j}^{n} \left(\xi_i^\mu - a\right)\left(\xi_j^\mu - a\right) \tag{3}$$

Here, a is a value between 0 and 1 representing the average sparsity of the memories in the network. The parameter m is the total number of memories (6 in our case) each denoted by a value of μ, and ξ denotes a pattern of activity, such that ξ_i^μ denotes the state of the ith unit in memory μ.

Encoding our phones in this way presupposes that they have already been defined as points in the state space, i.e. each phone must be represented by a specific pattern of activity in the network, before each pattern can be made into an attractor by manipulating the connections between units. The process of defining each phone was accomplished by first creating pseudo-random sparse patterns for each phonological feature (LAB,COR,DORS, and VOI), then superimposing the relevant features for each phone (e.g. /b/ is LAB+VOI, etc.). Each phone was also superimposed with a unique "noise" pattern, to ensure that each memory was distinct enough to enable correct retrieval.

This method of generating phones ensures that the overlap between the memories in the network reflects shared phonological properties. For example, both /p/ and /b/ will share a number of units, deriving from the shared feature LAB. This is essential for ensuring that when

the network cannot retrieve a voiced memory, it will instead default to the voiceless equivalent, rather than returning a random memory or mixed-state.

Finally, in order to control the effects of final and non-final position, we can add one additional unit to our network, which will be used to trigger the devoicing rule. This unit can be switched off or on by the experimenter. When active, it will inhibit those other units that are associated with the representation of voicing. This unit is a simplification loosely inspired by the hypothesis that prosodic information is transmitted across the brain, in the form of slower oscillations which can excite or inhibit local networks (Giraud & Poeppel 2012).

3.2.4 Results

If our grammar has been encoded successfully, then activation of the final-position unit should cause the attempted retrieval of a voiced phone, to instead result in the retrieval of its voiceless counterpart. This is indeed what we see in Figure 8, which demonstrates the retrieval of the coronal phones /t, d/ in various contexts:

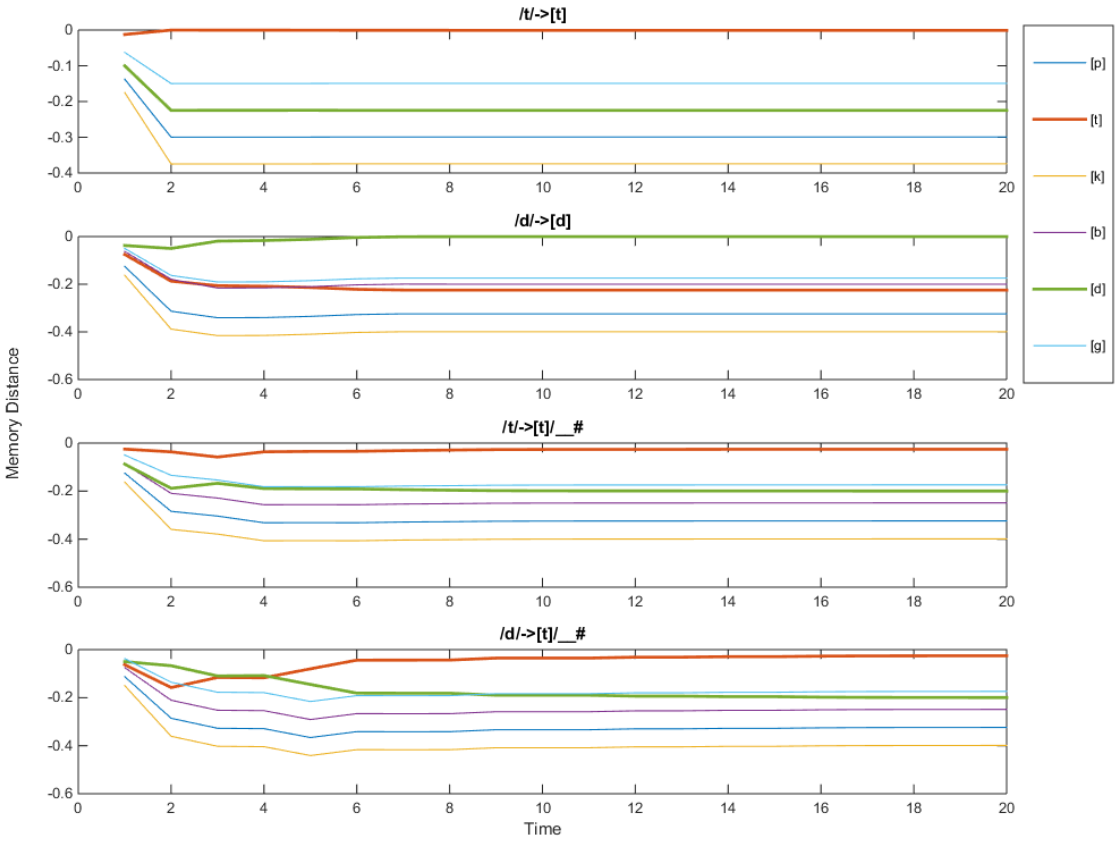

Figure 8: Network evolution during retrieval of coronals.

The horizontal axis shows time, while the vertical axis shows the distance of the network state from a given memory. When the memory distance is equal to 0, it means that the network has arrived exactly at an encoded attractor.

The four plots each show the retrieval of a given memory, which can be interpreted as deriving a surface form from a given input. In each case, the network is given an ambiguous or incomplete pattern of one of the memories in the system at t=1, and as the system evolves over time, it proceeds to retrieve a (near-)complete memory. In the first 3 plots, we see the system retrieving exactly the memory it is told to retrieve, i.e. the input and surface form are one and the same. In the fourth plot however, the network is asked to retrieve /d/, but instead returns a /t/. That is, the network has devoiced a /d/ because the "final position" unit is active.

The simulations demonstrate how the model can exhibit discrete behaviour: it clearly tends towards one of the encoded memories, even when it begins in an intermediate state. This is the effect of the attractor basins quantizing or "chopping up" the state space. Additionally, the model is able to implement a simple context sensitive rule: the inhibitory signal is enough to push the model away from a "voiced" memory into its voiceless equivalent.

However, the model can also give us some insight into the gradience observed in final devoicing, and how this gradience relates to the apparently discrete attractors. Although the model clearly tends towards one of the encoded memories, closer examination reveals that the memory retrieval is slightly different depending on the context.

To help us visualize this difference, we can perform a simple multi-dimensional scaling algorithm to the trajectories of each memory retrieval, allowing us to plot the trajectories onto two, abstract dimensions.

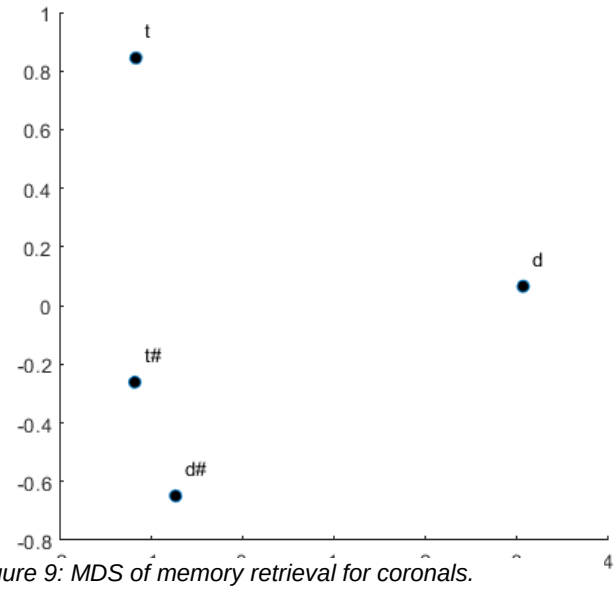

Figure 9: MDS of memory retrieval for coronals.

This gives us an abstract Euclidean distance between the four different phones. Crucially, we can see that the non-final segments are distinct from their coda equivalents, and while the coda position /t/ and /d/ are very similar to one another, they are in fact not the same. Thus, although the system clearly exhibits discrete-like behaviour, there is nonetheless a small amount of "wiggle room" which can emerge during memory retrieval. And while the network contains no notion of production factors such as voice onset time (etc.), it is easy to envisage how these small differences in the network activity could carry forward into production, thereby giving rise to phenomena such as incomplete neutralization.

Conceptually, the wiggle room itself can be understood as a natural consequence of implementing discrete categories in a continuous space. More specifically however, the wiggle room is caused by two different factors. Firstly, the activation of the inhibitory "final position" unit causes the basins of the different attractors to shift slightly, thus creating the large difference between the initial /t/ and the final /t/ and /d/.

Secondly, the trajectory of the individual memory retrievals is altered by the fact that each retrieval begins from a different point. By analogy, the journey to New York is different depending on whether one begins in Hong Kong or London. Thus, even if the attractor point is the same, the actual trajectory the network traces through the basin is different.

Of course, this second point requires us to understand the memory retrieval as being more than simply arriving at a static point in the state space. That is, memory retrieval is defined as moving through a basin attraction, rather simply arriving at fixed point. The significance of this point is not immediately obvious in the toy model under examination here. However, this becomes critical in more sophisticated attractor models which transition between individual memories over time (e.g. the Latching Network; Collins 2019).

3.2.5 Discussion of the Models

Simplistic though these models of final devoicing are, the comparison of the two does present a more general schema for relating the continuous and the discrete. While the network is fundamentally continuous, it is ultimately capturing the same facts as our formal rule. The fact that we can observe gradience within a discrete category (something not specified by our formal analysis) does not entail that the discrete category does not exist. To be sure, the discrete categories of the grammar are *emergent* phenomena, insofar as each unit in the network has no conception of what a discrete category could be. But the categories are unambiguously present, exactly definable, and indispensable to explaining the dynamics of the system. Therefore, it is meaningless to impose a total dichotomy between discrete and continuous systems, since the model can be viewed as both[25].

Finally, some critics might respond that the attractor and formal grammar are only loosely equivalent in an *extensional* sense. That is, the attractor network is only approximating the formal grammar, rather than *implementing* it. This sort of reasoning often appears in conjunction with the claim that formal linguistic theories should be understood as *intensional* theories of cognition (e.g. Hale and Reiss 2008; Odden 2013, etc.). I will respond briefly to this argument in section 3.4.1. In short, I argue that this reasoning misattributes the utility of a formal analysis. Rather than being a tool for directly deducing the intensional representations in speakers' heads, a formal analysis can be shown to carry more information about the causal structure of the grammar. This is the topic of the next section.

3.3 Effective Information and the Role of Formal Analysis

So far, this paper has argued that the formal model and the attractor network in section 3.2 can be understood as different analyses of the same system – in our case, a toy phonological grammar. However, even if we accept the equivalence of the two models in some sense, this doesn't entail that both models are necessary for understanding the system in question. To put it another way: if we can talk about grammar in neurobiological terms, what use then is a formal theory?

Clearly, the two models are of a radically different character. While the formal model stipulates discrete objects (phones, features) and a rule which acts over those objects (devoicing), the

[25] See also Gafos & Benus (2006) for a thorough discussion of phones-as-attractors.

attractor model attempts to derive the same behaviour from a system of elements which, individually, bear no real resemblance to the ontology of the formal model.

One might be tempted to argue that the attractor model should be interpreted as implying that the formal model is in some sense misleading, or not "real", and should therefore be dispensed with. For Port and Leary, for example, the fact that the formal analysis is technically correct (insofar as it makes correct predictions about the behaviour of the language in question) is not sufficient to make formal analysis a worthwhile endeavour, and propose instead that symbols be demoted to the status of "symboloids":

> *"We do not deny that the phonologies of languages exhibit symbol-like properties, such as reusable and recombinable sound patterns. A small inventory of segment-sized, graphically represented phonological categories can provide a practical scheme for representing most languages on paper. But what is in speakers' heads is apparently not symbols analogous to graphical letters. The term symboloid seems appropriate for these cognitive patterns." (p.950)*

That is, we may continue using our symbol(oid)s for descriptive purposes, but we shouldn't pretend that they offer some insight into how language works, and we certainly shouldn't entertain the notion that symbols actually exist in speakers' heads.

I contend that this final point, whether symbols exist in speakers' heads, is likely a philosophical goose chase that goes far beyond linguistics and neuroscience[26]. However, on the issue of whether formal linguistics actually improves our understanding of cognitive function, what can be shown is that Port & Leary's argument is precisely backwards. Somewhat counter-intuitively, in the case where discrete symbols are emergent from some underlyingly continuous system[27], those symbols are not merely descriptively useful, but are in fact a more revealing explanation of the causal structure of the system.

