SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

Neuroscience Area – PhD course in

Cognitive Neuroscience

Associative Transitions in Language Processing

Candidate:
Massimiliano Trippa

Advisor:
Alessandro Treves

Academic Year 2018-19

Contents

Chapter 1

Introduction

Language is the mental faculty that mostly characterizes humans from other living beings on Earth. We can communicate concepts, facts and ideas with each other by speaking, writing, typing or using gestures. Already these aspects show the incredible multimodality of language, involving a wide range of motor and sensory abilities. Communication is then supported by the even more exceptional cognitive ability to translate an abstract concept into a structured sequence of symbols (i.e., language production) able to elicit a similar concept in someone else's mind (i.e., language comprehension). We perform all these types of processing in our everyday life with minimal effort.

Language and the brain

Neurologists in the 19th century assigned the neural origin of language to two small areas in the left hemisphere [1][2] following the study of aphasic patients with narrowly defined brain lesions. Broca's area in the inferior frontal lobe, close to motor areas, was associated with the articulation of speech. On the other hand, Wernicke's region in the superior temporal gyrus, close to the auditory cortex, was thought to store the sound representations of words [3][4]. Further subdivisions of the two areas were then associated with different impairments, resulting in an even more confined localization of the language faculty in the brain [5][6][7]. Language, however, is a highly

complex function and we may argue if a cortical area of just a few square centimeters could really be the only actor in the play.

The increase in popularity and refinement of neuroimaging techniques in recent years have allowed for the study of the neural correlates of language in normal conditions. This shift from the investigation of impaired patients to healthy subjects revealed a widespread involvement of cortical regions in both production and comprehension tasks.

Authors in [8] were able to predict the cortical patterns of activation associated with newly presented words by combining the distributed activations resulting from the presentations of words with similar meaning. Similarly, data from healthy participants hearing meaningful stories while in an fMRI scanner were used to create an atlas of the semantic tuning of the human cortex [9]. In both studies, language processes appear to involve extensive portions of the cerebral cortex.

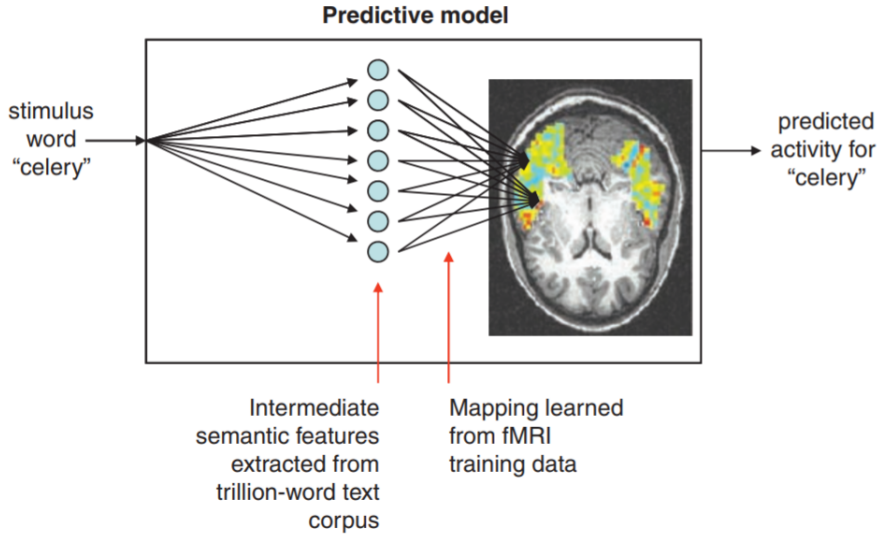


Figure 1.1: The prediction of fMRI signal is derived, in the study by Mitchell et al. in two steps. First the meaning of *celery* is defined by a set of semantic features collected from large corpora of text (e.g. "vegetable" and "edible" are proper features for *celery* but they are not for *airplane*). In the second step, the neural activation related to each feature is learned through the presentation of lists of words to the participant. Combining the features for the word *celery* results in a prediction for the fMRI activation pattern when the word is presented to the participant [8].

A distributed perspective on language

A possible reconciliation of the two opposite types of observations may be found in Hebb's concept of cell assemblies [10][11]. Hebb's theory states that functional units composed of many neurons, i.e., the cell assemblies, are formed in the cortex as a result of a frequent simultaneous activation that causes synaptic strengthening. Evidence for these associative connections between frequently co-active units was found in many studies [12][13][14][15].
Pyramidal cells, with their long axons entering white matter and their high variability in size, have been hypothesized to be the major means of communication between the different regions of the cortex, supporting the formation of large-scale activity patterns. The crucial feature of pyramidal cells, described in [16][17], is the typical bipartite branching of their dendritic trees. Basal dendrites, surrounding the soma, receive local connections from neighboring cells; conversely, apical dendrites, extending from the apex of the cell body towards the upper layers of the cortex, receive input from long-range cortico-cortical connections from other brain regions.
Braitenberg and Schuz envisioned a crucial role for this dual, local and global, nature of the cortex provided by the A (apical) and B (basal) systems [17][16]. They hypothesized the whole cortex to be an associative memory machine, in which the B-systems encode a set of memories as local attractors and the A-system encodes global attractors, by virtue of long-range connections. In the theory of dynamical systems, an attractor is defined as a set of values, or a state, towards which the system tends to evolve [18][19]. In simple terms, an attractor can be thought of as a large valley surrounded by mountains: in this sense, the dynamics of the system in consideration is comparable to the flow of a river descending the mountains to form a lake in the valley. This analogy also suits the description of a memory system like our brain: a strong memory of a past event (i.e., the valley) can be entirely recalled by merely hinting at small but relevant details of the scene (i.e., the flow of the river). Attractor states can thus be used to model the dynamics of the cortex in memory retrieval [20][21][19][18]. Furthermore, a crucial, and

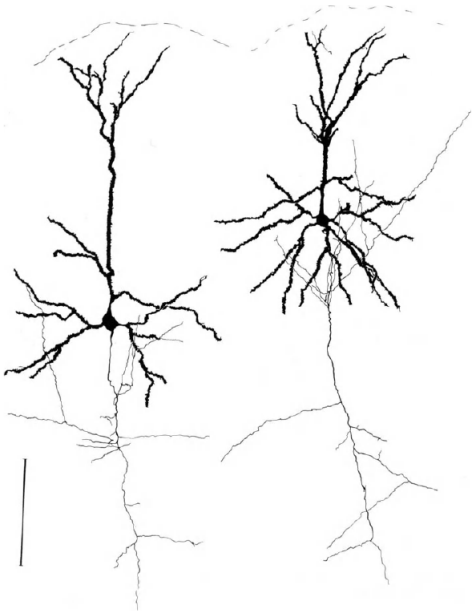


Figure 1.2: Camera lucida tracing of two pyramidal cells, taken from [17]. The dashed line represents the surface of the cortex. The basal dendrites surround the typical conic shaped soma, while the apical dendrites reach the first layers of the cortex. A long myelinated axon departs from the bottom of the soma and enters the white matter to create cortico-cortical connections.

often overlooked, feature of this theory is the ability to model the conversion from the analog computations performed by the neurons into discrete macroscopic neural states, suitable to describe the encoding of the atoms of language, namely words, syllables, phonemes, morphemes, etc. In light of these concepts, unified with the view of Braitenberg in [17], we can interpret the semantic cortical maps in [9] as a collection of many local attractors, encoding semantic features, that are recollected together, in a global attractor state, to form the neural representation of a word.

From memory to functions

Language cannot be reduced to only the description of how linguistic information is stored in the brain: considering its dynamics is, therefore, a fundamental step to grasp the mechanisms which allow us to communicate with each other. Related concepts tend to elicit one another in a possibly indefinite process, allowing for the production and expression of more com-

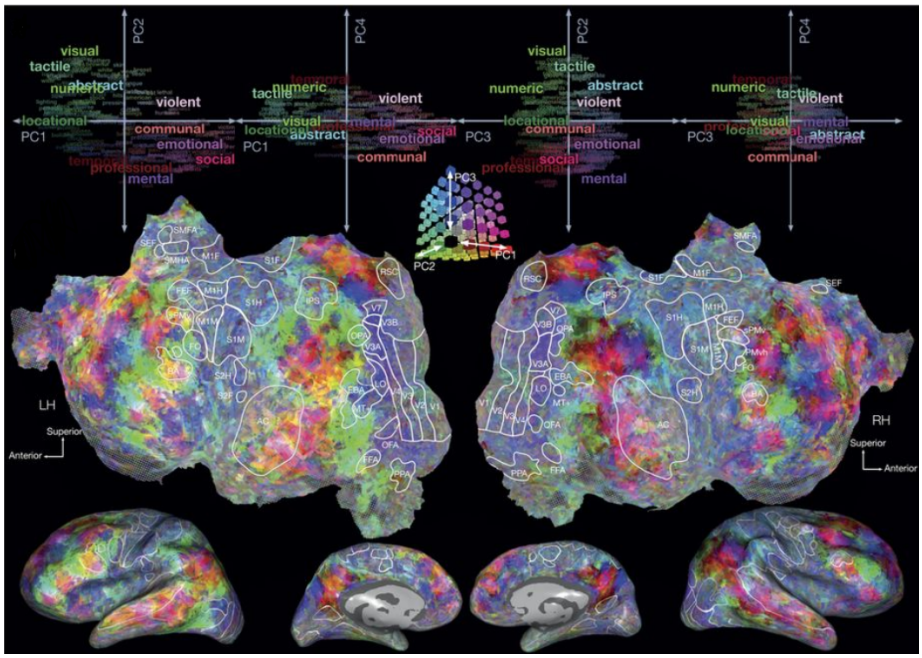


Figure 1.3: Map of the semantic tuning of the human cortex from Huth et al. in [9]. Each color represents the encoding for a broadly defined semantic category. For example, from green to purple the semantic tuning shifts from sensory-related concepts (visual, tactile, numeric) to abstract, human-related, entities (social, emotional, mental).

plex and profound thoughts [22]. This simple intuition already reveals the dynamic nature of language and pushes the investigation to a further level of analysis.

To achieve the level of complexity of language, simple cued retrieval, described in detail for autoassociative networks with the evolution towards an attractor [23][24][25] [26], cannot be the only mechanism at play in the cortex. Neuropsychological models usually describe specific cognitive functions in terms of sequences of specialized routines, expressed as box-and-arrows models, conceptually similar to the flow charts of computer code [27]. However, the neural implementation of these models often implies detailed assumptions on the functioning and connectivity of single neural networks, about which we have no clear evidence [27][28]. Associative mechanisms, on the other hand, are often described as an alternate path, a "lateral thinking" option [29], to perform the same task of a specialized function. Nonetheless, associative retrieval and synaptic plasticity are the only neural mechanisms

with strong empirical evidence [30]. We, therefore, envision the cortex as a memory machine able to achieve its complex behavior only by means of purely associative mechanisms [29].

A network model for the cortex

Following this idea, a Potts attractor neural network [31][32][33] model has been previously proposed [34][35][36], where Potts units, representing patches of cortex, are a key element to describe not only the storage of concepts but also the dynamics from one memory item to the next. In this model, the retrieval of a pattern follows the dynamics towards an attractor while the hopping between memory items, reminiscent of a free sequence of thoughts, is ruled by neural adaptation and inhibition, which both contribute to the destabilization of the current attractor state. We will refer to the jumping between attractors states as *latching dynamics* [37][38][39].

The purpose of this thesis is to take the first steps towards a purely associative model of cortical functions. To tackle this ambitious goal, we will exploit the tools offered by the Potts neural network previously developed in our lab and apply them to 3 specific examples of processes previously described in terms of box-and-arrows models.

In **Chapter 2** we will introduce a Potts neural network model of the cortex. After describing its essential functionality, we will introduce a method to teach instructions to the network. Instructions are coded in the network as *heteroassociative* connections that can be used to model rule-guided behavior, fixed temporal sequences or frequent associations.

In a first attempt to link our theoretical model with experimental data we also designed an exploratory study on word transitions by exploiting a popular Italian game on word associations, named "Il Bersaglio". A detailed description of the experiment is presented in **Chapter 3** while its implementation with our network model is treated in **Chapter 4**.

In the last two parts of this thesis we will introduce two other applications of our Potts model. **Chapter 5** will propose a latching dynamics model of

Mind Wandering, a high-level cognitive function that involves not only the posterior "semantic" cortex but also the hippocampus and the ventromedial prefrontal cortex. Finally, **Chapter 6** will focus on the challenge of modeling the *Phonological Output Buffer* (POB), a key short-term memory device for our language production faculty, enabling us, in its simplest role, to decode and translate the linguistic information coming from posterior cortical regions into sequences of syllables and phonemes.

Chapter 2

A Neural Network Model of the Cortex

The model described in this chapter, previously proposed and studied in [34][35][36][40], was inspired by Braitenberg in [16]. In his studies on the amount of white matter one would expect in a mammalian brain, Braitenberg hypothesized a parcellation of the cortex into $\sqrt{N}$ compartments each containing $\sqrt{N}$ neurons. Assuming each neuron in a compartment to have its axon entering the white matter and connecting to a different compartment, he was able to find good agreement with different quantities measured in the cortex [17]. The following model will approach the problem in a similar way: the cortex is first split into separate local networks, each with its own local attractors. Then the cortico-cortical connections are thought to mediate the interaction between such local networks and the resulting global network is thus proposed as a model of the cortex.

2.1 The Potts Attractor Neural Network

A Potts system of interacting units was first introduced in statistical physics in 1952 [41] and its neural network version was then studied by [31][32] as a generalization of a Hopfield binary network [20], in which units are allowed

to be in only two states, active or inactive. Potts was the name of the student encouraged to study systems of multi-state units, and the advantage brought by a Potts network is precisely to allow for more than a single active state. In modeling cortical dynamics, these additional states can be exploited to represent the different dynamical patterns of activity of a local network of neurons. In this sense, a Potts unit becomes a model for a local patch of cortex, and the network can be thought of as a model of the entire cortex, or at least of the large swaths comprised of its so-called association areas. One can regard such a generalization as an autoassociative network of N Potts units interacting through tensor connections.

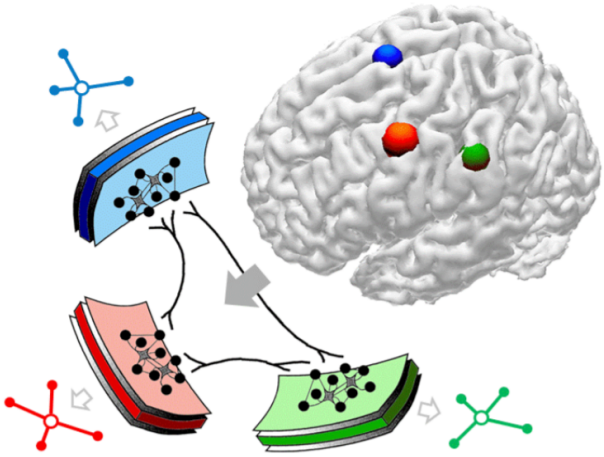


Figure 2.1: Schematic illustration of a small network of 3 Potts units with 4 possible active states. Each state represents the pattern of activity of the local network of neurons in the patch of cortex. Picture taken from [35].

2.1.1 Tensorial interactions between patches

Long-term memories are stored as global patterns through the weights of the connectivity matrix, reflecting a previous Hebbian learning phase.
Each memory μ is then defined as a vector of states taken in the overall activity configuration by each unit i: ξ_i^μ. We take each Potts unit to have S

possible active states, labeled e.g., by the index k, as well as one quiescent state, $k = 0$, when the unit does not participate in the activity configuration of the memory. Therefore $k = 0, ..., S$, and each ξ_i^μ can take values in such abstract categorical set. The simil-Hebbian tensor weights read

$$J_{ij}^{kl} = \frac{c_{ij}}{c_m a\left(1 - \frac{a}{S}\right)} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu k} - \frac{a}{S}\right)\left(\delta_{\xi_j^\mu l} - \frac{a}{S}\right)(1 - \delta_{k0})\,(1 - \delta_{l0}) \tag{2.1}$$

where i, j denote units, k, l denote states, a is the fraction of units active in each memory, $c_{ij} = 1$ or 0 if unit j gives input or not to unit i, c_m is the number of input connections per unit, and the δ's are Kronecker symbols. The subtraction of the mean activity per state $\frac{a}{S}$ ensures a higher storage capacity [31]. In a non-dynamical formulation, the units of the network are updated in the following way:

$$\sigma_i^k = \frac{\exp\left(\beta r_i^k\right)}{\sum_{l=1}^{S} \exp\left(\beta r_i^l\right) + \exp\left[\beta\left(\theta_i^0 + U_i\right)\right]} \tag{2.2a}$$

$$\sigma_i^0 = \frac{\exp\left[\beta\left(\theta_i^0 + U_i\right)\right]}{\sum_{l=1}^{S} \exp\left(\beta r_i^l\right) + \exp\left[\beta\left(\theta_i^0 + U_i\right)\right]} \tag{2.2b}$$

where r_i^k is the variable representing the input to unit i in state k within a time scale τ_1 and U_i is effectively a threshold. From **Eqs.**2.2, we see that $\sum_{k=0}^{S} \sigma_i^k = 1$, and note also that σ_i^k takes continuous values in the (0,1) range for each k, whereas the memories, for simplicity, are assumed discrete, implying that perfect retrieval is approached when $\sigma_i^k \simeq 1$ for $k = \xi_i^\mu$ and $\simeq 0$ otherwise.

As a result of the attractor dynamics, the model thus allows for the conversion from a collection of graded responses from the single patches $\{\sigma_i^k\}$ to a unique global state μ of the cortex. If the connectivity matrix c_{ij} is such that each Potts unit receives the influence of C other units, the quantities S and C (and the total number of units, N) are the main parameters that determine the storage capacity of the network. Global activity patterns, composed of local active and inactive states in the various units, can indeed be stored in

the Potts network by the plasticity model in **Eq.**2.1. They are then attractor states, and the network functions as an auto-associative memory, retrieving one of p stored global activity patterns from a partial cue. Cued retrieval is possible up to a limit $p = p_c$ which is roughly $p_c \simeq \frac{CS^2}{a}$ – very large, whatever the assumptions about C, S and a. A model, therefore, of long-term memory, which can hold, say, millions of items in a network of the size of the human cortex.

2.1.2 Latching Dynamics

When the Potts model is studied as a model of cortical dynamics, U_i is written as $U + \theta_i^0$, where U is a common threshold acting on all units, and θ_i^0 is the threshold component specific to unit i, but acting on all its active states, and varying in time with time constant τ_3. This threshold is intended to describe local inhibitory effects that tend to turn off the units, in keeping with the general observation that inhibition in the cortex is exerted only locally [42].
The time evolution of the network is governed by the equations

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t) \tag{2.3a}$$

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t) \tag{2.3b}$$

$$\tau_3 \frac{d\theta_i^0(t)}{dt} = \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^0(t) \tag{2.3c}$$

where the variable θ_i^k is a specific threshold for unit i in state k, varying with time constant τ_2, and intended to model adaptation, i.e., synaptic or neural fatigue specific to the neurons active in state k. The field h_i^k instead represents the input to the patch i, coming both from cortico-cortical connections

and the internal activity of its neurons, namely

$$h_i^k = \sum_{j \neq i}^{N} \sum_{l=1}^{S} J_{ij}^{kl} \sigma_j^l + w \left(\sigma_i^k - \frac{1}{S} \sum_{l=1}^{S} \sigma_i^l \right). \tag{2.4}$$

Here w is another parameter, the *local feedback term*, discussed in [35][40], that models the stability of local attractors. This helps the network to converge towards an attractor, by giving more weight to the most active states, and thus effectively deepening the attractors. The evolution of the network according to **Eqs.**2.3 can be described as a saltatory dynamics, called *latching* [37]: network activity is driven towards an attractor and remains in the same state until neural fatigue destabilizes it. At this point, the activity shifts, usually towards the closest stable attractor in what appears as a sequence of jumps from one state to the next. An example of latching dynamics can be seen in **Fig.**2.2, where the current state of the network is displayed by plotting, for each time point, the overlaps $\{m^\mu\}$ of the activity pattern with each of the stored memories. The overlap $m^\mu(t)$ of the state at time t with the stored pattern μ is defined by:

$$m^\mu(t) = \frac{1}{aN\left(1 - \frac{a}{S}\right)} \sum_{i=1}^{N} \sum_{k=1}^{S} \left(\delta_{\xi_i^\mu k} - \frac{a}{S} \right) \sigma_i^k(t) \tag{2.5}$$

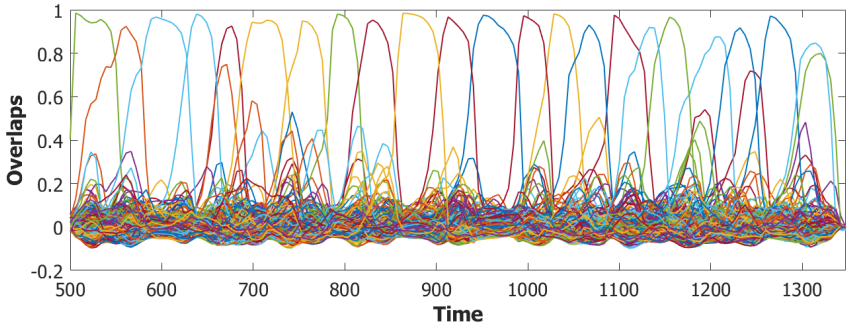


Figure 2.2: Example of latching dynamics. Colored lines represent the overlaps $\{m^\mu\}$ of the network with each stored pattern μ. The network is cued to a pattern at time $t = 500$ and the overlap with that pattern reaches a value close to 1. After some time, the initial pattern becomes unstable, due to the effects of adaptation and inhibition, and the next pattern is retrieved. Here the process goes on indefinitely.

With the correct set of parameters this latching process goes on indefinitely, mimicking an infinite recursion [22][37][43]. In a regime of slow inhibition (i.e., when τ_3 is a longer timescale than τ_1 and τ_2) such dynamics are mainly guided by correlations between the different memory attractors. Since latching transitions occur with very uneven probability among different pairs of patterns, latching statistics can define which long-term memories are readily accessible to the network from any given starting point. These statistics can then be seen as a metric in the space of memory patterns and compared with experimental measures on the objects of interest: semantic or orthographic similarity for words, associative strength for episodic memories, or in many other ways. Such types of correlations and the transitions they facilitate can be dissociated, once one introduces some internal structure in the so-far undifferentiated, homogeneous Potts network.

2.1.3 Slow and fast inhibition

In the previous section, we introduced the concept of latching dynamics and the set of equations (**Eqs.**2.3) that define it. However the evolution of the network strongly depends on its parameters and, in particular, on the time constants τ_1, τ_2 and τ_3, each associated with a particular neural mechanism: τ_1 is the typical time with which activity propagates in the network, τ_2 regulates the firing rate adaptation of the neurons in a patch and τ_3 represents the timescale of inhibition exerted by inhibitory interneurons on, mainly, pyramidal cells. While neuronal excitability, here represented by the timescale τ_1, is very well understood [44][45], adaptation and inhibition have a more complex and mysterious nature. Firing frequency adaptation has been shown to have a wide range of timescales, depending on stimulus type and history [46][47][48]. Local inhibition, instead, relies on the influence on pyramidal cells of, at least, three different classes of $GABA$ergic interneurons [42][49], each acting with its timescale.

For simplicity, in our future simulations, we will consider the case of only a single timescale for modeling adaptation, with the only exception of **Chap-**

ter 6, where the introduction of a second time constant will be discussed. However, we will always assume adaptation to act "slower" than neuronal excitability.

On the other hand, the effect of local inhibition can be both faster or slower than neuronal excitability, depending on the particular type of receptor by which it is mediated. In this thesis we will manly consider the slowly adapting regime ($\tau_1 < \tau_2 \ll \tau_3$), for which inhibition is mediated by only $GABA_B$ receptors. In this regime, correlations among patterns guide the latching dynamics, making this the optimal functioning mode to model semantic transitions, for example, in **Chapters 3-4**. One can also consider what may be called a rapidly adapting regime, ruled by $GABA_A$ receptors, when τ_3 is very short, $\tau_3 < \tau_1 < \tau_2$ (which also has the incidental computational advantage that rich dynamics unfold within limited CPU time). Analytical considerations derived for both regimes can be found in [35][36].

To model the more realistic case in which both slow and fast inhibition are taken into account we could replace the inhibitory or non-specific threshold θ_i^0 with the sum $\theta_i^A + \theta_i^B$ (to denote fast, $GABA_A$ and slow, $GABA_B$ inhibition, respectively) and writing separate equations:

$$\tau_3^A \frac{d\theta_i^A(t)}{dt} = \gamma_A \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^A(t) \tag{2.6a}$$

$$\tau_3^B \frac{d\theta_i^B(t)}{dt} = (1-\gamma_A) \sum_{k=1}^{S} \sigma_i^k(t) - \theta_i^B(t) \tag{2.6b}$$

with, instead of τ_3 either short or long, $\tau_3^A < \tau_1 < \tau_2 \ll \tau_3^B$ and γ_A determining the balance of the two. Note that a more realistic approach would be to consider inhibition at all time scales, in line with experimental findings [49]. First results with this combined inhibition show an improvement in latching quality and length, but a more analytical investigation on this topic will be addressed in [50]. An example application of this regime will be given in **Chapter 6**.

