Neural bases of variable binding in symbolic representations : experimental and modelling exploration

Martín Pérez-Guevara

UNIVERSITÉ PARIS DESCARTES

MEMBRE DE
USPC
Université Sorbonne Paris Cité

UNIVERSITÉ PARIS DESCARTES

École doctorale N° 158 (ED3C) Cerveau-Cognition-Comportement

Neurospin / Unicog U-992 / Équipe de neuroimagerie linguistique

Bases neuronales de binding dans des représentations symboliques

Exploration expérimentale et de modélisation

Par Martín Pérez-Guevara

Spécialité de doctorat : Neurosciences cognitives et informatiques

Dirigée par Christophe Pallier

Présentée et soutenue publiquement le 29 Novembre 2017

Devant un jury composé de :

Dr. Xavier ALARIO	Rapporteur	Directeur de recherche, Universite Centre Saint-Charles. Marseille, France.
Dr. Sander BOHTE	Rapporteur	Chargé de recherche, CWI. Amsterdam, The Netherlands.
Dr. Emmanuel DUPOUX	Examinateur	Directeur de recherche, École des Hautes Etudes en Sciences Sociales. Paris, France.
Dr. Marc DE KAMPS	Examinateur, Président du jury	Professeur, University of Leeds.
Dr. Christophe PALLIER	Directeur de thèse	Directeur de recherche INSERM/CEA/SAC/DSV/DRM/Neurospin center. Saclay, France.

UNIVERSITÉ PARIS DESCARTES

MEMBRE DE
USPC
Université Sorbonne Paris Cité

UNIVERSITY OF PARIS DESCARTES

Doctoral School N° 158 (ED3C) Brain-Cognition-Behavior

Neurospin / Unicog U-992 / Language neuroimaging team

Neural bases of variable binding in symbolic representations

Experimental and modelling exploration

By Martín Pérez-Guevara

PhD Specialty : Cognitive and Computational Neuroscience

Directed by Christophe Pallier

Presented and defended publicly the 29th of November 2017

Jury composition :

Dr. Xavier ALARIO	Reviewer	Research Director, Universite Centre Saint-Charles. Marseille, France.
Dr. Sander BOHTE	Reviewer	Research Associate, CWI. Amsterdam, The Netherlands.
Dr. Emmanuel DUPOUX	Examiner	Research Director, École des Hautes Etudes en Sciences Sociales. Paris, France.
Dr. Marc DE KAMPS	Examiner, Jury President	Professor, University of Leeds.
Dr. Christophe PALLIER	Doctoral advisor	Research Director INSERM/CEA/SAC/DSV/DRM/Neurospin center. Saclay, France.

Acknowledgements

I want to thank Christophe Pallier for giving me the opportunity to develop my scientific thinking under his supervision, for his intellectual advice and the great mood he brought to the lab everyday with his insights and jokes. Also a substantial part of this work was made possible thanks to an intense exchange with Marc De Kamps, I am deeply grateful for all the dedication, supervision and assistance he offered. I want to thank too the reviewers, Xavier Alario and Sander Bohte, and the examiner, Emmanuel Dupoux, for kindly agreeing to take the time to evaluate this work.

In addition I want to thank Bertrand Thirion and Gaël Varoquaux, that gave me the opportunity to grow as a software developer and scientist in the open source and open science community, it was a great complementary experience to my PhD. I thank Stanislas Dehaene, for welcoming me in the Unicog lab and for the several interesting discussions we had. I also want to mention Paul Smolensky for the brief but rich exchanges we had. Thanks to him the first projects of this PhD got kickstarted.

I can not forget to mention my partner Elodie Doger de Speville, that gave me an incredible amount of personal and professional support in the worst and best moments, specially stoically in the worst.

Several other colleagues in the lab, Florent Meyniel, Matthew Nelson and Thomas Hannagan, and several friends, Andres Hoyos, Darinka Trubutschek and Valentina Borghesani, kindly assisted this work as well and I am thankful to you all for your support. Considering the experimental part of this work, the Manip radio team were very helpful, I dont imagine how the acquisitions would have gone without their constant assistance. I thank you too.

Of course I also thank all other friends and colleagues in Unicog and Parietal, that in one way or another, contributed to this work or enriched with their company my PhD experience.

UNIVERSITÉ PARIS DESCARTES

Doctoral School N° 158 (ED3C)
Brain-Cognition-Behavior

Résumé – Bases neuronales de binding dans des représentations symboliques

Le travail présenté dans cette thèse fait partie d'un programme de recherche qui vise à comprendre comment le cerveau traite et représente les structures symboliques dans des domaines comme le langage ou les mathématiques. L'existence de structures composées de sous-éléments, tel que les morphèmes, les mots ou les phrases est très fortement suggérée par les analyses linguistiques et les données expérimentale de la psycholinguistique. En revanche, l'implémentation neuronale des opérations et des représentations qui permettent la nature combinatoire du langage reste encore essentiellement inconnue. Certaines opérations de composition élémentaires permettant une représentation interne stable des objets dans le cortex sensoriel, tel que la reconnaissance hiérarchique des formes, sont aujourd'hui mieux comprises[120]. En revanche, les modèles concernant les opérations de liaisons(*binding*) nécessaires à la construction de structures symboliques complexes et possiblement hiérarchiques, pour lesquelles des manipulations précises des composants doit être possible, sont encore peu testés de façon expérimentale et incapables de prédire les signaux en neuroimagerie.

Combler le fossé entre les données de neuroimagerie expérimentale et les modèles proposés pour résoudre le problème de *binding* est une étape cruciale pour mieux comprendre les processus de traitements et de représentation des structures symboliques. Au regard de ce problème, l'objectif de ce travail était d'identifier et de tester expérimentalement les théories basées sur des réseaux neuronaux, capables de traiter des structures symboliques pour lesquelles nous avons pu établir des prédictions testables, contre des mesures existantes de neuroimagerie fMRI et ECoG dérivées de tâches de traitement du langage.

Nous avons identifié deux approches de modélisation pertinentes. La première approche s'inscrit dans le contexte des architectures symboliques vectorielles (VSA), qui propose une modélisation mathématique précise des opérations nécessaires pour représenter les structures dans des réseaux neuronaux artificiels. C'est le formalisme de Paul Smolensky[172], utilisant des produit tensoriel (TPR) qui englobe la plupart des architectures VSA précédemment proposées comme, par exemple, les modèles d'Activation synchrones[170], les représentations réduites holographique[158], et les mémoires auto-associatives récursives[35].

La seconde approche que nous avons identifiée est celle du "Neural Blackboard Architecture" (NBA), développée par Marc De Kamps et Van der Velde[187]. Elle se démarque des autres en proposant une implémentation des mécanismes associatifs à travers des circuits formés par des assemblages de réseaux neuronaux. L'architecture du Blackboard repose sur des changements de connectivité transitoires des circuits d'assemblages neuronaux, de sorte que le potentiel de l'activité neurale permise par les mécanismes de mémoire de travail après un processus de liaison, représente implicitement les structures symboliques.

Dans la première partie de cette thèse, nous détaillons la théorie derrière chacun de ces modèles et les comparons, du point de vue du problème de *binding*. Les deux modèles sont capables de répondre à la plupart des défis théoriques posés par la modélisation neuronale des structures symboliques, notamment ceux présentées par Jackendoff[99]. Néanmoins, ces deux classes de modèles sont très différentes. Le TPR de Smolenky s'appuie principalement sur des considérations spatiales statiques d'unités neurales artificielles, avec des représentations explicites complètement distribuées et spatialement stables mises en œuvre par des vecteurs. La NBA en revanche, considère les dynamiques temporelles de décharge de neurones artificiels, avec des représentations spatialement instables implémentées par des assemblages neuronaux.

Dans la deuxième partie de la thèse, nous testons empiriquement le *principe de superposition* qui stipule que l'activité

associé à une structure est la somme des activités de ses parties. Ceci est une des hypothèses les plus cruciales du TPR de Smolensky. Afin d'obtenir un ensemble de données pertinent pour tester ce principe, nous avons créé une expérience IRMf dans laquelle les participants lisaient ou entendaient des pseudomots composés de deux syllables CV. Nous avons employé un approche de décodage de l'activité BOLD afin d'analyser comment ces bisyllabes sont encodées dans diverses régions cérébrale. Nous avons obtenu de bon scores de classification dans certaines régions sensorielles et nous avons reproduit des effets connus, tel que les représentations semi-locales superposées induites par la rétinotopie. Dans le cas des régions auditives, nous avons trouvé un faible évidence en faveur de la superposition dans les zones supérieures dans la hiérarchie de traitement auditif. Nous avons montré que la classification des items bi-syllabiques dans les régions 44 et 45 de Broca etait significative et que l'ensemble de ces régions montrait des preuves en faveur de la superposition.

De plus, nous avons trouvé des résultats qui militent contre l'existence de représentations superposées dans la zone de la forme visualle des mots (VWFA), ce qui est cohérent avec les recherches antérieures sur la représentations de mots entiers dans cette région[75]. Nous avons également vérifié qu'il était possible de décoder les représentations auditives dans la VWFA, suggérant que cette région est impliquée aussi bien dans le traitement de la parole que des mots écrits[205]. Toutefois, un résultat surprenant a été l'absence totale de généralisation des modèles de décodage utilisés d'une modalité sensorielle à une autre. Ce maque de généralisation pourrait être interprété comme un manque de sensibilité dû à la variabilité du signal des représentations, ou encore comme l'absence de représentations amodales pour un pseudo-mot bi-syllabique simple. En dehors des zones sensorielles, nous avons observé dans la plupart des régions avec des scores de classification significatifs, une variabilité extrême des scores de précision pour des items individuels, de sorte que peu d'entre eux avaient des scores particulièrement élevés, alors que la plupart restaient de façon uniforme à un niveau de chance. Ce pattern particulièrement précis pourrait s'expliquer par le manque de parcimonie et la faible variabilité dans la distribution spatiale des valeurs des vecteurs neuraux sous-jacents aux représentations neuronales, pour lesquels nous n'avons par chance, capturé quelques segments déviants. Au regard de ces résultats, nous pensons qu'il serait intéréssant dans une perspectives future de tester le principe de superposition avec des signaux BOLD, en utilisant des résolutions spatiales plus élevées comme celles obtenues par des techniques récentes telles que l'IRMf laminaire[111].

Nous nous sommes également intéressés à la dynamique temporelle des liaisons qui pourrait être détectée dans les mesures de neuro-imagerie IRMf et ECoG. Etant donné que le TPR de Smolensky n'a pas de prédictions particulières sur la dynamique temporelle neurale ou sur les décharges neuronales biologiques, nous nous sommes focalisés sur les prédictions de la NBA. Dans la deuxième partie de la thèse, nous avons créé une nouvelle implémentation de la NBA basée sur les techniques de densité de population, qui nous a permis de faire des prédictions temporelles de haute résolution de la dynamique neurale liée au processus de liaison. Une partie importante de ce travail a été réalisée en collaboration avec Marc De Kamps.

Nos simulations s'appuient sur la dynamique des modèles de point de décharges des neurones : Les neurones qui Leaky-integrate-and-re (LIF) et adaptive-exponential-integrate-and-re (AdEx). Plutôt que de simuler des milliers de neurones en décharges, nous avons utilisé des techniques de densité de population (PDT) pour modéliser la dynamique au niveau de la population. Bien que liée aux modèles basés sur les taux de décharge, pour les PDTs la correspondance avec les quantités de population moyennées de neurones en décharge peut être montrée rigoureusement. En particulier, nos simulation montrent que les dynamiques transitoires sont capturées avec plus de précision par les PDT que par les modèles basés sur les taux de décharge. Le contraste entre les modèles LIF et ADEx nous ont permis de démontrer que, bien qu'ils ne soient pas différenciés par la dynamique moyenne, leur paramétrisations ont de fortes implications pour le timing et le contrôle des événements de traitement des phrases.

Nous montrons que notre implementation de l'architecture NB, avec des paramètres réglés unqiauement pour répondre à des contraintes opérationnelles, reproduit qualitativement les profils d'activités neuronales de deux

expériences de neuro-imagerie, utilisant l'EcoG[141] et l'IRMf[153], et mettant en oeuvre des opérations de *binding* linguistique. En même temps que la flexibilité partiellement explorée de la NBA pour représenter des structures d'arbres binaires arbitraires et des schémas d'analyse, ces résultats en font un outil prometteur pour l'exploration des hypothèses linguistiques et une prise en compte quantitative subtile des mesures de neuroimagerie multi-échelles.

Publications

[1] Modelling the neural dynamics of binding in language with the Neural Blackboard Architecture, Martín Pérez-Guevara, Marc De Kamps and Christophe Pallier. *In Preparation.*

[2] Relationships between Regional Radiation Doses and Cognitive Decline in Children Treated with Cranio-Spinal Irradiation for Posterior Fossa Tumors. Doger de Speville, E., Robert, C., Perez-Guevara, M., Grigis, A., Bolle, S., Pinaud, C., ... & Grill, J. (2017). Frontiers in Oncology, 7, 166.

Communications orales

[1] Neural Blackboard Architecture simulation with simple biological networks captures the behavior of diverse neuroimaging measurements during language processing. Perez-Guevara M, De Kamps M, Pallier C. (2016). Oral communication and poster at NIPS workshop (MLINI).

[2] Experimental explorations on Smolensky's superposition principle with double digit numbers. Perez-Guevara M, Pallier C. (2016). Poster at Neuroscience Workshop Saclay (NEWS).

Doctoral School N° 158 (ED3C)
Brain-Cognition-Behavior

Summary – Neural bases of variable binding in symbolic representations

The aim of this thesis is to understand how the brain computes and represents symbolic structures, such like those encountered in language or mathematics. The existence of parts in structures like morphemes, words and phrases has been established through decades of linguistic analysis and psycholinguistic experiments. Nonetheless the neural implementation of the operations that support the extreme combinatorial nature of language remains unsettled. Some basic *composition* operations that allow the stable internal representation of sensory objects in the sensory cortex, like hierarchical pattern recognition, receptive fields, pooling and normalization, have started to be understood[120]. But models of the *binding* operations required for construction of complex, possibly hierarchical, symbolic structures on which precise manipulation of its components is a requisite, lack empirical testing and are still unable to predict neuroimaging signals.

In this sense, bridging the gap between experimental neuroimaging evidence and the available modelling solutions to *the binding problem* is a crucial step for the advancement of our understanding of the brain computation and representation of symbolic structures. From the recognition of this problem, the goal of this PhD became the identification and experimental test of the theories, based on neural networks, capable of dealing with symbolic structures, for which we could establish testable predictions against existing fMRI and ECoG neuroimaging measurements derived from language processing tasks.

