UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

HSL-ISK

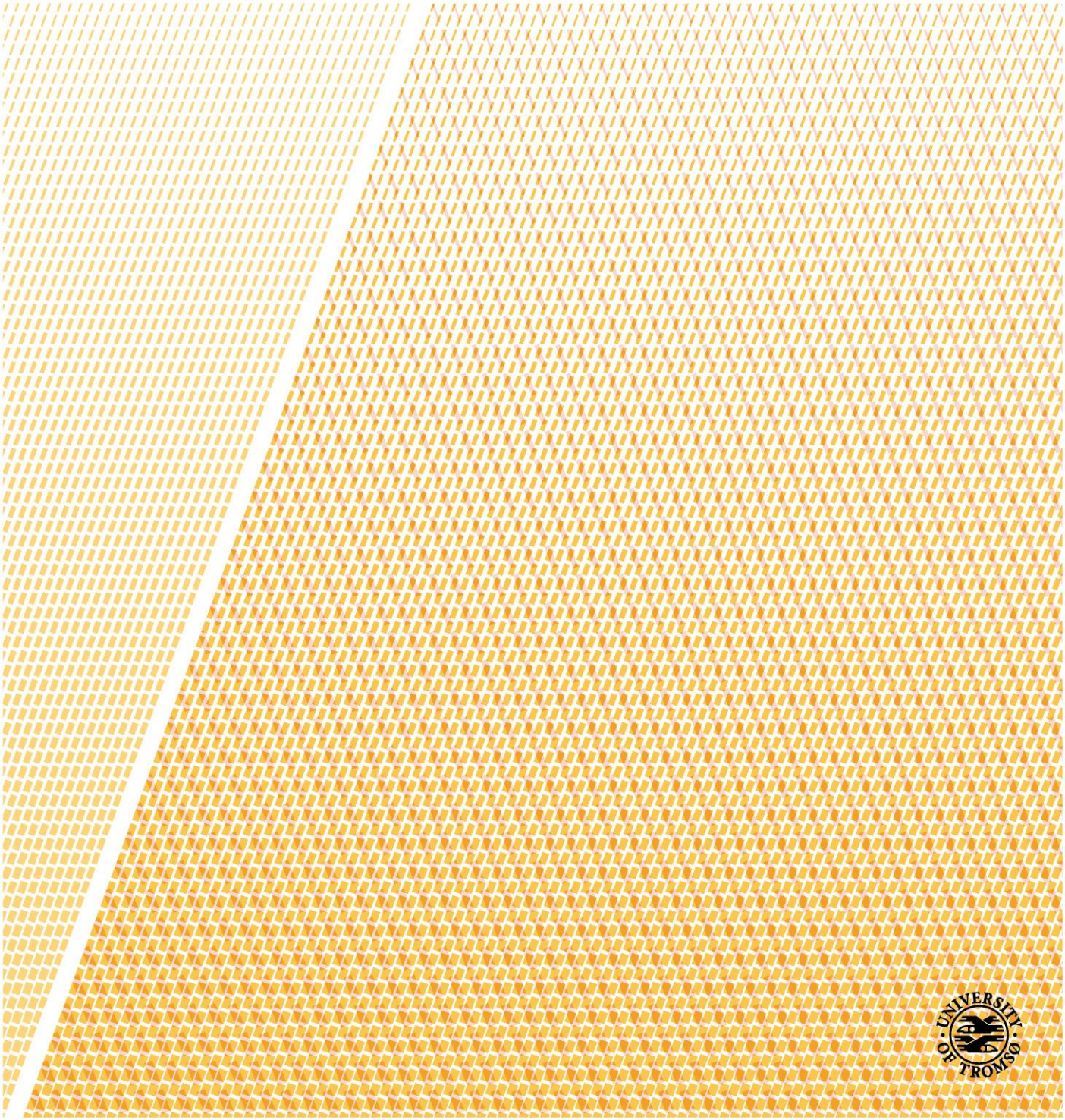
Neural Attractors and Phonological Grammar

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

—

Joe Collins

A dissertation for the degree of Philosophiae Doctor – June 2019



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.