2.2 Instructing the Potts network

The previous definitions are the foundations of a very simple model of the cortex. However complex brain functions may need the introduction of *rule-based* memories (e.g. frequent associations, idioms, fixed sequences of actions, schemas) that cannot be simply described by a purely autoassociative network. Thus we can consider the pairing of configuration μ to configuration ν, which is instructed to succeed it in time, $\mu \to \nu$. These can be partial configurations, defined only over a specific subnetwork, and their heteroassociation may coexist with several other ones, $\mu \to \nu$, $\mu \to \rho$, $\mu \to \psi$, ...
One may regard the long-term memory for a transition $\mu \to \nu$, stored in a subnetwork of the cortex, as a schema, that favors its repetition, with different content in the complementary portion of the network which does not express the schema.
A conceptually distinct situation is when the pairing is only held in short-term memory, to remember a specific sequence for a short time. In this case the favored transition $\mu \to \nu$ is taken to be unique, and reproducing it corresponds to successful remembering in the short term.

2.2.1 Patch-level implementation

Both these situation can be construed to involve the pairing of the complete or incomplete sets of adaptive thresholds $\left\{\theta_i^k\right\}$ that have been raised by the activation of configuration μ to the state variables $\left\{\sigma_j^l\right\}$ that have to be activated in configuration ν.
If μ is not a steady configuration of activity by an underlying extended cell assembly, in fact, but rather it represents a continuous attractor which at the microscopic, intrapatch level keeps changing in time, expressing the pairing in terms of the thresholds $\{\theta_i^k\}$ instead of the activity variables $\{\sigma_i^k\}$ implies that the transition is only favored once the continuous attractor has largely run its course, and is close to be destabilized (by the very same $\{\theta_i^k\}$ thresholds).

Focusing for now only on the long-term heteroassociation we can write the following expression for the couplings:

$$J_{ij}^{kl,het} = \frac{c_{ij}\lambda}{c_m a\left(1-\frac{a}{S}\right)} \sum_{\mu=1}^{p} \sum_{\nu\neq\mu}^{p} G^{\mu\nu} \left(\delta_{\xi_i^\mu k} - \frac{a}{S}\right)\left(\delta_{\xi_j^\nu l} - \frac{a}{S}\right)(1-\delta_{k0})\,(1-\delta_{l0}) \tag{2.7}$$

where λ modulates the strength of the heteroassociation and $G^{\mu\nu} = \{0,1\}$ defines the activity patterns associated one to the other. Plasticity is taken to have been refined over many repetitions of learning the rule, and so the coupling to be optimized for the long-term storage of these transitions. At this point the new field h that unit i in state k experiences is

$$h_i^k = \sum_{j\neq i}^{N} \sum_{l=1}^{S} \left[J_{ij}^{kl}\sigma_j^l + J_{ij}^{kl,het}\theta_j^l\right] + w\left(\sigma_i^k - \frac{1}{S}\sum_{l=1}^{S}\sigma_i^l\right). \tag{2.8}$$

This heteroassociative contribution will be further investigated in the next sections and will be developed according to the specific needs for modeling different cortical functions, without changing its core character defined here.

2.2.2 Measuring the effects of instructions on latching

A purely autoassociative Potts attractor neural network, undergoing latching dynamics, hops from a discrete activity configuration to the next in a sequence of spontaneous transitions. In a slowly adapting regime, such transitions mainly occur between correlated patterns. Heteroassociative instructions may be exploited to link together patterns independently on their correlation. However, to better understand the performance of a heteroassociative network in hopping between randomly correlated patterns, we first need to define both what we mean for correlation between Potts patterns of activity and a measure for the quality of latching.

Correlation between patterns

In a binary Hopfield network, where units can either be *on* or *off*, the correlation between two patterns is proportional to the number of units that are active in both configurations. In a Potts network, however, each unit can either be inactive or active in one of its S possible states. This produces two types of correlations, which we call C_1 and C_2, defined for two configurations μ and ν by

$$C_1 = \frac{1}{aN} \sum_{i=1}^{N} \left(1 - \delta_{\xi_i^\mu 0}\right) \delta_{\xi_i^\mu \xi_i^\nu} \tag{2.9a}$$

$$C_2 = \frac{1}{aN} \sum_{i=1}^{N} \left(1 - \delta_{\xi_i^\mu 0}\right) \left(1 - \delta_{\xi_i^\nu 0}\right) \left(1 - \delta_{\xi_i^\mu \xi_i^\nu}\right). \tag{2.9b}$$

Here C_1 measures the fraction of units that are active and in the same state in both patterns. Conversely, C_2 measures the fraction of units active in both patterns but in different states. Following these definitions, we consider as highly correlated those pairs of configurations with high C_1 and low C_2 values, based on our assumption that related memories should elicit similar representations in the brain. An example of the correlation space C_1-C_2

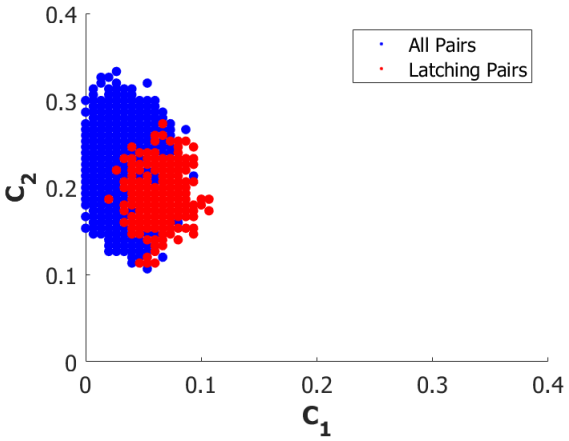


Figure 2.3: Correlation space spanned by a set of $p = 200$ randomly generated patterns with $a = 0.25$, $N = 600$ and $S = 7$. After cueing the network, pairs of patterns for which a spontaneous transition occurred are highlighted in red. The slowly adaptive regime allowed for a hopping between pairs laying in the correlated region of the plot.

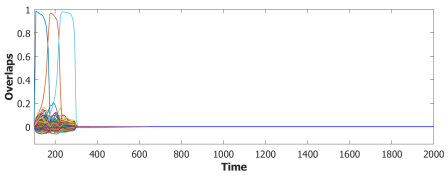
spanned by a correlated latching dynamics can be found in **Fig.**2.3, where latching transitions occured between pattern pairs having high C_1 and low C_2 values.

Quality of latching

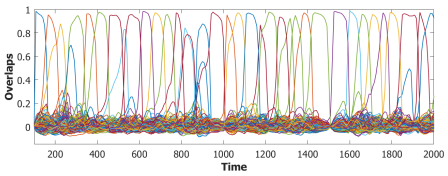
Latching sequences can look quite diverse depending on the parameters of the simulation. The principal dimensions in which the dynamics can differ are the length of latching sequences and quality of latching. For measuring latching length we will consider both the length of the simulation, i.e., the time after which activity in the network dies out, and the number of latching steps performed in the dynamics. A latching step from μ to ν is counted only if the overlaps before and after the switch reach the threshold $m = 0.5$. For latching quality, instead, we will consider the average difference d_{12} between the highest overlap at time t, say m_1, and the second highest, say m_2:

$$d_{12} = \frac{1}{t_2 - t_1} \langle \int_{t_1}^{t_2} (m_1(t) - m_2(t))\, dt \rangle_{cue} \tag{2.10}$$

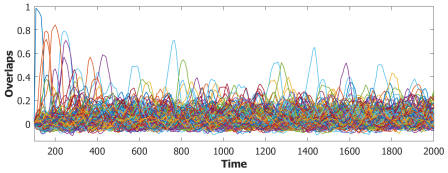
Three very different examples of latching are shown in **Figs.**2.4, where measures of length and quality can be clearly visualized.



(a) Short latching.



(b) Long sequence with good quality.



(c) Low quality and number of steps.

Figure 2.4: Examples of three very different latching dynamics. (a) Sequence with high d_12, but very short and only 3 latching steps. (b) Long sequence with good quality of latching. (c) The network remains active for a long time but both quality, d_{12}, and the number of latching steps are very low.

2.3 Instructed Latching: Results

The previous section focused on the possibility of teaching instructions to a Potts neural network. The mechanism proposed is a heteroassociative component, introduced in the tensorial couplings between units, that links one memory stored in the network with one, or more, other memories, that may follow the first. To avoid the interference between the memories, stored in the network through σ-σ connections, and the instructions, we suggested implementing the heteroassociation on a separate set of connections. Therefore, we introduced a new type of connection between units, mediated by a θ-σ interaction in the field h_i^k (**Eq.**2.8).

In this section we will compare this definition of heteroassociation with one based on a σ-σ interaction. After that, we will consider the case in which instructions are encoded by only a subset of connections, namely by dividing the Potts network into a purely autoassociative subnetwork and a heteroassociative one.

All simulations will be performed with the following set of parameters: $N = 600$, $S = 7$, $p = 200$, $c_m = 90$, $a = 0.25$, $U = 0.1$, $\beta = 12.5$, $w = 0.45$, $\tau_1 = 3.33$, $\tau_2 = 100$ and $\tau_3 = 10^6$. For each batch of simulations, we will compare the effect of varying both λ, i.e., the heteroassociative strength, and the number of instructions D associated with each pattern.

The performance of the heteroassociative couplings will be compared based on the fraction of followed instructions f defined by

$$f = \frac{T_{instructed}}{T_{tot}} \tag{2.11}$$

where $T_{instructed}$ is the number of transitions, with overlap above 0.5, that follow the instructions, and T_{tot} is the total number of latching steps in the sequence. f is 1 if the network always follows the instructions, and 0 if it never does.

2.3.1 Simulation results: Single network

Simulations of both models with θ-σ and σ-σ heteroassociative couplings showed high sensitivity to changes of the parameter λ and to the number of instructions D. Each (D, λ) pair was simulated 50 times, each time by cueing the network with a different pattern. The D instructions were randomly chosen for each pattern.

Latching dynamics

Latching length increased both with the number of instructions D and the strength λ due to the increased input to units they bring through the field h_i^k. The number of steps instead increased up to a certain value of λ, after which the activity of the network becomes so high that no real latching can occur, being all units constantly receiving a high input from hetereoassociative connections. The two types of interaction show similar behaviors when

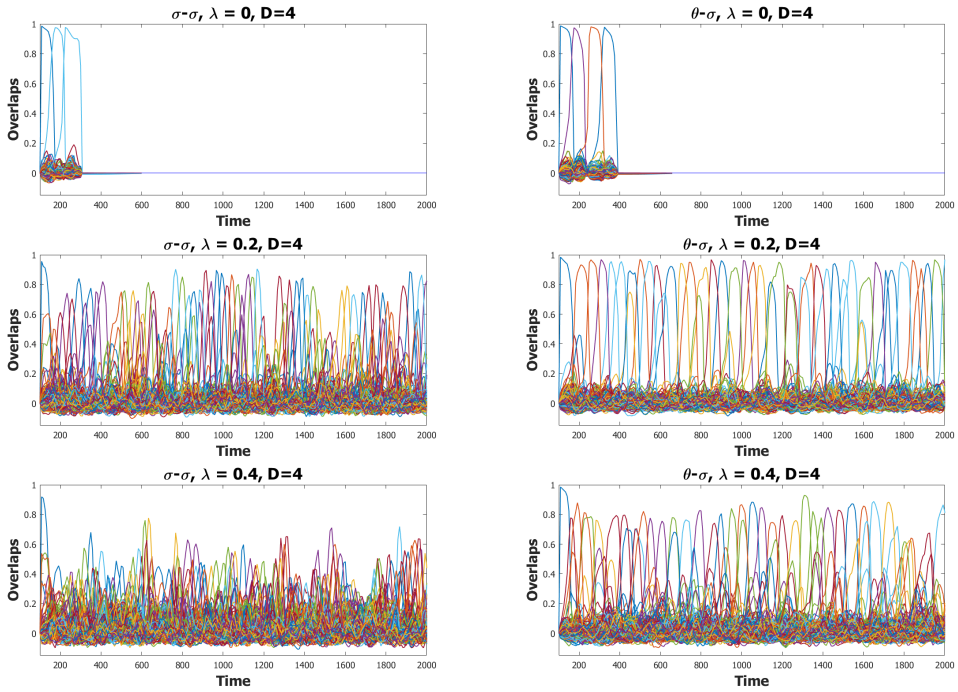


Figure 2.5: Comparison of latching sequences with σ-σ (left column) and θ-σ (right column) interactions for different values of λ. Latching with a θ-σ heteroassociation results less noisy even at high values of λ compared to the σ-σ one.

increasing λ and D, however latching with θ-σ heteroassociation appears less noisy with intermediate values of λ.

Heteroassociative input with the θ-σ coupling becomes comparable to the autoassociative component only when the $\{\theta_i^k\}$'s are high, i.e., when the current attractor is becoming unstable and the network has to jump to the next attractor state.

On the other hand, the σ-σ interaction is always active, both when the network is reaching an attractor state and when it has to make a transition, causing constant interference between stored patterns and instructions. This leads to the rapid decrease of latching quality and to the high number of latching steps, compared to the θ mediated interaction, as we can see from **Figs.**2.6-2.7. For higher values of λ (e.g. $\lambda \gtrsim 0.5$) latching quality deteriorates for both types of interactions.

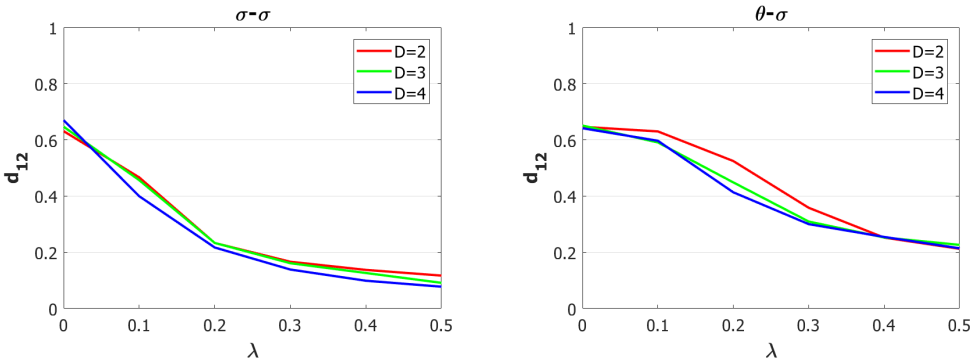


Figure 2.6: Quality of latching as a function of λ for the two heteroassociative mechanisms. d_{12} decreases drastically as λ increases. The steepest decrease occurs for the simulations wih a σ-σ coupling.

Followed fraction

Besides the latching quality, another important factor to consider is the performance of the network in following instructions. When considering latching sequences, the main feature that distinguishes the two heteroassociative mechanisms is the memory effect introduced in the system by the θ-σ coupling. Indeed, as we can notice from the comparison in **Fig.**2.8, if we would only consider as successfully followed instructions the ones given at the pre-

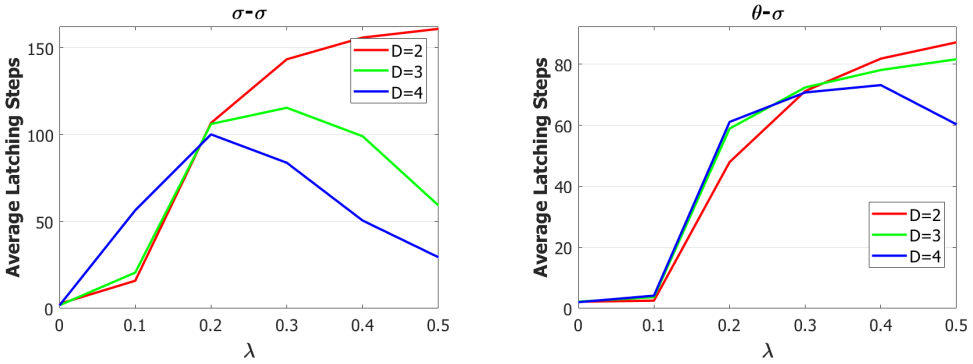


Figure 2.7: Average number of latching steps as a function of λ for the two heteroassociative mechanisms. The maximum number of latching steps in the plots depends on the value of D. The larger input to units given by the σ-σ coupling greatly increases the number of latching steps performed by the network while keeping fixed the total simulation length.

vious step, the simulations with the θ-σ coupling, would have a value of f around 0.5. This would mean that roughly half of the transitions would be

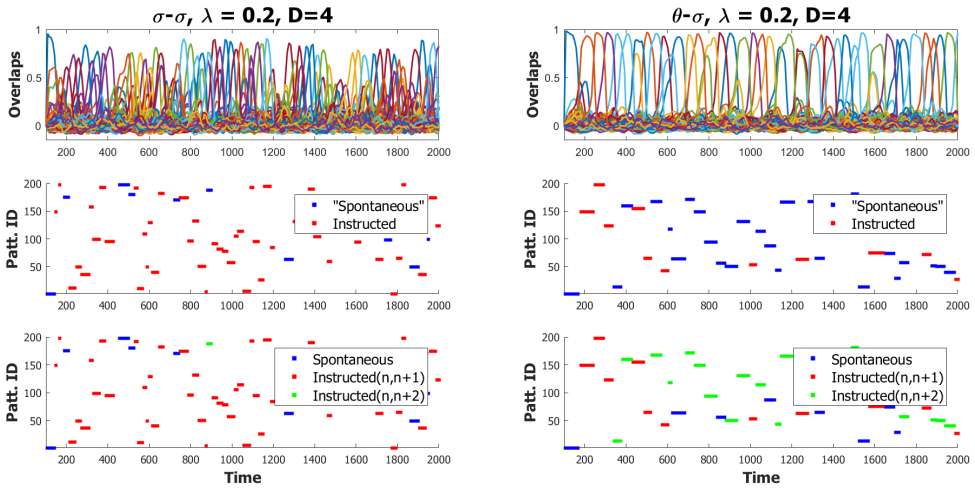


Figure 2.8: Examples of latching sequence for the two heteroassociative couplings with highlighted information on the type of each latching step. The first row shows the overlaps while the second and third rows show the indices of the patterns retrieved by the network. In red are displayed the transitions that follow an instruction from the step before. In green instead, are displayed the transitions that follow an instruction from a pattern two steps before them.

spontaneous. However, if we consider what appears as a spontaneous step in relation to the pattern retrieved two latching steps before it, we can notice that a very high fraction of those "spontaneous steps" can be counted as a "2-step" followed transition ($n \to n+2$ followed fraction plotted in **Fig.**2.9).

On the other hand, with a high enough value of λ, simulations with the σ-σ interaction almost always follow the instructions of the step before, with a value of f approaching 1.

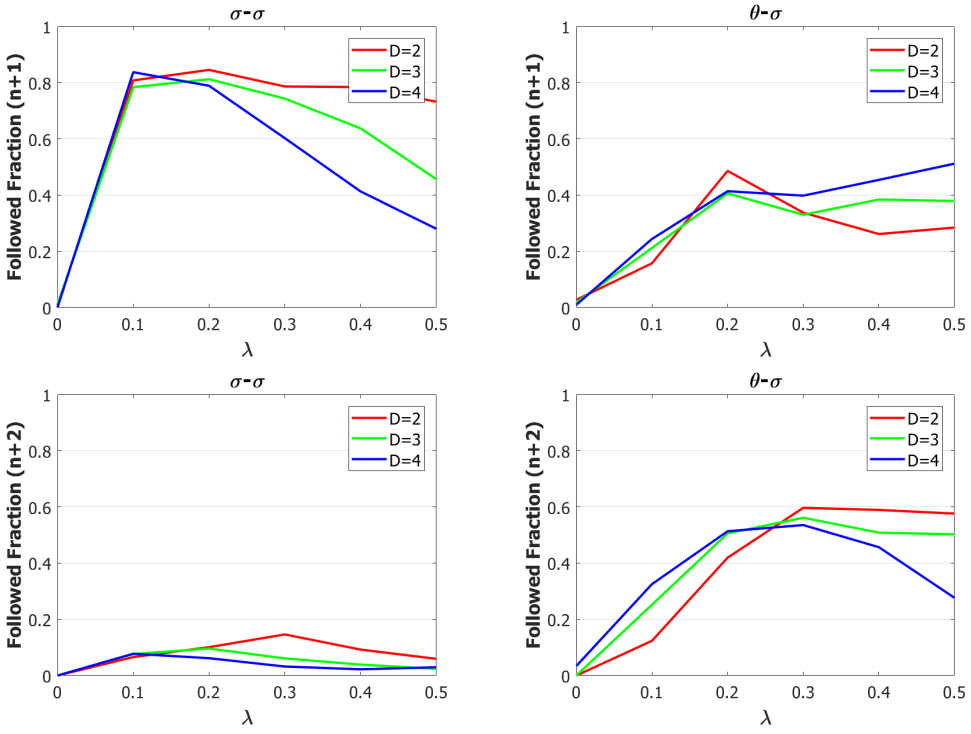


Figure 2.9: Comaparison of $n \to (n+1)$ vs. $n \to (n+2)$ followed fractions for the two types of heteroassociative couplings. First-order followed instructions are higher in number, at low λ, for σ-σ simulations but rapidly decrease as λ increases. Almost no second-order followed pairs can be found with this heteroassociative interaction. First and second-order transitions in θ-σ simulations not only have similar values of f but also are almost constant in number as a function of λ.

The fast decaying quality of latching for simulations with the σ-σ interaction prevents the network from following the instructions when both D and λ increase. Conversely, the θ-σ coupling produces an almost constant value of f when λ varies. The lower values of f, in this case, are compensated by the fraction of transitions that follow the instruction from the pattern retrieved two steps before. This longer memory in the θ-σ network is linked to the slow evolution of the $\{\theta_i^k\}$'s that allows keeping the information on previously retrieved patterns for a brief period of time.

Correlations

Spontaneous latching, in the regime defined by our set of network parameters, occurs between correlated pairs of patterns. Does correlations influence latching also when it is instructed? To answer this question we analyzed the cumulative distribution of correlations between latching pairs, divided by the type of latching transition occurred.

For convenience, we introduce some abbreviations: FP denotes a pair of patterns that follows the instructions, SP a spontaneous transition and AP any possible pair, whether occurring in a latching sequence or not.

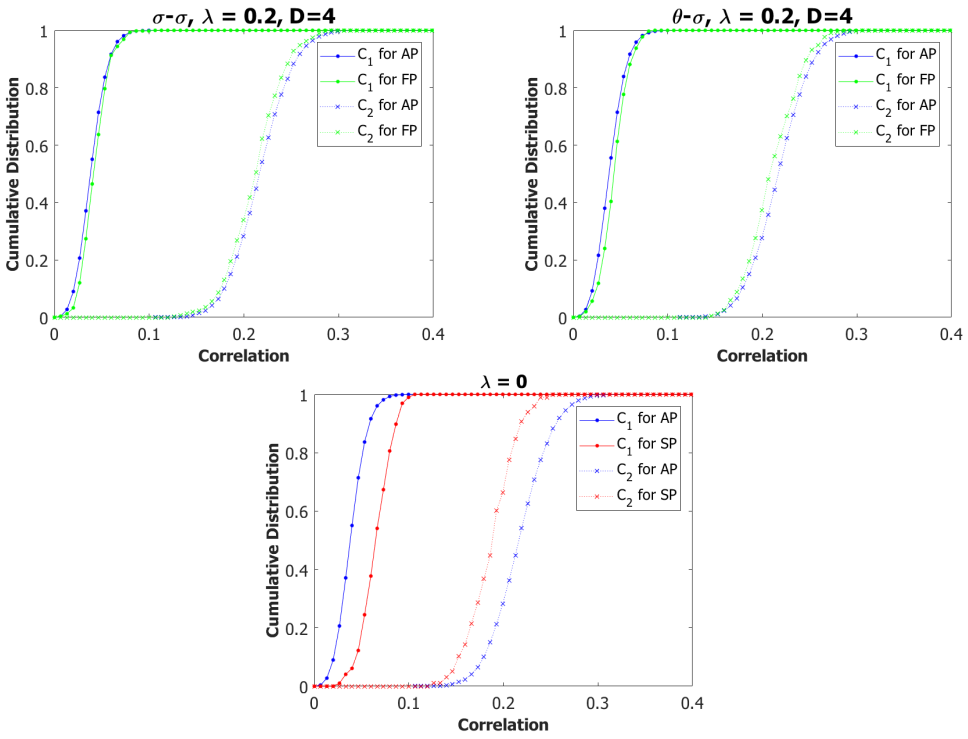


Figure 2.10: Cumulative distributions of correlations between patterns. SP correlations plotted in the second row panel for $\lambda = 0$ show that spontaneous latching occurs between patterns with high values of C_1 and low of C_2. On the other hand FP transitions have a distribution of correlations that approaches the one of AP, meaning that followed transitions are less sensitive to correlations.

Fig.2.10 shows the comparison between the cumulative distributions of correlations for spontaneous ($\lambda = 0$) and instructed latching pairs. While spontaneous latching forces jumps between correlated pairs (i.e., high C_1 and low C_2), instructed latching is less influenced by correlations, as it can be seen

from the FP curves approaching the AP distribution.
The simulations on a single network show crucial results for the implementation of rule-based memories in our network model of the cortex:

- Instructions stored trough **Eq.**2.7 are effective in guiding the dynamics (high values of f);
- θ-σ coupling introduces a memory effect that allows the network to follow instructions from previous latching steps ($n \to (n+2)$ transitions);
- θ-σ interaction produce less noisy dynamics by reducing the interference of instructions with stored memories;
- FP transitions occur even for instructions between uncorrelated pairs of patterns.

Possible improvements to the current model will be discussed in the net section.