We identified two powerful but very different modelling approaches to the problem. The first is in the context of the tradition of Vectorial Symbolic Architectures (VSA) that bring precise mathematical modelling to the operations required to represent structures in the neural units of artificial neural networks and manipulate them. This is Smolensky's formalism with *tensor product representations* (TPR)[172], which he demonstrates can encompass most of the previous work in VSA, like Synchronous Firing[170], Holographic Reduced Representations[158] and Recursive Auto-Associative Memories[35].

The second, is the *Neural Blackboard Architecture* (NBA) developed by Marc De Kamps and Van der Velde[187], that importantly differentiates itself by proposing an implementation of binding by process in circuits formed by neural assemblies of spiking neural networks. Instead of solving binding by assuming precise and particular algebraic operations on vectors, the NBA proposes the establishment of transient connectivity changes in a circuit structure of neural assemblies, such that the potential flow of neural activity allowed by working memory mechanisms after a binding process takes place, implicitly represents symbolic structures.

The first part of the thesis develops in more detail the theory behind each of these models and their relationship from the common perspective of solving the binding problem. Both models are capable of addressing most of the theoretical challenges posed currently for the neural modelling of symbolic structures, including those presented by Jackendoff[99]. Nonetheless they are very different, Smolenky's TPR relies mostly on spatial static considerations of artificial neural units with explicit completely distributed and spatially stable representations implemented through vectors, while the NBA relies on temporal dynamic considerations of biologically based spiking neural units with implicit semi-local and spatially unstable representations implemented through neural assemblies.

For the second part of the thesis, we identified the superposition principle, which consists on the addition of the neural activations of each of the sub-parts of a symbolic structure, as one of the most crucial assumptions of Smolensky's TPR. To obtain a relevant dataset to test this principle, we created an fMRI experiment where participants perceived bi-syllabic CVCV pseudoword items in auditory and visual modalities, looking for sensory independent representations,

and used decoding techniques to analyse how these were encoded in diverse brain regions. We achieved high accuracy scores in our decoding models for representations in sensory areas and reproduced known effects like the superposed semi-local representations induced by retinotopy. In the case of auditory regions we found weak evidence in favour of superposition in areas higher in the auditory processing hierarchy. We show that bi-syllabic item classification is significant in regions 44 and 45 of the Broca's complex and that the whole complex portrays evidence in favour of superposition.

Moreover we found evidence against superposed representations in the visual word form area (VWFA), which is coherent with previous evidence of whole word representations in that region[75]. We also verified that it was possible to decode auditory representations form the VWFA, providing additional evidence to the literature body claiming that this region can be modulated by speech as well as reading[205]. We were surprised by a global lack of generalization from decoding models trained in one sensory modality to the other, which can be either interpreted as a lack of sensitivity due to variability of the representations signal or as the absence of amodal representations for simple bi-syllabic pseudowords. We observed in most regions with significant classification scores, outside of sensory areas, extreme variability in the accuracy scores of individual items, such that few had particularly high scores while most remained uniformly at chance level. This particular accuracy pattern could be explained by lack of sparsity and low variability in the spatial distribution of values of the neural vectors underlying the neural representations, for which we captured only some deviant segments by chance. From this we still think that it would be worth to further test the superposition principle with BOLD signals but only if taking advantage of higher spatial resolutions as those offered by recent techniques like laminar fMRI[111].

We were also interested in the temporal dynamics of binding which could be reflected in fMRI and ECoG neuroimaging measurements. As Smolensky's TPR do not have particular predictions on neural temporal dynamics or biological neural spiking, we decided to focus on predictions of the NBA. So for the second part of the thesis we created a new implementation of the NBA based on population density techniques, that allow us to make temporal high resolution predictions of neural dynamics linked to the binding process. A large amount of work, done in collaboration with Marc De Kamps, was needed to actually implement the NBA.

Our simulations are based on the dynamics of spiking point model neurons: leaky-integrate-and-fire (LIF) and adaptive-exponential-integrate-and-fire (AdEx) neurons. Rather than simulating thousands of spiking neurons, we use population density techniques (PDTs) to model dynamics at the population level. Although related to rate based models, for PDTs the correspondence to population-averaged quantities of spiking neurons can be shown rigorously. In particular transient dynamics are captured more accurately than by rate based models. Contrasting LIF and AdEx models allowed us to demonstrate that, although they are not importantly differentiated by average dynamics, their parametrization have strong implications for the timing and control of phrase processing events.

We demonstrate that an NBA implementation, only tuned to operational constraints, qualitatively reproduces the neural activity patterns of at least two neuroimaging experiments involving linguistic binding at different spatio-temporal scales. With the sole implementation of the binding mechanism we qualitatively reproduce temporal segments of the neural dynamics of sentence comprehension from intracortical recordings (ECoG) patterns[141]. Our model also replicates sub-linear patterns of hemodynamic responses caused by phrase constituency manipulations[153] and produces an alternative hypothesis to explain it, based on the number of binding operations executed during phrase processing. These results, alongside the partially explored flexibility of the NBA to represent arbitrary binary tree structures and parsing schemes, makes it a promising tool for linguistic hypothesis exploration and future refined quantitative accounts of multi-scale neuroimaging measurements.

Publications

[1] Modelling the neural dynamics of binding in language with the Neural Blackboard Architecture, Martín Pérez-Guevara, Marc De Kamps and Christophe Pallier. *In Preparation.*

[2] Relationships between Regional Radiation Doses and Cognitive Decline in Children Treated with Cranio-Spinal Irradiation for Posterior Fossa Tumors. Doger de Speville, E., Robert, C., Perez-Guevara, M., Grigis, A., Bolle, S., Pinaud, C., ... & Grill, J. (2017). Frontiers in Oncology, 7, 166.

Oral communications

[1] Neural Blackboard Architecture simulation with simple biological networks captures the behavior of diverse neuroimaging measurements during language processing. Perez-Guevara M, De Kamps M, Pallier C. (2016). Oral communication and poster at NIPS workshop (MLINI).

[2] Experimental explorations on Smolensky's superposition principle with double digit numbers. Perez-Guevara M, Pallier C. (2016). Poster at Neuroscience Workshop Saclay (NEWS).

Contents

Part I

The binding problem: theoretical and computational models

1 Theories of variable binding

In this chapter we introduce the binding problem in neuroscience. We also explain two main modelling approaches to the problem, namely Smolensky's tensor product representations and the Neural Blackboard Architecture (NBA).

1.1 Approaching the binding problem in language neuroscience

The binding problem

We want to understand how the brain computes and represents symbolic structures, such like those encountered in language. The existence of parts in structures like morphemes, words and phrases has been established through decades of linguistic analysis and psycholinguistic experiments. Nonetheless the neural implementation of the operations that support the extreme combinatorial nature of language remains unsettled. Some basic *composition* operations that allow the stable internal representation of sensory objects in the sensory cortex, like hierarchical pattern recognition, receptive fields, pooling and normalization, have started to be understood[120]. But models of the *binding* operations required for construction of complex symbolic structures on which precise manipulation of its components is a requisite, lack empirical testing and are still unable to predict neuroimaging signals.

The term *binding* was introduced into the neuro-scientific community by von der Malsburg[196] during the first explorations of neural phase synchronization. At this first stages of the study of *binding*, the term was really being used to study "feature binding", which just consists on association of concepts to form an object internal representation that will not have its properties confused with another object. An example would be to not confuse the colors of a "blue square" and a "red circle" presented together on a screen. Binding was also motivated by the empirical discovery of the distributed and segmented encoding of features along the cortex. For example color and shape, in the case of vision, are robustly integrated during perception, but can be independently impaired by brain damage, which implies that the two features

are represented independently in the cortex, even though we perceived them in unity.

If we consider the *binding* problem in generality, as presented by Feldman[65], it has several sub-problems from which "feature binding" is one of them. The current work is motivated instead by the "variable binding" sub-problem. Feldman[65] presents "variable binding" as an abstract high level cognitive faculty, mainly required by symbolic thought. As explained by Marcus *et al.*[120], it consists on creating a transitory link between two pieces of information: a variable (like Z in a equation, or a placeholder like *noun* in a phrase) and an arbitrary instantiation of that variable (like a number to replace Z in the equation, or a word that corresponds to the *noun* placeholder). It goes beyond the extensively studied sensory, attention and short-term memory phenomena of "feature binding", that only require appending features to a bag or set, to avoid confusion with other simultaneous representations.

The need for "variable binding" is to run logical inference on data structures that encode relationships between their items. For example the sentence "*Mary owns a book*" allows to establish a relation of the type *own(Mary, book)* that implies *owner(book, Mary)*, such that we can later ask the question "*Who owns this book?*", which would not be answerable under a simpler "feature binding" mechanism that would just confuse the three words in a bag as just belonging to the same group. To implement this in language, most linguistic theories propose that there are types of words, named grammatical categories, like 'noun' and 'verb', that are instantiated during sentence comprehension to be combined under a finite set of constraints. These instantiated word types would point to each other to form a graph data structure, a tree, on which query and join operations can be performed, and they would also point to their corresponding specific words. Then solving "variable binding" in language, requires a biologically feasible implementation of a pointer mechanism that can link instantiated grammatical categories and their corresponding words. For the rest of this work, whenever we use the term *binding* for simplicity, we will really be referring to the more specific "variable binding" sub-problem.

Additional challenges for the neural implementation of language processing

In "Foundations of Language", Jackendoff presents four important challenges that any proposal for the neural implementation of language processing must face[98], from which "variable binding" is only one of them. These challenges are the massiveness of binding, the problem of 2, the problem of variables ("variable binding") and the short and long term encoding problem.

The massiveness of binding is related to the combinatorial explosion that is encountered in symbolic structures like language, suggesting the impossibility to store in advance all combinations in memory. The problem of 2 is related to the representation of the same component, for example the same word, in the

same structure but with a different purpose or meaning, for example to denote two different objects. A concrete example would be the word "ball" in "the blue ball and the red ball". The problem of variables is to propose a mechanism to manipulate a symbolic structure to extract partial information from it, for example to ask "where did the children go?" and extract "the park" from the sentence "the children went to the park". The short and long term memory encoding problem is related to the fact that the brain has to be able to represent in short term memory transitory new formed structures to perform certain cognitive operations, as well as structures that will be stored and retrieved from long term memory. It is necessary to explain how both mechanisms operate together to completely account for the encoding of symbolic structures.

The basic properties of any model considered must at least be able to answer Jackendoff's challenges, besides providing the neural mechanism to instantiate symbolic representations and perform binding.

Summary of models identified to approach the binding problem

We identified two powerful but very different modelling approaches to the problem. The first is in the context of the tradition of Vectorial Symbolic Architectures (VSA) that bring precise mathematical modelling to the operations required to represent structures in the neural units of artificial neural networks and manipulate them. This is Smolensky's formalism with *tensor product representations* (TPR)[172], which he demonstrates can encompass most of the previous work in VSA, like Synchronous Firing[170], Holographic Reduced Representations[158] and Recursive Auto-Associative Memories[35].

The second, is the *Neural Blackboard Architecture* (NBA) developed by Marc De Kamps and Van der Velde[187], that importantly differentiates itself by proposing an implementation of binding by process in circuits formed by neural assemblies of spiking neural networks. Instead of solving binding by assuming precise and particular algebraic operations on vectors, the NBA proposes the establishment of transient connectivity changes in a circuit structure of neural assemblies. The potential flow of neural activity allowed by working memory mechanisms after a binding process takes place, implicitly represents symbolic structures.

Both modelling approaches considered in this work, namely Smolensky's tensor framework and the Neural Blackboard Architecture, satisfy Jackendoff's challenges[98].

1.2 Smolensky's tensor product representations

The integrated connectionist/symbolic cognitive architecture (ICS)

In the Harmonic Mind[172], Smolensky presents an integrationist view of the current theoretical approaches to model cognition. On one hand, the brain architecture seems to be best represented by a purely connectionist approach, in which interconnected neural units parallely process vectorial representations. On the other hand, symbolic architectures and computation has been behind the most successful models to explain the mind and its related behaviors[160; 84; 170; 158; 35]. These two different approaches have been put at odds by the *eliminativists*, that claim we do not need anything besides purely connectionist models to account for cognition. On the other hand the *implementationalists* claim we only need symbolic computation to develop cognitive theories. Smolensky argues instead for what he calls a *split-level* architecture, in which the highest symbolic computational provides functionally relevant structure, while the lowest connectionist computational level provides physically relevant structures.

Similar to a previous proposal of Marr[122] called the Purely Symbolic Architecture (PSA), Smolensky provides a framework on which, with tensor algebra in his case, the gap between the connectionist and symbolic levels is filled to explain all aspects of symbolic thought in cognition. This is accomplished by establishing an equivalence or isomorphism between the constituents in symbolic and vectorial representations. Also a correspondence is established between tensor algebraic operations and algorithms implementable in feed-forward and symmetric recursive neural networks. This isomorphism is then codified in what Smolensky refers to as *tensor product representations*.

Representations Principle of ICS and implementation of basic tensor product representations

The main assumption of the representation principle in ICS is that cognitive representations are implemented by widely distributed neural activity patterns (activation vectors), which have a global structure that can be described with the discrete data structures of symbolic cognitive theory. Three basic structural operations are proposed to act on the symbols or constituents of symbolic structures: combination by superposition, variable binding by tensor products and embeddings with recursively defined role vectors.

Combinations by superposition mean that parts of a structure are represented by vectors with the same dimension, that are then simply added together to create the complete structure vector, as illustrated for the phonemes of the word "cat" in Figure 1.1.

This addition operation raises the question of how complete information about individual components can be extracted from the final vectorial structure. In particular there is an issue to determine order of the constituents, because

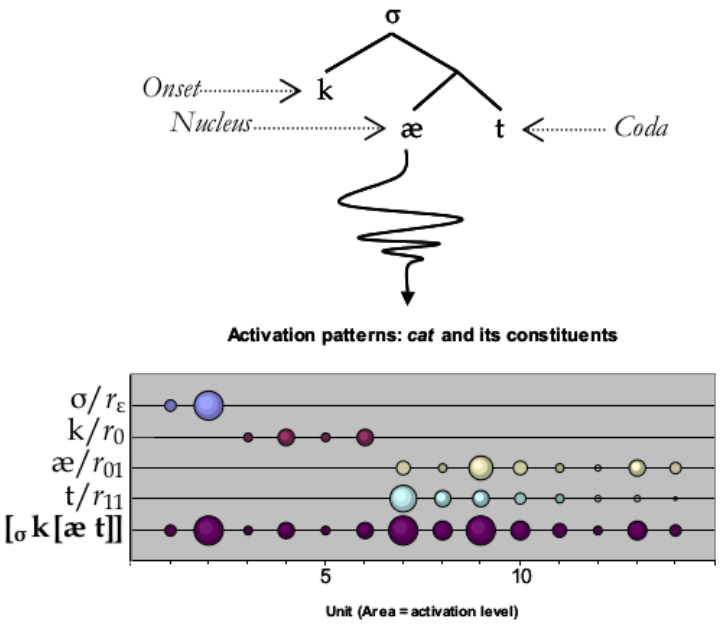


Figure 1.1: **Superposition illustration from Smolensky's Harmonic Mind:** This plot was taken from Smolensky's Harmonic Mind. It illustrates the neural activation vectors corresponding to the bindings of each of the phonemes of the word "cat", such that their sum would constitute the activation vector of the word according to the superposition principle. Phonemes are considered as *Fillers* and node positions in the structure tree as *Roles*

addition is a commutative operation. To address this issue Smolenky proposes that each constituent is formed by the binding, through a tensor product, of a symbol or content vector, called a *Filler*, with a slot of the complete symbolic structure called a *Role*.