The crux of this argument depends on the concept of *Effective Information* (EI), proposed by Tononi and Sporns (2003) as a measure of information regarding causation within a system. As

---

[26] Indeed, Port & Leary's argument here is a logical corollary of the "implementational view" mentioned in section 3.2.5 and 3.4.1.

[27] As opposed to elements in a discrete-time, symbol-processing device.

Hoel et al. (2013) note, there are certain cases where the EI of an emergent, macro-level abstraction is higher than the micro-level description. Hoel et al. argue that such cases are example of causal emergence.

Causal emergence is counter-intuitive, since it seems to contradict the assumption that lower levels of explanation are, in some sense, truer or deeper than a macro-level generalization. For example, explaining the behaviour of a gas in terms of the kinetic energy of individual molecules *seems* perhaps more insightful than an equation which simply predicts changes in pressure as a function of heat. According to an EI-analysis however, the simple thermodynamic equation reveals more about the causal relationship between pressure and heat than the molecular explanation, implying that causation itself is a property which emerges only at the macro-level.

Hoel (2017) resolves this apparent puzzle by relating causal emergence to the concepts of compression and bandwidth. In the simplest possible terms, the standard, reductionist view of macro-level abstractions treats them as something akin to *lossy compression*, i.e. a way of discarding information while keeping the general gist for practical purposes, but ultimately an incomplete version of the data (see "symboloids" above). However, causal emergence is something closer to *denoising*, i.e. it provides clearer version of the true signal, where the "signal" is understood to be information about the underlying causal structure of the system.

In the case of our toy phonological system then, the formal description ultimately gives us the clearest view of why the system behaves the way it does (i.e. something devoices *because* it is in a final position), while the attractor model ultimately gives us the same information with a lot of extra noise.

Crucially, this claim is not just a philosophical speculation. Using EI, we can actually quantify the amount of information both the attractor network and our formal analysis, thereby proving the causal emergence of the macro-level.

3.3.1 Defining Effective Information

EI is defined in terms of *Mutual Information* (MI), which is a core concept in Information Theory. MI is centred around the idea that, if two variables are correlated then each variable implicitly contains information about the other. For example, if we know that British people drink tea more often than Americans do, then knowing whether someone is British or American improves our ability to infer whether or not they also drink tea. That is, information about one

variable (nationality) is implicitly also information about another variable (propensity for tea drinking). The intuition here is fairly straightforward, MI is simply the measure of *how much* two variables tell us about each other. What Hoel's method shows is that this concept can be levied to help us understand how much a scientific theory actually elucidates the causal structure of the system under examination.

This method depends on the idea that any well defined model can be recast as a (potentially infinite) number of possible states, (i.e. the state space) and some principle(s) which determine which states are actually possible at a given point in time, i.e. the *interventions*. The notion of "state space" here is identical to the one in the attractor network. However, these concepts also have direct translations in phonological models: for example, in Optimality Theory, the state space and the interventions are equivalent to *GEN* and *EVAL* respectively, while in a rule-based system they would be equivalent to the set of all possible representations and the rules.

With these concepts in place, we can give an exact definition of EI, namely, it is the MI between the possible states of a system, and the possible interventions on that system. Therefore, a high EI means that knowledge about the interventions on the systems also gives us a lot of information about the subsequent state of the system (and vice-versa). Given that the interventions and states are determined by the model under examination (e.g. a formal grammar), EI can be understood as a measure of how much a model actually captures the causal structure of the underlying system. Moreover, because the same system can be analysed at different levels of abstraction, we can use EI to quantify how much causal information is conveyed at each level of abstraction.

Because our attractor model and formal model have radically different state spaces, our comparison will depend on a component of EI, termed *effectiveness*, which is normalized to the size of the state space. Formally, this is defined as the EI of a model divided by the entropy of the intervention distribution. However, since the MI between two variables is highest when each variable perfectly determines the other, *effectiveness* can be defined wholly in terms of the *determinism* and *degeneracy* of the system (Hoel 2017):

$$effectiveness = [determinism] - degeneracy \quad (4)$$

Therefore, the *effectiveness* of a system is equal to 1 in the case where the system is wholly deterministic and non-degenerate, i.e. where each intervention on the system always produces exactly one outcome and where each outcome is always the possible result of only one

intervention. The *effectiveness* will tend to 0 in the cases where all outcomes are equally probable for any intervention (maximum indeterminacy), or where all interventions produce the same outcome (maximum degeneracy).

3.3.2 Effectiveness of the Formal Phonological Grammar

To see how this measure can apply to linguistic model, we can now turn to our toy phonological system from section 3.2.

Normally, calculating the *effectiveness* of this system would entail calculating both the *determinism* and the *degeneracy* of the system (eqn 4 above). Conveniently however, because the system is strictly deterministic, the *determinism* of the system is equal to 1. Therefore it remains only to calculate the *degeneracy* of the system. This is defined in terms of the Kullback-Leibler divergence (D_{KL}) between two probability distributions: the *Intervention Distribution* (I_D) and the *Effect Distribution* (E_D), with regard to the size of the system *n*:

$$degeneracy = \frac{D_{KL}(E_D|I_D)}{log_2(n)} \tag{5}$$

First, we can determine the size of the state space, i.e. the total number of states in the system. In the case of the formal phonological model, there are 6 possible phones – 3 places of articulation, each with a voiced and voiceless variant – and the capacity to distinguish coda and non-coda positions. Therefore, the model has 12 possible states ()Table 5:

Figure 10: Toy Phonological System

	Coda		**Non-coda**	
	Voiced	**Voiceless**	**Voiced**	**Voiceless**
Labial	b#	p#	b	p
Coronal	d#	t#	d	t
Dorsal	g#	k#	g	k

To define I_D and E_D, we must evaluate the system's behaviour in terms of the possible interventions over every state of the system, and the effects of those interventions. This is made simpler in our case as we need only consider a single intervention, namely our final-devoicing

rule(number). We proceed then by specifying, for every state in the system, what the effect of the devoicing rule would be. This is shown in Table 5:

Table 5: Interventions and Effects

Interventions (time = t)	**Effects (time = t+1)**
devoice(b#)	p#
devoice(d#)	t#
devoice(g#)	k#
devoice(p#)	p#
devoice(t#)	t#
devoice(k#)	k#
devoice(b)	b
devoice(d)	d
devoice(g)	g
devoice(p)	p
devoice(t)	t
devoice(k)	k

Here, Hoel's notation of $do(s_i)$ is used to denote an intervention over a given state s_i. Of the 12 interventions, 9 of them produce no change in the state of the system. In the case of the 3 voiced codas however, the intervention results in a change of state, namely: the voiced segment devoices (marked in red).

The left column in Table 5 contains the elements of the *Intervention Distribution* (I_D). Since *effectiveness* is always calculated in the maximum entropy case (i.e. all interventions are equally probable), the probability of every $do(s_i)$ in our system is simply $n^{-1} = \frac{1}{12}$. This, then, gives us our *Intervention Distribution*.

To determine the E_D, we turn to the right hand column in Table 5, the effects of each intervention on the system. This allows us to calculate the probability for all states in the system, after the intervention has been applied. Owing to the simplicity of our toy phonological system this is relatively trivial, and gives us the values in Table 6.

Table 6:

s_i	I_D	E_D
b#	1/12	0
d#	1/12	0
g#	1/12	0
p#	1/12	2/12
t#	1/12	2/12
k#	1/12	2/12
b	1/12	1/12
d	1/12	1/12
g	1/12	1/12
p	1/12	1/12
t	1/12	1/12
k	1/12	1/12

Once we have the values for both I_D and E_D, we can calculate the *degeneracy* using equation (5), and subsequently the *effectiveness*, which come out as ~0.07 and ~0.93 respectively. The value of the *effectiveness* comes very close to a perfect score of 1. But in this case, the system has a small amount of *degeneracy* because at time *t+1*, the states p#, t# and k# can each be reached from two different states at time *t*.[28]

3.3.3 EI of the Attractor Network

Having determined the *effectiveness* of our formal phonological model, we can now turn to the micro-level model of the same system, i.e. our attractor network. Once again, we need to determine the *determinism* and *degeneracy*. And like our formal model, the attractor network is strictly deterministic and therefore has a *determinism* of 1, leaving us only to calculate the *degeneracy*.

In fact, it is trivially true that attractor networks have a high *degeneracy*. This is because the number of attractors must be significantly smaller than the total state space of the model. This follows from the observation that, if every state in the system were stable, then the system could never move from one state to another, ergo attractor dynamics (or indeed *any* dynamics) would be strictly impossible. Indeed, Hoel himself notes that "*A high degeneracy is a mark of attractor dynamics.*" (2017:4).

This point becomes even more obvious when we begin consider the scale of the state-space (i.e. the value of *n*) in an attractor implementation of a symbolic system such as our toy grammar. Of course, the state space of our attractor network is continuous rather than discrete, so strictly speaking *n* is not defined. However, for the sake of calculating the *effectiveness*, we can treat the individual units in the model as if they were binary. This simplification is possible because the memories themselves are defined in terms of units which either active or inactive (i.e. either

[28] In phonological terms, they could be derived either from underlyingly voiced or voiceless segments. This entails that neutralization in the phonology will always result in a non-zero *degeneracy*.

a 0 or a 1). Thus, although the continuous values between 0 and 1 do play a role in individual unit updates, they aren't in fact necessary to define the memories (i.e. attractor states) in the system. Bearing this in mind, if our model contains 41 binary units, the state space of our attractor model contains 2^{41} individual states. That is, n=2,199,023,255,552 and is massively larger than the state space of the formal model ($n=12$).

Of course, computing the effect distribution over 2 trillion states is utterly intractable. Nonetheless, we can approximate the *effectiveness* of the system by employing a simple trick, which is to treat the timestep between an intervention and its effect as the complete retrieval of a memory, rather than the transition between each of the 2 trillion individual states. This simplification is possible because most of the 2 trillion states are not attractors but simply transitory states which the system will tend away from over time. Thus, by treating t+1 as convergence on an attractor, our effect distribution only contains 9 non-zero values (i.e. 6 possible onset segments and 3 coda segments). Therefore our effect distribution will in fact resemble the effect distribution of the formal system, but with an additional 2 trillion or so zero-states (i.e. states which cannot be reached after an intervention):

s_i	I_D	E_D
1	2^{-41}	2/12
2	2^{-41}	2/12
3	2^{-41}	2/12
4	2^{-41}	1/12
5	2^{-41}	1/12
6	2^{-41}	1/12
7	2^{-41}	1/12
8	2^{-41}	1/12
9	2^{-41}	1/12
10	2^{-41}	0
…	…	…
2^{41}	2^{-41}	0

Table 7:

Recall that the three zero-states in the *Effect Distribution* of the formal model were enough to bump the *effectiveness* from a perfect score of 1 down to 0.93. Thus, we should expect the huge increase in zero-states in the attractor model to have a much larger detrimental effect on the *effectiveness*.

This is indeed the case. Given the simplifications discussed above, we can calculate that our attractor model is *effectiveness*≅0.174, which is considerably lower than the value of 0.93 for the formal model (recall: the scale is between 0 and 1).

However, *effectiveness* does not take into account the different sizes of the two models. To see which gives us the most causal information, then the size of the systems needs to be taken into account. Specifically, a greater *effectiveness* will only result in a greater *EI* if

the *effectiveness*-difference is greater than the difference between the size of the macro and micro models:

$$EI(macro) > EI(micro)\ iff\ \frac{eff(micro)}{eff(macro)} > \frac{n(macro)}{n(micro)} \quad (6)$$

In our case, $\frac{0.174}{0.93}$ is considerably larger than $\frac{12}{2^{41}}$, meaning that the formal linguistic model has a higher *EI* than the attractor model, and is therefore a clear case of causal emergence. In other words, the formal model carries more information about the causal structure of the underlying system.

3.3.4 Discussion of the EI Analysis

While the mathematics of the *EI* analysis are relatively straightforward, it is worth spelling out some of the implications of causal emergence in this case.

Firstly, as one can infer from equation (6), the higher *EI* of the formal model does not arise trivially from a simplification of the underlying neural model. It is not enough to merely subsume the many states of the micro model into a smaller number of states in the macro model because, if all else is equal, a larger state space will contain *more* information than a smaller state space. Rather, causal emergence occurs only when the information lost from shrinking the state space is outweighed by the information gained from a decrease in randomness or degeneracy.