2.3.2 Simulation results: the double network

Single network simulations have shown that the heteroassociative mechanisms introduced successfully drive network choices when latching occurs. However, latching quality is deteriorated when both the number of instructions and the heteroassociative strength increase. To address this problem,

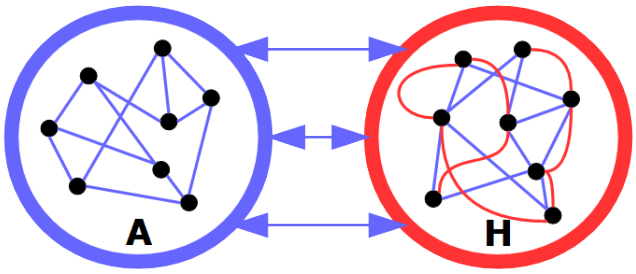


Figure 2.11: Schematic representation of a double network with a purely autoassociative subnetwork A and a heteroassociative subnetwork H. Red connections represent heteroassociative couplings. Blue lines instead represent autoassociative connections between units.

we performed the same simulations on a bipartite network: subnetwork A is a purely autoassociative network while H mixes both an autoassociative and a heteroassociative character. The goal of the simulations was to compare the performance when only a subset of the connections encode instructions. The only changes in parameters with respect to the previous simulations are in the number of units and in the connectivity. Networks A and H have $N_A = N_H = 300$ units and $c_m = 90$ input connections per unit. The c_m connections are equally divided into internal and external ones.

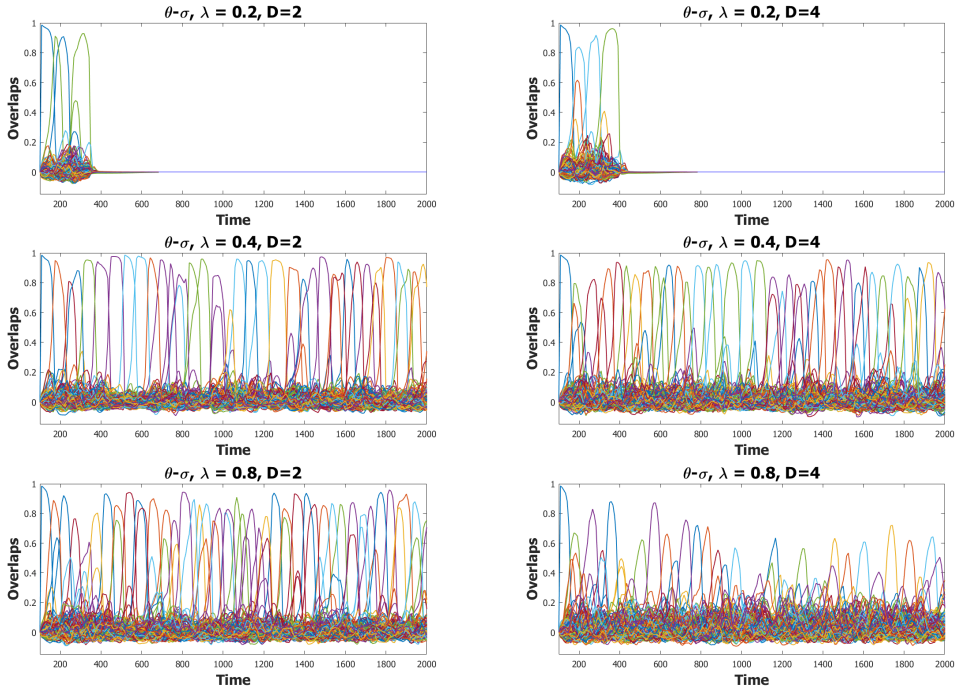


Figure 2.12: Comparison of θ-σ latching sequences with $D = 2$ (left column) and $D = 4$ (right column) for differnt values of λ. Heteroassociative instructions are stored in c_m^H connections only. Latching quality improves at low D or with low enough values of λ. Latching length increases with λ even with a lower amount of heteroassociative connections.

In this way A has $c_m^A = 45$ input connections from its units and $c_m^{HA} = 45$ from units in network H. The same structure is set to H with $c_m^H = 45$ and $c_m^{AH} = 45$. The heteroassciative couplings in **Eq.**2.7 are assigned to the internal connections of network H, i.e., $\{c_{ij}\} = \{c_m^H\}$.

Since no qualitative difference from previous simulations has been found in the comparison between σ-σ and θ-σ couplings, we will show results only for

the θ-σ mechanism.

Fig.2.12 shows examples of latching with $D = 2$ and $D = 4$ for $\lambda = \{0.2, 0.4, 0.8\}$. Compared to the single network implementation, these examples show that higher values of λ ($\lambda \gtrsim 0.2$) are needed to reach the infinite latching regime. However the fraction of followed instructions, past the $\lambda \approx 0.2$ threshold, remains similar to the single network case (**Fig.**2.13, first row). These results suggest that relegating the heteroassociative cou-

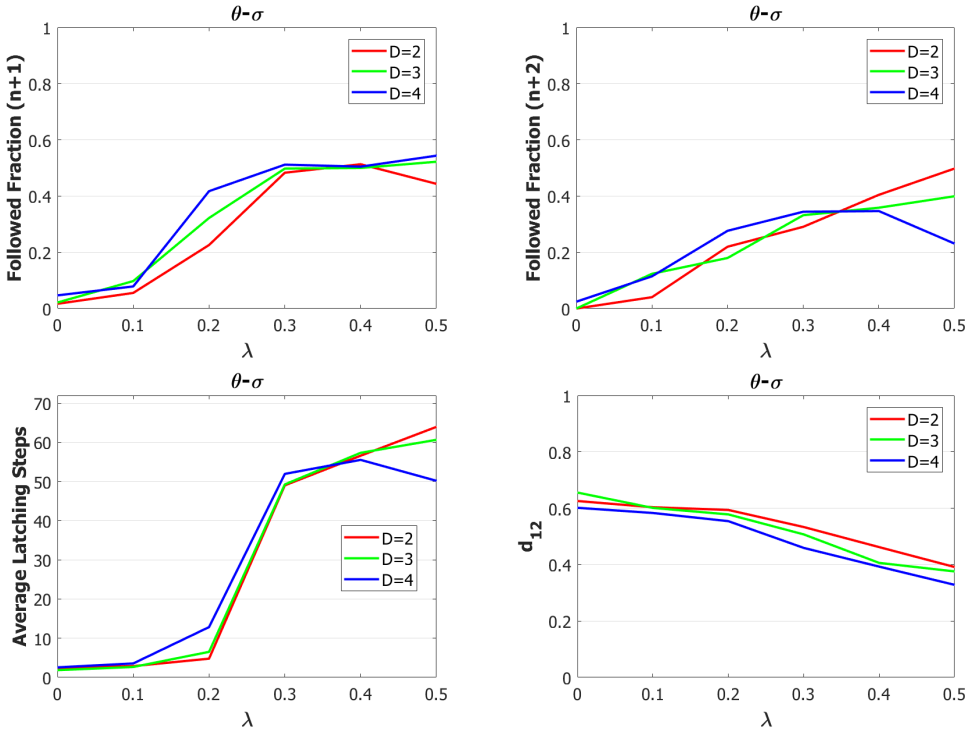


Figure 2.13: Summary of the performance of a bipartite network with a heteroassociative component. In the first row are plotted the fraction for the two orders of followed transitions. The behavior does not deviate from the single network case. In the second row are plotted the average number of latching steps (left) and the quality of latching (right). Latching steps dramatically increase for $\lambda \gtrsim 0.2$ while d_{12} slowly decreases with λ.

plings to a subset of connections does not prevent the network from following instructions or rule-based associations, allowing for the introduction of more structured models of the cortex, like the ones developed in the next chapters.

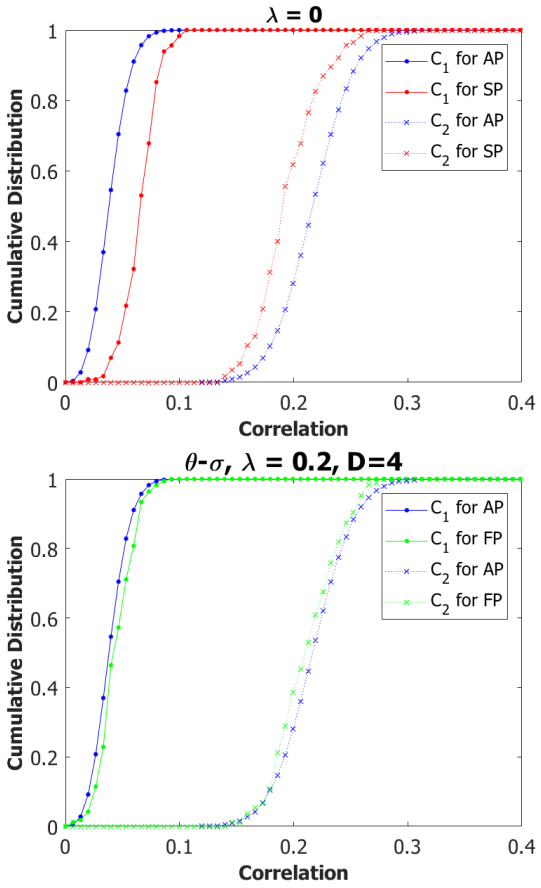


Figure 2.14: Cumulative distributions of correlations between patterns in the two network architecture. SP correlations plotted on the top panel for $\lambda = 0$ show that spontaneous latching occurs between patterns with high values of C_1 and low of C_2. On the other hand FP transitions on the bottom panel have a distribution of correlations that approaches the one of AP, similarly to the single network case.

Chapter 3

An Experimental Investigation of Latching

Many studies on the neural substrates of language measure responses to unnatural stimuli such as semantic or orthographic violations, non-existent words, strings of random letters, or very fast serial word presentations. These highly controlled conditions have the advantage of isolating the linguistic processes of interest, at the expense of a more ecological understanding of language. Latching dynamics, instead, can be thought of as a primitive model of natural language production, where each latching step represents a concept, or even a word, in a hypothetical sentence. Thus, an experimental investigation of latching would naturally develop into a study on spontaneous speech production.

The human mind can flexibly combine the meanings of individual words to construct structured sentences. However, an experimental investigation of this processing is challenging. For example, it is still a matter of debate how our brain encodes sentence meaning or, even, how we could quantify such a thing. While vectorial semantic models proved to be predictive in many linguistic tasks [8][9][51], their conversion to models of sentence meaning is still at the initial stages. Anderson et al. [52], for example, proposed to model sentence meaning with simple additive or multiplicative operations on

the space defined by word vector models of semantics. A sentence, however, can have a very different meaning than the simple summation of the words that compose it: the sentences *a dog bites the man* and *the man bites a dog* share the same words but have opposite meanings, resulting in different brain activations [53]. Therefore syntax must be the first ingredient to be included for a proper investigation on this topic. However, how our brain builds and interprets syntactical structure is still an area of open discussion in both linguistic theory and neuroscience [54][55].
From our modeling perspective, implementations of sentence meaning and syntax are currently being developed for the Potts neural network [56][57], with similar mechanisms as the ones described in the previous chapter, but the generated utterances are still not comparable with natural human production.
For all the previous reasons we decided, for now, not to focus on spontaneous speech, but to try, instead, to reduce the problem to its core and most simple mechanism, namely the interaction between single words. Similar words tend to elicit one another, a process which our Potts network represents in terms of its latching dynamics. In this experimental investigation, we will induce the participants to link words through specific types of association, allowing us to study the neural trajectories of activity behind word transitions.

3.1 Experimental design

Early attempts

Our focus in this work is on word transitions in ecological conditions. For this reason, we designed a task that could be entertaining for participants by using only existing Italian words presented at a slow pace. To do so, we took inspiration from a game, named *Il Bersaglio*, literally *The Target*, published weekly on a popular Italian journal. The game appears as a set of words arranged in random order on a shooting target. The goal of the player is, by starting from a highlighted word, to find the only correct sequence that leads

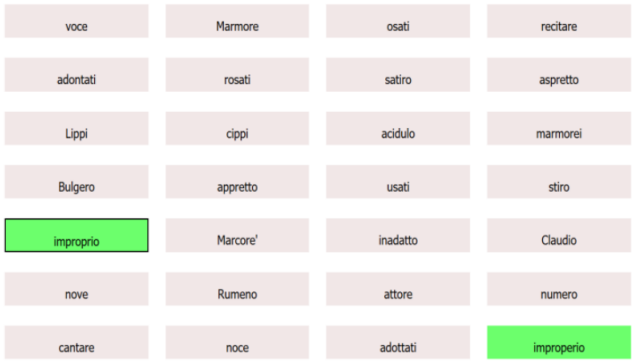


Figure 3.1: First implementation of the experiment. Subjects had to play 12 rounds of the game by associating all the words on the screen, starting from the word highlighted in the bottom right corner.

to the center of the target and includes all the words in the set. The player, at each step, has to find the next word in the sequence, among those on the target, which satisfies one of a well-defined set of associations. The final sequence thus resembles a latching process on the given word list, guided by the rules of association imposed by the game. However, the early version of the experiment in **Fig.**3.1, mimicking the original game, was too complex to constrain into a proper experimental setting.

Revised design

In our final experimental transposition of the game, participants were asked to select words (Targets), appearing one after the other in the center of the screen, which satisfied precise rules of association with a reference word (Prime), allowing us to investigate the neural signatures of latching between stored memory items.

The stimuli were Italian words that could be associated with each other, either orthographically or semantically, as shown in **Fig.**3.2. The list of associations for this pilot study was taken from an online database of a computer version of the game. A more strictly controlled set of stimuli will be used for the confirmatory experiment presented later in this thesis.

Figure 3.2: List of the seven rules of association with relevant examples. Prime words are in capital letters.

Both classes of allowed associations could be further divided into different subtypes: *addition*, *omission* or *change* of one letter for the word form (orthographic) class and *synonym*, *antonym*, *semantic* (encyclopedic) or *association* of ideas (collocations, words often co-occurring) for the semantic class. Participants were instructed to recognize these seven subtypes. The 60 *word form* trials were subdivided into 20 trials per subtype of transition. For the semantic class, instead, 15 trials were assigned to each of the four categories. Each trial started with the presentation of the prime word, in capital letters, followed by a maximum of four words, only one of which could be a correct association. Trials with no correct association were also included as a control condition and the total 180 trials were equally divided for the three types of transition, **"Semantic"**, **"Word Form"** and **"No Association"**. Each trial ended when the correct associate was selected or after the presentation of all four words if no association was found. Both trial and word orders were randomized for each participant.

3.2 Behavioural Experiment

3.2.1 Methods

The first experiment was a reaction time (RT) task. The response phase overlapped with the presentation of the word. If the word appearing on the screen was thought to be related to the prime, the participant had to press a button, as fast as possible, to select it as an associate. Otherwise, the participant had to wait until the presentation of the next word, without any button press.

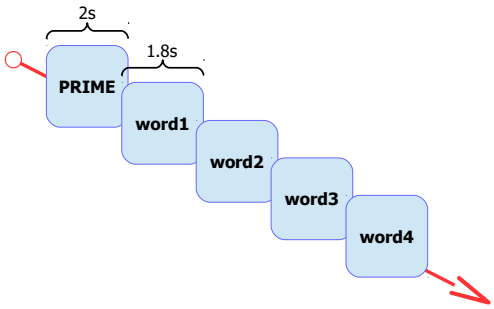


Figure 3.3: Trial structure. Red arrow represents time. The correct association (Target) could be in any of the four words position. Words appeared on the screen one at a time in the center of the screen.

Prime words were presented in capital letters for 2s, test words for 1.8s, and a response (button press) was allowed for as long as the word was displayed on the screen. Reaction times were then calculated from the first frame of the monitor displaying the selected word. Participants received feedback on the button press according to the correctness of the response.

3.2.2 Results

For the behavioral experiment, 22 subjects were tested for a total of 2362 correct trials ($\sim$ 89% accuracy, $22 \times 120 = 2640$ total trials). Semantic associations are, on average, more difficult to recognize ($\sim$ 84% accuracy) than word form ones ($\sim$ 95% accuracy), as shown in **Fig.**3.4. Nonetheless, the fraction of correct responses is still very high for both classes of transitions,

proving that participants could easily accomplish the task.

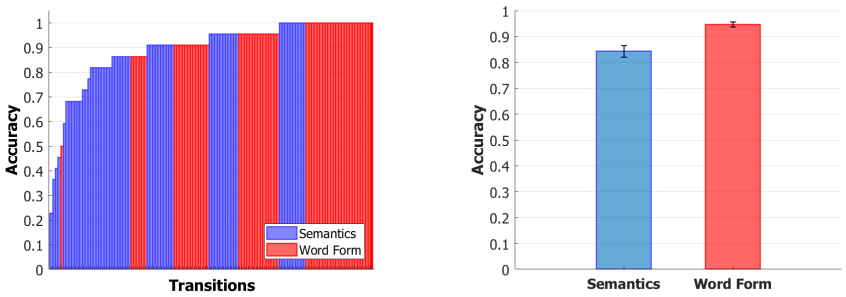


Figure 3.4: Distribution (left) and average proportion (right) of correct responses for the two categories of transitions. Semantic associations are more difficult to recognize than word form ones. A Wilcoxon rank-sum test on the distributions of correct responses proved this difference to be statistically significant with $p < 0.001$. Error bars represent the standard error of the mean.

As expected from the analysis of the accuracy, the distributions of reaction times also show average faster processing for word form associations. A Wilcoxon rank-sum test proved the difference of reaction times of $\sim$ 70ms to be highly significant with $p < 0.001$ (**Fig.**3.5).

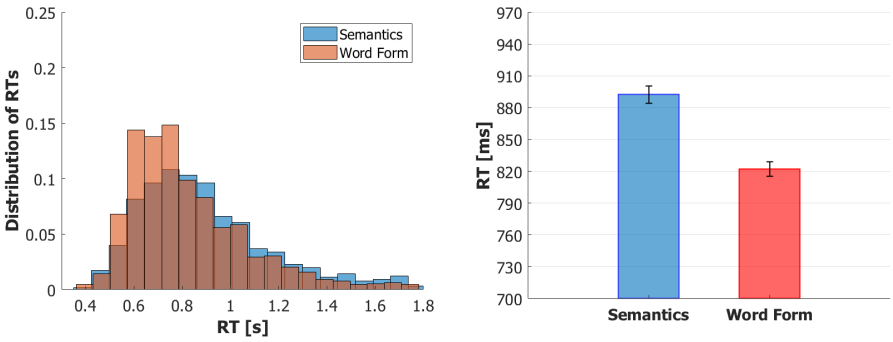


Figure 3.5: (Left) Distribution of Reaction Times for both conditions. (Right) Mean Reaction Times. Error bars represent the standard error of the mean.
Semantic transitions are recognized slower than word form ones, reflecting the increased difficulty of the first condition.

Analysis of transition subtypes

By looking at the single subtypes of association in **Fig.**3.6, we can notice that overall the three word form subtypes are associated with shorter average reaction times. However, *omission* transitions show consistently lower reaction

times. This could be explained as a word-length effect; however, such interpretation would suggest longer reaction times for *addition* with respect to *change*. For this reason, an effect of letter identities should be investigated in future experiments, since the *omission* is the only transition type which does not use different letters than the ones in the prime.

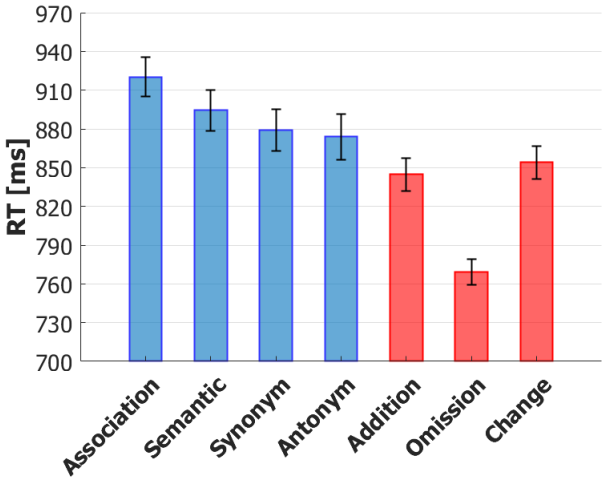


Figure 3.6: Mean reaction times for the seven types of allowed transitions. Error bars represent the standard error of the mean.

For what regards semantic transitions, the main difference in reaction times appears between *association* and *synonym/antonym* conditions. *Semantic* transitions instead lie in between and the differences in reaction times between *semantic* and both *association* and *synonym/antonym* are not statistically significant (see **Fig.**3.7 for statistical analysis results), indicating that encyclopedic transitions may share both the semantic and the frequency aspects of the *synonym/antonym* and of the *association* types of transition, respectively. Stricter and more controlled definitions of the classes of transitions are needed to further investigate their supposedly different representations. Therefore, in the following EEG experiment we will mainly consider the **semantic** and **word form** categories, while a more controlled stimulus set will be used for the confirmatory experiment.

	semantic	synonym	antonym	addition	omission	change
association	-	*	**	***	***	***
semantic		-	-	**	***	*
synonym			-	*	***	-
antonym				-	***	-
addition					***	-
omission						***

Figure 3.7: Significance results from comparing the reaction times between all pairs of conditions with a Wilcoxon rank-sum test. * stands for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$.

3.3 EEG Experiment

Given the encouraging results of the behavioral experiment, we decided to investigate the neural signatures associated with our word transitions through an electroencephalographic (EEG) experiment.

3.3.1 Methods

Experiment design

For this second experiment, we kept a similar design as in the behavioral version but with a different modality for the response. Participants this time had to judge the relation with the prime after each word, presented in the center of the screen for 1.5s. The response phase was signaled on the screen with the presentation of a 'YES' on one side of the screen and a 'NO' on the other. Participants were instructed to reply 'YES' if the previous word was related to the prime and 'NO' otherwise. The answer was selected by pressing 'f' or 'j' on the keyboard. Each participant saw the 'YES' always on the same side of the screen, which was initially assigned at random, so that 18 participants had it on the right while the other 11 participants had it on the left.

The 'YES/NO' question allowed us to collect the EEG signatures for the *no association* condition, defined as the set of trials with correct 'NO' responses. Participants were never asked which rule they followed when a 'YES' response was given; therefore, *semantic* and *word form* conditions were distinguished only according to the stimulus set.

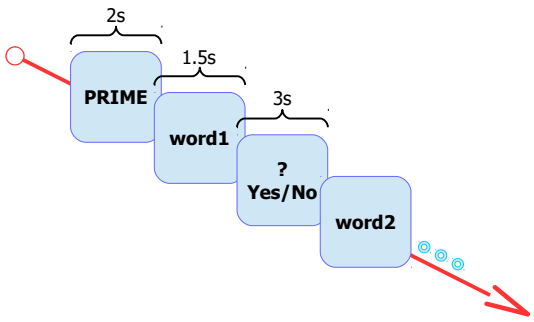


Figure 3.8: Trial structure. The red arrow represents time. The correct association (Target) could be in any of the four words position. Words appeared on the screen one at a time in the center of the screen. The response phase started after the 1.5 seconds of word presentation.

Data collection

EEG data were collected in a sound-proof booth for 29 subjects. The brain activity was recorded with a 64 channel BioSemi ActiveTwo system (BioSemi Inc., Amsterdam, Netherlands) at a sampling rate of 1024Hz. A Common Mode Sense (CMS) active electrode was used as the reference, and a Driven Right Leg (DRL) passive electrode was used as the ground. Two external electrodes placed on the right and left of the outer canthi, and one external electrode placed under one eye were used to obtain horizontal and vertical electrooculograms (EOG). Two additional electrodes were placed on the left and right mastoids, and their average was used as reference in the analysis. Individual electrode offsets were kept between $\pm 30\mu$V. Participants were requested to minimize movement throughout the experiment except when they had a break.

Data preprocessing

Preprocessing and analysis were done in MATLAB with the *eeglab* toolbox [58]. Collected data underwent different cleaning procedures, described in the following pipeline:

- *Downsample*: the number of data points was first reduced to achieve a sampling frequency of 256Hz.

- *Filtering*: to downsampled data were then applied a high-pass filter (lower bound at 0.1Hz) to remove the continuous component and a low-pass filter (upper bound at 40Hz) to remove the 50Hz electrical noise.
- *Segmentation*: we epoched the continuous EEG data into short time-series of 1.5 seconds around the onset of the test word presentation. Each epoch started 100ms before word presentation to have enough timepoints for baseline removal.
- *Channel rejection*: bad channels were removed with the *eeglab* function `pop_rejchan`, for which we used the three available methods. *Kurtosis* threshold was set to 4σ, *joint probability* threshold was set to 4σ, and *abnormal spectra* was checked between 1 and 30 Hz, with a threshold of 3σ.
- *Independent Component Analysis*: ICA was performed on epoched data to remove eye blink artifacts [58][59].
- *Baseline*: data 100ms before word onset was averaged and the resulting value was subtracted to align all trials.
- *Trial rejection*: trials containing extreme values ($\pm 200\mu$V) and improbable trials (4σ threshold on the trial occurrence probability distribution, calculated from the total distribution of values in the set) were removed.

After the cleaning procedure, each epoch was assigned to its condition according to the response given by the participant, including only correct responses. The datasets of each condition were pruned by randomly discarding trials to ensure the same number of trials per condition. Data from different participants were then merged in a single 'super subject' dataset. All the epochs belonging to the same condition were then averaged together for a first inspection at Event-Related Potentials (ERP).

Statistical analysis

Statistical analysis of EEG data was performed with a nonparametric clustering method exploiting the spatiotemporal evolution of EEG signals on the scalp [60]. The algorithm compared the difference between 2 conditions, for every time point and electrode, with a nonparametric permutation *t test*. Adjacent spatiotemporal points with a statistically significant difference ($p<0.05$) were clustered together. The candidate clusters found with this procedure were then statistically tested with a nonparametric permutation *t test* to assess their significance [61][62].