The idea of *Role* vectors is similar to the notion of "frame" introduced by Minsky in 1975[135], which corresponds to the assignment of a fixed set of atomic elements to a fixed set of atomic roles. The nature of the *Role* vectors could be based on positional roles that denote absolute coordinates of a graph structure, like a vector representing the second node of the left branch in a tree. Alternatively they could be based on contextual roles, such that properties are bound together, like if we had the tensor product of an Adjective and a Noun to denote that the Adjective modifies the specific Noun. How we define the roles that will be part of the binding of a symbol is an open question. Currently positional roles are considered as a plausible explanation for the tree node positions of syntactic trees, while contextual roles are considered plausible to bind semantic concepts to relevant semantic contexts.

By assuming linear independence between the *Filler* vectors and between the *Role* vectors, it is possible to secure perfect recovery of a *Filler* vector by computing the inner product of the corresponding *Role* vector with the complete structure vector. It is also possible then to recover *Role* vectors by the inner product of their bound *Filler* vectors. Nonetheless if the same *Filler* is bound to more than one Role, like the word "star" in the sentence "The big star above the small star", the linear combination of all the respective *Roles* would be retrieved instead of a specific one.

Enforcing linear independence importantly restricts the amount of neural units necessary to be greater than the number of concepts encoded and not enforcing it would create intrusion, where the extracted *Filler* vector will also contain a linear combination of all other *Filler* vectors. Nonetheless there is a graceful degradation of the encoded representations with the degree of dependency of the *Role* or *Filler* vectors, that degrades as the square root of N for the N dimensional space given by N neural units. The expected intrusion (EI) has the form given in Equation 1.1. This graceful degradation also implies a graceful saturation of a connectionist network of fixed size with N neural units, such that the exact most conservative estimate of the expected total magnitude of intrusions for m bindings also grows as the square root of N.

$$EI = \sqrt{\frac{2}{\pi(N-1)}} \tag{1.1}$$

The mathematical form of a tensor product representation is provided in Equation 4.1. In Figure 1.2 we illustrate the tensor product of a *Filler* and a *Role* vector, which operates in a similar way to an outer product, multiplying each item of the first vector by each item of the second vector to determine the value of the neural units.

$$Structure = Filler_1 \otimes Role_1 + ... + Filler_n \otimes Role_n \tag{1.2}$$

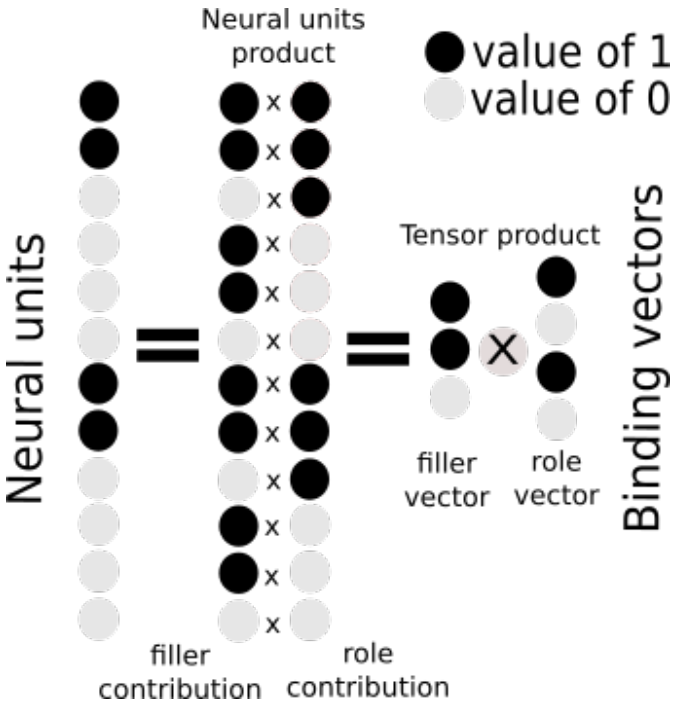


Figure 1.2: **Tensor product illustration:** The tensor product operates like the outer product of a *Role* and a *Filler* vector, of dimensions 4 and 3 respectively in the figure. Then each neural unit in the resulting binding neural activation vector, of dimension 12 in this case, encodes the multiplication of one component of the *Role* by one component of the *Filler*. The neural activation vectors of multiple bindings would be summed according to the superposition principle.

another important property of *Role* vectors is that they permit the definition of recursive embeddings. Hierarchical tree structures, as those proposed by

phrase grammars in language, require definition of roles at each level of the proposed trees and need to have the flexibility to implement as many levels as the faculty of language allows. Nonetheless there are several ways to implement such hierarchies from which Smolensky emphasize two. First the possibility to have local representations, with dedicated neural units, for each level of the tree. Second to have completely distributed representations that use all neural units for all levels, by binding tree level *Role* vectors to their corresponding upper level nodes in the tree hierarchy. In the case of asymmetric branches that would create a dimensionality difference in the *Roles* outer products, a dummy *Role* vector is introduced to rebalance the tree branches.

Local, semilocal and distributed representations

An important property of the ICS tensor product representations is that they have the flexibility to accommodate any degree of locality, which means that they can be made local, semilocal or completely distributed. The locality of a representation consist on the amount of neural units that are employed by the different *Filler* and *Role* vectors. Representations that correspond to a one-to-one mapping between possible elements represented and neural unit sets are purely local representation. If only the *Role* vectors have a local structure, then these would be role register or semilocal representations, for which an example would be roles modelling the position of an image with respect to the eyes, since there are inverse hemispheric projections in primary visual areas of the two eyes. Finally in fully distributed representations all neural units can be recruited for any representation.

There are three important examples in the previous literature of fully distributed representations, supporting the idea of Parallel Distributed Processing (PDP): the coarse coding representations of Hinton McClelland and Rumelhart[169], that focus on the many-to-many relation between visual positions and the activation of receptive fields; the conjunctive coding of McClelland and Kawamoto[128] that consist on three-way conjunctions of the learned features of nouns, verbs and semantic roles; and the wickelfeatures of Rumelhart and McClelland[168] that employed the 1-neighbour context decomposition to learn the binding of phonetic segments as *Fillers* to phonetic contexts as *Roles* to represent the past tense of english verbs.

It is important to understand which is the degree of locality of representations in a cognitive domain, because local and distributed networks differ in several properties. In the case of linear networks there is a transformation from any local representation to its distributed version and vice-versa, but this is not the case with non linear activation functions like those describing saturation and adaptation phenomena in neurons. Neural damage would have different effects depending on network locality since distributed representations are more resilient to local damages. Learning of

distributed patterns by networks could be more challenging and take more time, due to the interference of synergy of the concepts representations. There is better generalization of representation patterns in the case of distributed representations due to the similarity that can be established with unseen patterns, while in a local network representations must be orthogonal. Finally there is an important difference in the representational capacity of the network, since "N" neural units can only support "N" local representations, but a distributed network can maintain a higher number of representations for which exactness decrease gracefully.

Generalization of tensor product representations to accomodate previous vectorial symbol architectures (VSA)

One of the most powerful features of Smolensky's tensor product representations is that he can encompass most of the previous work in vectorial symbol architectures (VSA), like Synchronous Firing[170], Holographic Reduced Representations[158] and Recursive Auto-Associative Memories[35].

In Chapter 7 of the Harmonic Mind, Smolensky performs an in-depth analysis of the typology of previous vectorial symbol architectures (VSA) in the literature to show how they can be accommodated by tensor product representations. Some models, like the parietal cortex model of Pouget and Sejnowski[160] and the propositional information models of Halford, Wilson and Phillips[84], are simply equivalent to tensor product representations. Other important models, including Synchronous Firing[170], Holographic Reduced Representations[158] and Recursive Auto-Associative Memories[35], can be considered as tensor product representations if we generalize them by inclusion of postprocessing operations from tensor algebra.

The Synchronous Firing[170] model became important for its biological plausibility and the efficiency of employing time as an additional neural resource. It is also the simplest model to accommodate, since it does not require additional tensor algebra operations, but only reconsidering conceptually the neural resources and the nature of *Role* vectors. Using time as a neural resource simply requires that we define time slot *Role* vectors alongside semantic Role and Filler vectors. Shastri *et al*[170] proposes to implicitly bind a semantic role like "giver" to a semantic filler like "John", by explicitly binding both of them to a common *formal role* representing a time slot, which differs from previous considerations of contextual/semantic roles formulated to bind directly "giver" to "John". The roles distinction is portrayed in equations 1.3 and 1.4, that correspond to contextual and formal role considerations respectively. Formalizing this model with tensor product representations facilitates its comparison to other models and makes its extension from local to completely distributed representation almost trivial.

$$giver \otimes John, \quad \text{contextual/semantic role} \tag{1.3}$$

$$giver \otimes timeslot_1 + John \otimes timeslot_1, \quad \text{formal role} \tag{1.4}$$

In the case of Holographic Reduced Representations[158], developed to model human memory, they are of interest because they predict empirical results on how people relate structured elements. In this model *Filler*/*Role* bindings are achieved by employing a vector operation called circular convolution instead of a tensor product. For *Filler* and *Role* vectors of dimension n, this operation is attractive because the dimensionality of the output vector remains as n, while a traditional tensor product would produce an output vector with dimensionality n^2. Since the requirements of tensor products grow exponentially with the depth of trees in hierarchical structures, circular convolution is a more economical operation in terms of neural resources, at the cost of renouncing to exact or general-purpose representations to have instead inexact or special-purpose representations. To accommodate this model and others based on vector reduction operations, Smolensky introduces the tensor contraction linear operator from tensor calculus, to be applied to the final symbolic representations, and proves that circular convolution is just a particular case of tensor contractions.

In the case of the autoencoder model of Recursive Auto-Associative Memories[35] (RAAM), it is of interest because of its capacity to learn which *Role* vectors allows the relevant structures received as input to be encoded, while displaying in some cases the same fully parallel processing implementable with standard tensor product representations. Smolenky demonstrates that the encoded representations in the middle layer of the RAAM model can be reproduced by tensor product representations by applying a squashing (sigmoidal) function element-by-element to a contraction of the superposition of the bindings performed with the RAAM input vectors.

We display the extension of the basic tensor products of Equation 1.5, with the contraction operator in Equation 1.6, followed by the element-by-element application of a function in Equation 1.7. Then the generalized tensor product is the element-by-element application of some function to the contraction over some pair of indices of the (superposition) addition of the tensor products representing the bindings of *Filler* and *Role* vectors. The basic tensor product representations are then just the specific case where the function is the identity and the contraction is the trivial contraction that do not perform a dimensionality reduction. Generalizing tensor product representations to allow post-processing by contraction and/or squashing allows to subsume under one formalism all alternatives in the literature, while keeping the principles of binding by tensor product and superposition of symbolic representations intact, since the generalization only add post-processing steps.

$$\sum_i Filler_i \otimes Role_i, \qquad \text{Basic Tensor Products} \tag{1.5}$$

$$C[\sum_i Filler_i \otimes Role_i], \qquad \text{with Contraction} \tag{1.6}$$

$$F[C[\sum_i Filler_i \otimes Role_i]], \qquad \text{with element-by-element Function} \tag{1.7}$$

How Jackendoff's problems are answered by tensor product representations in ICS

First, "The massiveness of binding" is addressed by the binding operation defined with tensor products alongside the graceful saturation of inexact representations. Second, "The problem of variables" is handled by the linear independence assumption between *Filler* vectors and between *Role* vectors that permits unbinding with the inner product, with a graceful degradation of information when the linear independence assumption is violated. Third, "The Problem of 2" is managed by binding the same *Filler* vector to different *Role* vectors, nonetheless if we were interested in querying the *Role* of a repeated *Filler* we would have problems, since we would recover the linear combination of all the corresponding *Role* vectors. Finally, learning the *Filler* and *Role* vectors in neural networks is analogous to a long term memory mechanism, while implementing the tensor product operations would permit instantiating in short term memory new symbolic structures from the binding of *Filler* and *Role* vectors. Moreover the generalization of tensor products to account for memory related models like Holographic Reduced Representations and RAAM, demonstrates its flexibility to model diverse memory related mechanisms.

1.3 The Neural Blackboard Architecture (NBA)

Neural models of language

To understand how the cognitive faculty of language operates, we need to take into account, not only the underlying supporting structures, but also their dynamics. This means that we have to consider simultaneously the grammars given by linguistic theory and a temporal component to give birth to computational mechanisms, like automaton models, capable of explaining behavior[83]. To extend this into neuroscience we have to go even further and also provide reasonable implementation models, corresponding to the biological components of the brain. This implementation is necessary to be able to go beyond behavioral measurements and ultimately test computational hypotheses directly against the currently available spatio-temporal neural measurements.

A good example of success in this direction is the computational theory of visual receptive fields[113] which has made impressively accurate predictions

about the shape of the biological visual fields in the retina. Knowledge of these basic units of visual perception has even recently allowed to correlate the mechanisms behind deep convolutional neural networks to visual pathways[80; 58] and has influenced our understanding of higher-level visual phenomena such as visual illusions[57]. Although expecting at the moment something similar in the case of language might sound overambitious, we must note that basic phonetic features have already been decoded in the Superior Temporal Gyrus from electrocorticography (ECoG)[133].

Numerous Artificial Neural Networks (ANNs) have been implemented, motivated by biological principles in the brain[18; 39; 134; 200; 173], to model particular aspects of brain language function or to reproduce behavior in specific language tasks. Nonetheless they lack dynamic biological considerations necessary to match their output with neuroimaging measurements, and except for Vector Symbol Architectures (VSA)[172], they are difficult to integrate into a general framework for the implementation of complete language functions.