Thus, not all systems can be abstracted in a way that conveys more information about the causal structure. For example, a micro-level system which is already strictly deterministic and non-degenerate already has an *effectiveness* of 1, and thus already gives us the maximum amount of casual information. Similarly, macro-level abstractions can only beat the *EI* of non-deterministic and/or highly degenerate systems in the case where the micro-level also exhibits asymmetric causal relationships (Hoel 2017:9). So it is noteworthy, and perhaps convenient, that the attractor model in section 3.2 is amenable to a macro-level abstraction, since not all systems will be. And on the occasion that they are, it seems foolhardy to reject formal methods given their clear utility.

3.4 Conclusion

The thesis of this paper is relatively simple: that linguistic formalisms are commensurable with the prevailing, "analogue" view of the mind/brain, and that such formalisms are indispensable if our ultimate goal is a complete account of the language faculty.

Variations on this general idea have been advocated many times before (Marr 1982; Gafos & Benus 2006; Smolensky & Legendre 2006; Edelman 2008, etc.). Nonetheless, it is clearly not universally accepted (Port and Leary 2005). Therefore, I have attempted to reiterate the argument by relating, in exact terms, a specific model of grammar with a specific model of neural function, and demonstrating, with an exact mathematical method, what the linguistic formalism gives us that the neural model does not, namely, a more informative account of the underlying causal structure of the system.

Of course, the exactness of this argument is only made possible by the use of models which are a drastic simplification of reality, both from the point of view of linguistics and neuroscience. However, there seems to be no reason to doubt that the general conclusion will hold even as the models increase in complexity. Indeed, if anything we should expect that the higher *EI* of the formal model would become even more indispensable as the complexity of the system increases, and the neural model becomes increasingly opaque.

Nonetheless, there are certainly related issues which these simplified models cannot address. For example, it could be argued that, for the toy grammar in section 3.2, only the attractor model can account for incomplete neutralization, and is therefore the better model. And this is true in some sense – *EI* should not be regarded as the absolute measure of a scientific theory, since there are many other factors which define a good theory (empirical coverage, parsimony, legibility, etc). However, as noted in section 3.2.2, there are many attempts to account for incomplete neutralization effects even within formal models (van Oostendorp 2008; Iosad 2012), and my goal here is not to suggest that these approaches are on the wrong track. Exactly what should be regarded as the proper domain of explanation for formal linguistics is rightly an open question. Moreover, it is a question which is largely orthogonal to my argument, since I am not claiming that formal linguistics should explain *everything* about the language faculty. Clearly, many topics will remain firmly in the domain of neuroscience, or phonetics, or psychology and the social sciences more generally. My claim is merely that formal linguistics can provide something which those other domains cannot: a proper treatment of the causal structure of grammar.

3.4.1 Implications for Formal Linguistics

Although this paper is positioned as a defence of formal linguistics, it is nonetheless at odds with various other conceptions of linguistic formalisms. As already noted (section 3.2.5), the implementational view of linguistic formalisms naturally leads to a very different approach to

relating neuroscience and linguistics. For example, consider the following quote from Bale & Reiss (2018):

> *"[W]e have attempted to build phonological theories out of a general logical and mathematical toolbox containing functions, sets, set operations, and variables. Theories built with these tools can be easily translated to precise algorithms. Precise algorithms, in turn, should ultimately make it easier to associate neurological states and activity with phonological cognition. […] Once we state our phonological patterns in such general terms, neuroscientists can try to figure out how functions, sets, set operations, and variables can be implemented in biological systems more generally." (p.4)*

This quote captures the implied utility of a formal linguistic model under the implementational view: it provides a computational parts list for neuroscientists (c.f. Poeppel and Embick 2005; Poeppel 2012).

However, we can note that the approach Bale & Reiss are describing doesn't seem applicable to the models presented in section 3.2, where some elements of the formal model can be easily translated into some phenomenon in the attractor model (e.g. phones are attractors) but others cannot - for example the rule in the formal model does not map onto any distinct component or algorithm in the attractor network (only when the global behaviour of the network is considered can we see the parallel between the two levels of abstraction). This demonstrates an important obstacle for the implementational approach: a macro-level abstraction can be decomposed in any number of *extensionally equivalent* ways, each of which can lead to *intensionally* different assumptions about the underlying neural reality (for example, Bale & Reiss' choice of set-theoretic notation could never lead us to deduce anything like the attractor model). And of course, from the linguist's perspective, there is no way of knowing *a priori* which method of decomposition will ultimately lead to the correct *intensional* model of the brain. This forces linguists to make unprovable claims about the *intensional* properties of their own models. For example, a phonologist might be forced to claim that phonological features are either privitive or binary (c.f. Odden 2013) – on the grounds that this distinction corresponds to some important difference at the neural level – long before they have even the vaguest notion of what that neural difference could possibly be.

By contrast, under the causal emergence view, the linguistic model is related to the neural model by the behaviour of the whole system. In effect, we are mapping a class of *extensionally equivalent* linguistic formalisms onto a class of *extensionally equivalent* neural models. Thus, the linguistic model still delimits the class of possible neural models, but the claim is a slightly weaker one than under the implementational view. Of course, this doesn't entail that there is no way of distinguishing between the equivalent models in each class – it might still be true that features really are either binary or privitive at the neural level – but we can postpone answering such questions until we at least have some coherent way of formulating them in neural terms. In the meantime, linguists can adjudicate between models using all the usual scientific criteria (empirical coverage, parsimony, legibility, etc), and those models can still be used to make non-trivial predictions about the underlying neural mechanisms (e.g. Collins 2019)

Ultimately, exactly how linguistic models relate to the neural reality is an unanswered question. However, the *EI* analysis does give linguists one final motivation to be cautious of putting all their eggs in the implementational basket: recall from section 3.3.4 that causal emergence cannot occur in the case where the micro-system is already deterministic and non-degenerate. This implies that, if linguistic models can be neatly decomposed into (e.g.) neural algorithms implementing set-theoretic operations, then it is unlikely that the linguistic model could give us any more causal information than the neural model. In other words, the implicit goal of the implementational program is one of *reductionism*: a complete neural account would make linguistic models scientifically redundant. Conversely, if the linguistic model is causally emergent (as in the case of the attractor network) then formal linguistics will persist as the best account of the causal structure, irrespective of how sophisticated the neural model becomes.

3.5 Bibliography

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. New York, NY: Cambridge University Press.

Bale, A., Reiss, C. (2018) *Phonology: A Formal Introduction.* MIT Press.

Collins, J. (2019) The Phonological Latching Network [this volume].

Conklin, J. and C. Eliasmith (2005). An attractor network model of path integration in the rat. *Journal of Computational Neuroscience*. 18: 183-203

Crutchfield, J. P. (1998). Dynamical embodiments of computation in cognitive processes. *Behavioral and Brain Sciences, 21*(5), 635-635.

Dale, R., Spivey, M. J. (2005). From apples and oranges to symbolic dynamics: a framework for conciliating notions of cognitive representation. *Journal of Experimental & Theoretical Artificial Intelligence, 17*(4), 317–342. doi:10.1080/09528130500283766

Edelman, G. M. (1992). *Bright air, brilliant fire: On the matter of the mind*. New York, NY, US: Basic Books.

Edelman, S. (2008). On the nature of minds, or: Truth and consequences. *Journal of Experimental & Theoretical Artificial Intelligence, 20*(3), 181-196.

Fodor, Jerry A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Massachusetts: MIT Press

Gafos, A. I., Benus, S. (2006). Dynamics of phonological cognition. *Cognitive science, 30*(5), 905-943.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience, 15*(4), 511-7.

Hale, M., Reiss, C. (2008). *The phonological enterprise*. Oxford University Press.

Hoel, E. P. (2017) When the Map Is Better Than the Territory. *Entropy* 19:188.

Hoel, E. P., Albantakis, L., Tononi, G. (2013) Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences, 110*(49).

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences of the USA*, vol. 79 no. 8 pp. 2554–2558.

Iosad, Pavel. (2012) Final devoicing and vowel lengthening in Friulian: A representational approach. *Lingua* 122, no. 8:922.

Kropff, E. and Treves, A. (2008), The emergence of grid cells: Intelligent design or just adaptation?. *Hippocampus*, 18: 1256-1269.

Lakoff, G. (1993). Cognitive phonology. *The last phonological rule*, 117-145.

Marr, D. (1982) *Vision*. MIT Press.

Odden, D. (2013). Formal phonology. *Nordlyd, 40*(1), 249-273.

O'Connor, T,. Wong, H. Y., (2015) Emergent Properties, *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.)

van Oostendorp, Marc, (2008) Incomplete devoicing in formal phonology. Lingua 188, no. 9:1362.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. OUP Oxford.

Poeppel, D. (2012) The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology Volume 29*

Poeppel, D., Embick, D. (2005). The relation between linguistics and neuroscience. In *Twenty-first century psycholinguistics: Four cornerstones*. Lawrence Erlbaum Associates.

Port, R.F. Leary, A.P. (2005) Against formal phonology. *Language 81*.

Port, R. F., & Van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition*. MIT press.

Roettger, T.B., Winter, B., Grawunder, S., Kirby, J., Grice, M. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics*, vol 43 pp 11-25.

Smolensky, P., Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar*. MIT press.

Tononi, G., Sporns, O. (2003). Measuring information integration. *BMC neuroscience, 4*(1), 31.

Paper 3

"Speech is the only window through which the physiologist can view the cerebral life."

(Fournie 1887 via Lashley 1951)

4 On the Language Specificity of Vowel Maps

Zeynep Kaya, Joe Collins, Alessandro Treves

4.1 Introduction

The vowels of natural languages can be largely mapped onto a 2D Euclidean space, whose dimensions span the frequency of the first (F1) and second (F2) vowel formants. These two acoustic parameters are in remarkable, although not quite linear correspondence with the way each vowel is produced, that is, how open is the vocal tract (F1) and how much the occlusion is moved towards the front of the tract.[29]The striking phenomenon, in fact, is that although this space is inherently continuous, individual languages employ the space as if it were divided into a small number of quasi-discrete basins, commonly referred to as the *vowel inventory* of a given language, or the set of vowels that a fluent speaker can reliably produce and discriminate. This partitioning of the continuous space can be regarded as a natural consequence of Ferdinand de Saussure's observation that the relation between sound and meaning in human language is arbitrary (de Saussure 1966). For example, a sound exactly half-way between the words "hat" and "hit" cannot obtain a meaning half-way between these words, but rather must be parsed as meaning one, the other, or neither.

Another consequence of this arbitrariness is that languages vary from one another both in terms of how they divide the acoustic space, and how the divisions are treated by the grammar. This is true even though the outer boundaries of the space itself are relatively invariant cross-linguistically, owing to consistent physiology between speakers of different languages (barring individual differences in vocal tract length, etc.).

This high degree of cross-linguistic variance has led phonologists to consider the sound patterns of languages in terms of abstract cognitive categories, called phones, rather than purely acoustic objects. Indeed, there are many attested phonological patterns which appear to contradict perceptual or articulatory considerations (Bach & Harms 1972, Buckley 2000, Collins & Krämer 2016), suggesting that the relation between phones and measurable acoustic quantities is complex and strongly language-dependent.

[29] "Largely" alludes to other features of vowel production, such as rounding, which also play a role in vowel discrimination.

This distinct nature of acoustic objects and cognitive categories necessitates the existence of language-specific processes for mapping the "physical" acoustic space onto a more abstract cognitive representation of that space. Native speakers of different languages, whose vowel inventories differ, should then perceive the similarity between sounds not in agreement with their acoustic Euclidean distance, but rather with a language-specific perceptual distance. With two different experiments, we aim to assess this hypothesis quantitatively, by measuring language-specific deformation of the common acoustic space.

4.2 Background: Categorical Perception as Attractor Dynamics

Minimal word pairs such as "hat" and "hit" have parallels in other domains such as bistable percepts in visual perception (e.g. Necker cubes; Necker 1832), as well as various other auditory phenomena (Warren & Gregory 1958; Deutsch 1974). This suggests that categorical perception in the vowel space may not be a language specific capacity *per se,* but perhaps a specific instance of a more fundamental neural mechanism.