3.3.2 Results

Clustering analysis

The clustering analysis of the EEG data returned the four significant clusters reported in **Fig.**3.9. Two clusters have been found for the comparison between *semantics* and *word form* conditions, and one for each of the comparisons with the *no association* condition. No significant difference was found from the comparison at the single subtypes level so we will focus only on the three main conditions.

Potentials peaking around 170ms (N170) and 200ms (P200) after stimulus onset are usually related to automatic processing of visual stimuli [63]. In our experiment, these components appear in all clusters but no significant difference can be found for our three conditions.

An N400 component (negative deflection of the signal peaking at 400ms after stimulus onset), which the dominant view interprets as a correlate of lexical retrieval and semantic memory access [64][65], appears for all conditions but with different modulations. Clusters 3 and 4 in **Fig.**3.9 show a highly significant difference around 400ms between both classes of association and the *no association* control condition, with a more negative deflection for the last compared to the first. However, no significant difference can be found in the N400 between *semantic* and *word form* transitions.

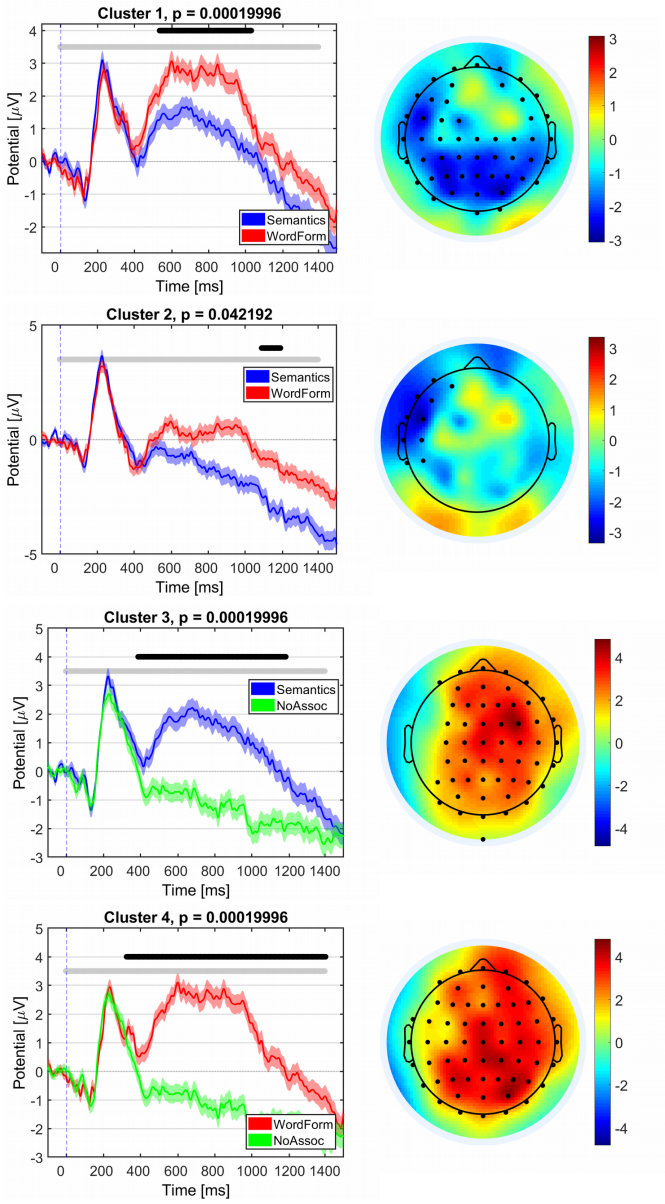


Figure 3.9: Clusters found by the clustering algorithm. (Left columns) EEG traces for the pair of conditions considered, calculated as the averaged signal coming from all the electrodes in the cluster. Grey bar is the time window considered in the analysis while the black bar highlights the statistically significant region. (Right columns) Spatial distribution of the clusters. Electrodes included are highlighted by black dots. Color represents the difference between the signals of the two conditions in the analysis. Note that the blue (lower potential) in clusters 1 and 2 to the right matches, in this case, the unrelated color coding on the left, where blue (semantic transitions) show a significantly lower potential than red (word form) ones. For cluster 3 the two color codes do not happen to match, while they do again for cluster 4.

A slow positive component peaking around 600-800ms, commonly referred to as P600, can be found for both types of associations compared to the *no association* control (clusters 3 and 4). Historically, P600 has been studied in sentence comprehension tasks, leading to its interpretation as the reflection of a process of integration and re-analysis of a word in its context [65][66]. In our experiment, this component not only appears when there is an association between the current word and its prime but also it shows a modulation of amplitude that distinguishes *word form* (higher peak) from *semantic* transitions. Note that cluster 1 includes most posterior electrodes as well as left and right frontal ones, differently from the usual centro-parietal distribution reported in the literature.

Finally, a left-anterior cluster (cluster 2) highlights a late slightly significant difference between *semantics* and *word form*. Note that unlike the broad P600 distribution over cluster 1, this late significant difference around $1.1s$ (well after the average reaction time in the behavioral experiment) is expressed in a small cluster of electrodes concentrated over the left frontal cortex. This may suggest a confirmatory process, perhaps the repetition or inversion of the semantic link, or the active mental execution of the orthographic change required to match prime and target stimulus. Interestingly, a late Left Anterior Negativity (LAN) component has been found in studies on morphosyntactic agreement [67][68], suggesting a processing of reanalysis of the prime-target pair.

Visual inspection of grand-averages

The statistical analyses of our exploratory study resulted in four spatiotemporal clusters with two main significant effects: an N400 component that distinguishes between recognized associations and the *no association* control, and a P600 potential higher for *word form* transitions compared to *semantic* ones. Nonetheless, further interesting effects may be hidden in the row distribution of ERPs, to be tested in future experiments.

The ERP traces in **Fig.**3.10 show two additional effects, not captured by the

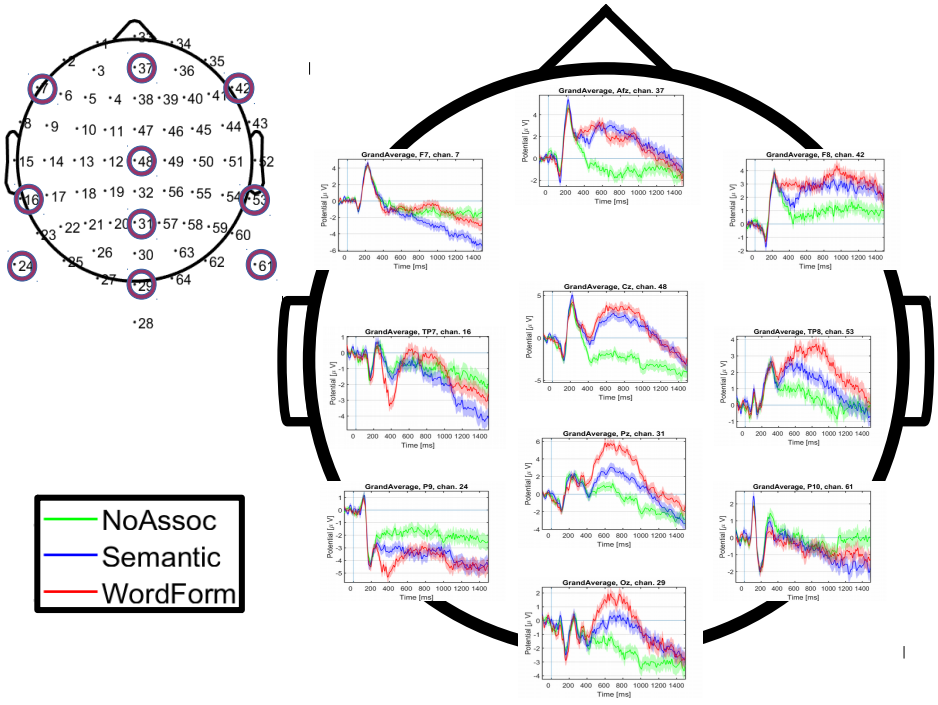


Figure 3.10: ERP traces for the 3 main conditions for 10 selected channels (electrodes) broadly spanning the whole scalp, represented here as flattened in 2D. The ERPs of each condition are obtained by merging together the datasets from all the subjects and averaging the signals of each electrode. Shaded area represent the standard error of the mean at each timepoint.

statistical analysis, which may be worth investigating in a second experiment. The first effect is a modulation of the N400 in right-anterior channels (electrode 42, F8, in the example traces) that differentiates our three conditions; negative deflections are more accentuated respectively for *no association* and *semantics* compared to *word form*. The second is again another modulation of the N400 appearing in left-posterior channels. This time the main effect regards *word form* transitions, with a more negative potential, compared to the other two conditions. The absence of a cluster for this left-posterior N400 could be due to a possible issue with the construction of the adjacency matrix for the electrodes in this region of the scalp, not fitting the spatial distribution of the ERP.

3.4 A new version of the experiment

3.4.1 Experimental design

The success of the exploratory experiment leads us to a deeper and more controlled investigation of word transitions. The first experiment showed no difference in the EEG traces of the different subtypes belonging to the same category. For this reason, we planned a second experiment with better-defined conditions in order to highlight the possible variability inside the two macro-conditions of *semantics* and *word form*. To do so, we kept the same experimental tasks described in **Figs.**3.3 and 3.8.

The new experiment, currently ongoing in the EEG phase, is designed to enhance the possible difference between semantic transitions and collocations (earlier included as *association* transitions) by considering as collocations mainly words co-occurring in frequent Italian idiomatic expressions. A further condition then aims to distinguish between noun-noun and noun-adjective (or adjective-noun) transitions, as the first step towards a more comprehensive and ecological study on language, including both semantics and syntax.

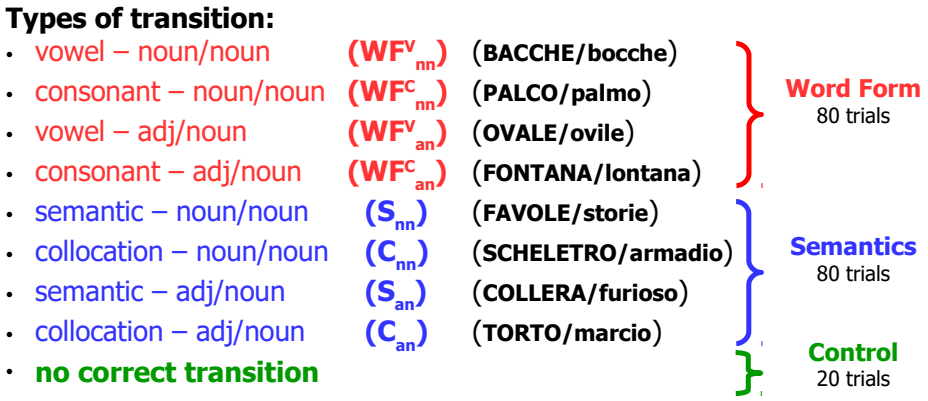


Figure 3.11: List of the eight conditions of the second experiment, with relevant examples. Prime words are in capital letters.

On the other hand, *word form* conditions have been restricted to single letter

change transitions. In this case, the noun-noun/noun-adjective distinction is added to a vowel-vowel/consonant-consonant condition, to check for a possible effect of letter identity. A summary of the eight total conditions with relative examples in Italian is shown in **Fig.**3.11.

3.4.2 Stimuli

Semantic and orthographic transitions for this second experiment have been collected from a dictionary of Italian words, extracted from the `subtlex-it` annotated corpus [69]. The dictionary was first cleaned by removing anomalous entries (e.g., wrong spelling, proper names, extremely long words, etc.) and by including only words tagged as nouns or adjectives. A random set of 2000 words was then chosen in the range of frequency (i.e., number of occurrences in the corpus that ranges from 1 to roughly $2 \cdot 10^6$ counts) between 10^2 and 10^5 counts. With the resulting set, we calculated the orthographic and semantic distances for all the possible pairs of words. We used the

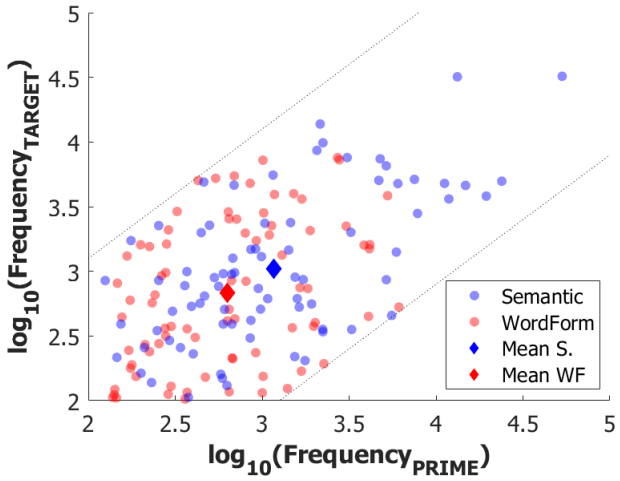


Figure 3.12: Scatter plot of the frequencies of each transition pair included in the final stimulus set. *Semantics* pairs are plotted in blue while *Word Form* ones are in red. The diamond points represent the mean frequencies for the two categories of association. Dotted lines represent the frequency range used for selecting transitions.

Levenshtein measure [70] for orthographic distance and the cosine distance, obtained from the *Snaut* [71] website, for evaluating semantic similarities.

This last measure was obtained from training a `word2vec` model with 200 dimensions and a 5 words window on the *ItWaC* corpus to derive a vectorial space of word meaning.
Pairs of associates were built for the eight conditions based on the measures previously described. Additional words were later included to balance the number of trials per condition. As shown in **Fig.**3.12, transitions were built between words whose frequency satisfied the following relation:

$$|log_{10}(f_{W_1}) - log_{10}(f_{W_2})| < 1.1, \tag{3.1}$$

where f_W is the number of counts mentioned above.
For collocation trials we included as associates words appearing in Italian idioms or that compose together the name of an object, book, etc. (e.g., CODA/paglia, DENTE/giudizio, CORDE/vocali, DIVINA/commedia).
Collocation pairs have on average greater semantic distances than semantic ones, at least for what regards noun/noun pairs. Adjective/noun pairs on average are estimated by the vectorial model as farther apart in meaning than noun/noun pairs. To check that transitions in the *collocation* conditions were different than the purely semantic ones, we defined a measure of collocation distance in our set by taking inspiration from the *pointwise mutual information* (PMI) [72]. From *Google Ngram Viewer* [73] we collected the frequencies (i.e., the raw counts) of the single words and of their collocations in a 3 words window and we calculated the PMI for each transition pair with the following formula:

$$\mathrm{PMI}(W_1, W_2) = log_2\left(\frac{f_{W_1,W_2}}{f_{W_1} f_{W_2}}\right) \tag{3.2}$$

where f_{W_1,W_2} is obtained by summing the frequency of all the collocations found in the search. The PMI values were then normalized by the maximum value in our set and negative values were set zero (i.e., words occurring together less than chance were set to the chance level). A collocation distance (actually, a quasi-distance, as it is not guaranteed to satisfy the tri-

angle inequality) for the couples (W_1, W_2) was then obtained by calculating $1 - \mathrm{PMI}(W_1, W_2)$. After the inclusion of a collocation distance, semantic distances were re-evaluated with a model less sensitive to syntactic structure, with 400 dimensions and trained with a 9 words window.
The final stimulus set showed different distributions of the defined distances, allowing for the categorization into the eight total conditions with 20 trials for each condition (**Fig.**3.13).

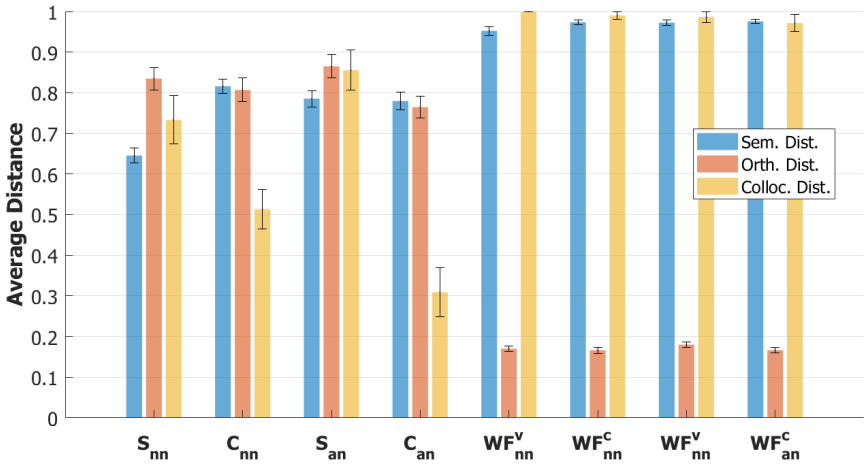


Figure 3.13: Bar plot with the distribution of average distances for each of the eight conditions. Error bars represent the standard error of the mean. The orthographic distance was normalized by the length of the longest word in the pair. Note that, among the 4 categories on the left, collocations have on average a lower collocation distance than purely semantic ones. Labels on the *x-axis* represent the experimental conditions defined in **Fig.**3.11, where **S** stands for semantic, **C** for collocation, **an** for adjective/noun, **nn** for noun/noun, **WF** for word form and superscript letters **c** and **v** for consonant and vowel respectively.

3.4.3 Reaction Times: results

A reaction times experiment was performed with the same modalities of the previous behavioural experiment using the new stimulus set. 15 participants performed the task with a 93% accuracy. Again, statistical analyses show higher accuracy for *word form* transitions (~ 97%) compared to *semantics* ones (~ 89%), as shown in **Fig.**3.14.
The ease of recognition is also reflected in the distribution of reaction times for the two classes of association, with a significant difference ($p < 0.001$) of

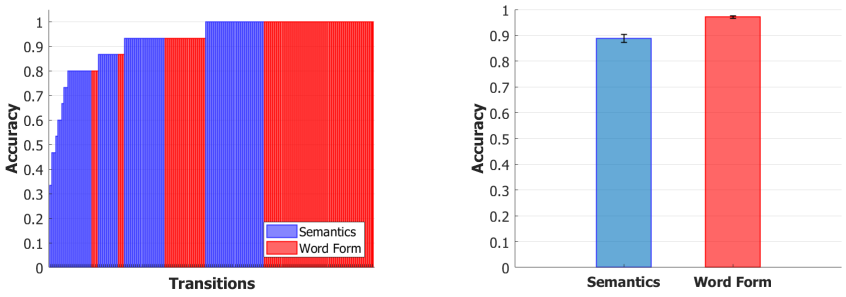


Figure 3.14: Distribution (left) and average proportion (right) of correct responses for the two categories of transitions. *Semantic* associations are slightly more difficult to recognize than *word form* ones, consistently with the previous experiment. A Wilcoxon rank-sum test on the distributions of correct responses proved this difference to be statistically significant with $p < 0.001$. Error bars represent the standard error of the mean.

~ 130ms. The shorter reaction times of this second experiment compared to the first are mainly due to a decrease in the response times for the *word form* category. The average RT of 730ms for *word form* transitions is the same average value measured for the *omission* subtype in the first experiment, even if this time transitions involve only the *change* of one letter. Since the subconditions that we considered were not explicitly told, participants in this new task had to check only for one rule of orthographic change, leading to the possible reduction in reaction times.

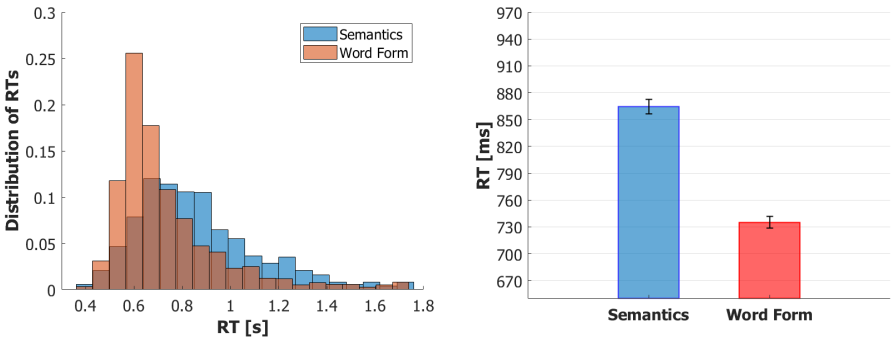


Figure 3.15: (Left) Distribution of Reaction Times for both conditions. (Right) Mean Reaction Times. Error bars represent the standard error of the mean. Semantic transitions are recognized slower than word form ones.

Association subtypes

Reaction times for the eight conditions are plotted in **Fig.**3.16.*Semantics* conditions show different trends between *semantic* (S) and *collocation* (C) transitions when comparing noun/noun (NN) with noun/adjective (AN) subtypes. However no parallel interaction is present in *word form* trials, where it would involve consonant/vowel changes .

In the NN comparison, *semantic* (S_{NN}) transitions are faster (~ 70ms) than *collocations* (C_{NN}), implying a possibly stronger effect of semantic rather than collocation links in driving the reaction times. The opposite occurs in the AN comparison, with shorter reaction times (~ 90ms difference) for the collocation (C_{AN}) subtype. With adjective/noun pairs, the semantic distances are balanced between the two conditions (see **Fig.**3.13) and only the difference in average collocation distance defines the two classes.

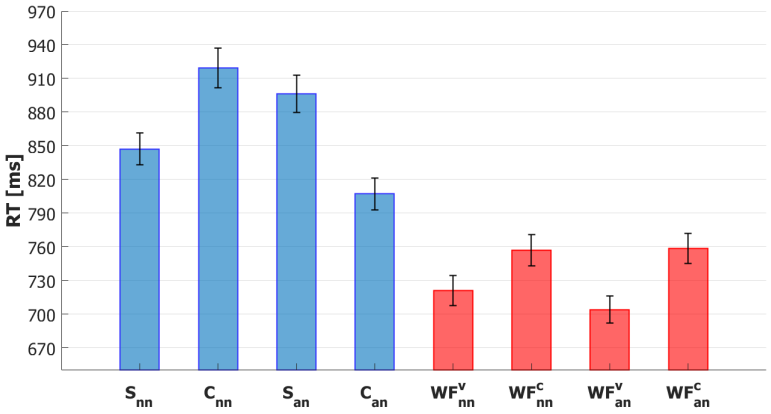


Figure 3.16: Mean reaction times for the eight types of allowed transitions. Error bars represent the standard error of the mean.

Word form conditions, as already noted, do not show a modulation for the NN/AN comparison. Interestingly, however, a clear modulation appears for the vowel/consonant comparison with an average difference of ~ 40ms, showing a faster processing for vowels.

The table in **Fig.**3.17 reports the significance results of a Wilcoxon rank-sum test for all pairs of conditions.

	C_NN	S_AN	C_AN	WF_V_NN	WF_C_NN	WF_V_AN	WF_C_AN
S_NN	***	*	–	***	***	***	***
C_NN		–	***	***	***	***	***
S_AN			***	***	***	***	***
C_AN				***	***	***	***
WF_V_NN					*	–	**
WF_C_NN						**	–
WF_V_AN							***

Figure 3.17: Significance results from comparing the reaction times between all pairs of conditions with a Wilcoxon rank-sum test. * stands for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$.

A currently ongoing EEG experiment, with the same design of the previous ERP experiment, will hopefully shed light on the neural correlates of the word transitions described in this section.

Chapter 4

Potts Model Implementation

The encouraging results of the experiments described in **Chapter 3** pushed us towards the development of a network architecture able to perform the same task.
In this chapter, we will propose an approach to model priming tasks in the framework of a latching process, in a similar fashion to the model of spreading activation suggested by Lerner et al. in [39]. In this work, the authors showed how a latching process can replicate some of the semantic priming effects that challenged the attractor network approach, such as mediated and asymmetric priming. The modeling of priming by Lerner et al., however, was indirect, being based on the results of a simulated process of spreading activation, starting from the *prime* node, on a small semantic network.
On the other hand, in our implementation we will explicitly model both the encoding of the prime and the process of recognition of the correct target, in a network with the minimal structure imposed by the task.

4.1 Modeling a priming task

The Potts model in [34] [36] was taken as a starting point for our network implementation of a priming task. Each Potts unit represents a patch of cortex, thought of as a local network of many neurons. Each local network

is supposed to store S items, which can be thought of as semantic or orthographic features. The composition of these distributed features generates the representation of words and concepts in a similar fashion to [8]. The same composition of features can thus be implemented in our network by building patterns of co-active Potts units. The patterns of activity generated in this way are then stored in the connections between the units as stable attractors of the network by means of **Eq.**2.1. Once the network is cued with an external field, its activity converges to the closest global attractor. However, the introduction in this model of the time-dependent thresholds defined in **Eqs.**2.3, representing inhibition and adaptation, allows destabilizing the activity, making it jump from an attractor state to another close to the first.

4.1.1 Network architecture

As a first, primitive, model of our behavioral task, we have limited the experimental conditions to the only *semantic* and *word form* transitions. In addition, for the *semantic* condition, *collocations* have been included as rule-based memories, as an application of the instructed latching described in **section** 2.2.

Semantic and orthographic networks

To model the task described in **Chapter 3**, we constructed a network divided into two components: one subnetwork storing orthographic information and the other encoding the semantic content of words. The first word form network was the only one to receive the external cue representing the visual stimulus in the experiment. The visual cue in this implementation elicits, in the word form subnetwork, the representation of a target word, which in the global network will interact with the representation of the previously stored prime word.
Each network was initialized with independent sets of randomly correlated patterns. Each word representation in the global network was then built by linking one pattern of the semantic network (S) with one of the word form

network (WF). Thus, there was a 1 to 1 correspondence between the patterns of the two subnetworks. A schematic view of the full network model is presented in **Fig.**4.1.