More relevant to our work are previous efforts to model language function with more biologically plausible Spiking Neural Networks (SNNs)[94; 166; 18; 121; 56; 162; 161; 72; 123], that would eventually allow to establish a mechanistic link between neural measurements and computational linguistic hypothesis. Contrary to the VSA and the Neural Blackboard Architecture (NBA)[187], these do not follow a general theoretical framework, to address all the neural challenges of a complete language function implementation, that can also provide a mechanistic explanation for the most basic computational components and behaviors.

In most models, biological details necessary to match high temporal resolution in-vivo neural patterns of language processes have been kept out of scope. This has been a reasonable strategy considering the computational cost of building circuits with detailed neural models based on simulations of each neuron. Nonetheless recent developments like population density techniques[47] now permit to simulate state-of-the-art temporally detailed dynamics of circuits of neural populations.

In this work we will go beyond previous SNN simulations that were limited in scope to describe language function and temporal resolution of the neural dynamics. We will implement a temporally detailed spiking neural network circuit inspired by the Neural Blackboard Architecture[187]. The circuit implementation will be capable of realizing the binding operation for any level of language processing and for any grammar theory and parsing scheme, but we will focus on its application to the syntactic structure of phrases.

Introduction to the Neural Blackboard Architecture

Van der Velde and De Kamps[190] argue in favour of a small world network model that, thanks to transient changes in its connectivity, allows the

formation of complex structures. Binding takes place in the Neural Blackboard Architecture by conditionally co-activating neural assemblies representing grounded concepts and instances of variable types, which is a process driven by a control mechanism. The co-activation of the neural assemblies activates a working memory mechanism that last for a short period of time, to permit future activation of one bound neural assembly by its pair.

In this framework, working memory acts as a control that reduces inhibition on paths of neural flow necessary to maintain the bindings established by the initial transient co-activation, such that pointers have been declared implicitly between the co-activated concepts. Data structures are implicitly encoded by the short lived reinforced paths of neural activity flow. Then query operations are possible by reactivating nodes - included in the query - that induce co-activation of answer nodes, thanks to the reinforced connectivity. This successive co-activation of neural assemblies referred as "binding by process", leads to a short-term lived graph that implicitly encodes the final data structure.

The level of abstraction of the NBA allows to apply it to several cognitive functions like motor control, attention and symbolic thought. In the case of syntactic parsing during language comprehension, one needs a grammar to specify the necessary variable type relations and some parsing scheme to determine the bindings' timing. The NBA provides a circuit with nodes that can be readily interpreted in terms of spiking neural populations. This can be conceptually linked to the notion of cell assemblies, whose existence and functional relevance, as computational units, is supported on substantial biological evidence[95].

Circuits of the architecture

A complete illustration of the blackboard architecture is provided in Figure 1.3. Nodes in Figures 1.3.A and 1.3.B represent neural cell assemblies that can be interpreted as linked spiking neural populations. There are several previous implementations of sub-circuits of the NBA with varying degrees of biological plausibility, the latest relying mostly on Wilson Cowan population dynamics[52]. Some of the previous simulations attempted to address diverse aspects of language processing, such as ambiguity[67] and learning control from syntactic stimuli[188]. Other simulations addressed circuit implementation issues like how to develop a connectivity matrix with randomly connected networks[189] and how to implement a central pattern generator sub-circuit for sequential activation [191].

We will focus on providing a summary of the Neural Blackboard Architecture operation from a perspective relevant to variable binding. For a deeper review of the NBA circuit and mechanisms we recommend reading a recent paper with a circuit design and examples that focus on sentence processing[48], as well as the original framework proposal introducing abstract

combinatorial structures[187].

A "Gating Circuit", illustrated in Figure 1.3.A, is the most basic component of the NBA, from which all other circuits are built. The main idea is that neural activity would flow from the assembly X to the assembly Y, but is blocked by the Gate Keeper (GK) assembly, which is also excited by assembly X. So to allow directional activity flow from X to Y, a Control (Ctl) assembly has to inhibit the GK assembly. Notice that it is trivial to extend the gating circuit for bidirectional control of activity flow as illustrated in Figure 1.3.B. Introducing bidirectional conditional control signals is what gives the NBA the possibility of implementing separately queries like 'what follows X?' or 'what follows Y?'.

Another basic mechanism of the NBA is a proposal for working memory (WM). Persistent neural activity in response to stimuli is considered to be the neural process underlying active (working) memory, and its implementation is hypothesized to be based on excitatory reverberation[199]. Based on this, the NBA considers a Delay Activity[45] mechanism as a biologically plausible implementation of WM. It consists on a neural assembly, that after being excited beyond a certain threshold, achieved by the coactivation of input populations, will maintain a constant amount of activation for a short period of time. By maintaining its activity, WM acts as a short lived bidirectional link between two assemblies. This process can be equated to the creation of an implicit pointer from one assembly to the other, such that future reactivation of one assembly can be driven from the other to perform query operations. The respective "Memory Circuit" is shown in Figure 1.3.B.

Two bidirectional "Gating Circuits" connected by a "Memory Circuit" form a "Compartment Circuit" capable of implementing variable binding and query operations. The key point of this circuit is that Main assemblies (MA), representing grounded concepts or instances of variables types, activate Sub assemblies (SA), if a control signal driven by another mechanism allows it. Then co-activation of SAs is what realizes a temporary binding of MAs by activating WM. So one "Compartment Circuit" models specifically the neural activity of a variable binding operation. It is operated by a mechanism that drives control signals simultaneously in multiple "Compartment Circuits" to instantiate binary tree like data structures on which query/unbinding operations can be performed later.

Finally, a "Connection Matrix", portrayed in Figure 1.3.C, allows the implementation of a complete "Blackboard". It contains variable type relations learned by the "Blackboard" as sets of mutually inhibitory "Compartment Circuits" that enable the selection of the "Compartment Circuits" requested by the control mechanism. We portray the "Blackboard" as a regular grid for illustrative purposes, although there is already a proof of concept implementation with randomly connected networks[189]. Also implementing a general syntactic control mechanism should be feasible, as suggested by the Feed-forward artificial neural networks employed in previous

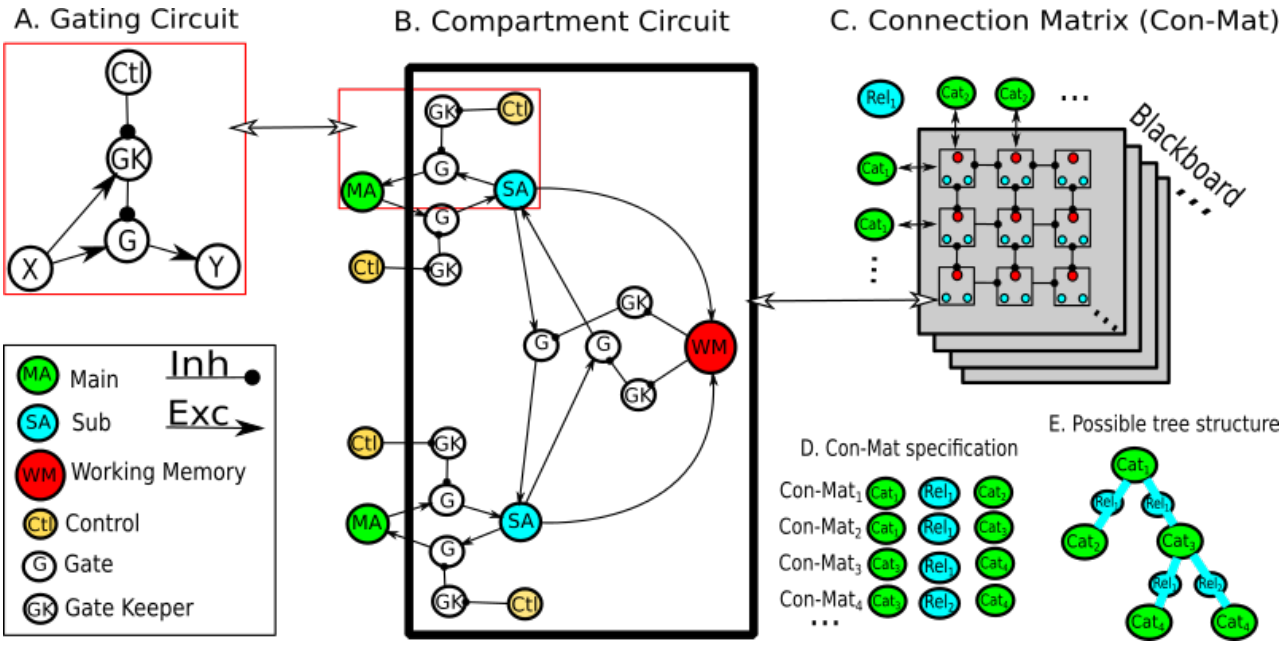


Figure 1.3: **The Neural Blackboard architecture: A.** Gating circuit that allows the implementation of conditional neural activity transfer between Neural assemblies X and Y through a gate assembly. The gate keeper assembly (GK) is activated by the X assembly and then inhibits the gate assembly (G). To let information flow through the gate assembly, a control assembly (Ctl) must therefore inhibit the gate keeper assembly. **B.** Architecture of a single compartment circuit of a connection matrix. Six gating circuits are arranged in a way that makes conditional bidirectional neural activity flow between two main assemblies possible. Control assemblies regulate the direction of information flow and allow the activation of sub assemblies. The two sub assemblies excite the working memory assembly which, once activated, encode the binding of the main assemblies and allow activation to flow between them if the controls allow it too. **C.** Each connection matrix contain n by m compartment circuits that encode the same relationship type between the same pair of assembly categories. There are m available assemblies for one category and n available assemblies for the complementary category and only one cell circuit can activate its working memory assembly to link two particular assemblies due to mutual row and column inhibition of cells in the connection matrix. The size of the connection matrix effectively represents memory limitations. A blackboard is composed of an arbitrary number of connection matrices that encode different relationship types for a pair of assembly categories. **D.** A blackboard is composed of multiple connection matrices, where each of them is defined by two node categories and a relationship type between them. **E.** Example of a possible tree structure that can be represented based on the specified connection matrices.

NBA simulations [188] and recent state of the art feed-forward network architectures that have shown top performance for diverse language parsing tasks [6]. Moreover a more recent proposed extension of the NBA, that imitates the motor circuit of the marine mollusk Tritonia diomedea, shows how to generate patterns for sequential activation control[191]. Simulating these higher level mechanisms is a task out of the scope of this work, since we focus specifically on reproducing the neural signatures of variable binding operations.

Instantiation of symbolic representations with the NBA

The level of abstraction of the NBA allows to apply it to several cognitive functions like motor control, attention and symbolic thought. In the case of syntactic parsing during language comprehension, one needs a grammar to specify the necessary variable type relations and some parsing scheme to determine the bindings' timing. In contrast to VSA, the NBA provides a circuit with nodes that can be readily interpreted in terms of spiking neural populations. This can be conceptually linked to the notion of cell assemblies, whose existence and functional relevance, as computational units, is supported on substantial biological evidence[95].

Applying the NBA to syntactic processing in language consists of two simple assumptions. First, equating the parsing mechanism to the control mechanism that coordinate binding events of words and word types and phrase types. Second, determining the number of compartment circuits necessary to instantiate a complete syntactic structure and the content of MA nodes from a grammar theory. The NBA has the flexibility to test any arbitrary parsing mechanism and an important variety of alternative theories of grammar based on binary trees. For example dependency grammars that assume multiple direct word bindings instead of the hierarchical phrase bindings modelled in this work have been employed in previous simulations[188].

To understand how a sentence is processed in the NBA, let us consider first the simplest case of binding two words, like "Sad student", belonging to grammatical categories instantiated in the MAs of one "Compartment Circuit", such that one MA is an "Adjective" corresponding to "sad" and the other one is a "Noun" corresponding to "student". The MAs activate with timings corresponding to word presentation, reflecting processing of the word grammatical category. Then an assumed parsing mechanism determines that a link operating on "Adjective" and "Noun" types is necessary in the blackboard, driving activity in several "Compartment Circuits" from which only one, that we consider as the recruited "Compartment Circuit", completes co-activation of SAs to drive WM and realize binding between the word types. To process a complete phrase this process is repeated by recruiting more "Compartment Circuits", realizing an implicit representation in the cortex of the whole phrase through the activation of the Working Memory neural assemblies.

How Jackendoff's problems are answered by the Neural Blackboard Architecture

First, "The massiveness of binding" is addressed by instantiation of variable types as assemblies that are bound to grounded concepts and other variable types instances, allowing the creation of combinatoric structures on demand. Second, "The problem of variables" is handled by the previously explained co-activation mechanism capable of creating pointers from grounded concepts to variable type instances. Third, "The Problem of 2" is managed by having multiple neural assemblies that instantiate the same variable type in the architecture but that can occupy different parts of the same data structure. Finally a working memory mechanism is provided, that allows transient short-term co-activations of concepts to be maintained without interfering with the possibility of storing related data structures in the long term in other parts of cortex with other mechanisms.

1.4 Summary and comparison of the modelling approaches

On one hand Smolensky proposes that the brain employs explicit active encodings, in neural units, of "unified" data structures produced by tensor products acting as binding operations on spatially stable, unique and linearly independent neural unit vectors. These data structures can be later queried with inner products acting as unbinding operations. The latter are resilient to squashing functions, like those proposed by Plate, that can importantly decrease the number of neural units necessary for the final representation as the tensors increase in dimensionality with more complex structures. Representations in this model can be completely distributed and nothing is clarified about the encoding of parallel representations in memory. Smolensky offers in great detail implementations of VSA with feedforward and symmetric recursive ANNs[172] and has recently shown how to extend the framework with an optimization scheme to instantiate input representational vectors[173]. Nonetheless, no important operational consideration is given to time, although it is possible to employ it as a tensor for vector encoding purposes, as is done for Synchronous Firing. This limits the neural dynamics predictions of the framework and its interpretation with SNNs.

On the other hand the Neural Blackboard Architecture proposes that the brain encodes complete symbolic structures implicitly, encoded by the activity of short term memory mechanisms. A circuit of neural assemblies on which neural activity flows conditioned by control and memory mechanisms allows both binding and query operations. Since the NBA explicitly defines the architecture and operation of the circuits, it is straightforward to implement them with SNNs. By representing the bound concepts as specific neural assemblies the NBA induces local representations and by allowing arbitrary selection of mutually inhibitory competing sub-circuits (Compartment circuits)

makes the representations themselves dynamic and spatially unstable.