One likely candidate for such a mechanism is attractor dynamics, whose application to cognition has been studied within theoretical neuroscience since the 1980s (Amit 1989; Daelli & Treves 2010). Attractor dynamics can be characterized as those of a dynamical system, which evolves towards a small subset of possible configurations. This is often visualized as a landscape of valleys and peaks, with the stipulation that the system will always roll downhill towards a valley floor. If we conceive of the valley floors as being memories, such as lexical items like "hat" and "hit", then attractor dynamics will be enough to ensure that the system will always evolve towards the memories themselves, and away from the ambiguous zones in between.

Various researchers have then posited attractor dynamics as a fundamental mechanism for categorizing speech sounds. Attractor dynamics have been used in this context for reconciling theoretical frameworks of grammar (Gafos & Benus 2006), but also as motivation for experimental work in speech perception. For example, Tuller *et al* (1994) investigated English speakers' ability to distinguish the pair *say*/*stay* by inserting a silence between the /s/ and the vowel portion of the stimuli. By increasing or decreasing the length of the silence, Tuller *et al* found that participants could be biased towards perceiving the stimulus as the word *stay,* even when the formant structure of the vowel does not suggest a preceding /t/. Importantly, the relationship between the length of the silence and participants perception was dependent on the presentation order of the shorter and longer silences. Consequently, the bifurcation of *say/stay*

does not take the form of a sharp, context independent boundary, but rather is suggestive of a complex non-linearity as one attractor slowly weakens and gives way to another.

Others have pursued a similar line of enquiry motivated by the so-called "perceptual magnet" effect[30]. For example, Iverson and Kuhl (1995) studied the warping of the space around the high, front vowel /i/. In one experiment, participants were played synthesized pairs of vowels, between /i/ and /e/, and asked to judge whether they thought the vowels were the same of different. They found that the closer two vowels were to a prototypical /i/ or /e/, the more likely they were to be confused by participants. And conversely, the closer two vowels were to a category boundary, the more likely they were to be discriminated.

Our experiments employ a similar vowel-confusion paradigm to Iverson and Kuhl (1995). Our first experiment uses stimuli that span the continua between four different acoustically defined vowel pairs, to contrast the perception by subjects in four different L1 groups: Italian, Spanish, Turkish, and Scottish English. Our second experiment extends the comparison between L1 groups to 2 dimensions, by using stimuli that cover the whole vowel space, in neighboring pairs. This allows us to visualize vowel perceptual space as a deformed map of the acoustic space for L1 speakers of Italian, Turkish, and Norwegian. In a variant of the second experiment, we probe the stability of the vowel maps for Norwegian (late-)bilinguals in L1 vs L2 contexts.

4.3 First Experiment

Stimuli

Stimuli were created by first recording pairs of CV syllables, and using software to generate new CV syllables with intermediate vowel qualities. For example, given the recorded CV pair [fu] and [fy], with approximately equal first formant for the vowels [u] and [y], the software would produce two new CV pairs with the same initial consonant and F1 as the recorded pair, but with intermediate F2 values. The result in this case is a total of four CV syllables, which are spaced roughly "equidistant" along a single continuum, namely the F2 of the vowel (Figure 12). Equidistant is in inverted commas, because even for pure tones equal distances in Hz are not perceived as equally distinct, an issue often addressed by using the "bark" scale mentioned below, which is roughly linear from 0 to about 600 Hz, and then roughly logarithmic (Zwicker 1961). We refer to each set of four stimuli, generated from the same two recordings, as a CV-

[30] Essentially a qualitative descriptor for attractor dynamics.

quartet. Any two vowels from a quartet can be related in terms of their "distance" from one another, on a discrete scale from 0 (same vowel) to 3 (original recorded pair).

The following CV syllable pairs were used to generate each quartet: [vu]~[vy]; [fu]~[fy]; [bo]~[bœ]; [po]~[pœ]; [dæ]~[dɛ]; [tæ]~[tɛ]; [gɔ]~[gʌ]; [kɔ]~[kʌ]). These stimuli were chosen to satisfy two constraints: i) all the syllables in each quartet should have the same initial consonant, and ii) each quartet has a "voice counterpart" quartet with the same vowels but with an alternate voicing specification in the initial consonant (e.g. [vu]~[vy] is the voice counterpart of [fu]~[fy], etc). This allows us to manipulate the presentation of the stimuli so as to minimize confounds from the initial consonant.

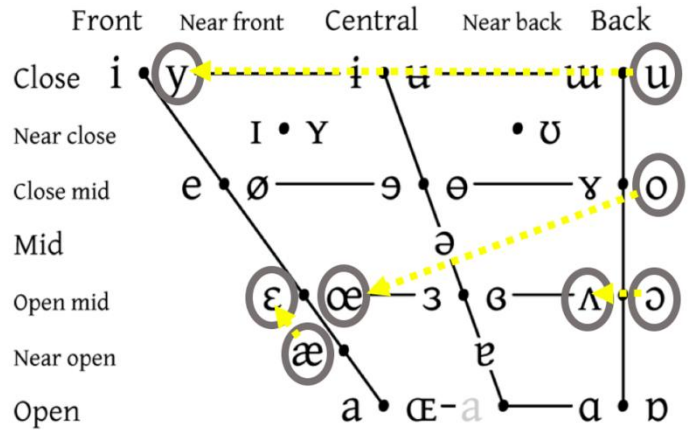

Figure 11: Recorded vowels (grey circles) and continua for morphs shown on standard vowel parallelogram.

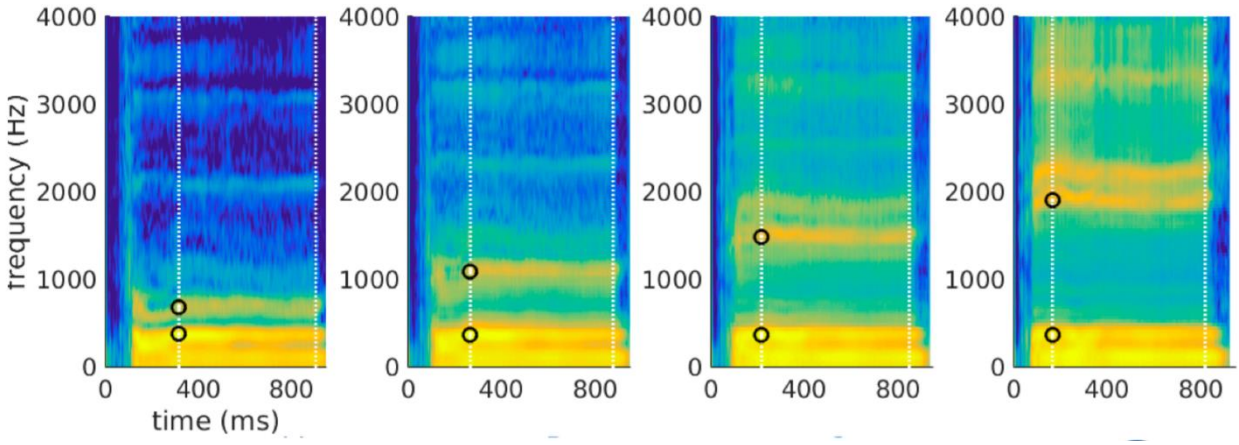

Figure 12:Spectrograms of a single CV-quartet from [fu] (leftmost) and [fy] (righmost) recordings, with intermediate morphs (middle two). The circles show the frequencies of the first and second formants, which form anchor points for the morphing algorithm.

CV syllables were produced by recording a phonetically trained native speaker of British English. Each syllable recording was trimmed to 400ms.

Participants

A total of 64 subjects participated, comprised of four equally sized groups of native Italian, Spanish, Turkish, and Scottish English speakers. All participants exhibited sufficient competence in English to interact fluently with the researcher.

Method

For each trial, participants were played two CV syllables in succession, with a 300ms pause in between, from a pair of voice-counterpart quartets – in other words, the consonant was always different (voiced/non-voiced or vice versa), while the vowel sound could be the same. After the second CV syllable, they were asked to press a button within a 2 second time window if they perceived the two vowels as the same. Conversely, if they perceived them as different they were asked not to respond. Responses given up to 100ms after the elapsed 2 second window were counted as a "same" response for the purpose of the analysis.

Each participant was presented with 160 trials. The trials were presented in a random order, with a different order for each participant, in order to minimize the order effects described by (e.g.) Tuller *et al* (1994).

Results

By counting "same" responses to each vowel pair, we can assess the participants ability to discriminate between ambiguous vowels within each CV-quartet. First, we can sketch possible curves for the proportion of "same" responses as a function of the distance between the vowels. The extent to which these responses deviate from linear gives some indication of the topography of the attractor space. Figure 13 shows idealized psychophysics curves for hypothetical "narrow" and "broad" attractors. In the case of "narrow" attractors, we expect participants to confuse only those vowels which are close to a prototypical vowel. In the case of "broad" attractors however, we expect higher confusion between vowels which are further apart, as both vowels fall within the same basin of attraction. The ill-defined notion of equidistance along the x-axis makes the comparison between results for speakers of different languages more interesting than the actual curves for a single group of subjects *per se*.

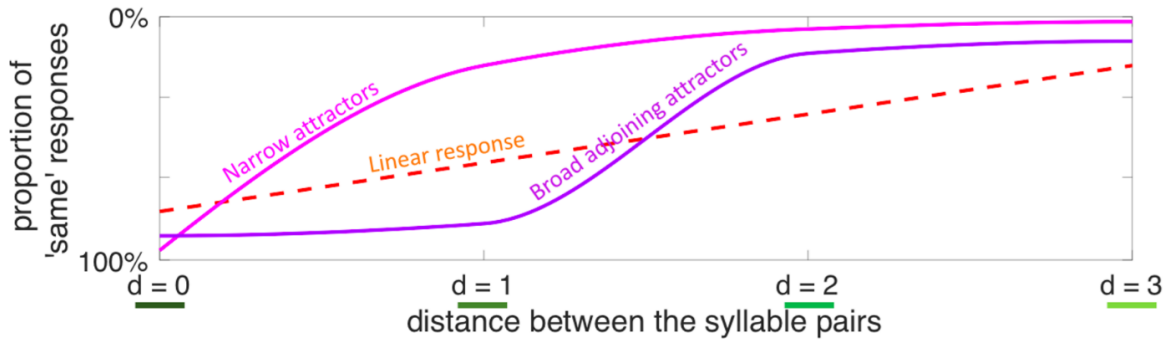

Figure 13: Idealized diagram showing psychophysics curves for hypothetical "narrow" or "broad" attractors, as compared to a strictly linear response.

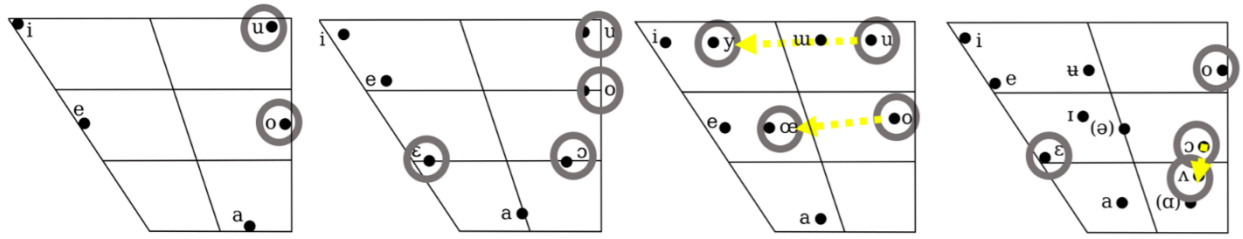

Figure 14: Native vowels for Spanish, Italian, Turkish and Scottish English (left-to-right). The circles and dotted lines denote those which coincide with the recorded stimuli and morph continua (respectively).