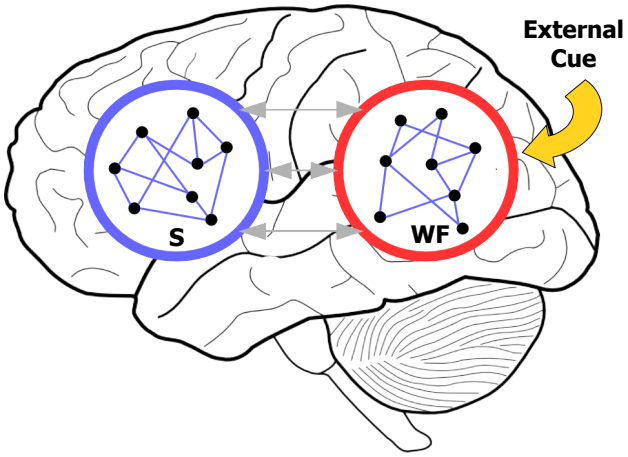


Figure 4.1: Network architecture for simulating the word-transitions experiment. The global network is splitted into a *Semantic* and a *Word Form* network. The two networks are linked by autoassociative connections such that each word is represented by a unique global pattern of activity. The external visual cue affects only orthographic units.

4.1.2 Encoding of the prime

A priming experiment, like the one in the previous chapter, involves two main types of processing, namely the encoding of the prime word in short-term memory and the recognition of its possible relation, based on specific similarities in long-term memory, with the current target word.
For the encoding of the prime word, we exploited the local parameter w, which appears as a positive feedback term acting on Potts states (**Eq.**2.1.2), helping the global network to converge to an attractor. This term is usually kept constant for all states and all units; however, to produce a short-term memory of the prime, we changed the w of all the states active in the pattern representing the prime, effectively increasing the depth of its attractor.

The change to the w parameter is defined by the following equation:

$$w_i^k = w + \tilde{w}\ \delta_{k,\xi_i^{\text{prime}}} \tag{4.1}$$

where $\tilde{w}$ is a small positive number, ξ_i^{prime} is the state of unit i in the activity pattern representing the prime word and δ is the Kronecker function. In our case, no time-dependence has been included in the short-term memory, but a degradation of the memory of the prime word could easily be included for modeling longer tasks, with a decaying value for $\tilde{w}(t)$.
The mechanism in **Eq.**4.1 is only one of the available alternatives for modeling short-term memory in our Potts attractor neural network, but a detailed analysis on this topic will be treated in [50].

4.1.3 Structure of the word-space

Word transitions based on similarity

To test the associative capability of the network in the two conditions of the experiment, we generated different sets of randomly correlated patterns to store in the two subnetworks. Then, for each condition, we selected the five global patterns in the full network with the highest correlation with a chosen prime pattern, and we manually changed them such that the source of their correlation could be restricted to only one of the two subnetworks. This manipulation allowed us to construct models of semantic and word form associates of the prime.
The prime was chosen among all patterns, as the one having the highest number of highly correlated patterns with it. However, since for Potts patterns we do not have a unique definition of correlation, each unit being active in any of its S possible states, we defined, as in **section** 2.2, two types of correlation: C_1 equal to the fraction of units active in the same state in the two patterns considered; C_2, instead, equal to the fraction of shared active units but in different states. Since similar words share common features and features in our model are represented by the active states of the Potts units,

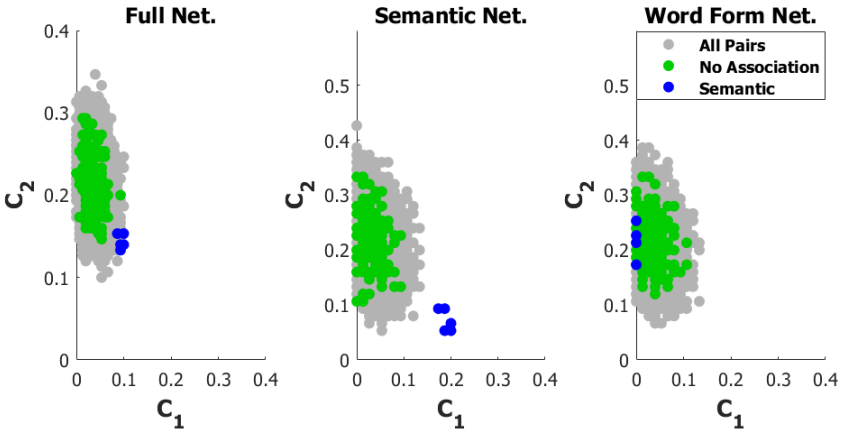


Figure 4.2: Distribution of correlations in the full network and its two subnetworks. Grey dots represent the correlations between all pairs of patterns, in green are highlighted only the pairs including the prime while in blue are plotted the five patterns changed to mimic, in this case, a semantic relation with the prime. C_2 axes for the two subnetworks, in this and future plots, have a different scale than for the full network to reduce the spacing between dots, since distances in this correlation space are sensitive to network size.

we considered as highly correlated patterns those pairs with high values of C_1 and low values of C_2.

According to the previous definitions, an example of semantic associations with a prime word is shown by the $C_1 - C_2$ plot in **Fig.**4.2. While the grey cloud in the figure represents the correlations between all pairs of patterns, the green cloud shows the correlations between all pairs in which one of the patterns is the prime word. As we can notice, green dots span a region of relatively low correlation with the prime and thus the relative pairs are used to effectively model the *no association* control condition of the experiment. Conversely, blue dots represent the five patterns modified to be highly correlated with the prime, having the main source of correlation, in this case, in the semantic network.

Word transitions based on collocations

Collocations are words often occurring one after the other in a text or speech. Both experiments showed a modulation of reaction times sensitive to this type of transition, therefore we included it as a third association mechanism. This kind of relations can be thought and implemented in our model as an

instruction that forces a jump from the first word to the second. Instructions can be stored in the Potts network as a heteroassociative term in the coupling tensor J_{ij}^{kl}, as in **Eq.**2.7, linking one pattern to another.
For the simulations in this chapter, we chose the five patterns with the lowest C_1 correlation with the prime, and we manually changed them to set their C_1 values to zero in both subnetworks. The association with the prime was then implemented as an instruction of the type *target* $\rightarrow$ *prime*. In order to evaluate the effect of the change of w in guiding the dynamics towards the prime, we added for each target word a further instruction pointing to a non-prime pattern, *target* $\rightarrow$ *non-prime*.

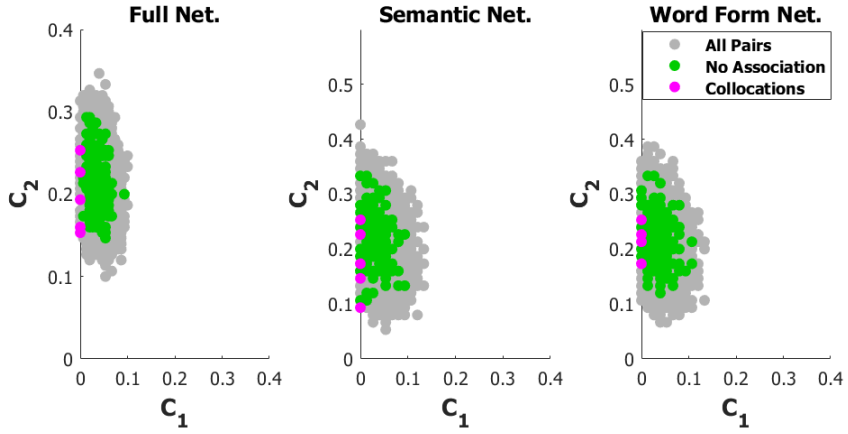


Figure 4.3: Distribution of correlations in the full network and its two subnetworks. Grey dots represent the correlations between all pairs of patterns, in green are highlighted only the pairs including the prime while in pink are plotted the five patterns for which an instruction towards the prime was implemented through an heteroassociative term.

Fig.4.3 shows the distribution of correlations between the five patterns which form a collocation with the prime. Purple dots, representing collocation pairs, lie on the $C_1 = 0$ axis to avoid the possibility of spontaneous transitions to the prime.

4.1.4 Latching back to the prime

In the analysis of our experimental task, we focused on the behavioral and neural responses to the target words. Similarly, in this simulated experiment, we have put our attention on the process of recognition of a correct

associate. For this reason, after storing the prime in the short term memory, we consider a successful recognition as a latching process coming back to the representation of the prime. Therefore, simulations started from cueing the network with a target word, waiting for the latching process to recall the original prime word. The number of latching steps required to reach the prime pattern is considered as an analog of the reaction time for recognition of a correct association.

An example of a successful recognition of a prime-target pair is given in **Fig.**4.4, where pattern *0* is initially cued and the first latching step drives the network in the representation of the prime word (red latching step in the second row of the figure), related to the target through a semantic association.

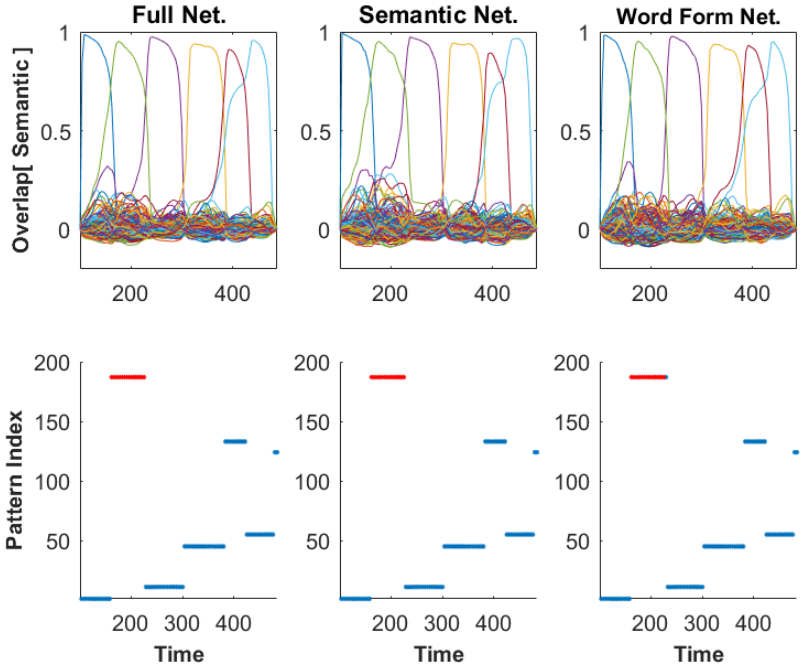


Figure 4.4: Example of a trial with a latching transition from a semantically associated pattern (*pattern index 1*) to the prime (*pattern index 187*, in red) occurring as the first latching step. (Top row) Evolution of the overlaps for the full network and its subnetworks. (Bottom row) Plots of the index of the pattern with the highest overlap with the state of the networks in time. Note that the two subnetworks proceed in nearly synchronous latching.

4.2 Simulation Results

4.2.1 Parameter setting

For our simulations of the model, we considered a network with $N = 600$ units, equally divided in the two subnetworks, with $S = 7$ possible active states. The total number of randomly correlated patterns was $p = 200$. Each unit received an input connection from $C = 90$ other units, equally distributed between the two subnetworks. For the positive feedback term w, we chose a value of 0.45, which puts us in a region of correlated latching transitions (**Fig.**4.5). For the units and states active in the prime, instead, we chose a value $\tilde{w} = 0.1$. All the other parameters of the model were kept as in the simulations presented in **Chapter** 2.

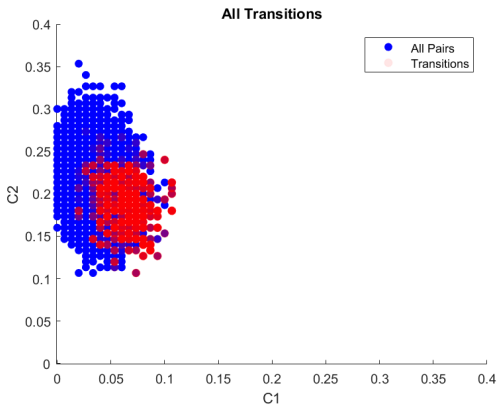
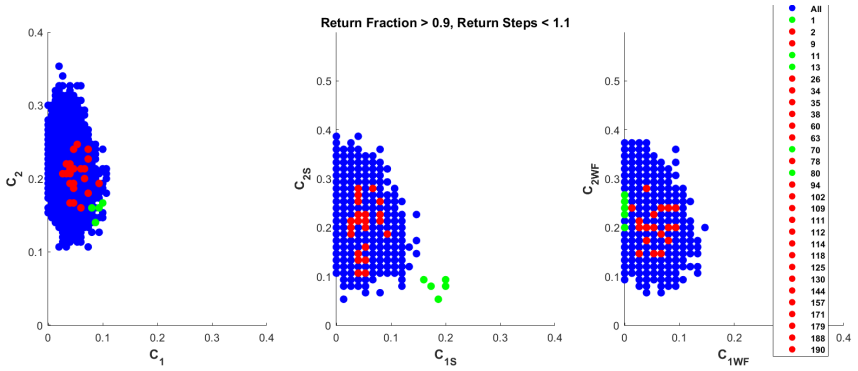


Figure 4.5: Distribution of correlations for all latching transitions observed in a simulation with $w = 0.45$ for all states. The red cloud covers a region of high C_1 and low C_2 values (correlated latching).

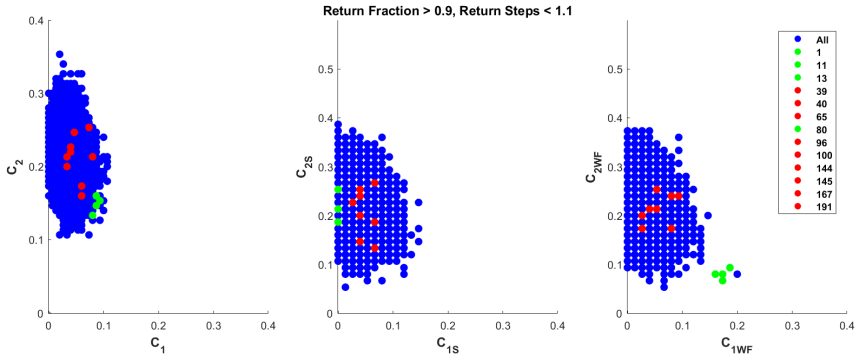
4.2.2 Effectiveness of priming

In a first batch of simulations, we checked that our manipulations on correlation distributions and on the encoding of the prime could allow the network to recall the prime after being cued with a correct association target. For the effectiveness of collocations transitions we instead refer to **section** 2.2 on instructed latching.

For these simulations, we generated a set of p patterns for which pattern 187 was chosen as prime and patterns $\{1, 11, 13, 70, 80\}$ were selected and modified to implement semantic or word form transitions. The two types of transitions were simulated separately by changing only the five selected patterns. The network was cued, through its word form part, ten times for each of the p stored patterns, for a total of 2000 independent simulations.



(a) Semantic associations



(b) Word form associations

Figure 4.6: Distribution of correlations for the patterns for which the first latching step was to the prime in at least 90% of the trials. (First row) Simulation with strong semantic relations and (Second row) with strog orthographic relations in the 5 modified patterns, namely patterns $\{1, 11, 13, 70, 80\}$. Green dots represent immediate returns of these five modified patterns. Red dots show all the non modified patterns returning to the prime in the first latching step.

In **Fig.**4.6 we focus our attention on those patterns for which, in all ten trials, the first latching step was on the prime. By highlighting these patterns in the C_1-C_2 graph we can then check their relation with the prime. We can notice that the five modified patterns almost always lead to the prime in the first latching step, with the only exception of pattern 70 in *word form*

associations. However, they are not the only ones. Indeed various degrees of correlation with the prime may lead other patterns to latch to it. This result shows how the tuning of w with **Eq.**4.1 effectively broadens the attractor of the prime, facilitating transitions towards it. The large number of patterns that follow this type of transitions in the current implementation pushes towards the definition of a more structured word-space in future versions of the model, where associates will not be artificially constructed from a randomly correlated set of patterns. An ecological way to implement structured relations between stored memories has been proposed for the Potts model in [74].

4.2.3 Simulating reaction times

In a second batch of simulations, we included trials of association pairs linked by a collocation relation modeled with heteroassociative couplings stored in the internal connections of the orthographic network. We used in this case a θ-σ interaction with $\lambda = 1$ (see **section** 2.2 for the details on heteroassociation).

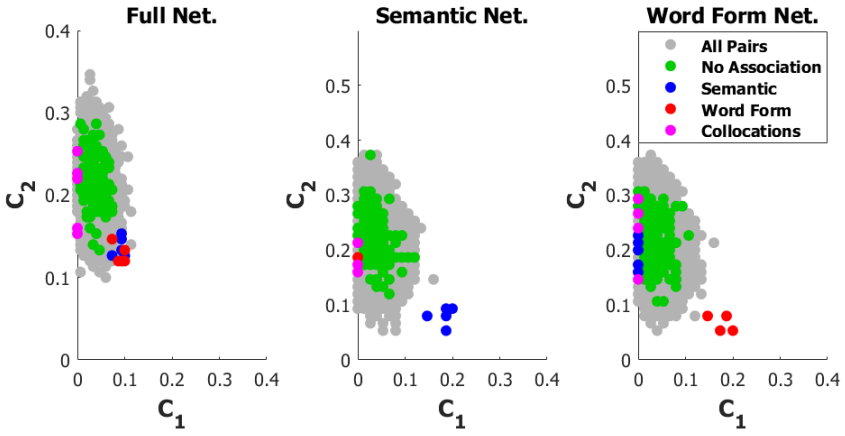


Figure 4.7: Distribution of correlations for the patterns with all three types of transition: semantic (blue), word form (red), and collocation (pink). Non-associated patterns are in green while all pairs of patterns are in grey.

We first generated five independent sets of p patterns; for each set, we then found the pattern with the highest number of highly correlated patterns

with it and we used it as prime. Once we defined a prime pattern for each of the sets, we changed its 15 mostly correlated patterns to model the tree rules of association, as described in **section**4.1.3. This search for highly correlated patterns was used to remove as many unstructured correlated pairs as possible, by transforming them into *semantic*, *word form* and *collocation* associations. With this procedure, each set of patterns had a unique prime and five associates for each type of transition. **Fig.**4.7 shows the distribution of correlations for one of the five sets of patterns.
For this new set of simulations, we measured the distribution of reaction times for each type of transition, including the ones with no relation, by considering the average number of latching steps required to return to the representation of the prime. Each of the $p-1$ target patterns was cued ten times for each of the five sets of patterns.

Results

From this last simulation, we can notice different distributions of reaction times for the three types of transition, whose averages can be qualitatively compared with our behavioral data.
As we can notice from **Fig.**4.8, "no association" transitions have longer reaction times compared to the three classes of proper associations. Returns to the prime in this control condition, as already discussed in **section** 4.2.2, are due to the artificial structure of the word-space, where semantic representations are manually built from randomly correlated patterns. Furthermore, in these simulations we considered complete latching trajectories, rather than only their first steps, making it more likely for the simulation to eventually reach the broad attractor state of the prime. Nonetheless, the difference between the distributions indicates the efficacy of our manipulations.
By looking at the three main conditions, we can see that collocation transitions are slower than similarity-based associations, in analogy with the noun-noun comparison between semantic and collocation transitions in the second experiment. On the other hand, no clear difference in the reaction

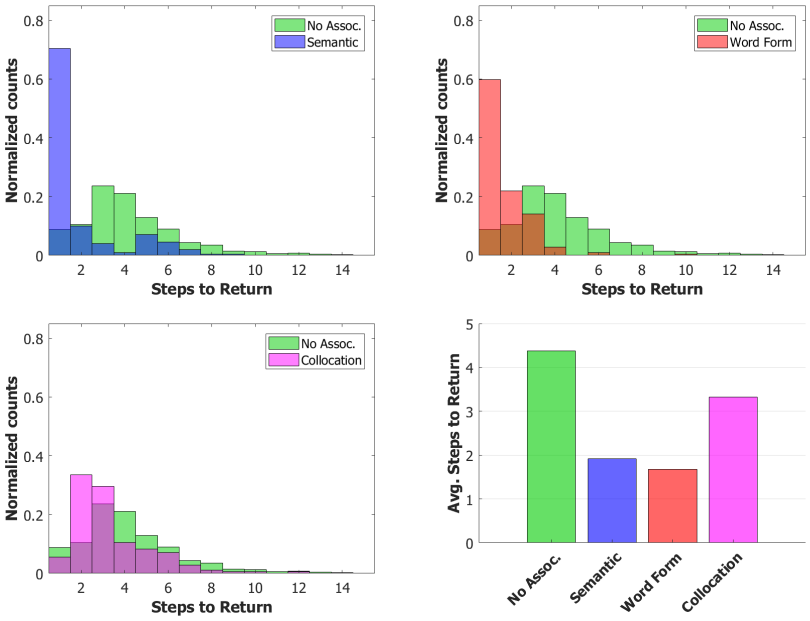


Figure 4.8: Distributions and means of reaction times for the three types of transition, compared to "no association" transitions

times can be seen between semantic and word form associations.

Conclusions and Outlook

Orthographic transitions may take advantage, at the level of human reaction times, from the visual similarity of the stimuli, even before reaching word-level processing. However, as suggested by the results of the second experiment, a distinction between the processing of vowels and consonants could be implemented in the network, bearing possibly different reaction times even at the word level. Furthermore, in the current implementation, the two subnetworks are connected by autoassociative connections which yield an almost instantaneous information transfer from the orthographic network to the semantic one. Further segregation between the two networks would then account for faster processing of visual stimuli compared to accessing semantic memory. In such a model, the average activity of the units in the two subnetworks may be considered as an *in-silico* correlate of the evoked potentials found in the EEG experiment, similarly to the results from the ar-

tificial neural network by Cheyette and Plaut [75], with the main difference, in our case, of being the result of a general cortical model, rather than an *ad hoc* model of semantic priming.

Chapter 5

A Spontaneous Stream of Thoughts: Mind-Wandering

In the previous chapters, we studied one of the most basic mechanisms in language production and comprehension, namely the associative relations that link one word to another. To do so we developed an experiment on word transitions with a structure simple enough to be easily implemented with our Potts neural network. This approach allowed us to avoid experimental and computational complications that arise when studying language in ecological settings. Although we limited as much as possible the use of unnatural stimuli, experimental constraints often shape our understanding of the neural processes underlying different human mental faculties. In this chapter, we will review the main results from one of the most ecological approaches to the investigation on human memory and language processes, namely the research line on *mind-wandering*. In the second part of this chapter, we will present a model for mind-wandering by means of our Potts attractor neural network, able to replicate some of the interesting results from hippocampal and prefrontal patients, somehow impaired in their ability to mind-wander.

5.1 Mind-Wandering

In our everyday life, our conscious experience continuously ebbs and flows between the current task and unrelated thoughts and memories. This shift of attention from external to internal contents is spontaneous and is a very common phenomenon. It has indeed been shown that people spend from 25% to 50% [76] of their waking time thinking about something different from their "here and now".

The intimate and spontaneous nature of mind-wandering makes this mental process one of the most mysterious and fascinating products of the human brain. However, because of its almost inaccessible nature, MW has been historically studied through *Experience Sampling*, the most common method to access the contents of spontaneous thoughts. Across task, participants are intentionally interrupted and probed to report whether their thoughts have been fully on-task or "elsewhere", and, in case, about the contents of such off-task thoughts. Recently, the investigation on mind-wandering and self-generated thoughts has seen an explosion, mainly due to the advent of sophisticated imaging techniques that led to the discovery of a consistent activation of certain brain regions during rest. What is now known as the "Default Network" has been linked to mind-wandering and it involves a set of interconnected brain regions, including the medial temporal lobes (MTLs), ventromedial prefrontal cortex (vmPFC), posterior cingulate cortex, and the angular gyrus [77][78][79][80].

5.1.1 Impairments of spontaneous thinking

The role of the different brain regions in MW is still unclear but experimental investigations, in particular on vmPFC and hippocampal patients, have shed light on the possible different influence of these two regions on spontaneous thought generation.

VmPFC patients show a reduced frequency of mind-wandering, but, when they do mind wander, their thoughts are mostly about the present and never

about the future [81]. Interestingly, vmPFC damage does not change the frequency with which participants claim they are unaware of the content of their off-task thoughts, suggesting it causes impaired construction, not meta-awareness, of mind-wandering contents [82].
On the other hand, hippocampal patients experience mind-wandering with the same frequency as controls. However, while off-task thoughts are rich in details in healthy controls, in hippocampal patients they are semanticized, not episodic and mainly present-oriented [83].
We tentatively propose, therefore, that during mind-wandering vmPFC initiates the construction of events alternative to direct perceptual experience, by coordinating the activation of relevant schemata [84], which the hippocampus uses to build a rudimentary sketch of the event.

5.1.2 A latching perspective on MW

Mind-wandering, a spontaneous train of thoughts unfolding in a rather unconstrained fashion, is thus reminiscent of a latching process, in which some of the transitions appear random while others are more guided by local schemata. Therefore we propose that vmPFC participates in the mechanics of neocortical latching by facilitating congruent consecutive retrieval of stored memories, while their content is boosted by the hippocampus.
Mathematizing the psychological concept, a schema may be conceived, in the framework of our Potts model, as the association of attractor state k in local network i with the subsequent attractor state l in local network j, an association extracted over multiple similar occurrences of the same sequence of states [84]. This view can thus be easily implemented in the Potts network as an heteroassociative contribution described in **section** 2.2.
On the other hand, a Potts model connected with a hippocampal model may utilize it as an "episodic content booster," reinvigorating streams of thoughts in the cortex without completely been driven by it if hippocampal output representations are activated not too frequently relative to the sequence of neocortical states.