From the description of these models we can appreciate that they approach the problem very differently, which motivates experimenting with both of them. They employ different practical neural implementations and simulations. Also they assume different properties of the internal neural representations of concepts. In Table 1.1 we present together all the different aspects of both modelling approaches. Smolenky's TPR relies mostly on spatial static considerations of artificial neural units with explicit completely distributed and spatially stable representations implemented through vectors, while the NBA relies on temporal dynamic considerations of biologically based spiking neural units with implicit semi-local and spatially unstable representations implemented through neural assemblies. Another difference between models is how they handle multiple parallel representations in memory. Smolensky do not propose any particular mechanism, although using the same neural units for this would work with the creation of memory slot roles. The NBA handles parallel representations in memory explicitly, by keeping separate neural assemblies assigned to each structure, but then its capacity is limited by the size of blackboard and the dynamics introduced by the mutual inhibition of compartment circuits in a connection matrix.

Aspect	Smolensky's TPR	NBA
About modelling:		
Neural simulation	Artificial NN	Spiking NN
Temporal dynamics	Not included	Included
Representation	Neural unit vectors	Neural assemblies
Parallel repr model	Memory slot roles?	Separate neural assemblies
Representation properties:		
Declaration	Explicit	Implicit
Spatial stability	Static (temporally stable)	Dynamic (temporally unstable)
Locality	Distributed or local	Local
Operation implementation:		
Composition of bindings	Superposition (addition)	Compartment recruitment
Binding	Tensor product	Working memory assembly activation
Unbinding	Inner product	Reactivation of bound neural assemblies

Table 1.1: **Modelling approach comparison:** We present all binding related aspects studied in this work about Smolensky's tensor product representations and the Neural Blackboard Architecture.

2 Methodological background

In this chapter we provide a quick summary of methodological details useful to better understand the superposition experiment analysis (BOLD-fMRI related methodology) and the implementation of the Neural Blackboard Architecture (neural simulation related methodology).

2.1 BOLD-fMRI

The BOLD-fMRI signal

The first studies of BOLD-fMRI, that showed how sensory stimulation modulated a blood oxygenation level dependent contrast date back to 1992[146]. BOLD-fMRI is one of the most common neuroimaging techniques, that captures non-invasively indirect measures of neural activity in a whole brain volume, with a high spatial resolution (1-3mm^3) and a low temporal resolution (1-3 seconds). This technique takes advantage of the fact that "ferrous iron on the heme of deoxyhemoglobin is paramagnetic, but diamagnetic in oxyhemoglobin"[36]. This means that a strong magnetic field can detect changes in the concentration of oxygen in the blood stream, which is modulated by neural activity. The shape of the modulation of oxygen concentration due to changes on neural activity is called the hemodynamic response function (HRF). Boynton *et al.*[22] showed that a double gamma basis function, applied with a linear regression model, could capture well the HRF. Although the HRF can take up to 30 s to completely develop, it was shown that the response of two stimuli add linearly if their presentation is separated by at least 2 seconds[30]. We show an example double gamma basis function[76] in Figure 2.1.

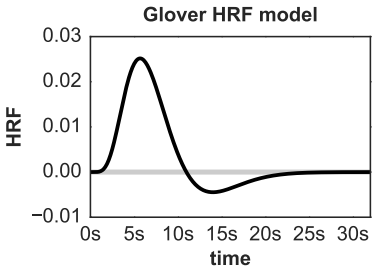


Figure 2.1: **Canonical Double gamma basis function (HRF)[76]:** The HRF first shows a quick increase from 1 to 5.2s, then "undershoots", drops below baseline, from 5.2 to 12.2s, and finally comes back to baseline from 12.2 to 30s.

BOLD-fMRI preprocessing

To acquire brain images, a subject is introduced in an MRI scanner and from the pulse sequences of an acquisition protocol, images, formed by "voxels" of a certain volume, for example 1.5mm^3, are reconstructed. The obtained datasets are preprocessed with a variety of pipelines, some of which have been extensively evaluated[177]. Common pipeline steps are: slice timing

correction, motion correction, spatial co-registration, spatial normalization and spatial smoothing. Since the brain volumes are acquired by slices in different time points, it is necessary to extrapolate to a common time point the measurements of all slices which leads to slice timing correction. Subject movements during the acquisition have to be taken into account to build a voxel time series that correctly represents spatial location, so motion correction is implemented. Functional images that contain the BOLD signal are commonly co-registered with a T1 anatomical scan of the subject to be able to extract voxels corresponding to anatomical structures like gray matter and allow normalization. Then anatomical images from subjects are projected into the space of a reference image, such that group level activations can be estimated by compensating an important portion of inter-subject variability[85]. Finally and optionally, the resulting images are smoothed with a Gaussian kernel to increase the local signal-to-noise ratio (SNR), due to spatial correlation of voxel activations. More details on different preprocessing steps can be consulted in Lindquist review[1].

[1] M. A. Lindquist. The statistical analysis of fmri data. *Statistical Science*, pages 439–464, 2008

Effects estimation in univariate analysis

After preprocessing, traditionally BOLD time series are analysed with a General Linear Model (GLM)[71]. This practice remained because it was demonstrated that BOLD responses to stimuli add approximately linearly if the stimuli presentation is separated for at least 2 seconds[30]. To fit the GLM, a design matrix is produced in which different conditions are modelled with different regressors. In each condition, the onsets and durations of the corresponding events are modelled as a stepwise constant (boxcar) signal, that is then convolved by an HRF like the one shown in Figure 2.1. In Figure 2.2 we show the construction of an example design matrix with four conditions. A GLM model, described by Equation 2.1 and illustrated in Figure 2.3, is applied separately to each voxel.

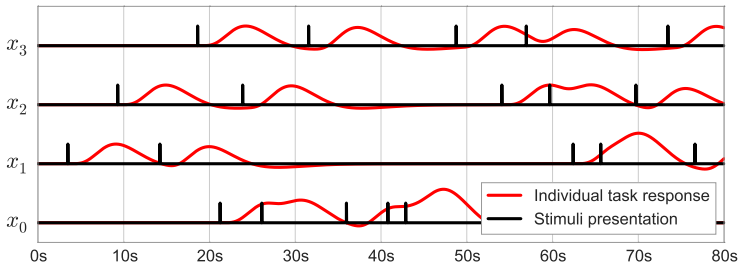


Figure 2.2: **Illustration of a design matrix:** Event onsets from four fMRI experimental conditions are convolved with the HRF to approximate the BOLD response.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

The design matrix corresponds to the $\mathbf{X}$ in the GLM estimation and the $\boldsymbol{\beta}$ (betas) corresponds to the estimated amplitude of the BOLD response of each condition in a voxel. The betas of all voxels considered together are called

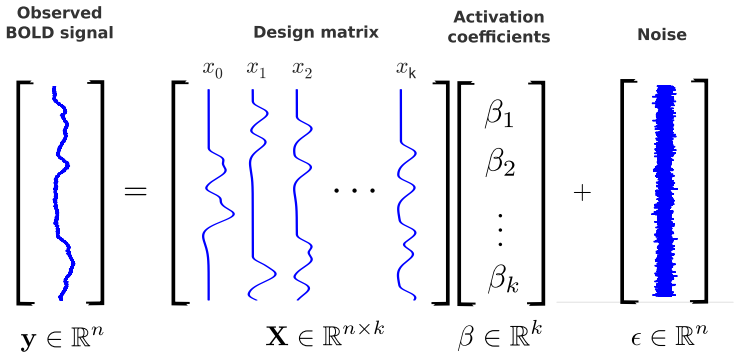


Figure 2.3: **GLM model:** The signal in each voxel is modelled as a linear combination of the time series given in the design matrix plus noise.

"beta maps", which are then employed to compare conditions across the brain. The comparison is done by testing if a particular linear combination of betas, called a contrast, is different from 0. Staying with the four conditions example, a contrast vector $\mathbf{c} \in \mathbb{R}^4$ would be defined as $\mathbf{c} = [+1, 0, -1, 0]$ to test in which voxels there is a significant positive difference between the first and third conditions[114]. From this contrast, a t, z or F statistic map, normally called statistical parametric maps, will be computed and then thresholded at some level of p-value significance to interpret the surviving spatial clusters of activations in the brain. We illustrate the computation of the p-value for a z-test in Figure 2.4.

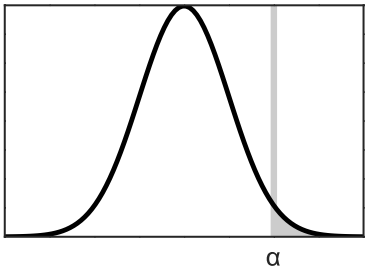


Figure 2.4: **Classical z-test:** In a z-test a gaussian distribution provides the reference for which we can estimate the accumulated probability of a particular value α, such that we can compute its p-value.

Alternatively to this way of computing p-values, it is also possible to employ non parametric approaches in which, under some theoretical constraints, we can estimate the empirical distribution of the contrast of permuted condition labels and observe the probability of the real labels on that distribution. More details on the statistical analysis of fMRI data can also be consulted in Lindquist review[2].

[2] M. A. Lindquist. The statistical analysis of fmri data. *Statistical Science*, pages 439–464, 2008

Decoding of activation maps

The GLM mass univariate fMRI analysis is a forward model. Forward models, also called encoding models, model brain responses following a stimulus. Inverse models, also called decoding models, go in the opposite direction, they predict stimuli from brain images. A scheme of these concepts is shown in Figure 2.5.

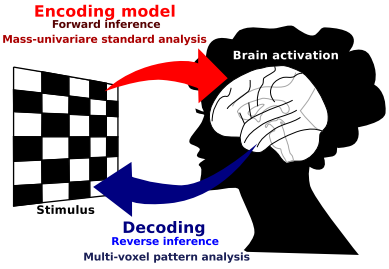


Figure 2.5: **Encoding and decoding scheme:** We provide a scheme showing how decoding and encoding models relate to brain activations and stimulus.

With decoding models we explore the possibility that the spatial neural activity patterns, reflected in the amplitude of estimated BOLD responses in voxels, carry distributed information beyond the overall activity of individual voxels. This type of multivariate approach, has been very influential in the analysis of fMRI data [192]. It was named initially as "multivoxel pattern analysis" [143] and later as “multi-variate pattern analyses” [88]. It has been shown that the relationship between stimuli and beta maps can be captured appropriately by linear models, considering that non-linear models tend to have a similar performance as the linear ones[136]. Moreover employing

linear models can give better insights into how each individual feature (voxel) contribute to the final prediction[87].

A commonly used decoding model in fMRI is linear support vector machines (SVMs), that from a set of "support vector" points draw an hyperplane to maximize the "margin" distance between the hyperplane and the nearest data points from two classes of points. The SVM expresses the hyperplane as the coordinates of a vector orthogonal to the hyperplane, such that the absolute magnitude of each coordinate or "coefficient" related to a feature (voxel) can indicate how important the feature is for the separation of classes. In Figure 2.6 we illustrate a particular mathematical formulation of the SVM called NuSVM, in which the number of support vectors selected by the algorithm is controlled by the "*nu*" parameter in the model.

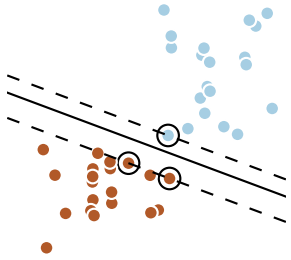
A decoder is evaluated by its capacity to predict correctly a stimulus or condition from a given set of voxel activations (from beta maps). In the case of classification of balanced conditions, the typical evaluation metric is accuracy, computed from the number of correctly classified samples over all samples. Accuracy of a trained model should be evaluated on left out unseen data to secure we correctly capture true generalization performance of the model. This is necessary due to the risk of overfitting or over-learning the particularities of the samples selected to train the model instead of the general trend.

A common procedure to select the best model, to optimize generalization accuracy, is to perform K-fold cross-validation[7]. This procedure consists on dividing the dataset in "K" data segments, such that iteratively a segment will be left out as unseen data to evaluate the accuracy of a model, which was trained on the rest of the data. The selection of model parameters (hyperparameters), like *nu* in a NuSVC classifier, should also be cross-validated to avoid the introduction of a positive bias in the generalization accuracy of the model, with a nested cross-validation scheme[33].

After we have estimated the generalization accuracy of the model, it is desirable to be able to assess its significance, in particular considering the possibility of finding accuracy scores slightly better than chance. This verification is important due to the possible biases and fluctuations that can be introduced in the accuracy scores by noise in fMRI data and the small sample sizes normally available. A typical procedure to achieve this is to randomly exchange condition labels on the data points, to obtain permuted labels, and train a new model on the permuted labels. The empirical distribution obtained from the accuracy on the "N" permuted label sets allows to compute a p-value, by assessing how extreme is the accuracy of the model trained on the real labels.

Another problem we face with fMRI data is that of feature (voxel) selection. Considering the curse of dimensionality, which explains that we need an amount of samples that grow exponentially with the number of features considered, and the small sample sizes commonly available, we are encouraged to diminish the amount of features (voxels) considered by a model as much as

a) small *nu*



b) large *nu*

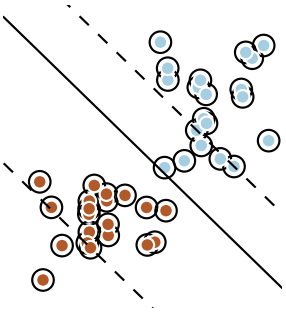


Figure 2.6: **NuSVM example:** We present two classes as blue and brown points. The NuSVM learns a hyperplane, a line in this two dimensional case, to discriminate the two classes. The *nu* parameter permits to regularize the algorithm by controlling the number of support vectors selected. For small value of *nu*, few observations are selected as support vectors, while for a large value of *nu*, all observations are selected.

possible.

There are two ways to deal with this problem. The first is by "filtering", which consists on selecting features based on some procedure unrelated to the accuracy score of the trained model. A typical way of doing this in fMRI is by ranking voxels according to scores obtained from univariate tests, like an F test to detect any difference between all conditions in a voxel. The second is by "wrapping", for which we consider subsets of features as an hyperparameter of the model and then decide on the best subset in the same way that we would select other hyperparameters of the model, by nested cross-validation.

2.2 Neural simulation

We assume that the Neural Blackboard Architecture (NBA) lives in the cortex, and seek a good compromise between realistic modelling of the cortical dynamics and the tractability of the simulation. State-of-the-art simulations of larger cortical structures are based on point model neurons that allow the inclusion of biological details such as synaptic dynamics and adaptation, but are restricted to about the size of a cortical column [159]. For larger scale networks, such as ours, a population-based approach is currently the only feasible approach. The two choices are: rate based models or population density techniques (PDTs). In rate based models, the population is described by a single variable, usually related to the population firing rate or average membrane potential of neurons in the population. A prominent example is the Wilson-Cowan equation [201], which describes the dynamics of the population activity as a first order linear differential equation driven by inputs. Another example is the Jansen-Rit model [100], which is primarily motivated by phenomenological considerations. In both examples, the relationship with the underlying neural state is unclear. We have opted for PDTs, also a population based approach, but one where the relationship with the dynamics of a group of spiking point model neurons can be made rigorous. Although they are computationally more expensive than rate based models, they are easier to manage than a full-blown model using spiking neurons, which would need hundreds of thousands of neurons at the scale of the cortical network considered here. We will briefly set out the assumptions that we use in modelling populations and describe the numerical methods involved.