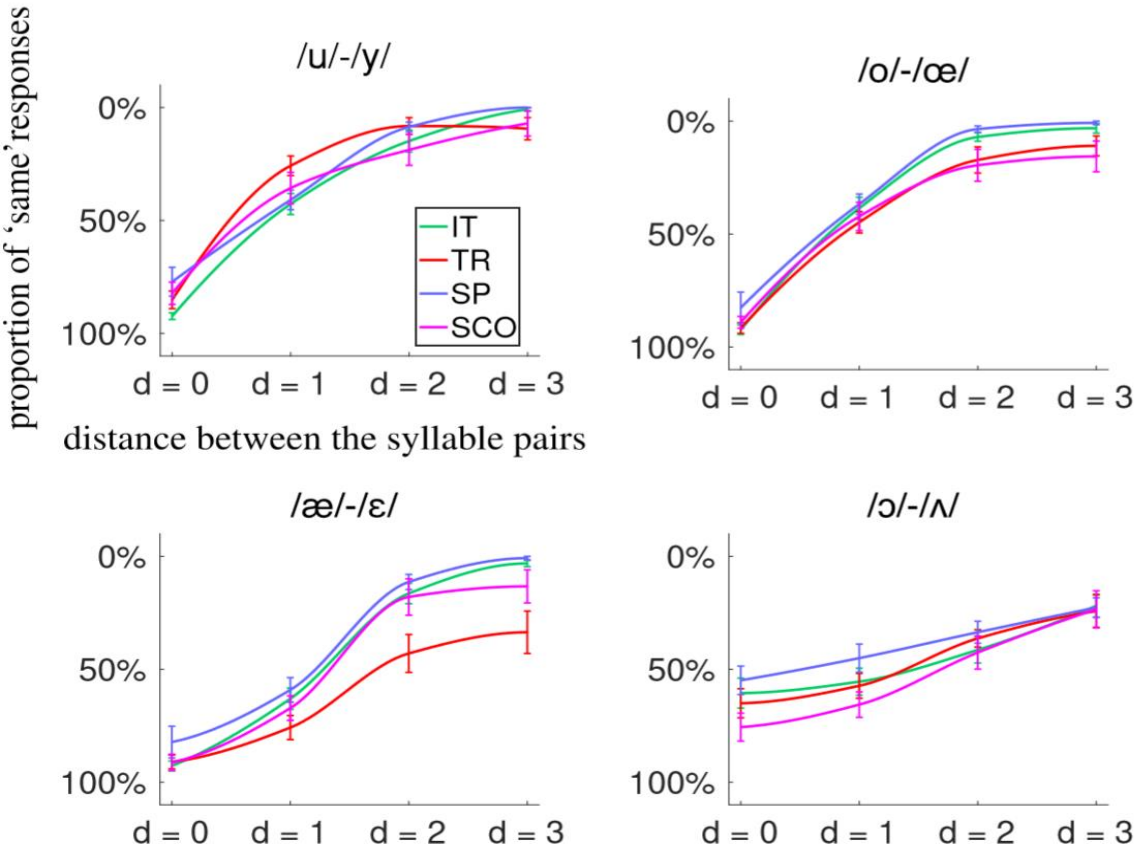

Figure 15: Psychophysics curves for each CV-quartet. The different colour lines correspond to the different language groups.

The results in Figure 15 show some differences between language groups and vowel-pairs. First, across language groups it appears that the [u]-[y] "continuum" tends to fit better the narrow attractor concept (narrow relative to the perceptual distance between the extremes), relative to the [æ]-[ɛ] continuum, where the extremes are perceived as closer (though still easily discriminable, except by Turkish speakers). The [o]-[œ] continuum falls somewhat in between the narrow and broad attractor models. Finally, the [ɔ]-[ʌ] continuum appears close to linear, and none of the language groups seemed to perceive two distinct attractors along it. Turning to language-specific effects, in the case of the [u]-[y] quartets, Turkish seems to show the narrowest basins of attraction, presumably owing to Turkish' high, unrounded [ɯ] vowel which falls approximately in the [u]-[y] quartets. Conversely, Turkish appears to have the broadest attractors in the [æ]-[ɛ] continuum, likely owing to the lack of native vowels in this portion of the vowel space. Speakers of Scottish English share some of the perceptual peculiarities of Turkish speakers, but to a reduced extent. However, overall the differences between the language groups are perhaps smaller than we might expect. It is possible that these psychophysics curves mask some differences due to averaging across morph pairs, therefore in Figure 16 we plot the mean frequency of 'same' responses between individual vowel morphs for each language group.

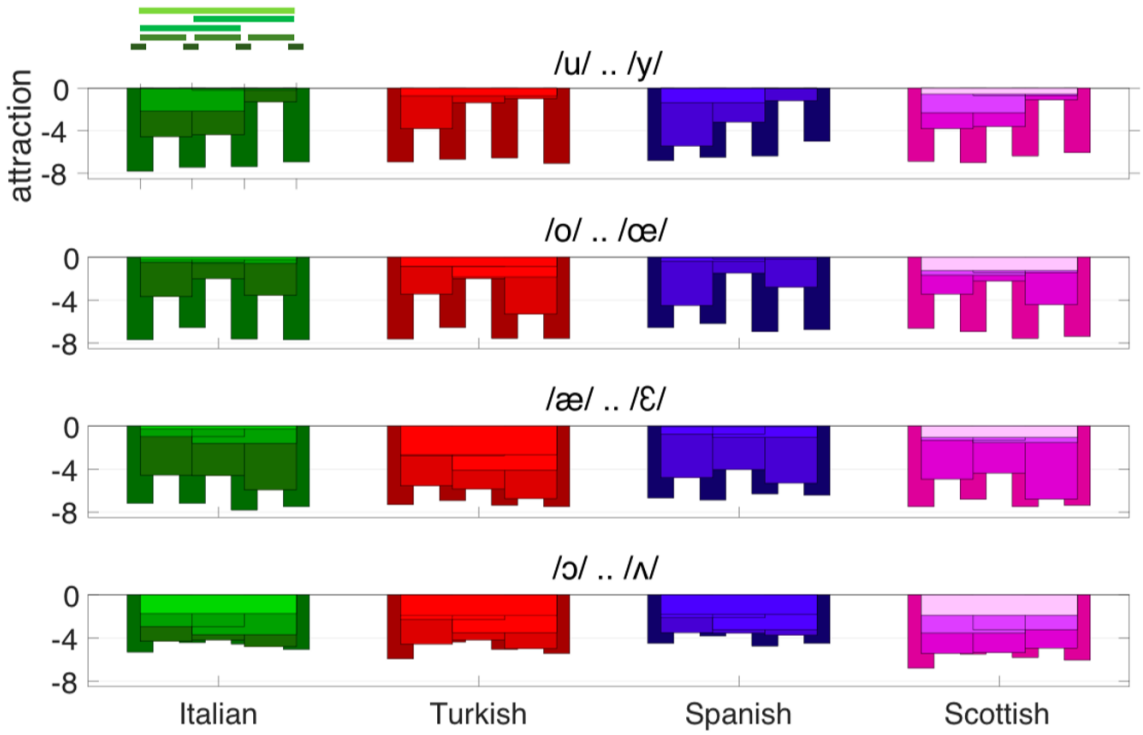

Figure 16: Mean frequency response for each CV-quartet by language group. Within each language, the colour shade corresponds to the vowel distance, such that the darkest shade represents distance=0 while the lightest shade represents distance=4.

The plots in Figure 16 reveal a few more details. For example, the Turkish response to the [u]-[y] CV-quartet shows a generally lower degree of confusion between more distant vowels (wider bars). However, the plot in Figure 16 also tells us that all language groups showed greater confusion on the [u]-[y] continuum further back (closer to [u]), and that Turkish is distinguished specifically by a lower confusion in the middle of this CV-quartet, which again suggests the presence of another distinct standard vowel, the Turkish [ɯ], along this morphing continuum.

4.4 Second Experiment

The first experiment shows some differences between language groups, but the differences are nonetheless smaller than we expected. We speculated that this is because each CV-quartet only spans a unidimensional segment, in two instances quite short a segment, in the vowel space. Therefore, for the second experiment we wanted to extend the analysis to a potential language-specific distortion of the full 2D maps, by using stimulus pairs that still probe local perceptual distances (at larger separations, all stimuli are discriminable, so these experiments would not be informative) but allowing for a complete "triangulation" of the vowel space in each language.

Stimuli

The second experiment employs most of the same CV syllable recordings as the first experiment, in particular the four morphs in each of the two longer [u]-[y] and [o]-[œ] segments, and only the two extremes in the shorter [æ]-[ɛ] and [ɔ]-[ʌ] segments, for a total of 12 vowel sounds. We then added 4 additional intermediate vowels at relatively empty locations in vowel space (positions 6,7,8,12 in Figure 17) and, most importantly, we contrasted each sound with all its immediate neighbors, thus obtaining a triangulation based on the 35 segments in Figure 17. Each subject was tested on each contrast 4 times, the 4 combinations of which vowel sound and which consonant comes first.

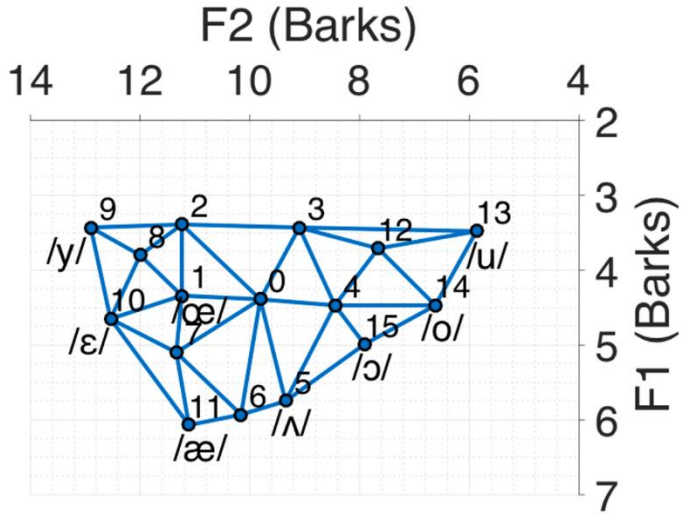

Figure 17: All 16 vowel stimuli plotted in "triangulated" vowel space (frequency in barks).

Participants

60 subjects participated, comprised of three equal-size groups of Italian, Turkish, and Norwegian speakers.

Method

The method for each trial was the same as the first experiment, however the number of trials increased to 256. Of those 256 trials, 64=16x4 presented two stimuli with identical vowels, 140=35x4 presented two stimuli which are adjacent nodes in Figure 17 (i.e. joined by a single line with no intervening nodes), and the remaining 52 presented stimuli in which the vowels were separated by more than one node, but still members of the same CV-quartet.

For the analysis, only the responses to adjacent or identical vowel-pairs were included. We defined a "perceptual distance" between two vowels, *i* and *j*, as the negative log transformation of the symmetrized matrix representing the proportion of "same" response between *i* and *j*.

$$d(i,j) = -\ln\left(\frac{S(i|j)S(j|i)}{S(i|i)S(j|j)}\right)$$

This value is calculated for each adjacent vowel-pair in each language group. Then a gradient descent algorithm (see appendix), adapted from Kohonen (1982), finds a best fit for mapping the relative perceptual distances within each language group onto a 2D plane.

Results

The average perceptual distance for each vowel pair, across all language groups, is shown in Figure 18, as well as any language group outliers ($p<0.001$; calculated from 1000 ensembles of 20 participants). In general, we see that the Norwegian group has much better perception than the Italian and Turkish groups.

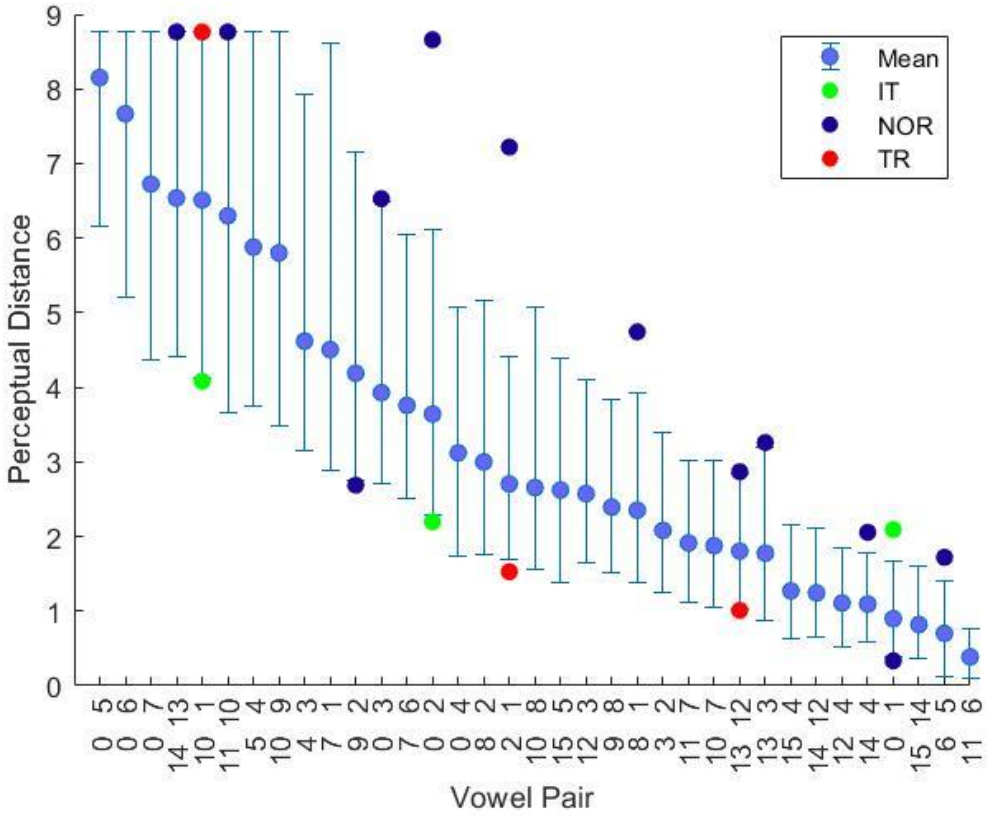

Figure 18: Mean perceptual distance for each adjacent vowel pair. In the case where a language group were outliers (p<0.001), the perceptual distance for that language group is also plotted.