Model predictions

In such a network a lesion to the hippocampal component is expected to result in reduced episodic content boosts, with preserved schema-driven transitions: latching is thus long-lasting but constrained to schematic contents.
Conversely, a lesion to vmPFC is expected to disarticulate mind wandering, leaving it over-dependent on the hippocampal content booster: ephemeral, inconsequential, short-lived mind wandering is now triggered by the infrequent hippocampal output and poorly assisted by schema-guided construction processes.

5.2 A Potts model of MW

5.2.1 Network architecture

Our network model for MW aims to capture the functioning of the main components involved in the Default Network in the human cortex. For this reason, we considered the tripartite network sketched in **Fig.**5.1.

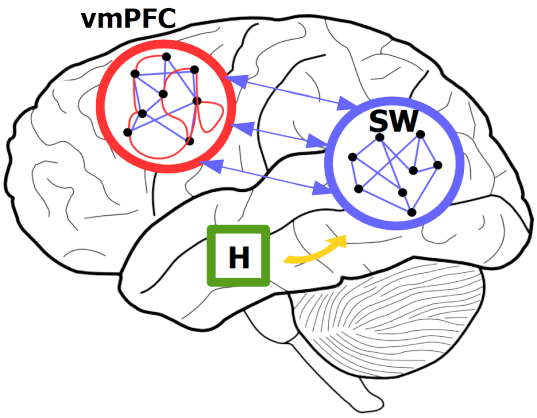


Figure 5.1: Schematic representation of the network model of the Default Network considered as a latiching model of mind-wandering. *vmPFC* is an heteroassociative network storing schema associations between stored patterns. *SW* (i.e., *semantic workspace*) is an autoassociative network that stores as activity patterns the semantic representations of words and concepts. *H* represents instead the hippocampal input to the neocortical network *SW*.

The *SW*, i.e., *semantic workspace*, network can be regarded as the core component of the model, storing the semantic representations of the single concepts which will be concatenated during mind-wandering simulations. This memory component is implemented by an autoassociative Potts neural network described in **Chapter 2**. The latching dynamics occurring in such a network will be considered as a simulated mind-wandering process. The different latching properties of the network in **Fig.**5.1 will then be related to the properties of the *vmPFC* and *H* components.

Hippocampal input

The hippocampus in this implementation is not directly modeled with a Potts network, instead only its input to the semantic network is considered. We propose the role of the hippocampus in mind-wandering to be that of reinvigorating the activity of the semantic network by recalling episodic memories. These memories are modeled as static collections of four patterns that are elicited all together for a brief period. For each of the p stored patterns in network *SW*, we chose at random three other patterns that would form with it an episodic memory. The distribution of correlations between any chosen pattern and its three other components in its episodic memory are highlighted in red in **Fig.**5.2.

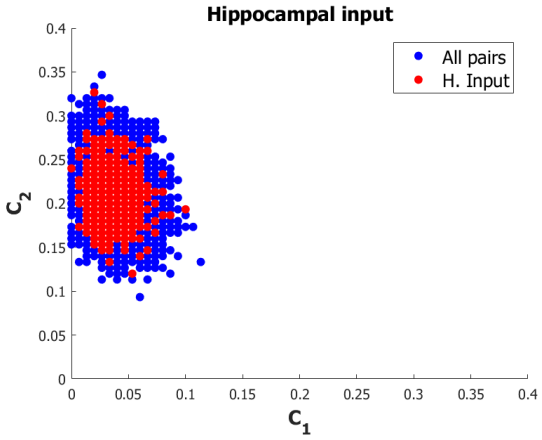


Figure 5.2: Distribution of correlations between all pairs of patterns (blue) and of pairs occurring in the same episodic memory (red).

During simulations, the hippocampus cues network SW for five times at times $t^{HI} = \{700, 900, 1200, 1600, 2100\}$. The cue is modeled by an additional term in the field h_i^k of units in SW, namely:

$$h_i^{k,H} = gW\left(t, t^{HI}\right) \sum_{\mu \in \{\mu\}^{episode}} \frac{1}{4} \delta_{\xi_i^\mu k} \tag{5.1}$$

where g is the strength of the cue, $W\left(t, t^{HI}\right)$ is a window function equal to 1 when $t^{HI} \leq t \leq t^{HI} + 100$ and 0 otherwise, and $\{\mu\}^{episode}$ is the collection of four patterns that compose an episodic memory. At times t^{HI} the cued episodic memory is the one associated the pattern with the highest overlap m in network SW.

Schemata in the vmPFC

Schemas in our model are considered as frequent ordered associations which can be implemented as heteroassociative instructions (**section** 2.2) stored in the internal connections of the vmPFC network. To enhance the heteroassociative role of vmPFC, favouring the encoding of schemas rather than single memory items, we diminished the influence of the autoassociative component by a term $(1 - \lambda)$. In this way **Eq.**2.8 becomes:

$$h_i^k = \sum_{j \neq i}^{N} \sum_{l=1}^{S} \left[(1 - \lambda) J_{ij}^{kl} \sigma_j^l + J_{ij}^{kl,het} \theta_j^l \right] + w \left(\sigma_i^k - \frac{1}{S} \sum_{l=1}^{S} \sigma_i^l \right) \tag{5.2}$$

where i is a unit in the vmPFC subnetwork and λ is the strength of heteroassociative connections that appears as a multiplicative factor in the definition of $J_{ij}^{kl,het}$ (see **Eq.**2.7).
Schemas are then constructed between correlated patterns in network SW. Therefore, for each pattern three instructions were added towards the three most correlated patterns with it, as shown by the distribution in **Fig.**5.3.

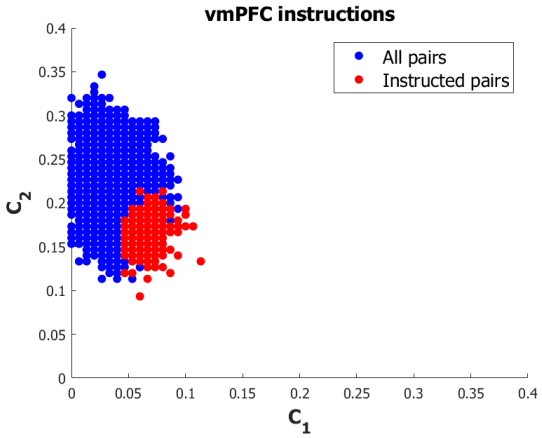


Figure 5.3: Distribution of correlations between all pairs of patterns (blue) and of all pairs of instructed transitions (red).

Parameter setting

The parameters used for simulating mind-wandering are the following: $N = 600$ of which $N^S = 480$ and $N^{vmPFC} = 120$, $a = 0.25$ $c_m = 90$ of which $c_m^S = 72$ (internal connections in network SW) and $c_m^{vmPFC} = 18$, $S = 7$, $g = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, $\lambda = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, $p = 200$, $U = 0.1$, $\beta = 12.5$, $w = 0.5$, $\tau_1 = 3.33$, $\tau_2 = 100$ and $\tau_3 = 10^6$.

5.2.2 Simulating spontaneous thoughts: Results

For each (g, λ) pair we generated 30 latching sequnces. The influence of the parameters g and λ was used to simulate not only MW in healthy subjects but also in hippocampal or ventromedial patients.

Two examples of latching sequences are given in **Fig.**5.4. As we can firt notice, higher values of λ seem to increase the latching length. When activity dies out in the network, hippocampal input, if strong enough, may reactivate the network from its last "thought".

In **Fig.**5.5 we can observe the two effects of vmPFC in our model. When increasing λ latching transitions tend to follow schematic instructions, mixing them with spontaneous transitions and, if g is high enough, with episodic content. The second effect is the average increase in latching length. This

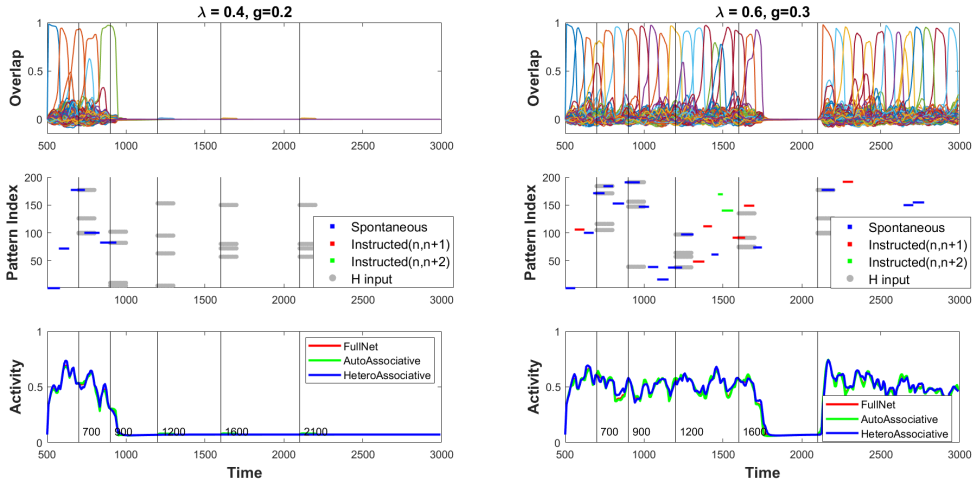


Figure 5.4: Examples of latching sequences for two choices for (g, λ). The first row shows the overlaps of the p patterns with the state of the network. The second row highlightes the types of transitions. In red and green are plotted the followed instructed transitions and in blue the spontaneous ones. In grey instead is shown the cued patterns from the hippocampus. The third row shows the activity of the full network and of its autoassociative (i.e., SW) and heteroassociative (i.e., $vmPFC$) components. High values of λ increase latching length while high values of g reactivate the network when latching dies out.

result is in analogy with mind-wandering in vmPFC patients, modeled by low values of λ, for which it has been found an impairment in the construction of mind-wandering contents. The reduced frequency of mind-wandering in these patients may be explained by our model by the difficulty in spontaneously generating long sequences from both an external or a hippocampal input. In this case, latching sequences resemble more a retrieval of single episodes rather than an extended mental travel.

When considering the contribution of the hippocampus on simulated thoughts, we notice again two main effects. **Fig.**5.6 shows a jump in the average number of latching steps that occurs for $g \simeq 0.3$. Below this value, the hippocampus has not enough strength to reactivate the cortical network, leaving only to the vmPFC the role of carrying on the latching sequence.

The second role of the hippocampal input is instead described in **Fig.**5.7, where we considered how much neocortical latching follows the episodic contents received from the hippocampus. As we can notice, for $g \gtrsim 0.3$ latching is mainly driven by hippocampal inputs for all values of λ but with a crucial

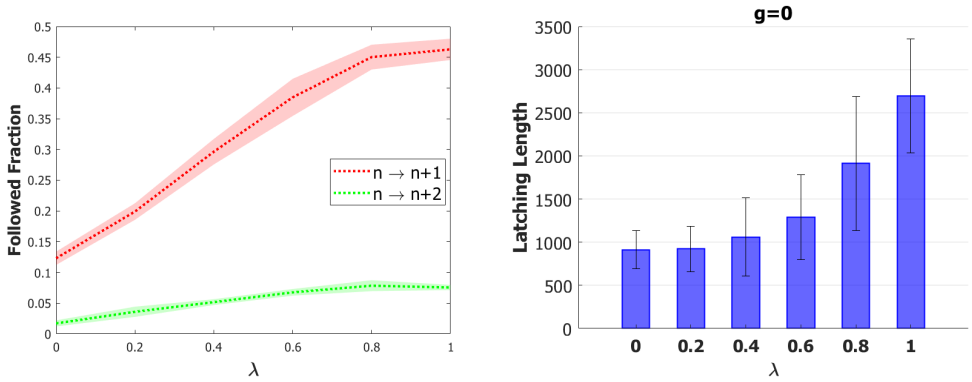


Figure 5.5: Effect of λ (i.e., the strength of schema instructions) on latching. (Left) Followed fractions are averaged over the different values of g (dotted lines). In red, the fraction of transitions that follow one of the schema instructions of the step before and, in green, of two steps before. (Right) Average latching length when there is no hippocampal contribution. Shaded areas (left plot) and error bars (right plot) represent the standard error of the mean. The two plots show an increase in followed fractions and latching length for increasing values of λ.

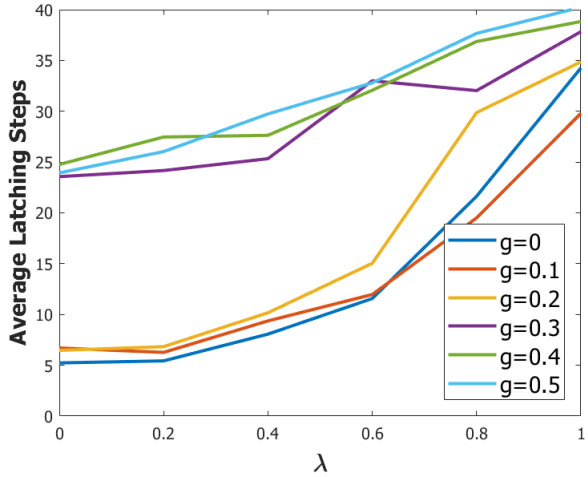


Figure 5.6: Effect of g on the average number of latching steps. When g reaches the threshold value of 0.3, its influence becomes strong enough to reactivate the network if latching ends. This effect is less relevant for high values of λ, when latching is long-lasting.

difference in the number of latching steps between the different memories the compose an episode. Indeed, for lower values of λ, the network tends to jump more between the items of an episodic memory.

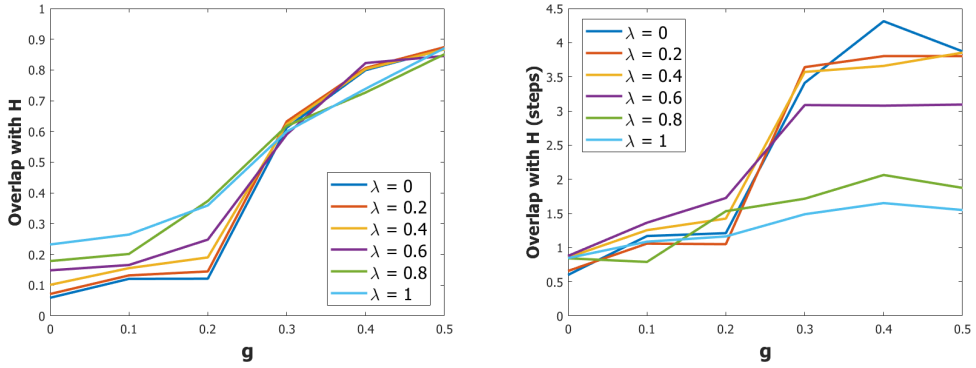


Figure 5.7: Effect of g in driving the contents of neocortical latching. (Left) Average overlap between latching steps and the suggested hippocampal patterns. This overlap is equal to one if latching steps are completely aligned with one or more of the cued patterns by H (evaluated only when the interaction is active). (Right) Average number of latching steps between the single patterns that compose an episodic memory.

The conjunct role of g and λ on latching length is outlined in **Fig.**5.8, where network activity at time t is calculated with

$$\text{Activity} = \frac{1}{aN} \sum_{i}^{N} \left(1 - \sigma_i^0(t)\right). \tag{5.3}$$

To summarize, high values of λ (i.e. $\gtrsim 0.6$), a synonym of a functioning vmPFC, generate long latching sequences with a balanced mixture of spontaneous and schematic transitions. Both types of transitions, however, occur in a regime of correlated latching, given by the chosen value of w and the set of schema instructions. Mind-wandering is thus driven by purely semantic associations in the absence of hippocampal input. A result that mimics the behavior of hippocampal patients.

On the other hand, for high values of g (i.e. $\gtrsim 0.3$), the hippocampus has a boosting effect on the network by both reactivating its units and driving its latching to sequences of episodic memories. Ventromedial patients that rely exclusively on hippocampal input can thus be modeled by the low λ and high

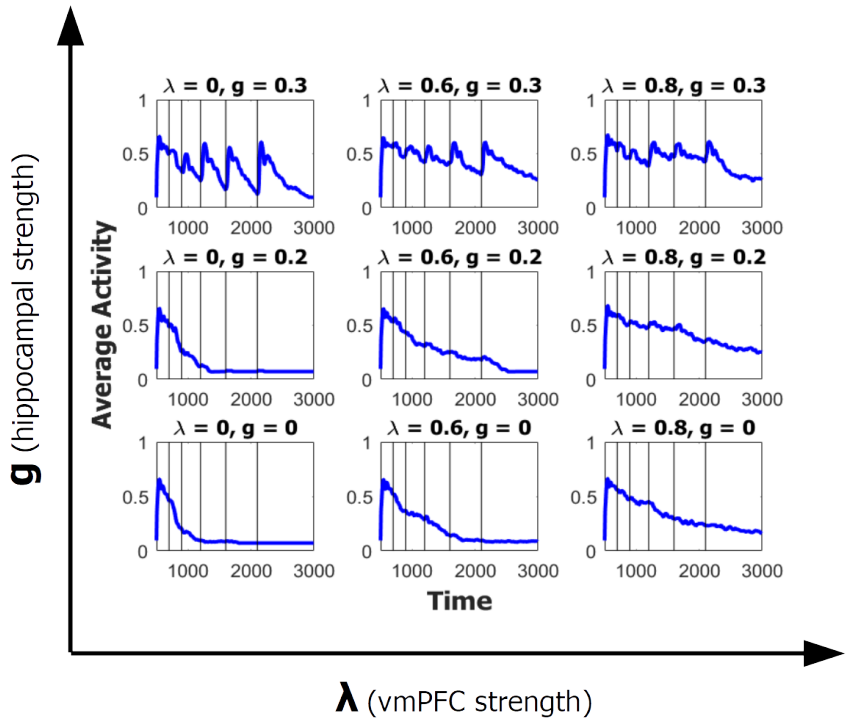


Figure 5.8: Network activity averaged over 30 simulations for nine (g, λ) pairs. A qualitative model of MW in healthy subjects is achieved with values of $g \gtrsim 0.3$ and $\lambda \gtrsim 0.6$. Different degrees of impairment are described by lower values of g and λ.

g parameter region, where latching is short-lived, driven by episodic contents and poor of semantic and schematic relations.

Chapter 6

The Phonological Output Buffer

The Potts neural network has been utilized in previous chapters to model large scale cortical networks and their behavior in performing high-level cognitive functions, like spontaneous thinking and word associations. In this chapter, instead, we will consider a Potts model implementation for a specific processing in the final stages of speech production, namely the phoneme concatenation at the level of the phonological output buffer.

6.1 What is the POB?

The Phonological Output Buffer (POB) is a conceptual construct which appears necessary to explain patterns of deficits observed in certain linguistic tasks, such as reading aloud. The occurrence in some subjects of phoneme omission, addition, replacement and misplacement, particularly when reading non-words, with concurrent preservation of the lexicon and of grapheme-to-phoneme conversion, leads to posit a stage downstream, or a device, where phonemes are assembled in the correct order, perhaps modified as required, and kept in short-term memory (STM) until the word or non-word is uttered by the subject [85][86][87][88].

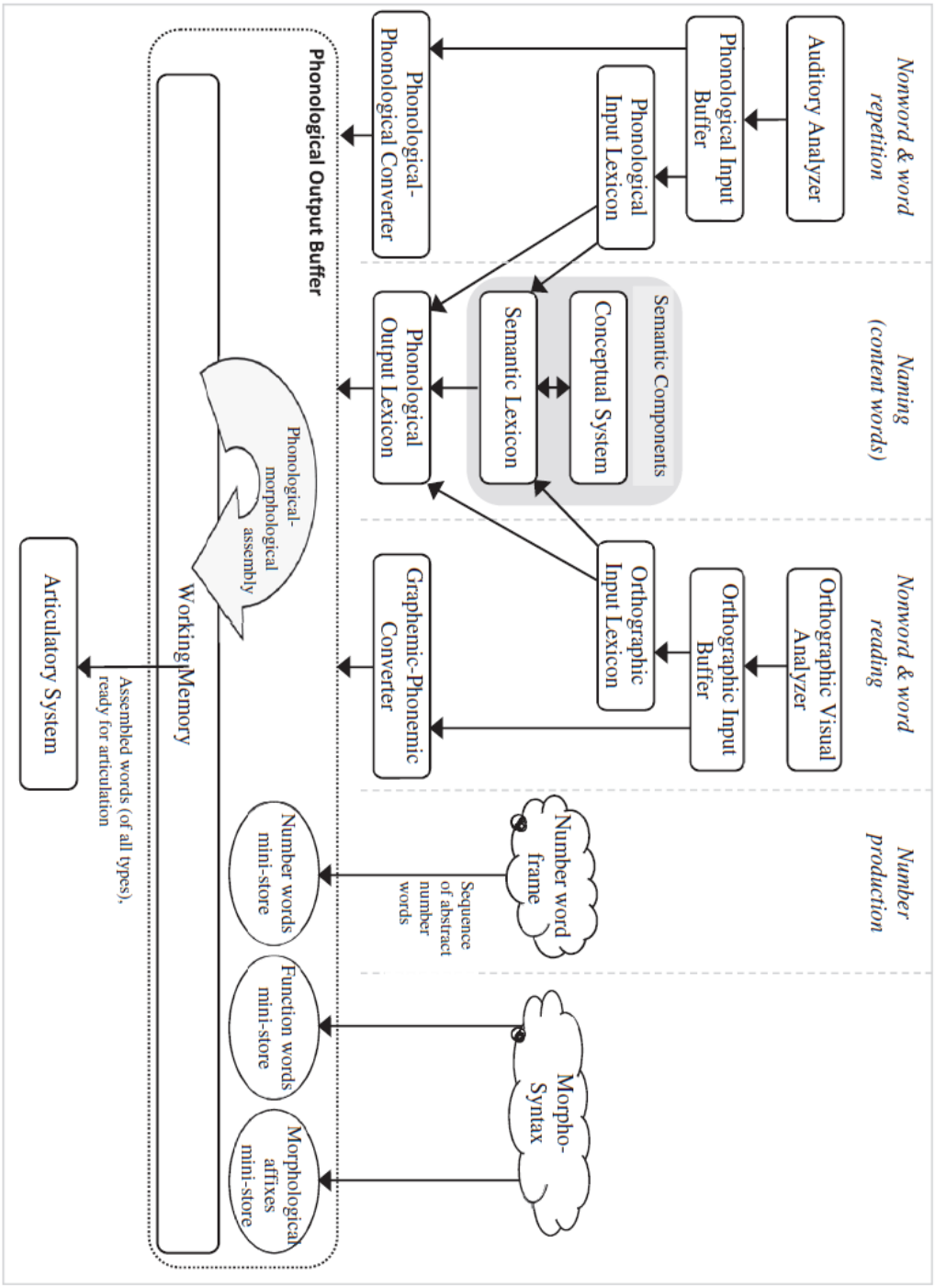


Figure 6.1: Several language production tasks can be modeled by different streams of computations, all passing through what appear to be the necessary operations performed by the POB, as discussed in Dotan and Friedmannn [89] from where this figure is taken.

It is therefore tempting to think of the POB as a specialized device, located somewhere in the brain, which has evolved to facilitate the human language faculty.

The notion of an *ad hoc* and unique phoneme assembly line is challenged, however, by the occurrence of similar deficits in the sequential production of other linguistic objects. Experimental data on patients with POB deficits have shown a tendency in committing errors specifically on elements belonging to the same linguistic category such as numbers, morphemes, function words, etc. [89]. This dysfunction for distinct linguistic material can appear both as phonological errors when uttering words (e.g. time $\rightarrow$ *tise*), or as semantic errors when uttering number or function words (e.g. eight $\rightarrow$ *nine*). Similar semantic errors can be observed for morphologically complex words, where a wrong, but existent, morpheme could replace the correct one. The discovery of these deficits led to the proposal of separate mini-stores specific for each linguistic object. The different mini-stores may be thought of as localized and physically separated, but close, networks with similar cortical mechanisms underlying their functioning. An alternative view to the multiple stores, however, can still not be excluded. The experimental evidence summarized above can also be compatible with a single memory store encoding different linguistic categories as separate clusters in its whole activity space. With this hypothesis mistakes by POB patients would be due to a failure of the network in specific regions of its activity space. Partial cortical localization of the different categories could still be implemented by assigning a preferred tuning to specific linguistic material, say numbers, onto a spatially defined subset of neurons of the network.

The main difference between the single and multiple stores hypotheses would then reside in their functional implications, one of which may be the storage capacity. Two mini-stores, one of which holding M_1 function words and the other M_2 numbers, should be able to operate concurrently and hold M_1 function words and M_2 numbers at the same time. A single network, in contrast, would be limited by a combined capacity determined by the degree of mutual

interference, down to $k\mathrm{M}_1 + (1-k)\,\mathrm{M}_2$ objects with full interference, with k set by the task at hand.
In the following sections we will propose, using a Potts network model of cortical dynamics, a first implementation of the phonological output buffer with the final purpose of evaluating its statistical constraints on memory performance.