Consider a leaky-integrate-and-fire (LIF) neuron, which is characterized by a single state variable: the membrane potential. If the neuron has a potential different from its equilibrium potential, or when it experiences an external drive, for example generated by a synaptic current, the potential evolves according to:

$$\tau \frac{dV}{dt} = -(V - V_{rev}) + I(t). \tag{2.2}$$

Here V is the membrane potential in V, τ the membrane time constant in s, V_{rev} the reversal potential and $I(t)$ and external current, which may comprise

contributions from other neurons in the form of spikes, and therefore may be stochastic. If the membrane is driven far above the equilibrium potential, at a potential V_{th}, the threshold, the neuron spikes. We assume it will be inactive for an absolute refractive period τ_{ref} and then finds itself reset to the equilibrium potential after that. This scenario is easy to simulate: using a simulator like NEST [74], or BRIAN [176], one can create populations of LIF neurons. In the simplest case a population is driven by synthetically generated input spike trains, where the spike train events are created by a random generators. The default assumption is that inter-spike intervals are Poisson distributed, although this can be extended to non-Markov processes [108]. It is clear that $I(t)$ in Eq. 2.2 now should be considered as a stochastic variable and that the threshold crossings of LIF neurons themselves are stochastic events as a consequence. Fig. 2.7 A demonstrates a simple scenario: a population of 10000 LIF neurons, driven by a stochastic input - Poisson generated spike trains, where each LIF neuron experiences about 800 input spikes per second. The simulation shows a spike raster of the population response: first nothing: although each LIF neuron receives input spikes and as a consequence has its membrane potential driven up, none of the neurons have reached threshold; then a spike volley: most neurons hit threshold at approximately the same time; followed by a period of relative silence: only interrupted by a few stragglers; at last a gradually achieved final neural state of asynchronous random firing. More complex networks can be formed by feeding the output spikes of one population into other populations.

This is a fascinating but unwieldy process and statistical methods have been used to describe it at the population level [175; 105; 150]. A population is described by a density function, which expresses how the population is distributed over state space. For LIF neurons this is a function $\rho(V)$, where $\rho(V)dV$ is the fraction of neurons with their membrane potential in interval $[V, V + dV)$ (when we integrate the density function over a certain state interval, we will refer to the result as the amount of *mass* in that interval). The initial distribution of the neurons in the population must be chosen, but the evolution of the density is tractable. It is clear that neurons move through state space due to the deterministic neural dynamics, Eq 2.2 for LIF neurons, and also go transitions due to the input spikes. The collective contribution of the stochastic process to the evolution of the density profile can be modelled using a Poisson master equation [73]; the contribution of the deterministic dynamics can be modelled using an advection equation (see [150] for a lucid explanation).

As a consequence, the process of simulating thousands of neurons is now replaced by modelling the evolution of a density which is given by a single equation:

$$\frac{\partial \rho}{\partial t} - \frac{1}{\tau}\frac{\partial}{\partial v}(\rho v) = \int dh p(h) \nu(\rho(v-h) - \rho(v)), \tag{2.3}$$

Here $p(h)$ is the distribution of synaptic efficacies, ν the frequency of the incoming spike trains, ρ the density function, t the time since start of simulation and v the membrane potential. Mass that is being pushed across threshold corresponds to neurons spiking; consequently the firing rate of the population can be calculated directly from the mass flux across threshold.

Efficient and stable simulation methods are available [145; 44; 46; 96], and remarkably, the process of solving Eq. 2.3 is computationally less expensive for LIF neurons than the direct simulation using NEST [145]. The process of keeping track of a single density function, and the communication between populations using firing rates rather than individual spikes, frees the modeller from keeping track of thousands of spikes per second and leads to simpler simulations. Figure 2.7 shows the very close correspondence between direct simulations of LIF spiking neurons and population density results. It shows, first, that the simulation results indeed are very close to that of the spiking simulation, and second, that Wilson-Cowan dynamics must be tuned in a way that PDTs do not: the correct steady state activation must be provided to the Wilson-Cowan dynamics in the form of a sigmoid, while in PDTs the correct steady state firing rate is calculated from first principles - input firing rate, synaptic efficacies and neural parameters - without any need for tuning.

The population density formalism can be extended to higher dimensional models. For example, the adaptive-exponential-integrate-and-fire neuron (AdEx) [27] is a two dimensional model that has the membrane potential and an adaptivity parameter as a variable. Consequently, the state space is two dimensional. The motivation behind this model is that first, it includes adaptation, and second that it is the effective approximation of the complex conductance-based processes that take place in a real neuron. The equations of the model are:

We consider the AdEx model as presented by Brette and Gerstner [27], which describes individual neurons by the following equations:

$$\begin{aligned} C_m \frac{dV}{dt} &= -g_l(V - E_l) + g_l e^{\frac{(V - V_T)}{\Delta_T}} \\ \tau_w \frac{dw}{dt} &= a(V - E_l) - w \end{aligned} \tag{2.4}$$

Where C_m is the membrane capacitance, g_l the leak conductance, E_l the leak potential (equivalent to the reversal potential for the LIF), V_T a threshold potential, Δ_T a shape parameter for the spike, τ_w the adaptation time constant, a the subthreshold adaptation parameter, V the membrane potential and w the adaptation parameter. Upon a spike, the neuron is undergoes a transition in w: $w \rightarrow w + b$, where b is the spike adaptation parameter. We use the parameters given by Brette and Gerstner (2005).

We illustrate the dynamics of the neuron in Fig. 2.8. The direction of the dynamics is shown by arrows, the speed of the dynamics by the size of the cells: big cells implies fast dynamics as the cells represent equidistant time steps. This shows that at $w = 0$ dynamics are leaky, i.e. towards the equilibrium,

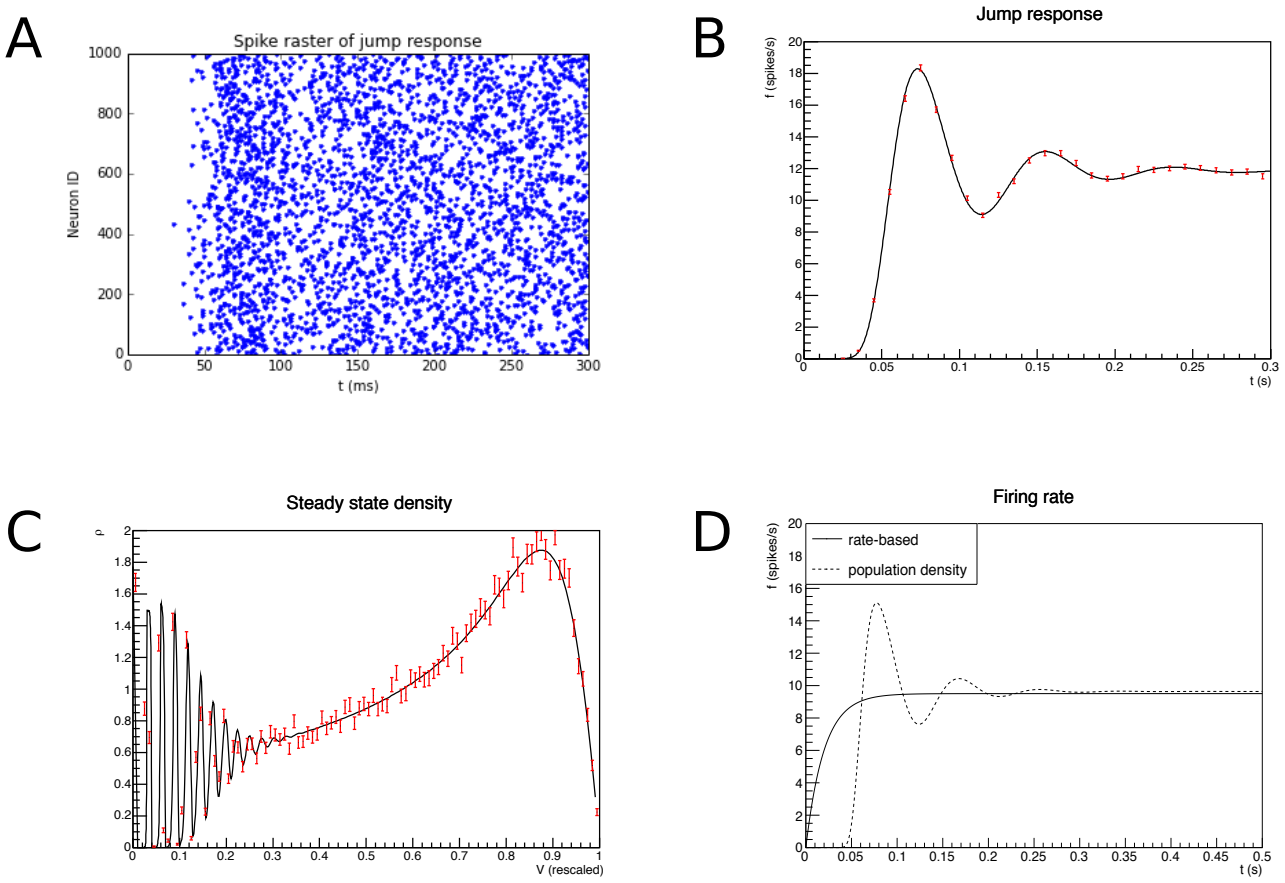


Figure 2.7: **LIF dynamics: A.** A spike raster showing an LIF population undergoing a jump response. Neurons are at equilbrium at $t = 0$. From $t = 0$ each neuron receives a Poisson distributed input spike train (λ = 800 Hz, $h = 0.03$, i.e. an input spike raises the PSP by 3% of the difference between threshold and equilibrium potential, $\tau = 50$ ms, following [150]). **B.** Firing rate calculated from the PDT method (solid curve), compared to firing rate from spiking neuron simulation (red markers). **C.** The density calculated by the PDT method (solid curve) at $t = 0.3$ s, compared to a histogram of the membrane potential over the population at the same time. **D.** Wilson-Cowan prediction for the firing rate, compared to PDT result. Importantly, Wilson-Cowan output must be tuned: the steady state value to which it converges is not predicted by the Wilson-Cowan equations, but must be provided as a sigmoid. In contrast, the PDT method calculates the firing rate from first principles, and agrees well with the spiking neuron simulation, within statistics.

except at high values of V, on the right, which corresponds to spike generation. At high values of w, there are two effects: stronger leak (larger cells) and a lower (more negative) equilibrium potential, which makes it harder for a cell at high w to be driven across threshold, precisely the effect one expects due to adaptation. At low w, the opposite happens: cells become more excitable. For very low w values, which can not be reached under cortical conditions, at least not for the parameters we used, there is the theoretical probability of a

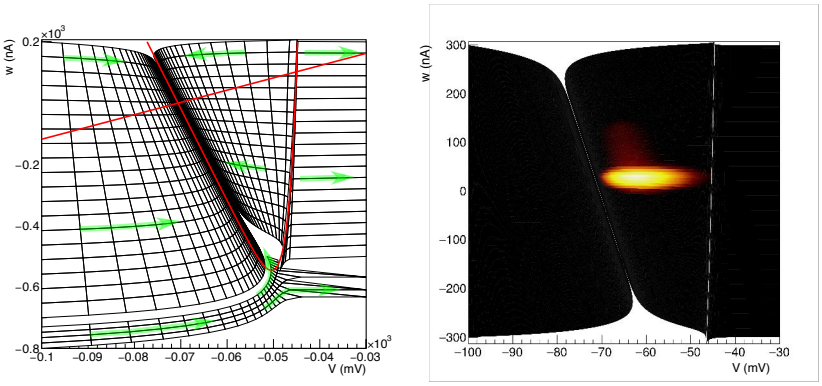


Figure 2.8: **AdEx dynamics:** Left: Overview of AdEx dynamics. Right: a heat plot of the density profile during simulation. On the horizontal axis the membrane potential, on the vertical axis the adaptivity parameter. Note that the right figure constitutes a considerable reduction of state space compared to left. For the connectivity parameters we use, the state space on the right is the part of state space reachable by dynamics.

rebound (neuron always spikes).

A density function now lives in this two dimensional space: $\rho(V, w)$. The evolution equation is a direct generalization of Eq. 2.3. For a model with n state variables $\vec{V}$, a point model takes the form:

$$\tau \frac{d\vec{V}}{dt} = \vec{F}(\vec{V}) \tag{2.5}$$

and the density equation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial \vec{V}} \cdot \frac{(\vec{F}\rho)}{\tau} = \int dh p(h) \nu(\rho(\vec{V} - \vec{h}) - \rho(\vec{V})) \tag{2.6}$$

, where $\vec{h}$ represents the effect of an input spike.

We represent the density function by a heat plot on state space: the highest values or white, low values are red. We are able to simulate the density function by a method analogous to that of [47; 96], generalized to two dimensions. In Fig. 2.8 we show the result of a simulation: the density function as a fixed point in time. As before, we can calculate the firing rate of the population by calculating the the flux across threshold (which is still given by $V = V_{threshold}$, i.e. the right hand side of the grid).

The simulation software, MIIND, is publicly available[3]. The LIF version of the algorithm has been available for some time [49], while the two dimensional version has become available recently.

[3] `http://miind.sf.net`

3 Objectives

Bridging the gap between experimental neuroimaging evidence and the available modelling solutions to *binding*, is a crucial step for the advancement of our understanding of the brain computation and representation of symbolic structures. From the recognition of this problem, the goal of this PhD became the identification and experimental test of the theories, based on neural networks, capable of dealing with symbolic structures, for which we could establish testable predictions against existing fMRI and ECoG neuroimaging measurements derived from language processing tasks.

We identified two powerful but very different modelling approaches to the problem: Smolensky's tensor product representations and the Neural Blackboard Architecture (NBA). In the case of Smolensky's tensor products, we considered the superposition principle to be one of its crucial assumptions, so we decided to acquire a new fMRI dataset to test it in different brain regions. In the case of the NBA, we built a new simulation to be able to perform predictions on the temporal dynamics and spatial patterns of binding observed in the neuroimaging literature.