Using the definition of perceptual distance above, we generated the deformed vowel maps for each language shown in Figure 20. In general, some information is lost during the deformation process, therefore the clusters generated by the algorithm do not align perfectly with the native inventories for each language (Figure 19). All three languages exhibit a clear difference between the perceptual space and the acoustic space. And all three languages show significantly more clustering in the back (low F2) portion of the space compared to the front. However, there are also significant differences between the languages, which become clearer when comparing the outliers in Figure 18 with the corresponding links in Figure 20.

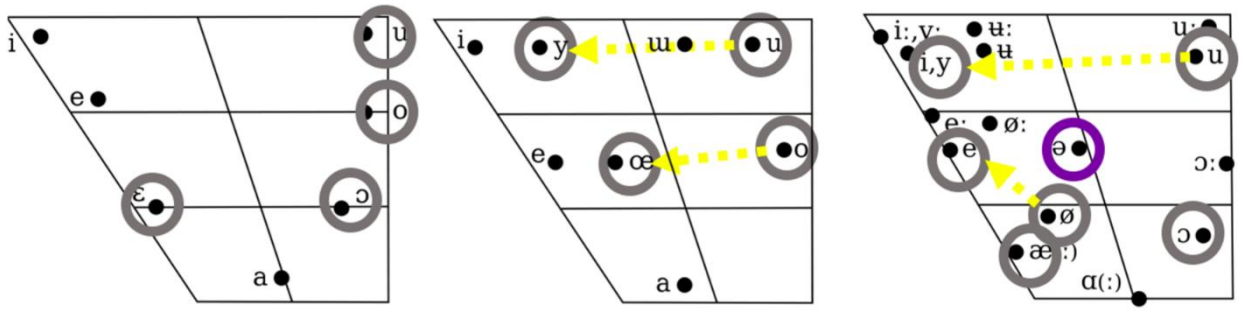

Figure 19: Vowel inventories of (left-to-right) Italian, Turkish, and Norwegian. The identification with the stimuli used in Exp.2 is somewhat arbitrary, particularly for Norwegian.

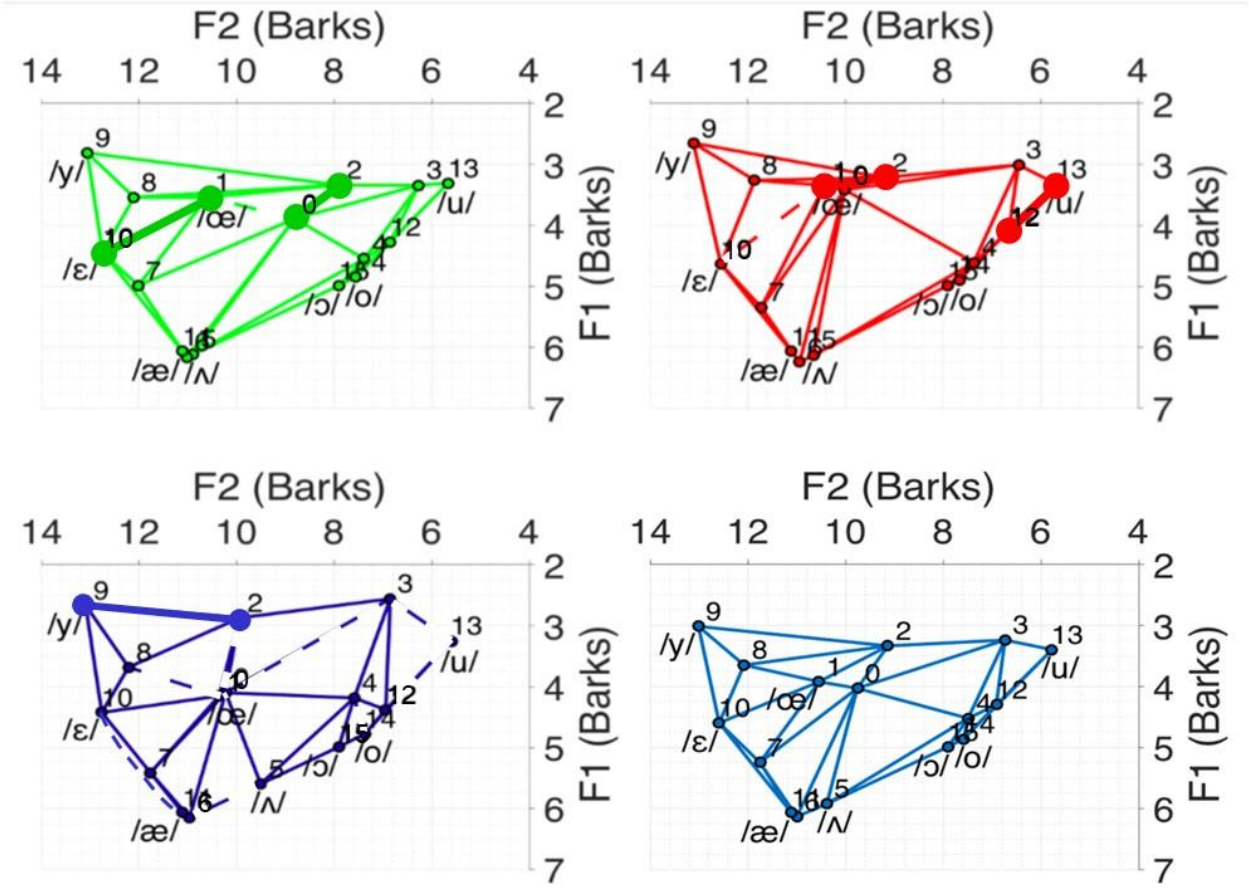

Figure 20: Deformed perceptual maps for Italian (green), Turkish (red), and Norwegian (dark blue), as well as the average map (light blue) that is created by feeding the algorithm with the perceptual distances of the three languages and then by averaging the three maps obtained for each language. High outlier links are indicated by dashed lines, while low outliers are indicated by thicker lines.

The Norwegians had showed better perception across a large number of vowel pairs, which results in less clustering in the deformed map. The two low outliers among Norwegian participants, 9-2 and 0-1, produce slightly different effects in the deformed map. While 9 and 2 are relative distant on the deformed map, there is a very tight cluster between sounds 0 and 1 (which had a very low perceptual distance in Figure 18). Given that this cluster is in the centre of the vowel space, we speculate that it is perhaps caused by the presence of a rounded central-

mid vowel in the native vowel inventory. Although the Norwegian inventory also has a schwa in this portion of the space, its distribution is heavily restricted by prosodic and morphological conditions, which our simple CV-stimuli do not consider. Therefore, it seems unlikely that any of the stimuli would have been perceived by Norwegian participants as containing a schwa.

Conversely, Italian participants were high outliers for the pair 0-1, which is reflected in a lack of clustering in the central portion of the space. This is likely because Italian lacks a true schwa, as well as any central vowel. However, Italian it does show a partial collapse of vowel 4 and 12 onto the nearby 15 and 14, even though we would have perhaps expected it to distinguish clearly between open [ɔ] (15) and closed [o] (14), given that these two vowels form minimal pairs in many dialects of Italian (including standard Italian; Bertinetto & Loporcaro 2005). However, not all dialects distinguish these vowels, and even in dialects which do maintain the distinction, it is restricted to stressed syllables (Krämer 2009). Our monosyllabic stimuli cannot capture all of these factors. Moreover, the Italian participants were not low outliers for any of these sounds, so the excessive degree of clustering is also likely a consequence of the algorithm.

Turkish shows a cluster in the high-centre portion of the space, sounds 1, 0 and 2, which appears to encompass both Turkish [ɯ] and [œ] – this is perhaps a consequence of the complex allophonic variation in Turkish vowel harmony, a consequence of which is that [ɯ] and [œ] cannot appear in the same word. Again, our monosyllabic stimuli cannot disentangle these factors[31].

Finally, the Norwegian group appeared to have generally better discrimination in the low/back portion of the space, particularly between vowels 6 and 5. This is arguably surprising given that Norwegian does not differ greatly from Turkish or Italian in the way this neighborhood is used. However, it is noteworthy that many varieties of English do distinguish more vowels in this space (Figure 14). This leads us to hypothesis that the increased Norwegian discrimination in this space may be caused by increased exposure to English (Ef.edu 2018). This is one of the questions we address in our next experiment, a variant in which we focus on bilingual subjects.

[31] Furthermore, there is evidence that many Turkish speakers lower mid-vowels in certain closed syllables (Gopal & Nichols 2016), whereas all our stimuli are open syllables. This might well lead to a perceptual-raising effect for our stimuli.

4.5 Bilingual variant of Experiment II

For the final experiment, IIb, we consider the role of L2 competence in the deformation of the vowel space. This experiment allows us to probe two questions: Firstly, do (late-)bilinguals employ separate vowel maps for each language, or do they merge their maps (c.f. Flege *et al* 2003) – which might indirectly explain the Norwegian results in second experiment as an effect of increased English competence?

And secondly, are our findings from the first two experiments confounded by the varying degrees of L2 competence among participants, and the differing linguistic contexts under which the experiments were conducted?

Stimuli

This experiment uses the same stimuli and paradigm as the second experiment, with the variation that participants were also presented with recordings of short stories in either English or Norwegian. The short stories were recorded by male (English) and female (Norwegian) native speakers.

Participants

Six native speakers of Norwegian, all with high self-reported competence in English. All participants were university students in fields where a high English competence would be either required or expected (e.g. English literature, etc). The stories were all trimmed to be of either 4 minutes or 2 minutes in length.

Method

The 256 trials were divided into four test sessions. Before each test session, participants were primed for an English or Norwegian language context by listening to short stories. After each testing session, participants would answer yes-no questions on the previous story, to assess their attention during the story segment. The yes-no questions were presented visually, in the same language as the short story. The questions also had a secondary function of leading participants to believe that they were taking a memory test, thereby diverting their conscious introspection from the true task.

Participants would listen to a 4-minute short story (in either English or Norwegian) before the first session, a 2-minute story before the second session, and before the third and fourth sessions participants listened to two halves of the same four-minute story.

Each participant completed a round of testing both with all Norwegian stories and all English stories, with a 15-minute break in between. The order of English vs Norwegian testing was evenly divided among participants.

Results

Accuracy rates for the yes-no questions were close to ceiling for the English stories (*mean 95%*) but lower for the Norwegian stories (*mean 46.67%*). Informal interviews after the experiments suggested that this is partially explained by one of the Norwegian questions being misunderstood, due to the use of an unusual wording in the question. We speculate that the remaining discrepancy is a consequence of participants paying more attention during the English stories, out of a desire to prove their English competence.

The perceptual distance from both conditions were once again mapped onto a 2D space using the gradient descent algorithm. A visual inspection of the deformed maps for the English and Norwegian contexts show no noteworthy differences between the two contexts (Figure 21). A McNemar's test (Cardillo 2007) for each vowel-pair over all trials suggests the priming condition had no significant effect ($p<0.05$) on all but 15 of the 64 vowel-pair combinations. The 15 pairs which passed the threshold for statistical significance all had low statistical power (*McNemar's Z-test<0.8, two-tailed*). We interpret these results as showing no evidence that the priming conditions had any effect on the participants perception. This is in accordance with earlier results suggesting that late-bilinguals do not develop new L2 vowel categories, but rather merge their L1 and L2 categories (Flege *et al* 2003).