6.2 Building a model of the POB

Language production models, as the one in **Figure** 6.1, typically consider the necessary processing stages needed to produce an utterance and which could vary depending on the task to be modeled. In this sense, the phonological output buffer can be considered as a special stage in which different streams of information have to converge to be transformed into a sequence of instructions to the Articulatory System [90]. For simplicity, we will focus on a task involving only real words, already known by the participants. This will allow us to consider only one type of input to the POB, namely the Phonological Output Lexicon (POL) [91][92]. For our purposes, we will consider the POL as a dictionary containing all the relevant phonological information for uttering a word. This simplification will be particularly useful to reduce the CPU time of our simulations by storing at once in the POL all the words that will be used to test the functioning of our POB model. However, our assumptions on the form of the input will hopefully be vague enough to be generalizable to other types of tasks (e.g. involving nonwords). The main goal for our POB model will then be to transform the compact package of phonological information coming from the POL into a temporal sequence of what we will refer to as syllables. In future implementations of the model further details and functions (e.g. phonotactic rules, morphological composition, *etc.*) may be added to the core mechanisms described in this thesis.

6.2.1 Network Structure

Our network model will consider the interaction between the POL and the POB, therefore we will focus our attention on 3 main components: 2 autoassociative subnetworks, modeling the behavior of the POL and the POB, and the connections between the two networks, responsible for the transfer of information from the POL to the POB. No feedback connection from the POB to previous stages of the language production model will be included for simplicity. Both subnetworks will be modeled by Potts attractor neural networks. The POL will receive its input in the form of an instantaneous cue to one of its stored patterns, namely the word to be uttered, from a previous stage and will transfer this information to the POB through heteroassociative connections as described in **Eq.**2.7.

Parameter setting

In all simulations we have modeled the word buffer as a Potts network of $N^{POL} = 600$ units, with $C_m^{POL} = 90$ internal connections and $p^{POL} = 200$ stored words. For both networks, the patterns stored were randomly generated without adding any sort of structured correlation. In order to prevent the buffer network from spontaneously latching, we chose the the following set of parameters: $w^{POL} = 0.45$, $\tau_1^{POL} = 3.33$, $\tau_2^{POL} = 33.3$, $\tau_3^{POL} = 10^6$, $S = 7$, $a = 0.25$, $\beta = 12.5$ and $U = 0.1$. On the other hand for the POB we set the parameters to allow it to be driven into a latching regime when instructed by the POL. For the POB we chose: $N^{POB} = 200$, $C_m^{POB} = 150$, $p^{POB} = 200$, $w^{POB} = 0.5$, $\tau_1^{POB} = 3.33$, $\tau_2^{POL} = 11.1$, $\tau_3^{POL} = 10^6$, $S = 7$, $a = 0.25$, $\beta = 12.5$ and $U = 0.1$. The choice of a three times faster adaptation, determined by τ_2, is motivated by our decision to model only 3 syllable words and thus to allow the POB to latch to all the 3 syllables while the word is still activated in the buffer. The number of heteroassociative connections between the two networks was fixed to $C_m^{het} = 150$. This has to be interpreted as the number of units in the POL influencing each unit in the POB. In the following paragraphs, we will illustrate the additional elements

that define our model of the POB.

Step 1: Cascade Input

In all simulations we have considered only 3 syllable words. To instruct the POB on the correct sequence of 3 syllables composing a word we worked on the shape of the heteroassociative matrix $G \in \mathbb{R}^{p^{POL} \times p^{POB}}$ in **Eq.**2.7. We assigned to each pattern stored in the word buffer a sequence of 3 syllables in a way that each syllable could be enrolled in composing only 3 words, each time in a different position. For instructing the sequential order of syllables for word W we set $G^{\mathrm{W,S_1}} = 1.0$, $G^{\mathrm{W,S_2}} = 0.9$ and $G^{\mathrm{W,S_3}} = 0.8$, where $\mathrm{S_1}$, $\mathrm{S_2}$, $\mathrm{S_3}$ are the indices of the first, second and third syllable of word W.

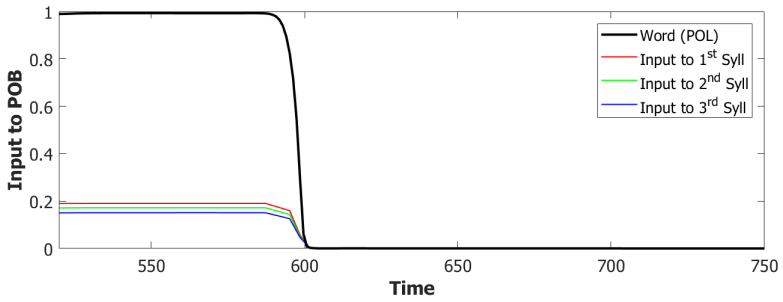


Figure 6.2: Input to the POB syllables (coloured lines) associated to the active word in the POL (black line).

We used a value of $\lambda = 0.2$ and a $\sigma - \sigma$ interaction to implement the heteroassociation. This type of mecanism was preferred to the $\theta - \sigma$ one to favour a synchronous dynamic of the two subnetworks. As we can see from **Fig.**6.2 the POL sends a constant input with different strengths to the 3 syllables to be produced. However the POB, even if sometimes it retrieves the correct sequence, as shown in **Fig.**6.3 by the *red*→*green*→*blue* color code, seems to enter a spontaneous latching phase where many wrong syllables are also retrieved. A possible origin of the problem can be seen by plotting the activity of the units encoding the 3 syllables (**Fig.**6.4). In an ideal scenario, the POB should be able to retrieve one syllable at a time, in the correct order and then turn off, waiting for the next utterance. What we observe here instead

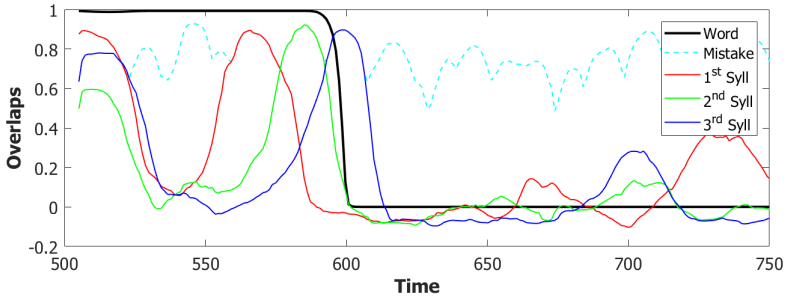


Figure 6.3: Example of the dynamics in the two subnetworks.

is sustained activity for all the units active in the 3 syllables, even after the end of the input.

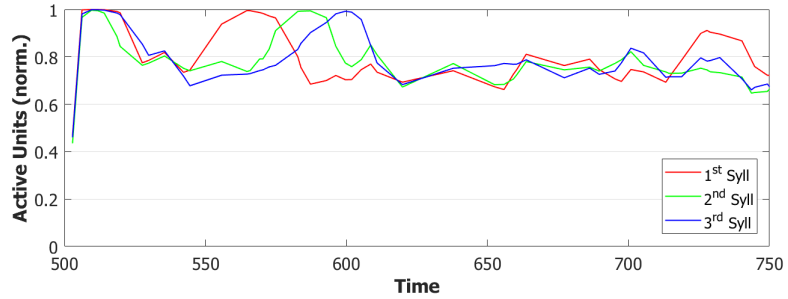
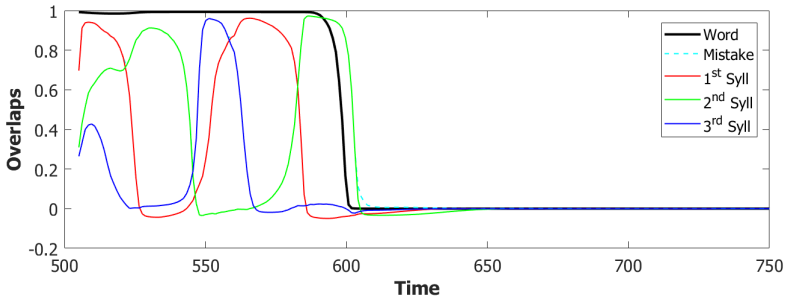


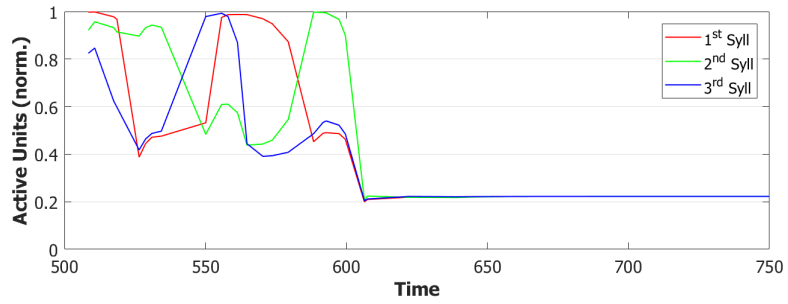
Figure 6.4: Normalized activity of the units encoding the 3 syllables. For each sullable this was measured as $\frac{1}{aN^{POB}} \sum_i^{N^{POB}} \left(1-\sigma_i^0\right)\left(1-\delta_{\xi_i^\mu 0}\right)$, where μ is the label of the syllable considered.

Step 2: Fast Inhibition

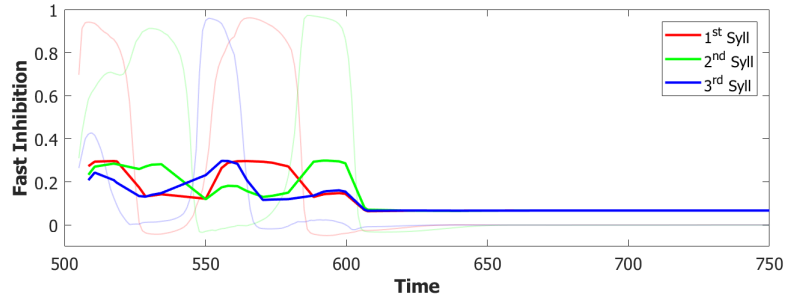
The constant signal from the POL induces an overactivation of the POB network which is effectively driven in a spontaneous latching regime, often preventing it to recover the correct sequence. A possible solution would be to artificially transform the constant input from the word buffer into a sequence of instantaneous cues. This approach would, however, shift the problem of serializing the phonological information to the POL. Another option would instead be to reduce the activity of the network by increasing the effect of inhibition on its active units. One way to achieve this is to introduce a fast inhibition component as the one treated in **section** 2.1.3.



(a) Example of latching dynamics.



(b) Activity of units encoding the syllables.



(c) Dynamics of the fast inhibition component.

Figure 6.5: Example of a simulation with the introduction of a fast inhibition component. The latching dynamics now ends with the deactivation of the word in the buffer and it is mainly restricted to the correct subset of syllables.

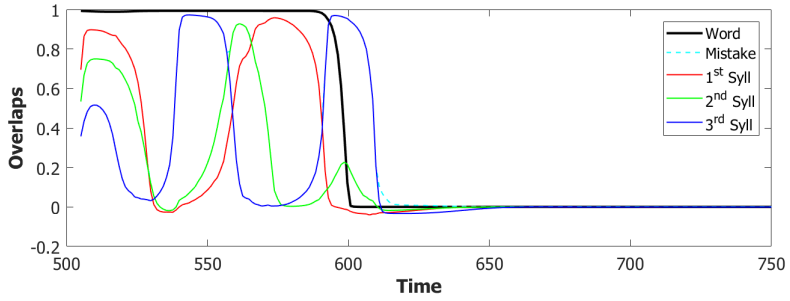


Figure 6.6: Example of a typical error induced by the co-activation of the 3 syllables.

For the simulations with this additional component, we used $\tau_3^A = 2$ and a proportion of fast inhibition $\gamma_A = 0.3$. As we can notice from the example in **Figs.**6.5a and 6.5b the overactivation problem induced by the constant input is resolved by the introduction of this new component. With this configuration, the network, over three batches of 50 simulations, retrieved the correct sequence in the first three latching steps around 55% of the times. However the POB still "speaks" for more than requested, even if it does that without adding new and unrequested syllables.

Coactivation of multiple syllables

A second and more subtle type of overactivation appears when comparing the three syllables with each other. Pairs of randomly correlated patterns share on average a^2N active units. For our sparsity $a = 0.25$ this corresponds roughly to a proportion of shared active units below 0.1. As we can see already in the example in **Fig.**6.5b, the minimum of this proportion of active units in our simulations fluctuates around a value of 0.4. This co-activation of syllables is indeed the main source of mistakes in this batch of simulations. **Figures** 6.5a and 6.6 show the effect of the overactivation on latching for $t < 540$, where all three syllables are simultaneously active in the network.

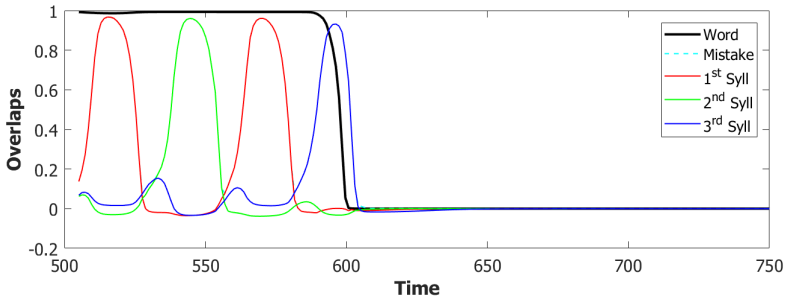
Step 3: Dynamic Global Threshold

The simultaneous activation of the 3 syllables leads the network into a mixed state where multiple patterns are active together. When this situation occurs the network has no immediately available pattern to latch to leading to false starts and spelling errors of the kind shown in **Fig.**6.6.
Co-activation of multiple patterns can be interpreted as a lack of competition between the syllables, mainly driven by a weak constraint on the total number of simultaneously active units. To increase the selectivity of our network, we need to introduce a mechanism that penalizes units not aligned with the most active syllable. This new type of inhibition was introduced in our simulations as a dynamic component added to the previously defined constant global threshold U:

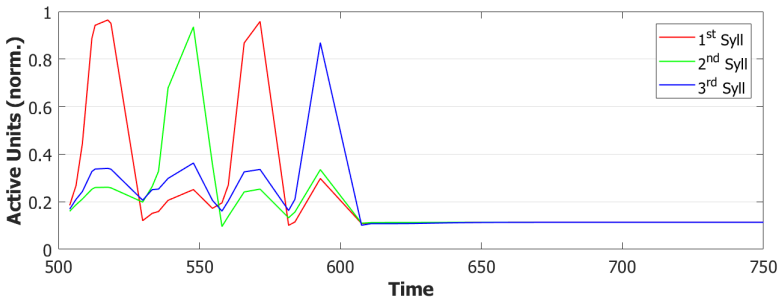
$$U(t) = U + \hat{U}(t) \tag{6.1a}$$

$$\tau \frac{d\hat{U}}{dt} = \frac{1}{aN^{POB}} \sum_{i \in POB} \left(1 - \sigma_i^0\right) - \hat{U}. \tag{6.1b}$$

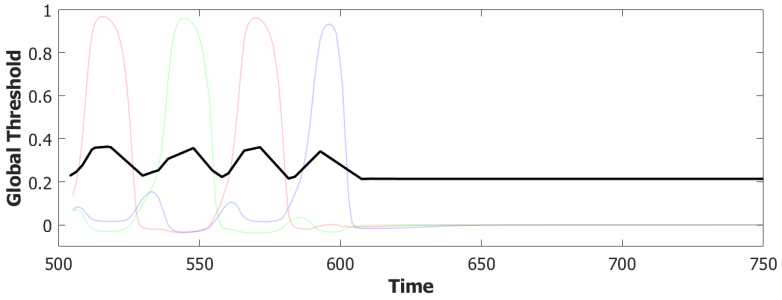
For our simulations, we set the value of τ equal to τ_3^A.
The introduction of the fast inhibition defined in **Eqs.**6.1 can be justified as a rough first-order correction to the problem of discretizing the cortex into units when defining our Potts network. This approximation, however, should be acceptable only for small enough networks, where it is reasonable to assume that the local inhibition to a unit, being this a fictitious discretization of a continuous substrate, may influence also other units close by.
Simulations performed with this new mechanism show a noticeable improvement in the quality of latching. Each syllabic utterance corresponds to an isolate latching step with no interference coming from other overlapping syllables. Nonetheless, as it will be discussed in depth in the next sections, the proportion of correct sequences decreased drastically in the way illustrated by the latching sequence in **Fig.**6.7a. The time dilation induced in the dynamics as a byproduct of the dynamic global inhibition introduced what we can



(a) Example of latching dynamics.



(b) Activity of units encoding the syllables.



(c) Dynamics of the new global threshold U.

Figure 6.7: Example of a simulation with the introduction of a fast global inhibition component. The latching dynamics is more polished compared to previous simulations. The effect of the new inhibition is to reduce the co-activation of multiple patterns.

define as a short-term memory issue in our simulations. A tentative solution will be addressed in the next, and final, step by modulating the adaptation component in the POB.

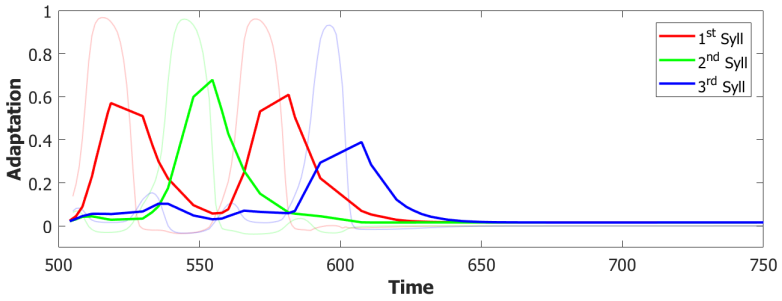


Figure 6.8: Dynamics of the adaptation for the 3 syllable for the simulation shown in **Figs.**6.7.

Step 4: Slow Adaptation

Adaptation, modeled by **Eq.**2.3, is the mechanism that forces active units to change their preferred state of activity once a certain amount has passed. **Fig.**6.8 shows the normalized amount of adaptation for each syllable. As we can notice from this example, at the time of the third utterance, the adaptation of the first syllable is already low enough to let it take advantage of its greater input and win the race against the third syllable.

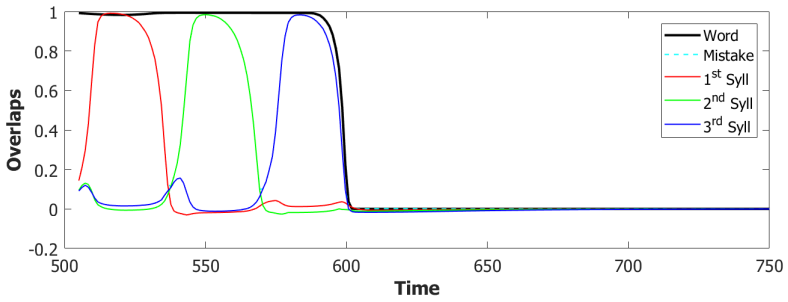
The time-constant τ_2 is then the parameter that regulates the memory of the network about the previously active patterns. The choice of a shorter time-scale to allow for a faster dynamics in the POB also corresponds to faster forgetting of the previous states. To correct this behavior, without altering too much the dynamics, we introduced a second term of adaptation, similarly to what has been done to merge slow and fast inhibition in **section** 2.1.3.

$$\theta_i^k(t) = \theta_i^{k(slow)}(t) + \theta_i^{k(fast)}(t) \tag{6.2a}$$

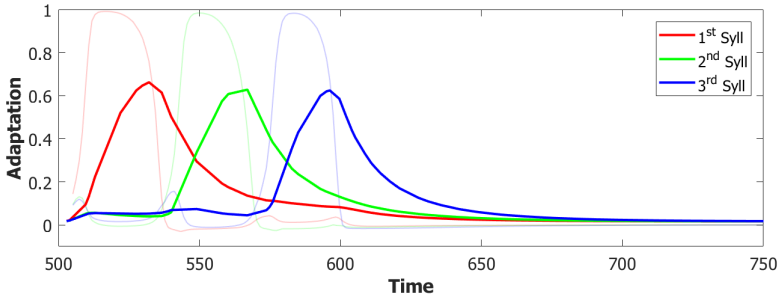
$$\tau_2^{(fast)} \frac{d\theta_i^{k(fast)}}{dt} = \gamma_2^{(fast)} \sigma_i^k - \theta_i^{k(fast)} \tag{6.2b}$$

$$\tau_2^{(slow)} \frac{d\theta_i^{k(slow)}}{dt} = \left(1 - \gamma_2^{(fast)}\right) \sigma_i^k - \theta_i^{k(slow)} \tag{6.2c}$$

To limit the number of parameters we set $\tau_2^{(fast)} = \tau_2^{POB}$ and $\tau_2^{(slow)} = \tau_2^{POL}$. For the proportion of slow and fast adaptation we set $\gamma_2^{(fast)} = 0.5$.



(a) Example of latching dynamics.



(b) Sum of slow and fast adaptation.

Figure 6.9: Example of a simulation with the introduction of a slow adaptation. The latching dynamics is more polished compared to previous simulations. The effect of the new inhibition is to reduce the co-activation of multiple patterns.

With this additional tool, many of the syllable repetitions were prevented (see **Figs.**6.9 for an example) and the performance of our model drastically improved to a value around 72% of correct sequences.

6.3 Simulation results

6.3.1 Performance of the POB model

In the previous section, we showed how we exploited the available mechanisms in our Potts neural network to build a basic model of the phonological output buffer. To analyze the performance of the final model, we run three batches of simulations, each having different sets of patterns for both the words and the syllables. For each set, we stored in the network 50 word – syllables associations which were all simulated by cueing the relative word in the POL. The resulting 150 simulations were then aggregated to analyze the performance of the network.

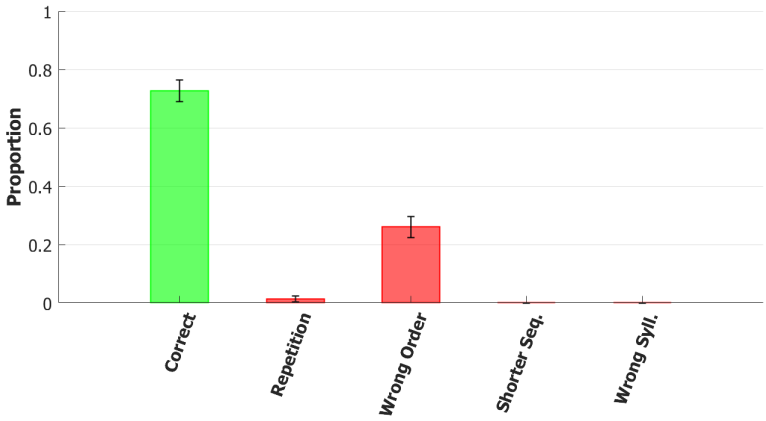


Figure 6.10: Report on the correctness of the first 3 utterances of the final POB model. Error bars represent the standard error of the mean.

In the final model the majority of utterances were correct (~ 72% accuracy), while the main type of errors committed by the network was to switch the position of two correct syllables.

6.3.2 Breaking the network: Analysis of errors

To better understand the role of the mechanisms included in the final model we ran, with the same procedure of the section before, other rounds of simulation each time removing a single one of the added components.

No slow adaptation

Slow adaptation was introduced in the third step to remove triplets of syllables with repetitions of the kind $\{1^{st} \rightarrow 2^{nd} \rightarrow 1^{st}\}$. The analysis of the performance in **Fig.**6.11 indeed confirms the prevalence of these repetition errors.

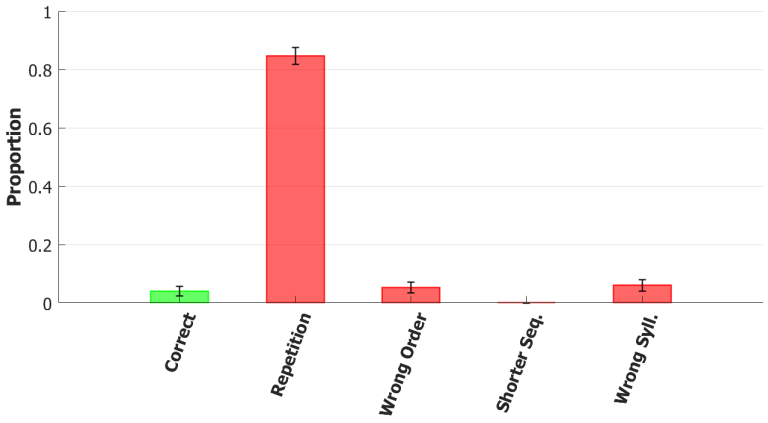
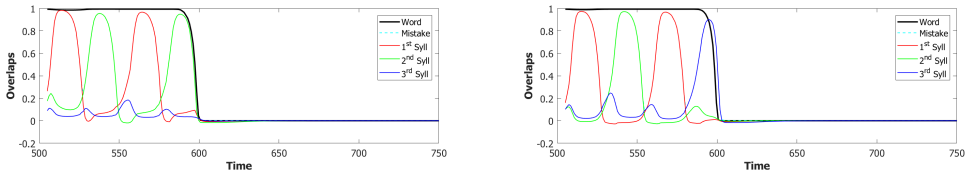


Figure 6.11: Report on the correctness of the first 3 utterances. The model considered has $\gamma_2^{(fast)} = 1$ while all other parameters are the same as in the complete model. Error bars represent the standard error of the mean.

Examples of repetition errors are shown in **Fig.**6.12. The type of repetition in **Fig.**6.12b could also be listed as an addition error, with the intrusion of the first syllable.