Objectives outline:

1. **Test the superposition principle of Smolensky's tensor product representations with BOLD-fMRI**
   (a) Design experimental manipulation for the acquisition of a two-syllabic pseudoword representations BOLD-fMRI dataset.
   (b) Extract pseudoword representation patterns with traditional univariate techniques
   (c) Develop tests with decoding algorithms to provide evidence in favour or against superposition in brain Regions of Interest and study the locality of those representations.
2. **Test the neural activity and temporal dynamics predicted by the Neural Blackboard Architecture**
   (a) Implement a compartment circuit simulation with spiking neural networks employing population density techniques

(b) Tune the implemented circuit only for correct binding operation

(c) Generate the neural activity of selected stimuli from fMRI and ECoG experiments

(d) Evaluate the qualitative similarity between the NBA circuit predictions and the results reported by the fMRI and ECoG experiments

Part II

Testing the superposition principle with bi-syllabic pseudowords

4 The superposition principle with BOLD-fMRI

In this chapter we introduce the problem of testing the superposition principle, that depend on Smolensky's tensor product representations, with BOLD-fMRI and how bi-syllabic pseudowords are modelled.

4.1 BOLD-fMRI interpretation of superposition and vectorial representations

The superposition principle: In Smolensky's Integrated Connectionist/Symbolic architecture (ICS)[172], the neural activation of a symbolic structure is given by the union of the *Filler*/*Role* bindings belonging to the symbolic structure. The set of *Fillers* that can be assigned, as well as the set of *Roles* will depend on the modelled stimuli. We could consider for example phonemes as *Fillers*, to be assigned to node positions in a tree structure as *Roles*, to finally form morphemes and words as the resulting symbolic structure. Smolensky proposes to employ the linear operation of addition as the union operator of the bindings, such that the neural activity of an abstract symbolic structure is given by Equation 4.1. We present a concrete example in Equation 4.2, where the word "cat" is formed by adding the bindings of the phoneme *Fillers* "k", "ae" and "t", with their respective positional *Roles* in a tree structure.

$$Structure = Filler_1 \otimes Role_1 + ... + Filler_n \otimes Role_n, \quad \text{Abstract representation} \tag{4.1}$$

$$S_{cat} = F_k \otimes R_0 + F_{ae} \otimes R_{10} + F_t \otimes R_{11}, \quad \text{Example word} \tag{4.2}$$

$$S_{figu} = F_{fi} \otimes R_{left} + F_{gu} \otimes R_{right}, \quad \text{Example pseudoword} \tag{4.3}$$

In this work we will consider pseudowords composed of the combination of two syllables of one consonant and one vowel (CVCV). We present in Equation 4.3 the modelled representation of the pseudoword "figu" as an example. Moreover in Figure 4.1 we show the BOLD-fMRI interpretation of neural vectors. The main idea is that BOLD activity in voxels is meant to represent

the aggregated neural activity of a set of neural units from the representations neural vectors. Aggregation of neural activity implies an important loss of information that could impede decoding of the representations if the values of neural activations are similarly distributed in different segments of the neural vector established for each voxel.

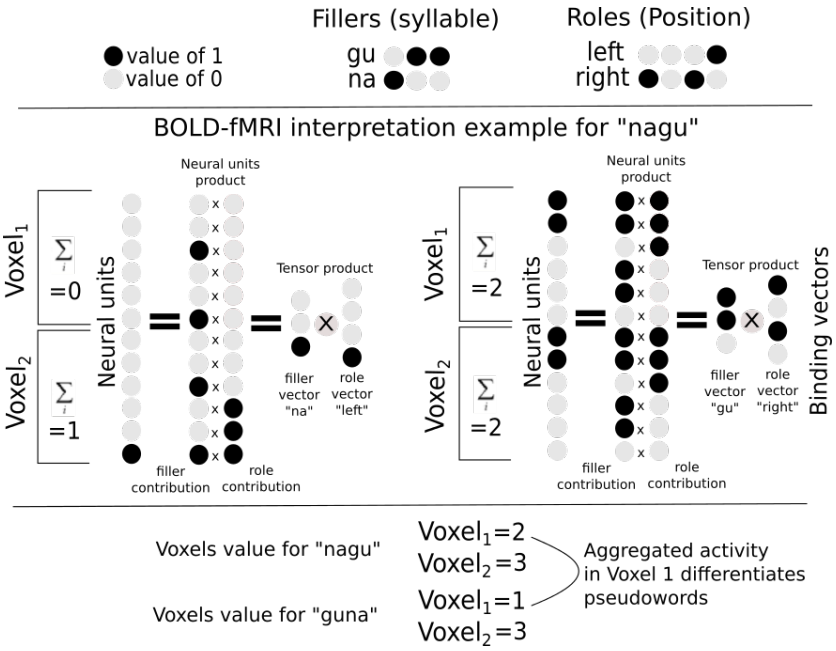


Figure 4.1: **Illutration of superposed tensor product representation in BOLD-fMRI:** We present the example neural vectors of the syllables "na" and "gu" bound to the left and right positions of a bi-syllabic pseudoword. We illustrate how the level of BOLD activity should reflect the aggregated activity of a segment of the neural units that form a representation. The superposition principle consists on the sum of the vector values from each binding, to obtain the final total activity in a voxel. The voxel values of the pseudoword "nagu" correspond to the plots of the neural vectors and those of the pseudoword "guna" were derived in a similar way. Due to the effect of aggregation, only one voxel in the example permits differentiating the two pseudowords.

Stability, uniqueness and intrusion of vectorial representations: An important assumption behind tensor product representations is that the hypothesized *Filler* and *Role* vectors have been learned by the cortex and are fixed to a specific set of neural units and values. Nonetheless there are some biological and theoretical factors that could go against this assumption. It is known that there is state-dependent adaptation in the cortex[103] and firing thresholds can be altered according to arousal state[129], which can also complexify the behavior of neurovascular coupling[101; 125]. Moreover there is evidence for the existence of cell assemblies in parts of the cortex, like the hippocampus, where neural spiking is importantly affected by local network interactions[86], and the formation and dissolution of dynamic cell assemblies have been demonstrated during cognitive processing[24].

These possibilities could increase importantly the variability of the unit neural activity or even imply the existence of more than one pattern assigned to a particular *Filler* or *Role* vector. An analogy of how this type of effects would operate against pattern identification is changing gaze position with respect to visual stimuli presentated on a screen. Not accounting for gaze position would give the impression of multiple representation patterns for

the same image, even though retinotopic representations are very stable and precise relative to other activations in the cortex, because the activation vectors would change spatial location (change neural units) from trial to trial.

Smolensky also proposes that the neural activation vectors of the *Fillers* and *Roles* should be linearly independent in the best case to allow for exact unbinding operations in the cortex, although linear independence is not a strict demand, because there is a graceful degradation as the correlation between vectors increases on a distributed representation. Nonetheless even if it was the case that underlying distributed neural unit representations were linearly independent, this do not imply that the aggregated activity of arbitrary segments of those neural units would remain independent, or even differentiable from each other to the necessary degree to detect it with the signal to noise ratio of the BOLD signal. For example in Figure 4.1 we illustrate the possibility of not being able to differentiate the pseudowords "nagu" and "guna" in their voxel activations, which was the case of Voxel 2 in the plot, even though their underlying *Filler* and *Role* neural vectors are linearly independent.

Locality and sparsity of vectorial representations: The tensor framework proposed by Smolensky allows for the possibility of completely distributed representations and encourages it, since distributed representations have several advantages in terms of pattern generalization and memory efficiency over local representations. From the neurobiological point of view, it seems likely that there are broadly distributed representations when we are able to find with coarse random sampling neurons tuned to specific experimental stimuli. Consider for example the work by [4], that characterizes the receptive field of a set of sampled neurons to moving dots. Spatially broadly distributed representations would be an advantage for BOLD-fMRI detection of neural patterns, since it increases the amount of voxels that would contain information about the patterns. Nonetheless this would only be the case if the spatial distribution of activation values across neural units is not uniform, such that we can capture higher random spatial differences between the aggregated activity patterns of the neural units.

Another property that would help pattern identification with BOLD-fMRI is having enough sparsity in the distributed representations to augment differences in the aggregated activity or even cause semilocal representations. A trivial example of semi local representations would be the inversed hemispheric retinotopic projections of the visual information shown to the different eyes. From the neurobiological point of view, it seems likely that there is certain degree of sparsity. Olshausen *et al.* shows how a coding strategy that maximizes sparseness is sufficient to account for important properties of the mammalian primary visual cortex, which are considered to be spatially localized, oriented and bandpass, comparable to the basis functions of wavelet transforms[149]. In the neuroscientific literature the actual degree

of sparseness related to neural representations is still debated, sometimes even only one neuron is found to be responsive to very specific stimuli, giving rise to the hypothesis of grandmother cells. An interesting debate on this account is developed by Bowers *et al.*[21], in which it is made clear that the degree of sparsity observed depends on the experimental stimuli defined and will vary across neural areas. The neural sampling methodologies employed so far in humans have not been able to completely characterize the degree of sparsity, because they are still not capable of capturing the separated neural activity of complete local neural populations.

Stimuli selected to test superposition on syntactic operations: There are few explorations in the neuroimaging literature about composition operations. Additive models of composition for sensory stimuli, similar to the superposition principle in tensor product representations, have been tested with multi unit neural recordings in monkeys. It seems that the composition operations employed by the brain depend importantly on the features considered. For example in the monkey's inferotemporal cortex, evidence was provided for conjunctive non additive models in the case of shapes composition[10], but when considering jointly shape and color in the same region, evidence for linearly additive composition was found instead[131].

More work has been done with sensory stimuli on other animal models, but testing specifically for symbolic representations is more complicated due to the limited measurement techniques that we have for the human brain. In the case of BOLD-fMRI there are already some studies employing a variety of machine learning techniques, that have tried to approach the problem in different cognitive domains. Decoding methods, classifying stimuli conditions from BOLD signals, have been used to demonstrate a compositional code similar to superposition for rule representations in the human prefrontal cortex[165]. In the case of language, Mitchell *et al.* tested an additive model of semantic features with encoding models, that predicted the BOLD brain images associated to English nouns[137], but no similar work has been done for syntactic features yet.

In this work we were interested in testing the additive model proposed by superposition on syntactic operations of language, which in most levels of language processing are hypothesized to depend on hierarchical tree structures. The idea of assuming positional *Roles* representing nodes of trees is relatively uncontroversial at the phonological and morphological level of language processing and previous work have been successful in characterizing the neural representations of isolated syllables with BOLD-fMRI[62]. Moreover several neural activity effects spread in the fronto-temporal language network, linked to phonological manipulations and pseudowords processing, have been reviewed in various metanalysis[195; 178]. Taking all this into account, we decided to test superposition of the syntactic representations of bi-syllabic

pseudowords with decoding techniques in BOLD-fMRI, for which modelled representation examples were given in Equation 4.3 and Figure 4.1.

4.2 From neural unit recordings to BOLD-fMRI measures of aggregated activity

In Smolensky's Integrated Connectionist/Symbolic architecture (ICS)[172], the implementation of symbolic representations is done through the activation of neural units that form part of a neural network. This means that Smolensky's tensor product representations have a straight interpretation on the spiking activity of Multi Unit Activity recordings (MUA) of neurons.

To test properties of Smolensky's proposal with other neuroimaging techniques like BOLD-fMRI, that reflect aggregated neural activity, it is important to verify that there is a linear mapping between the underlying neural activity and the aggregated activity. So we need a correspondence between the spatial location and neural activity values with respect to single neural unit measurements. Moreover, since in this work we want to test the additive model brought forward by superposition, it is important that the mapping from neural activity to BOLD remains approximately linear.

Regarding spatial localization of the signals, Siero *et al.* studied the spatial properties of the hemodynamic (BOLD) signal at 7T and reconfirmed its spatial correspondence with intracortical (ECOG) time series in the motor cortex for a finger tapping task[171]. They managed to decode spatially the tapping of different fingers and found that the spatial correlation between signals for the different fingers is high (on average R=0.54) and their maxima co-localized within 3 mm distance.

In the case of the mapping of neural activity values to aggregated activity, Cardoso *et al.* designed a visual task, in which drifting sine-wave gratings were presented passively to monkeys during 3-4 s while they fixated[32]. With this task they demonstrated high predictability (0.94 R squared) of a component of the BOLD hemodynamic response, the cerebral blood volume (CBV), from direct neural measurements (MUA and LFP). Nonetheless the BOLD signal itself is more complex, it depends on the coupling between cerebral blood volume (CBV), cerebral blood flow (CBF) and oxygen concentration measures (CMRO2), where the last two have been linked to adaptation and other non-linear effects[139]. For example Toyoda *et al.*[181], employing chequerboard visual stimuli with durations between 1 and 8 seconds, showed that the contribution of the oxygen extraction fraction (OEF) to the BOLD signal, which is a measure related to CMRO2, can be four to seven times greater than the contribution attributed to the CBV under the range of plausible parameters of neural activity and adaptation. But they also showed that the contribution ratio of OEF over CBV can be compensated with the experimental design, since the ratio decreases as the duration of the stimuli increases.

Despite this complexity of the BOLD signal, a consensus is emerging on a

linear relationship for long duration stimuli of enough intensity[8; 9; 29; 89; 110; 124; 152; 184]. Important exceptions exist, but often they are related to sensory stimuli of short-duration[5; 174; 193; 204], or low-intensity that do not overcome the activity threshold necessary to generate an hemodynamic response[193]. For example, strong evidence of a linear relationship between BOLD and MUA, for long-duration sensory stimuli with varying stimulation frequency, was provided by Devonshire *et al.*[53]. They studied regions inside and outside of the cortex and demonstrated the effect with electrical stimulation of the entire whisker pad on the left of a rat's snout, during 40 s with different pulse frequencies. All the mentioned evidence points to the idea that it is reasonable to interpret and test neural unit level representations with BOLD-fMRI, as long as temporal variables of the experimental design like length of stimulation or inter-stimuli intervals are manipulated to minimize the influence of BOLD non-linearities.

5 The syllables superposition experiment

In this chapter we present the two tasks of the experimental design, the Bold-fMRI data acquisition, preprocessing and processing, and the analysis methods employed to assess the likelihood of superposed representations in the Regions of Interest considered.

5.1 Experimental design

Participants: Five native French speakers participated in the experiment (two females with ages 22 and 32 and three males with ages 23, 26 and 36). All subjects had high school background from French universities (Bac) and were right handed with a Laterality Quotient (LQ) of at least 40 (mean 70, SD 20.98), as measured by the Edinburgh Handedness Inventory[147]. The experiment was conducted at the NeuroSpin center and all subjects came on four different days, for a total of four scanning sessions. The experiment was sponsored by the Unicog lab U-992 in NeuroSpin, and received ethical approval by the regional ethical committee (Comite de Protection des Personnes, hopital de Bicetre). All subjects gave written informed consent and received 80 euros for their participation.