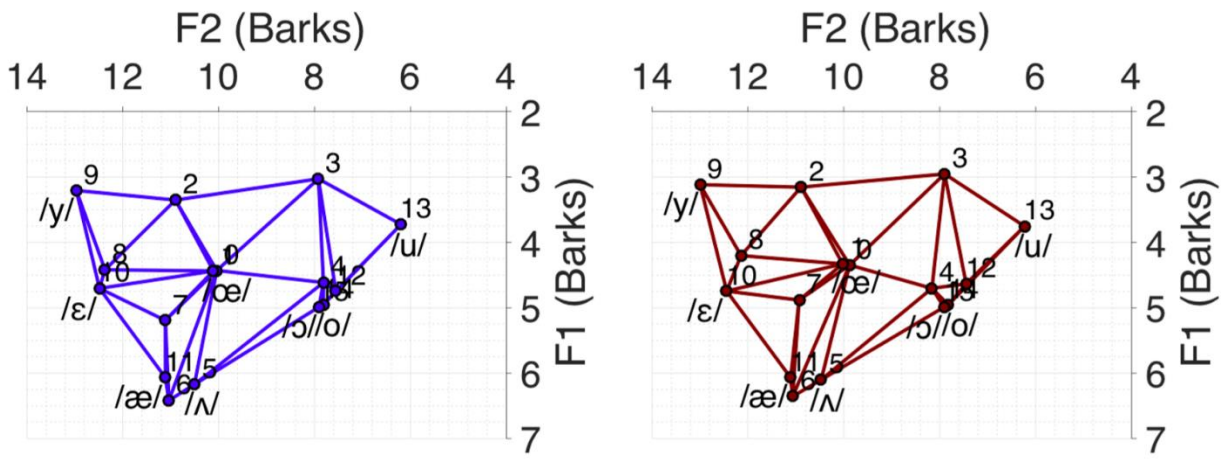

Figure 21: Comparison of the perceptual maps for the Norwegian (left) and English (right) priming conditions.

However, when comparing these maps to the Norwegian map from the second experiment (Figure 20), we do notice some differences, most notably the position of vowel 7, which is far more central in Figure 21 than in Figure 20. This result runs counter to our expectation that the results from these two groups should be the same.

First, we hypothesised that the unexpected difference could be a result of dialectal variation between the two groups, given that the two experiments were carried out in two different cities (Tromsø for the second experiment and Trondheim for the bilingual variation). However, while it is true that there are differences in this portion of the vowel space between the Trøndersk and Tromsø varieties of Norwegian (Vanvik 1979), it is not true that our participants were all speakers of these two varieties. The participants in both experiments spoke a wide variety of dialects, and most of them also had complex linguistic backgrounds (due to living in different cities, etc.) which precluded any neat categorization of the participants by dialect. Therefore, to explain the difference in perception, we would have to assume that merely living in Trondheim or Tromsø would be enough to alter the participants perception of the vowel space. Given that we could not find any significant effect from L2 exposure (Figure 21), it seems implausible that short-term exposure to dialectal micro-variation can nevertheless induce significant changes in speakers' perceptual maps.

Given this conclusion, we wanted to see if the different response from the two groups could be indirectly triggered by the different design of the two experiments. To do this, we compared the participants' responses in the Norwegian context condition to the Norwegian participants' responses in experiment II, for all trials involving vowel 7. This was the vowel whose perception clearly differed the most between the two groups. The other vowels with which

vowel 7 was most likely to be confused were vowels 6, 10 and 11. This was consistent for both groups of Norwegian speakers. However, the participants in experiment II were more likely to give a 'same' response for all three of these vowel-pairs (Table 8).

PERCENTAGE OF 'SAME' RESPONSES

VOWEL-PAIR	Second experiment	Norwegian-primed condition
7-6	23.75%	8.33%
7-10	35%	16.67%
7-11	43.75%	33.33%

Table 8: Comparison of responses to trials involving vowel 7. Note that the other 2 pairings with vowel 7 (7-1 and 7-0) both have a 'same' response rate below 0.02%.

In qualitative terms then, the responses of the two groups were the same, but the group in the second experiment were more likely to press the key to indicate that they perceived both vowels as the same.

Given these data, we might hypothesise that participants were fatigued or disinterested from the longer testing times required by the priming condition. However, the response rates across all vowel pairs do not support this hypothesis. On average, the group from the second experiment were actually *less* likely to press the 'same' button (*mean 34.61%)* compared to the group from the priming experiment *(mean 36.13%*). This observation holds even when comparing responses to trials involving identical vowels (*mean 89.61%* for 2nd experiment vs *mean 95.57%* for bilingual experiment). Therefore, there is no evidence that the participants who listened to the stories and yes/no questions were simply responding less due to fatigue. If anything, the mean response rates seem to suggest the opposite.

Finally, this leaves us with the hypothesis that the difference between the two groups is a sampling error, perhaps a consequence of the smaller sample size in the third experiment. Mann-Whitney U-tests comparing the two groups for each vowel pair suggests that the different response rates between the two groups is not statistically significant (vowels 7-6: *p=0.2467*, vowels 7-10: *p=0.1716*, vowels 7-11: *p=0.4285*). This suggests that a sampling error is the most likely explanation for the difference between the two groups.

4.6 Conclusion

Our first experiment extends the findings of (e.g.) Iverson & Kuhl (1995) to a wider portion of the vowel space, and provides some evidence for language specific perception of that space, perhaps governed by attractor dynamics.

By extending the stimuli in the second experiment to cover the whole space, we are able to approximate the deformation of the acoustic space by participant perceptions of that space. This method provides a much clearer perspective on the language-specific mapping from the acoustic space to the phonological space.

The null-result from experiment IIb suggests that the results from the first two experiments are unlikely to be confounded by the different linguistic contexts in which each language group performed the experiment. However, these results, combined with the higher Norwegian discrimination of English vowels, also supports the claims of Flege *et al.* that bilinguals merge their maps rather than developing separate maps for each language.

4.7 Bibliography

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks.* New York, NY: Cambridge University Press.

Bach, E., & Harms, R. T. (1972). How do languages get crazy rules. *Linguistic change and generative theory, 1*, 21.

Bertinetto, P. M., & Loporcaro, M. (2005). The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *Journal of the International Phonetic Association, 35*(2), 131-151.

Buckley, E. (2000). On the naturalness of unnatural rules. In *Proceedings from the Second Workshop on American Indigenous Languages. UCSB working papers in linguistics*(Vol. 9, pp. 1-14).

Cardillo G. (2007) McNemar test: perform the McNemar test on a 2x2 matrix.

Collins, J., Krämer, M. (2016) Crazy rules and grounded constraints. *Proceedings of CLS 51,* 99-113. Chicago Linguistic Society.

Daelli, V., & Treves, A. (2010). Neural attractor dynamics in object recognition. *Experimental brain research, 203*(2), 241-248.

Deutsch, D. (1974). An auditory illusion. *Nature, 251*, 307-309

Ef.edu. (2018). EF English Proficiency Index - A comprehensive ranking of countries by English skills. [online] Available at: http://www.ef.edu/epi/ [Accessed 26 Apr. 2086]

Flege, J. E., Schirru, C., & MacKay, I. R. (2003). Interaction between the native and second language phonetic subsystems. *Speech communication*, *40*(4), 467-491.

Gopal, D. & Nichols, S. (2016) Sonorant-conditioned mid vowel lowering in Turkish [talk]. *24th Manchester Phonology Meeting*, Manchester.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*(1), 553-562.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetic*s, 43:59–69

Kramer, M. (2009). *The phonology of Italian.* Oxford University Press.

Necker, L.A. (1832). Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid. *London and Edinburgh Philosophical Magazine and Journal of Science.* 1 (5): 329–337

de Saussure, F. (1966). *Course in General Linguistics*. McGraw-Hill, New York.

Tuller, B., Case, P., Ding, M., & Kelso, J. A. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human perception and performance*, *20*(1), 3.

Vanvik, A. (1979), *Norsk fonetikk*, Oslo: Universitetet i Oslo

Warren, R. M., Gregory, R.L. (1958). An auditory analogue of the visual reversible figure. *American Journal of Psychology, 71*, 613-621.

Zwicker, E. (1961), Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33, 248-248.

4.8 Appendix

The algorithm proceeds as follows. The initial fit distance between the sounds i and j, that is f(i, j), equals the Euclidean distance between their coordinates, given by the first and second formants F1 and F2:

$$f(i, j) = \sqrt{\{[F_1(i) - F_1(j)]^2 + [F_2(i) - F_2(j)]^2\}} \qquad (16)$$

In order to extract the perceptual distances in the same order of magnitude as the distances in the Bark space, we initially scale them by a scaling factor k:

$$k = \sum_{links} f(i, j) / \sum_{links} d(i, j) \qquad (17)$$

where both sums are over the 35 edges or links between the sounds. The scaling factor k of Italian, Turkish and Norwegian comes out as 0.503, 0.516, and 0.402 respectively. If perceptual distances are larger on average, we need a smaller scaling coefficient k, and therefore a lower value for k corresponds to a higher ability to discriminate sounds (which is the case for the Norwegian listeners).

At each iteration of the map adjusting algorithm, then, a sound i that is different from the previous one is randomly chosen. Its coordinates are then adjusted to minimize a cost function E, which is defined as:

$$E = \sum_{links} [\, f(i, j) - d(i, j)]^2 / [d(i, j)]^2 \qquad (18)$$

That is, we do gradient descent to find the local minima, with learning rate α:

$$\Delta F_{1,2}(i) = \alpha \sum_{links} [f(i, j) - d(i, j)][F_{1,2}(j) - F_{1,2}(i)]/[d(i, j)^2 f(i, j)] \qquad (19)$$

where we set $\alpha = 0.01$

Works cited

Alderete, J., & Tupper, P. (2018). Connectionist approaches to generative phonology. *The Routledge Handbook of Phonological Theory*. Routledge.

Alderete, J., Tupper, P., & Frisch, S. A. (2013). Phonological constraint induction in a connectionist network: learning OCP-Place constraints from data. *Language Sciences, 37*, 52–69.

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. New York, NY: Cambridge University Press.

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

Anderson, P. W., (1972) More is different. *Science. New Series, Vol. 177*, No. 4047. pp. 393-396.

Bach, E., & Harms, R. T. (1972). How do languages get crazy rules. *Linguistic change and generative theory*, *1*, 21.

Bale, A., Reiss, C. (2018) *Phonology: A Formal Introduction.* MIT Press.

Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G. E., & Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems* (pp. 9368-9378).

Bedau, Mark (1997). "Weak Emergence," Philosophical Perspectives, 11: Mind, Causation, and World, Oxford: Blackwell, pp. 375–399.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in neural information processing systems* (pp. 153-160).

Berkeley, I. S. (1997). A revisionist history of connectionism. *Unpublished manuscript*.

Bertinetto, P. M., & Loporcaro, M. (2005). The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *Journal of the International Phonetic Association, 35*(2), 131-151.

Block, N. (1995) The Mind as the Software of the Brain in E. E. Smith and D. N. Osherson eds. *An Invitation to Cognitive Science vol. 3: Thinking (2nd edition)*. MIT Press.

Boboeva, V., Brasselet, R., & Treves, A. (2018). The capacity for correlated semantic memories in the cortex. *Entropy*, *20*(11), 824.

Brunel, N., Hakim, V., & Richardson, M. J. (2014). Single neuron dynamics and computation. *Current Opinion in Neurobiology*, *25*, 149–155.

Bouchard, K. E., Mesgarani, N., Johnson, K., and Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature* 495, 327–332.

Buckley, E. (2000). On the naturalness of unnatural rules. In *Proceedings from the Second Workshop on American Indigenous Languages. UCSB working papers in linguistics*(Vol. 9, pp. 1-14).

Bybee, J. (1999). Usage-based phonology. *Functionalism and formalism in linguistics, 1*, 211-242.

Cardillo G. (2007) McNemar test: perform the McNemar test on a 2x2 matrix.

Chalmers, D. (1990). Why Fodor and Pylyshyn were wrong: The simplest refutation. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, Cambridge, Mass (pp. 340-347).