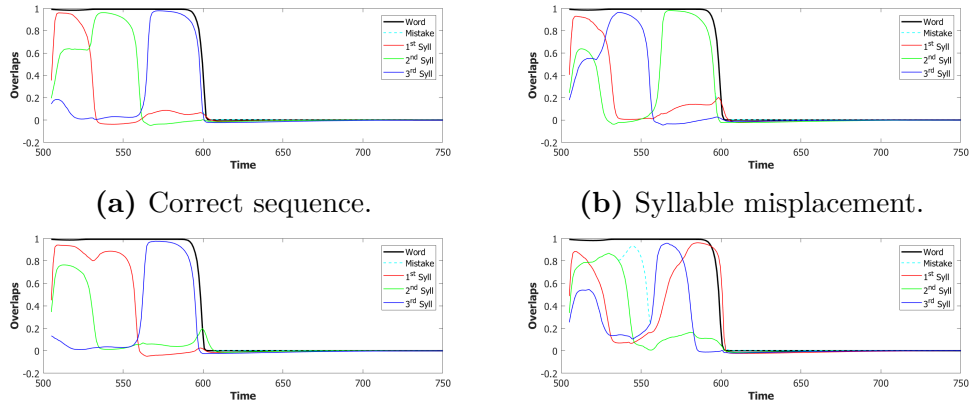


(a) Repetition of syll. "1" and "2". **(b)** Intrusion of syll. "1" before syll "3".

Figure 6.12: Exaples of typical errors in a network with no slow adaptation. Both types of errors are classified as repetition errors.

No dynamic global threshold

Simultaneous activation of multiple syllables lead us to the introduction of a competition mechanism in the model. In these simulations we removed from the complete model the time evolving component of the threshold U. For a fair comparison we assigned to U a higher value corresponding to the value assumed by **Eq.**6.1 for a network in the thermal state at a temperature $T = \frac{1}{\beta}$. In our case we set $U = 0.216$.



(a) Correct sequence. **(b)** Syllable misplacement.

(c) Syllable omission. **(d)** Syllable addition

Figure 6.13: Examples of a correct sequence and of three errors in a network with no dynamical global inhibition.

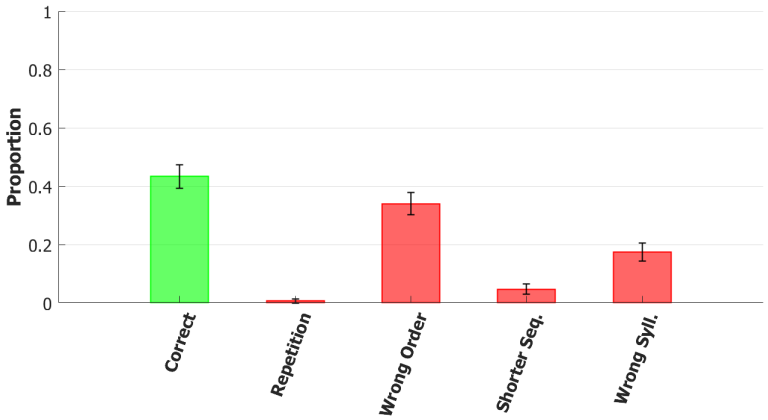
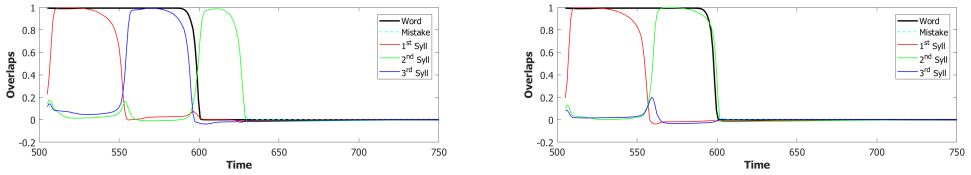


Figure 6.14: Report on the correctness of the first 3 utterances. The model considered has $U = 0.216$ and no temporal evolution for this parameter. Error bars represent the standard error of the mean.

This model showed a variety of types of errors and a low accuracy on correct sequences, as illustrated in **Figs.**6.13 and 6.14. For simplicity, we categorized omissions in "Shorter Sequence" errors, independently of the omitted syllable, and syllable additions in the "Wrong Syllable" class, to highlight the utterance of an intruded external syllable. All bisyllabic utterances involved only the correct syllables.

No fast local inhibition

Fast inhibition was the first "ingredient" added to the basic network. Simulations of the complete model but without the fast local inhibition showed the importance of this component. Omission errors were the predominant type of error in these simulations. Very few ($\sim 10\%$) trisyllabic words were indeed uttered by this network but almost never in the correct order.



(a) Inversion of syllables. **(b)** Omission of 3^{rd} syllable.

Figure 6.15: Examples of errors in a network with no fast local inhibition.

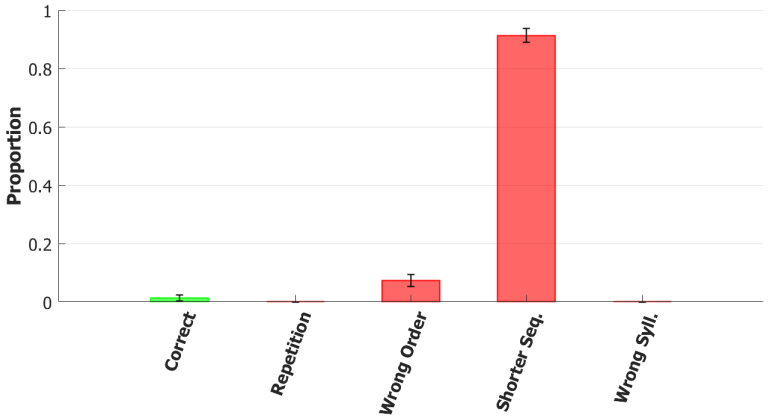


Figure 6.16: Report on the correctness of the first 3 utterances. The model considered has $\gamma_A = 0$, therefore only slow local inhibition is present. Error bars represent the standard error of the mean.

6.3.3 Discussion

Simulation results of the complete model of the POB show a good performance in producing trisyllabic utterances. The main purpose that this network achieved is to disentangle the compact information on the full word, stored in the POL, transforming it into an ordered sequence of syllables. While the role of the POB in human speech production is more complex than the one assessed in this chapter, possibly involving phonotactic and content dependent rules (i.e., numbers, plurals, irregular forms, etc.), this first-order description already captures interesting features of human performance.

Disrupting basic neural mechanisms in the model produces error patterns that may be compared with human subjects with specific POB impairments. Interestingly, each of the introduced mechanisms, namely adaptation, local and global inhibition, when removed produce each a different pattern of errors in the utterances, as shown in **Figs.**6.11, 6.14 and 6.16. Furthermore, the current implementation does not consider a feedback connection from the POB to the POL, which would be able to correct some of the wrong spellings reported in **Fig.**6.10 by modulating the strength of the inputs to the POB. A future version of the model will thus include recurrent connections, possibly increasing the accuracy of the utterances.

Nonetheless, the encouraging results of the current implementation suggest the suitability of our network for modeling not only higher cognitive functions but also specific low-level computations.

Chapter 7

Conclusion

In recent years, the advances of neuroimaging techniques have led to new insights on how our brain encodes information and on the dynamical interactions between different cortical regions. Nonetheless, our understanding of the role and functioning of the underlying neural mechanisms, of which we can only see the external, macroscopic, effects, is shaped by experimental constraints. Neuronal data not only are difficult to collect in human subjects but are also hard to compare with the macroscopic measures provided by the most popular neuroimaging techniques; the latter also having their specific issues, like the low space/time resolution or the often undetermined relation of the measured signal with the underlying neural computations.
From a methodological perspective, the neuropsychological investigation usually focuses its attention on the neural responses to external stimulations. In the literature of language processing, this is often done in non-ecological settings (e.g., unnatural stimuli, impaired patients, fast stimulus presentation, etc.), which allow to enhance the effects of interest. This kind of studies has led to detailed models for several neural functions, able to produce similar effects as the ones observed in the experimental setting. However, a fundamental question still remains unanswered: how does our brain use of these functions and their related encoded information during its spontaneous behavior? Or, how much can we extend the validity of these models to non-

experimental settings?

The ambitious goal of this thesis was to move the first steps towards a model of the cortex that could, in principle, link experimental results to spontaneous activity. To do so, we decided not to engineer models of specific functions to fit human performance; instead, we aimed to qualitatively replicate experimental results by including only biologically plausible mechanisms in a previously proposed network model of the cortex.

The Potts attractor neural network, described in the second chapter, has been proposed in recent years as a model of cortical dynamics. The model takes inspiration from Braitenberg's view of the cortex as an associative memory machine that stores memories as local (B-systems) and global (A-system) attractor states. Potts units, therefore, represent the activity of local neuronal networks while the full network stores memories as global attractors. This dual, local and global, nature can thus be considered as a model for the conversion from the analog computations performed by neurons into the discrete and macroscopic activity patterns observed with neuroimaging techniques.

Another strength of the model is its latching dynamics. This hopping process from one memory to another is introduced in the model by including three biologically inspired mechanisms, namely neuronal excitability, neural adaptation, and inhibition. In **Chapter 2**, we described how the dynamics is influenced by pattern correlations and rule-based associations. Within a certain region of the parameter space, latching transitions occur between correlated patterns (i.e., similar memories). The inclusion of a heteroassociative component allows for more complex latching sequences, which may include frequency-based transitions.

Together, the mechanisms described above define the cortical model that we used as a starting point for three different and more structured implementations.

In **Chapter 3** we have proposed a priming experiment that could test subjects on similarity- and frequency-based associations between words. Word similarity has been included as semantic or orthographic relations, while fre-

quency associations have been added as linguistic collocations. Orthographic relations proved to be faster to recognize with respect to collocations and semantic ones. Different ERP patterns have also been observed, which have allowed us to distinguish the different neural signatures related to semantic or orthographic associations and also with respect to a non-association control condition. A second experiment has highlighted, then, the different processing of vowel and consonants in orthographic associations. For semantic and collocation transitions, instead, an interaction of noun/noun vs. adjective/noun associations has shown the important effect of syntactic information, even at the single word level.

The priming experiment has then been used as a testing ground for our Potts neural network. In **Chapter 4**, we have considered a model with two networks, each encoding semantic or orthographic information. Collocations have been, instead, encoded as heteroassociative instructions between patterns. The recognition process has then been modeled as a spontaneous latching sequence elicited by the target word and, eventually, reaching the prime representation, stored in short-term memory, if an association is found. Simulated reaction times have shown a slower recognition of collocations, similar to experimental data in the noun/noun condition, but no difference has been found between semantic and orthographic relations. Future investigations including an improved definition of semantic and orthographic similarities will aim to find better agreement of the model with experimental data. Furthermore, a future implementation of syntax in the current model could be tested in the adjective/noun condition. Nonetheless, the current network can be regarded as a possible way of translating the processings involved in a structured experimental task into a spontaneous cortical process.

The experimental study of a truly spontaneous cortical process has then been considered in **Chapter 5**. The spontaneous thought generation during mind-wandering is a neural function that can be naturally described as a latching process. However, specific impairments of mind-wandering are observed in certain brain lesioned patients. A Potts model including a hippocampal

component, storing episodic memories, and a ventromedial prefrontal network, encoding schema representations, in addition to a semantic network, has been introduced to qualitatively replicate the observed impairments in hippocampal and ventromedial patients. In such a network, lesions to the hippocampus lead to purely semantic and schema-guided transitions. Latching length, however, has been shown to be similar to normal conditions, at least for models with a high enough contribution of the vmPFC network. Conversely, lesions to the vmPFC network lead to short latching sequences, mainly jumping between the memory items composing the episodic memories retrieved by the hippocampus.

Finally, in **Chapter 6**, we have shown the suitability of our cortical model in describing specific neural functions, such as the one of the phonological output buffer. Our basic network model suggests the importance of inhibitory and adaptive mechanisms in the final stages of speech production. Results from our implementation, indeed, relate different production errors when assembling syllable or phoneme sequences to specific dysfunctions of the inhibitory mechanism. Future investigations and more refined models, possibly including recurrent connections in the network, are, however, needed to extend the validity of the model to other linguistic materials and to confirm the current results.

Bibliography

[1] Paul Broca. Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique*, 6:330–57, 1861.

[2] Carl Wernicke. *Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis*. Cohn., 1874.

[3] Ludwig Lichtheim. On aphasia. *Brain*, 7:433–484, 1885.

[4] Norman Geschwind. The organization of language and the brain. *Science*, 170(3961):940–944, 1970.

[5] Katrin Amunts, Marianne Lenzen, Angela D Friederici, Axel Schleicher, Patricia Morosan, Nicola Palomero-Gallagher, and Karl Zilles. Broca's region: novel organizational principles and multiple receptor mapping. *PLoS biology*, 8(9):e1000489, 2010.

[6] Katerina Semendeferi, Kate Teffer, Dan P Buxhoeveden, Min S Park, Sebastian Bludau, Katrin Amunts, Katie Travis, and Joseph Buckwalter. Spatial organization of neurons in the frontal pole sets humans apart from great apes. *Cerebral cortex*, 21(7):1485–1497, 2010.

[7] Alfredo Ardila, Byron Bernal, and Monica Rosselli. How localized are language brain areas? a review of brodmann areas involvement in oral language. *Archives of Clinical Neuropsychology*, 31(1):112–122, 2016.

[8] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

[9] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453, 2016.

[10] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[11] Friedemann Pulvermüller. Hebb's concept of cell assemblies an the psychophysiology of word processing. *Psychophysiology*, 33(4):317–333, 1996.

[12] Ole Paulsen and Terrence J Sejnowski. Natural patterns of activity and long-term synaptic plasticity. *Current opinion in neurobiology*, 10(2):172–180, 2000.

[13] Natalia Caporale and Yang Dan. Spike timing–dependent plasticity: A hebbian learning rule. *Annual Review of Neuroscience*, 31(1):25–46, 2008. PMID: 18275283.

[14] Dominique Debanne, Beat H. Gähwiler, and Scott M. Thompson. Long-term synaptic plasticity between pairs of individual ca3 pyramidal cells in rat hippocampal slice cultures. *The Journal of Physiology*, 507(1):237–247, 1998.

[15] Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24):10464–10472, 1998.

[16] Valentino Braitenberg. Thoughts on the cerebral cortex. *Journal of theoretical biology*, 46(2):421–447, 1974.

[17] Valentino Braitenberg and Almut Schüz. *Anatomy of the cortex: statistics and geometry*, volume 18. Springer Science & Business Media, 2013.

[18] Walter J Freeman. The hebbian paradigm reintegrated: local reverberations as internal representations. *Behavioral and brain sciences*, 18(4):631–631, 1995.

[19] Daniel J Amit and Stefano Fusi. Learning in neural networks with material synapses. *Neural Computation*, 6(5):957–982, 1994.

[20] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[21] Daniel J Amit and Daniel J Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge university press, 1992.

[22] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[23] Dominic O'kane and Alessandro Treves. Why the simplest notion of neocortex as an autoassociative memory would not work. *Network: Computation in Neural Systems*, 3(4):379–384, 1992.

[24] Dominic O'Kane and Alessandro Treves. Short-and long-range connections in autoassociative memory. *Journal of Physics A: Mathematical and General*, 25(19):5055, 1992.

[25] D O'Kane and D Sherrington. A feature retrieving attractor neural network. *Journal of Physics A: Mathematical and General*, 26(10):2333, 1993.

[26] Alexis M Dubreuil and Nicolas Brunel. Storing structured sparse memories in a multi-modular cortical network model. *Journal of computational neuroscience*, 40(2):157–175, 2016.

[27] Edoardo Datteri and Federico Laudisa. Box-and-arrow explanations need not be more abstract than neuroscientific mechanism descriptions. *Frontiers in Psychology*, 5:464, 2014.

[28] Eric Hochstein. One mechanism, many models: a distributed theory of mechanistic explanation. *Synthese*, 193(5):1387–1407, May 2016.

[29] Athena Akrami, Eleonora Russo, and Alessandro Treves. Lateral thinking, from the hopfield model to cortical dynamics. *Brain research*, 1434:4–16, 2012.

[30] Henry Markram, Wulfram Gerstner, and Per Jesper Sjöström. Spike-timing-dependent plasticity: a comprehensive overview. *Frontiers in synaptic neuroscience*, 4:2, 2012.

[31] Ido Kanter. Potts-glass models of neural networks. *Phys. Rev. A*, 37:2739–2742, Apr 1988.

[32] Désiré Bollé, Patrick Dupont, and Jort van Mourik. Stability properties of potts neural networks with biased patterns and low loading. *Journal of Physics A: Mathematical and General*, 24(5):1065, 1991.

[33] Désiré Bollé, Patrick Dupont, and J Huyghebaert. Thermodynamic properties of the q-state potts-glass neural network. *Physical Review A*, 45(6):4194, 1992.

[34] Eleonora Russo, Vijay MK Namboodiri, Alessandro Treves, and Emilio Kropff. Free association transitions in models of cortical latching dynamics. *New Journal of Physics*, 10(1):015008, 2008.

[35] Eleonora Russo and Alessandro Treves. Cortical free-association dynamics: Distinct phases of a latching network. *Physical Review E*, 85(5):051920, 2012.

[36] Chol Jun Kang, Michelangelo Naim, Vezha Boboeva, and Alessandro Treves. Life on the edge: Latching dynamics in a potts neural network. *Entropy*, 19(9), 2017.

[37] Alessandro Treves. Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive neuropsychology*, 22(3-4):276–291, 2005.

[38] Emilio Kropff and Alessandro Treves. The complexity of latching transitions in large scale cortical networks. *Natural Computing*, 6(2):169–185, 2007.

[39] Itamar Lerner, Shlomo Bentin, and Oren Shriki. Spreading activation in an attractor network with latching dynamics: Automatic semantic priming revisited. *Cognitive science*, 36(8):1339–1382, 2012.

[40] Michelangelo Naim, Vezha Boboeva, Chol Jun Kang, and Alessandro Treves. Reducing a cortical network to a potts model yields storage capacity estimates. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(4):043304, 2018.

[41] R. B. Potts and C. Domb. Some generalized order-disorder transformations. *Proceedings of the Cambridge Philosophical Society*, 48:106, 1952.

[42] Jeffry S Isaacson and Massimo Scanziani. How inhibition shapes cortical activity. *Neuron*, 72(2):231–243, 2011.

[43] Michael C Corballis. *The Recursive Mind: The Origins of Human Language, Thought, and Civilization-Updated Edition*. Princeton University Press, 2014.

[44] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544, 1952.

[45] Jun Ma and Jun Tang. A review for dynamics in neuron and neuronal network. *Nonlinear Dynamics*, 89(3):1569–1578, Aug 2017.

[46] Shimon Marom. Neural timescales or lack thereof. *Progress in neurobiology*, 90(1):16–28, 2010.

[47] Nachum Ulanovsky, Liora Las, Dina Farkas, and Israel Nelken. Multiple time scales of adaptation in auditory cortex neurons. *Journal of Neuroscience*, 24(46):10440–10453, 2004.

[48] Asaf Gal, Danny Eytan, Avner Wallach, Maya Sandler, Jackie Schiller, and Shimon Marom. Dynamics of excitability over extended timescales in cultured cortical neurons. *Journal of Neuroscience*, 30(48):16332–16342, 2010.

[49] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. Gabaergic interneurons in the neocortex: from cellular properties to circuits. *Neuron*, 91(2):260–292, 2016.

[50] Kwang Il Ryom, Vezha Boboeva, and Alessandro Treves. Unpublished. 2019.

[51] Andrea Nadalini, Marco Marelli, Roberto Bottini, and Davide Crepaldi. Local associations and semantic ties in overt and masked semantic priming. In *CLiC-it*, 2018.

[52] Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395, 2016.

[53] Steven M Frankland and Joshua D Greene. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, 112(37):11732–11737, 2015.

[54] Jon Sprouse and Norbert Hornstein. Chapter 14 - syntax and the cognitive neuroscience of syntactic structure building. In Gregory Hickok and Steven L. Small, editors, *Neurobiology of Language*, pages 165 – 174. Academic Press, San Diego, 2016.

[55] Doug Merchant. *Idioms at the Interface (s): Towards a psycholinguistically grounded model of sentence generation*. PhD thesis, University of Georgia, 2019.

[56] Sahar Pirmoradian and Alessandro Treves. Encoding words into a potts attractor network. In *Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop*, pages 29–42. World Scientific, 2014.

[57] Serena Di Santo and Alessandro Treves. Unpublished. 2019.

[58] Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[59] Maximilien Chaumon, Dorothy VM Bishop, and Niko A Busch. A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of neuroscience methods*, 250:47–63, 2015.

[60] Edward T Bullmore, John Suckling, Stephan Overmeyer, Sophia Rabe-Hesketh, Eric Taylor, and Michael J Brammer. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images of the brain. *IEEE transactions on medical imaging*, 18(1):32–42, 1999.

[61] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190, 2007.

[62] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011:1, 2011.

[63] Albert Kim and Vicky Lai. Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from erps. *Journal of cognitive neuroscience*, 24(5):1104–1112, 2012.

[64] Manuel Carreiras, Blair C Armstrong, Manuel Perea, and Ram Frost. The what, when, where, and how of visual word recognition. *Trends in cognitive sciences*, 18(2):90–98, 2014.

[65] Marta Kutas and Kara D Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647, 2011.

[66] Harm Brouwer, Matthew W Crocker, Noortje J Venhuizen, and John CJ Hoeks. A neurocomputational model of the n400 and the p600 in language processing. *Cognitive science*, 41:1318–1352, 2017.

[67] Stefanie Regel, Lars Meyer, and Thomas C Gunter. Distinguishing neurocognitive processes reflected by p600 effects: Evidence from erps and neural oscillations. *PloS one*, 9(5):e96840, 2014.

[68] Nicola Molinaro, Horacio A Barber, Sendy Caffarra, and Manuel Carreiras. On the left anterior negativity (lan): The case of morphosyntactic agreement. *Cortex*, 66(156-159), 2015.

[69] Davide Crepaldi, Emmanuel Keuleers, Paweł Mandera, and Marc Brysbaert. Subtlex-it. *(http://crr.ugent.be/subtlex-it/)*, 2013.

[70] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[71] Marco Marelli. Word-embeddings italian semantic spaces: a semantic model for psycholinguistic research. (http://meshugga.ugent.be/snaut-italian-2/). *Psihologija*, 50, 10 2017.

[72] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

[73] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, and William Brockman. The google books team, joseph p. *Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*, pages 176–182, 2011.

[74] Vezha Boboeva, Romain Brasselet, and Alessandro Treves. The capacity for correlated semantic memories in the cortex. *Entropy*, 20(11):824, 2018.

[75] Samuel J Cheyette and David C Plaut. Modeling the n400 erp component as transient semantic over-activation within a neural network model of word comprehension. *Cognition*, 162:153–166, 2017.

[76] Matthew A Killingsworth and Daniel T Gilbert. A wandering mind is an unhappy mind. *Science*, 330(6006):932–932, 2010.

[77] Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124(1):1–38, 2008.

[78] Kalina Christoff, Alan M Gordon, Jonathan Smallwood, Rachelle Smith, and Jonathan W Schooler. Experience sampling during fmri reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences*, 106(21):8719–8724, 2009.

[79] Jonathan Smallwood, Kevin Brown, Ben Baird, and Jonathan W Schooler. Cooperation between the default mode network and the

frontal–parietal network in the production of an internal train of thought. *Brain research*, 1428:60–70, 2012.

[80] Kieran CR Fox, R Nathan Spreng, Melissa Ellamil, Jessica R Andrews-Hanna, and Kalina Christoff. The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *Neuroimage*, 111:611–621, 2015.

[81] Elena Bertossi and Elisa Ciaramelli. Ventromedial prefrontal damage reduces mind-wandering and biases its temporal focus. *Social cognitive and affective neuroscience*, 11(11):1783–1791, 2016.

[82] Elena Bertossi, Ludovica Peccenini, Andrea Solmi, Alessio Avenanti, and Elisa Ciaramelli. Transcranial direct current stimulation of the medial prefrontal cortex dampens mind-wandering in men. *Scientific reports*, 7(1):16962, 2017.

[83] Cornelia McCormick, Clive R Rosenthal, Thomas D Miller, and Eleanor A Maguire. Mind-wandering in people with hippocampal damage. *Journal of Neuroscience*, 38(11):2745–2754, 2018.

[84] Asaf Gilboa and Hannah Marlatte. Neurobiology of schemas and schema-mediated memory. *Trends in cognitive sciences*, 21(8):618–631, 2017.

[85] Max Coltheart. Disorders of reading and their implications for models of normal reading. *Visible language*, 15(3):245, 1981.

[86] Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204, 2001.

[87] Naama Friedmann and Max Coltheart. Types of developmental dyslexia. *Handbook of communication disorders: Theoretical, empirical, and applied linguistics perspectives*, pages 1–37, 2016.

[88] Tim Shallice, Raffaella I Rumiati, and Antonella Zadini. The selective impairment of the phonological output buffer. *Cognitive Neuropsychology*, 17(6):517–546, 2000.

[89] Dror Dotan and Naama Friedmann. Steps towards understanding the phonological output buffer and its role in the production of numbers, morphemes, and function words. *Cortex*, 63:317–351, 2015.

[90] Alfonso Caramazza, Gabriele Miceli, and Gianpiero Villa. The role of the (output) phonological buffer in reading, writing, and repetition. *Cognitive Neuropsychology*, 3(1):37–76, 1986.

[91] Max Coltheart and Elaine Funnell. *Reading and writing: one lexicon or two?* Academic Press, 1987.

[92] Jennifer R. Shelton and Michael Weinrich. Further evidence of a dissociation between output phonological and orthographic lexicons: A case study. *Cognitive Neuropsychology*, 14(1):105–129, 1997.