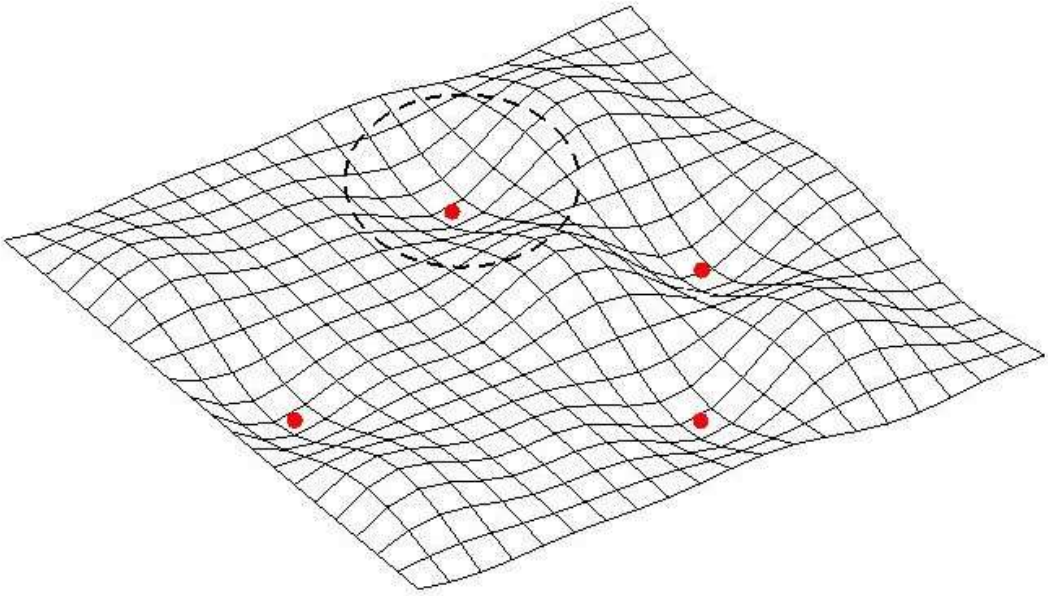


Figure 1: Conception of a network state-space. The z-axis corresponds to the free energy of the network. The red dots are attractors. http://www.scholarpedia.org/article/Attractor_network

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.

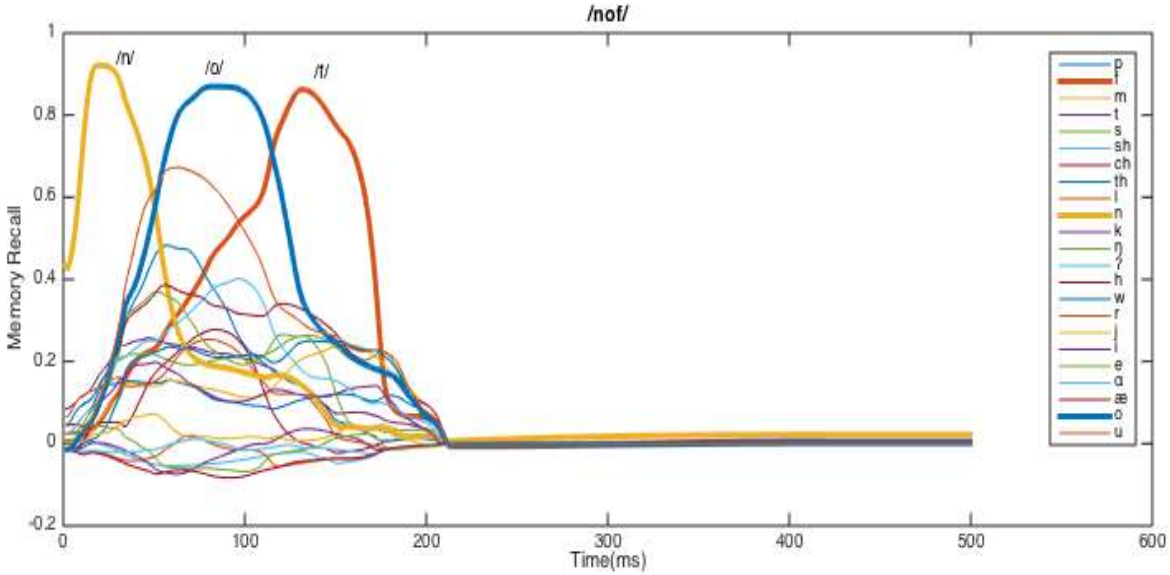


Figure 2: Example of a latching string. The PLN produces /nof/.

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.

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