Chomsky (1994) [interview] *Protosociology 6*, pp. 293-303

Chomsky, N. (2002) *On Nature and Language*. Cambridge University Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic inquiry*, *36*(1), 1-22.

Chomsky, N., & Guignard, J. B. (2011). Beyond Linguistic Wars. An Interview with Noam Chomsky. *Intellectica, 56*(2), 21-27.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. MIT press.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston and M. E. Beckman (eds.) *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press. 283-333.

Collins, J., Krämer, M. (2016) Crazy rules and grounded constraints. *Proceedings of CLS 51*, 99-113. Chicago Linguistic Society.

Conklin, J. and C. Eliasmith (2005). An attractor network model of path integration in the rat. *Journal of Computational Neuroscience*. 18: 183-203

Connors, B. W., & Gutnick, M. J. (1990). Intrinsic firing patterns of diverse neocortical neurons. *Trends in neurosciences, 13*(3), 99-104.

Copeland, B. J., & Proudfoot, D. (1996). On Alan Turing's anticipation of connectionism. *Synthese, 108*(3), 361-377.

Crutchfield, J. P. (1998). Dynamical embodiments of computation in cognitive processes. *Behavioral and Brain Sciences, 21*(5), 635-635.

Daelli, V., & Treves, A. (2010). Neural attractor dynamics in object recognition. *Experimental brain research, 203*(2), 241-248.

Dale, R., & Spivey, M. J. (2005). From apples and oranges to symbolic dynamics: a framework for conciliating notions of cognitive representation. *Journal of Experimental & Theoretical Artificial Intelligence, 17*(4), 317-342.

Dawson, M. R., Medler, D. A., & Berkeley, I. S. (1997). PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology, 10*(1), 25-40.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge, MA, USA: MIT Press.

Dennett, D. C., 1987, *The Intentional Stance*, Cambridge, MA: MIT Press.

Deutsch, D. (1974). An auditory illusion. *Nature, 251*, 307-309

Dewhurst, J. (2018) Computing Mechanisms Without Proper Functions. *Minds & Machines 28*: 569.

Dreyfus, H. L. (1972). *What computers can't do*. MIT Press.

Eberbach E., Goldin D., Wegner P. (2004) Turing's Ideas and Models of Computation. In: Teuscher C. (eds) *Alan Turing: Life and Legacy of a Great Thinker*. Springer, Berlin,

Edelman, G. M. (1992). *Bright air, brilliant fire: On the matter of the mind*. New York, NY, US: Basic Books.

Edelman, S. (2008). On the nature of minds, or: Truth and consequences. *Journal of Experimental & Theoretical Artificial Intelligence, 20*(3), 181-196.

Edelman, S. (2017). Language and other complex behaviors: Unifying characteristics, computational models, neural mechanisms. *Language Sciences, 62*, 91–123.

Ef.edu. (2018). EF English Proficiency Index - A comprehensive ranking of countries by English skills. [online] Available at: http://www.ef.edu/epi/ [Accessed 26 Apr. 2086]

Eliasmith, C. (1997). Computation and dynamical models of mind. *Minds and Machines, 7*(4), 531-541.

Elman, J. L. (1990). Finding structure in time. *Cognitive science, 14*(2), 179-211.

Feinerman, O., Pinkoviezky, I., Gelblum, A., Fonio, E., Gov, N. S. (2018) The physics of cooperative transport in groups of ants. *Nature Physics*: 1745-2481.

Flege, J. E., Schirru, C., & MacKay, I. R. (2003). Interaction between the native and second language phonetic subsystems. *Speech communication*, *40*(4), 467-491.

Fodor, J. A. (1975). *The language of thought*. Harvard university press.

Fodor, J. A. (1981) The Mind-Body Problem. *Scientific American*, *244*: 114–125.

Fodor, Jerry A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Massachusetts: MIT Press

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3-71.

Frisch, S. A. (2018) Exemplar theories in phonology. In: Hannahs, S. J., & Bosch, A. (Eds.). *The Routledge Handbook of Phonological Theory*. Routledge.

Gafos, A. I., & Benus, S. (2006). Dynamics of phonological cognition. *Cognitive science*, *30*(5), 905-943.

Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. John Wiley & Sons.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience, 15*(4), 511-7.

Gopal, D. & Nichols, S. (2016) Sonorant-conditioned mid vowel lowering in Turkish [talk]. *24th Manchester Phonology Meeting*, Manchester.

Haken, H. E., & Stadler, M. E. (1990). Synergetics of cognition: *Proceedings of the International Symposium* at Schlo–S Elmau, Bavaria.

Hale, M., Reiss, C. (2008). *The phonological enterprise*. Oxford University Press.

Hickok, G., Poeppel, D. (2007) The cortical organization of speech processing. *Nature Reviews Neuroscience* 8, 393-402.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hoel, E. P. (2017). When the Map Is Better Than the Territory. *Entropy* 19:188.

Hoel, E. P., Albantakis, L., Tononi, G. (2013) Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences, 110*(49).

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational properties. *Proc. Nat. Acad. Sci*. (*USA*) 79, 2554-2558.

Hopfield, J. J. (1984) Neurons with graded response have collective computational properties like those of two-sate neurons. *Proc. Nat. Acad. Sci. (USA) 81,* 3088-3092.

Hopfield, J. J. (2007). *Hopfield network.* Scholarpedia, 2(5):1977.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*(1), 553-562.

Iosad, Pavel. (2012) Final devoicing and vowel lengthening in Friulian: A representational approach. *Lingua* 122, no. 8:922.

Ising, E. (1925) Beitrag zur Theorie des Ferromagnetismus (Contribution to the Theory of Ferromagnetism). *Zeitschrift für Physik*. 31.

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. *Experimental approaches to phonology. In honor of John Ohala,* 25-40.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol 58*(6): 1233-1258.

Kang,C. J., Naim, M., Boboeva, V., Treves, A. (2017) Life on the edge: Latching Dynamics in a Potts neural network. *Entropy*, 19, p.468.

Kanter, I. (1988). Potts-glass models of neural networks. *Phys. Rev. A* 37, 2739.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*:59–69

Kramer, M. (2009). *The phonology of Italian.* Oxford University Press.

Krämer, M. (2012). *Underlying representations*. Cambridge University Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Kleene, Stephen C. (1956)[1951]. Representation of Events in Nerve Nets and Finite Automate. *Automata Studies, Annals of Math. Studies*. Princeton Univ. Press. 34.

Kropff, E. and Treves, A. (2008), The emergence of grid cells: Intelligent design or just adaptation?. *Hippocampus*, 18: 1256-1269.

Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G., & Smith, N. A. (2017). What Do Recurrent Neural Network Grammars Learn About Syntax?. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 1, Long Papers (pp. 1249-1258).

Lakoff, G (1988). A Suggestion for a Linguistics with Connectionist Foundations, in Touretzky, D ed. *Proceedings of the 1988 Connectionist Summer School*. UC Berkeley

Lakoff, G. (1993). Cognitive phonology. *The last phonological rule*, 117-145.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh*. New york: Basic books.

Lashley, K. S. (1951). *The problem of serial order in behavior* (Vol. 21). Bobbs-Merrill.

Luisi, P. L. (2002*) Foundations of Chemistry 4*: 183–200.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive psychology*, *37*(3), 243-282.

Marr, D. (1982) *Vision*. MIT Press.

Marr, D.; Poggio, T. (1976). From Understanding Computation to Understanding Neural Circuitry. *Artificial Intelligence Laboratory. A.I. Memo*. MIT.

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

McCoy, B. (2010) *Ising model: exact results*. Scholarpedia, 5(7):10313.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*. 343, 1006–1010.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford University Press.

Minsky, M. L. (1954). *Theory of neural-analog reinforcement systems and its application to the brain model problem* (PhD Thesis) .Princeton University.

Minsky, M., & Papert, S. A. (2017). *Perceptrons: An introduction to computational geometry*. MIT press.

Naim,. M., Boboeva, V., Kang., C. J., Treves, A. (2017) Reducing a cortical network to a Potts model yields storage capacity estimates. arXiv:submit/2036185 [q-bio.NC]

Nasrabadi, N. M., Choo, C. Y. (1992) Hopfield network for stereo vision correspondence, in *IEEE Transactions on Neural Networks*, vol. 3, no. 1, pp. 5-13.

Necker, L.A. (1832). Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid. *London and Edinburgh Philosophical Magazine and Journal of Science.* 1 (5): 329–337

von Neumann, J. (1951). *The general and logical theory of automata*. 1951.

von Neumann, J. (1956). Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies*, *34*, 43-98.

Nguyen, Noël & Wauquier, Sophie & Tuller, Betty. (2009). The dynamical approach to speech perception: From fine phonetic detail to abstract phonological categories. *Approaches to phonological complexity*, 5-31.

Newell, A. & Simon, H. A. (1963). GPS: A Program that Simulates Human Thought, in Feigenbaum, E.A.; Feldman, J. (eds.*), Computers and Thought*, New York: McGraw-Hill.

Newell, A. & Simon, H. A. (1976). Computer Science as Empirical Inquiry: Symbols and Search, *Communications of the ACM, 19* (3): 113–126.

O'Connor, T,. Wong, H. Y., (2015) Emergent Properties, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.)

Odden, D. (2013). Formal phonology. *Nordlyd, 40*(1), 249-273.

Olazaran, M. (1996). A Sociological Study of the Official History of the Perceptrons Controversy. *Social Studies of Science, 26*(3), 611–659.

van Oostendorp, Marc, (2008) Incomplete devoicing in formal phonology. *Lingua 188*, no. 9:1362.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language 95*(1), e41-e74. Linguistic Society of America.

Piccinini, G. (2004). The First computational theory of mind and brain: a close look at Mcculloch and Pitts's “logical calculus of ideas immanent in nervous activity”. *Synthese, 141*(2), 175-215.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. OUP.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics 2*(1), 33-52.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73-193.

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

Poeppel, D. (2012) The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology Volume 29*

Poeppel,D., Embick, D. (2005). Defining the relationship between linguistics and neuroscience. In A. Cutler ed. *Twenty-first century psycholinguistics: Four cornerstones*, Lawrence Erlbaum.

Port, R.F. Leary, A.P. (2005) Against formal phonology. *Language* 81.

Port, R. F., & Van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition*. MIT press.

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science, 275*(5306), 1604-1610.

Roettger, T.B., Winter, B., Grawunder, S., Kirby, J., Grice, M. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics*, vol 43 pp 11-25.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65*(6), 386.

Rumelhart, D. E. (1998). The architecture of mind: A connectionist approach. *Mind readings*, 207-238.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition, 1*(45-76), 26.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, D. E. & McClelland, J. L. (1987) Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach.* 3rd edition. Malaysia; Pearson Education Limited,.

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

de Saussure, F. (1966). *Course in General Linguistics*. McGraw-Hill, New York.

Schneider, W. (1987) Connectionism: Is it a paradigm shift for psychology? *Behavior Research Methods, Instruments, & Computers 19*.

Searle, J. R., 1992, *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Serra, R., & Zanarini, G. (1990). *Complex Systems and Cognitive Processes*.

Shagrir, O. (2006). Why we view the brain as a computer. *Synthese, 153*(3), 393-416.

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy, 26*(Supplement), 137-161.

Smolensky, P., Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar*. MIT press.

Tononi, G., Sporns, O. (2003). Measuring information integration. *BMC neuroscience, 4*(1), 31.

Treves, A. (2005) Frontal latching networks: A possible basis for infinite recursion. *Cognitive Neuropsychology*. 22(3-4).

Treves, A., & Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain?. *Network: Computation in Neural Systems*, *2*(4), 371-397.

Tsodyks, M. V, Feigelman, M. V. (1988) The Enhanced Storage Capacity in Neural Networks with Low Activity Level. *Europhysics Letters,* vol. 6, no. 2.

Tuller, B., Case, P., Ding, M., & Kelso, J. A. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human perception and performance*, *20*(1), 3.

Vanvik, A. (1979), *Norsk fonetikk*, Oslo: Universitetet i Oslo

Warren, R. M., Gregory, R.L. (1958). An auditory analogue of the visual reversible figure. *American Journal of Psychology, 71*, 613-621.

[Web of Stories - Life Stories of Remarkable People] (2016) Marvin Minsky - The problem with perceptrons [Video File]. Retrieved from https://www.youtube.com/watch?v=QW_srPO-LrI

Zwicker, E. (1961), Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, 33, 248-248.