Introduction to the experimental design: Two experimental designs were developed; a language localizer[118], to identify in each subject language processing regions, and a pseudoword representations design to obtain brain representations of the syntactic union of two syllable combinations devoid of semantic content. All experimental tasks were implemented with python scripts exploiting the capabilities of the Expyriment python library[107].

For both designs, visual and auditory sensory modalities were used for stimulation, since in language regions we aim to find abstract representations insensitive to sensory modality. Visual Stimuli consisted of text, projected one word at a time in rapid serial visual presentation (RSVP), on a translucent screen with a digital light processing projector (PTD7700E, panasonic, frame rate: 60 Hz, resolution of 1024 x 768), with a viewing distance of 89 cm. Auditory Stimuli were delivered through MRI-compatible headphones (MR confon), and the volume was adjusted for each participant to a comfortable

hearing level.

Language Localizer

Stimuli: The stimuli consisted in blocks of three phrases and blocks of three non language stimuli that varied in implementation with the sensory modality. These blocks were presented in an alternated fashion with the purpose of extracting brain areas processing language from the contrast of these block categories[118]. Visual stimuli was text presented in the screen with a fixed point Inconsolata font[1]. The text comprised 0.72 degrees of vertical visual angle and a maximum of 5.8 degrees of horizontal visual angle, with the longest word having 14 letters. The visual non language stimuli was formed by replacing words in the phrases with consonant strings, for example "the cat" could be replaced by "ztr pfg". Auditory stimuli consisted on the same phrases digitally recorded at 22.05 kHz in a quiet room by a male speaker. Phrase recordings had a mean duration of 2.33 seconds (SD, 0.41 s), giving a total average duration of 7 seconds for a block made of three sequences. To generate the control auditory non language stimuli, the phrase recordings were scrambled with the multiband approach suggested by Ellis and Lee[60], but with python code using the Brain Hears software[66].

[1] https://fonts.google.com/specimen/Inconsolata

Task and trial structure: The subjects were instructed to read or listen attentively to all stimuli presented. Each trial consisted on presenting one of the blocks designed, which were the grouped phrases or non language stimuli. Each block contained three phrases or three consonant strings, the first made of 9 units, the second of 10 units and the last of 9 units. A fixation cross was presented before the presentation of each phrase or string, for 500 ms, followed by a blank screen for 500 ms. In the visual case, each text unit was presented regularly for 200 ms, which is not the case in the auditory modality that has variable sequence duration. Between the presentation of the three stimuli a blank screen was presented for 600 ms. At the end of the presentation of the three stimuli a blank screen was presented for 7 seconds waiting for the next trial (the next block). There were 4 runs of acquisition and in each of them 90 trials were presented. In Figure 5.1 we show an example of a sequence in the visual modality.

CVCV Pseudowords presentation

Stimuli: The "CV" syllables "fi", "gu" and "na" were selected to form all possible "CVCV" pseudowords from their combinations: "fifi", "figu", "fina", "gufi", "gugu", "guna", "nafi", "nagu" and "nana". These syllables were selected under two constraints. The first was that all syllabic combinations could not lead to word formation, such that we could assume similar sensory and language processing of symmetric representations like "figu" and "gufi", expecting only syllable position effects. The second was that we wanted to

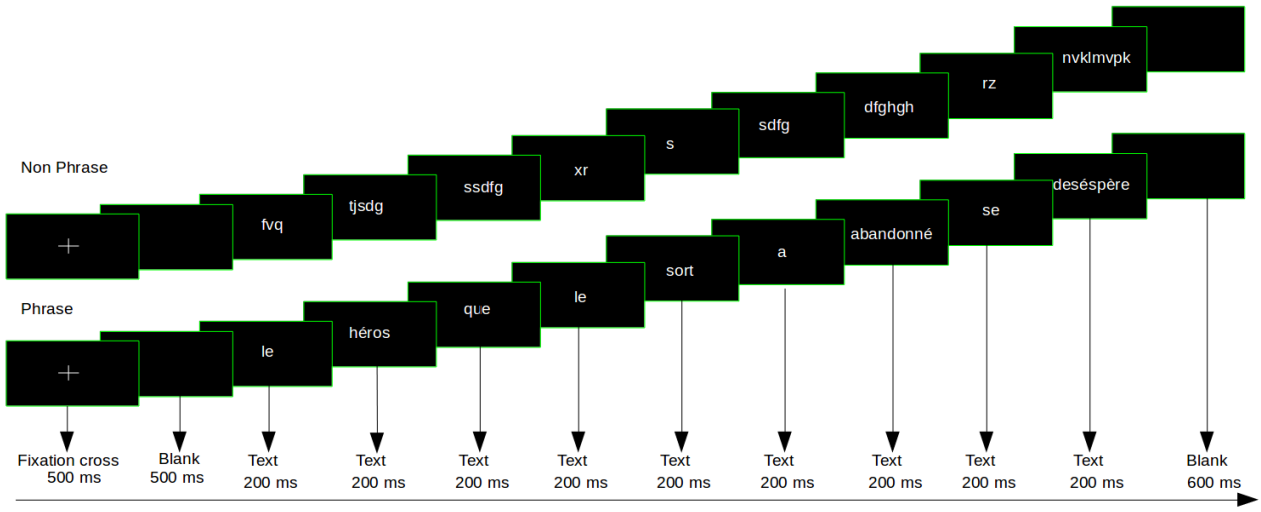


Figure 5.1: **Visual trial of the language localizer:** Each black square represents the screen at a different time point. Only one example phrase and consonant string is shown, which comprises only one third of a block stimuli.

improve auditory discriminability, so we maximized the featural distance between consonants and between vowels. We selected one velar consonant "gu", one labio-dental "fi" and one alveolar "na" with their respective high-back tongue "u", high-front tongue "i" and low-back tongue "a" vowels. The effect of featural distances in auditory representations was demonstrated in the cortex by the work on phonetic organization of spatial patterns of Bouchard et al[20].

The pseudowords were presented in a visual and auditory modality. In the visual case they were presented as text in the screen with a fixed point Inconsolata font. We decided to make the text as big as possible to increment expected retinotopic effects but also tried to avoid the stimuli perception to be too tiring for the subjects, so finally the pseudowords were presented as lower-case text centered on the screen, spanning maximum 2.39 degrees of vertical visual angle and maximum 5.05 degrees of horizontal visual angle. For the auditory stimuli, three tokens of the syllables 'gu', 'na' and 'fi' were recorded at 22.05 kHz in a sound-proof room by a male speaker. They were edited to have the same duration, by cutting some of the periods inside the vocalic part. They were then concatenated to generate the nine bisyllabic experimental stimuli. These stimuli all had a duration of 660ms. Probe stimuli, required by the task, consisted on smaller upper-case text spanning 0.6 degrees of vertical visual angle and 1.68 degrees of horizontal visual angle for the visual modality and modified recordings of the syllables with 10% higher pitch for the auditory modality.

Task and trial structure: The task consisted on keeping the pseudowords in memory for a possible comparison with a second pseudoword. The instruction given to the subjects was to fixate a green dot and to keep in memory a following pseudoword, until the arrival of a red dot that signalled

the end of the trial. From time to time the subjects had to make a comparison with a second pseudoword presented in the middle of the trial, in which case, confirmation of a positive match was indicated with a right hand button press and of a negative match with a left hand button press. We included the matching task to make sure subjects were complying with the task and paying attention to the stimuli, but we did not include a matching task on each trial to try to maximize the amount of stimuli presented in a session.

We show the structure of a trial from the visual modality in Figure 5.2. The green dot appeared for 0.5 seconds followed by a flashing presentation of the pseudoword, in the visual case, to be kept in memory for 3.2 seconds with a 0.25 seconds jitter. We decided to present the visual pseudowords for only 0.2 seconds to minimize the influence of saccades in the estimation of brain activations. The pseudowords were flashed twice for 0.1 seconds to increase visual response. In the matching task trials, the second pseudoword was presented for 0.5 seconds followed by a response and rest period of 6.5 seconds. At the end of the trial the red dot was presented for 0.5 seconds followed by a 2.5 seconds resting period. Each imaging run consisted of 45 trials (5 per pseudoword), where the order of presentation of the pseudoword conditions was shuffled. In total there were 8 runs in a session, with two auditory sessions and two visual sessions, for a total of 80 trials per condition per modality. Only nine trials were randomly selected to contain a second pseudoword to perform a matching task. In the auditory case the trial structure is identical except for the 660 ms duration of the pseudowords recordings, in which case the memory time was reduced to 2.8 seconds to have the same trial total duration as in the visual case.

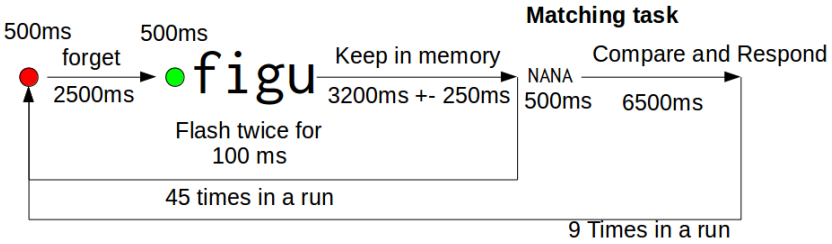


Figure 5.2: **Visual trial example of the pseudoword matching task:** A green dot is presented for 500 ms, followed by a pseudoword flashed twice for a total presentation duration of 200 ms. It has to be kept in memory for a period of 3200 ms with a 250 ms jitter. Nine times in a run a second upper-cased pseudoword is presented for comparison during 500 ms with a response period of 6500 ms.

5.2 Data acquisition and processing

Imaging: The acquisition was performed with a 3 Tesla Siemens Prisma Fit system equipped with a thirty two channels coil. Anatomical images were taken using a 3D Gradient-echo sequence and voxel size of 1x1x1 mm. Functional images were acquired as T2*-weighted echo-planar image volumes (Multi-Band EPI C2P from Minnesota University). The MultiBand EPI consisted on the parallel acquisition of 4 slices at a time, reconstructed by a parallel imaging reconstruction algorithm[34]. Eighty transverse slices covering the

whole brain were obtained with a TR of 1.5 s and a voxel size of 1.5 x 1.5 x 1.5 mm (TE = 26.8 ms, flip angle = 70, no gap). Moreover accurate timing of stimuli presentation relative to FMRI acquisition was achieved with an electronic trigger at the beginning of each run.

Acquisition sessions: Each subject had four sessions of scanning with a similar structure. The first two sessions included the visual version of the pseudoword matching task and the last two sessions the auditory version. Each scanning session lasted 78 min and 6 sec with an anatomical scan and 10 functional runs structured as follows:

1. Anatomical T1 (1 volume, 7 m 46 s)
2. Pseudoword matching task "Visual/Auditory" (253 volumes, 6 m 54 s)
3. Pseudoword matching task "Visual/Auditory" (253 volumes, 6 m 54 s)
4. Pseudoword matching task "Visual/Auditory" (253 volumes, 6 m 54 s)
5. Pseudoword matching task "Visual/Auditory" (253 volumes, 6 m 54 s)
6. Language localizer task "Visual" (435 volumes, 11 m 27 s)
7. Pseudoword matching task "Visual/Auditory" (253 volumes, 6 m 54 s)
8. Pseudoword matching task "Visual/Auditory" (253 volumes, 6 m 54 s)
9. Pseudoword matching task "Visual/Auditory" (253 volumes, 6 m 54 s)
10. Pseudoword matching task "Visual/Auditory" (253 volumes, 6 m 54 s)
11. Language localizer task "Auditory" (435 volumes, 11 m 27 s)

Data preprocessing: The OASIS-30 Atropos template atlas from Mindboggle[2] was used as reference for normalization and segmentation of the subjects anatomy. The methodology behind this atlas is based on state of the art algorithms from the Advanced Normalization Tools (ANTS) and a cohort of 101 manually segmented subjects, giving very precise probabilistic maps and anatomical ROIs[104]. A transformation between this template and one provided by ICBM in MNI space was also performed for MNI coordinate reports and visualization. The ICBM 2009a Nonlinear Asymmetric template was considered[42].

[2] http://www.mindboggle.info/data.html

After normalization and segmentation of each subject anatomy. The functional runs of all tasks were slice timed with SPM with reference to the 1st slice (default SPM behavior) and realigned with respect to the 3rd volume of the first acquired run of the first session. Realignment was performed with FSL MCFLIRT algorithm and co-registration was also performed with FSL but with the FLIRT algorithm employing a boundary based registration that takes into account previously performed white matter segmentation of the anatomy[79]. All preprocessing steps were implemented with the Nipype software[78].

Data processing: Two General Linear Model (GLM) estimations were performed, one on the non-smoothed, non-normalized and realigned functional images and the second on the smoothed version of the same

images, with a 6 mm Gaussian kernel. The non-smoothed beta maps derived were employed for decoding, while the smoothed beta maps were employed for parametric statistical tests. The GLM was implemented with the Nistats[3] software, which is part of the Nipy and Nilearn[2] ecosystem. A glover HRF was employed for the estimation with an additional cosine drift model to high-pass filter above 1/128Hz.

[3] https://github.com/nistats/nistats

The language localizer was modelled with two regressors for the block conditions, alongside motion regressors extracted from the realignment preprocessing step. Statistical estimation of a contrast between the two block conditions was performed on the smoothed images to extract the language network.

In the case of the pseudoword matching task, each pseudoword condition was modelled with one regressor, alongside left and right motor events derived from the behavioral responses and motion regressors extracted from the realignment preprocessing step. The condition beta maps corresponding to the smoothed images were employed for statistical estimation of motor contrasts and syllable position effects, for which a fixed effect model was considered across runs and sessions in each subject. To obtain statistical effects of syllable position, we modelled the conditions as two factors (left and right position), with three levels (syllables fi, gu and na). We estimated contrast vectors for the effect of left position, effect of right position and interaction of left and right positions, by employing the contrasts vector specification procedure of Henson and Penny[91].

It has been shown that taking into account trial-to-trial variability is desirable for multivoxel pattern analysis (MVPA)[1; 140]. As we wanted to look into the representational patterns of the different pseudowords, we decided to also estimate one beta map per trial, following the same methods employed for beta-series analysis[40]. This is also desirable to capture attention modulated variability in the voxel patterns of the pseudowords, since the task do not allow us to verify the processing integrity of each trial but only to motivate subjects engagement.

5.3 Data analysis

All data analysis was performed employing diverse Python scientific open source libraries[148]: Numpy[198], Pandas[130], Matplotlib[93], Ipython[156], Scikit-Learn[155] and the neuroimaging library Nilearn[2].

Regions of Interest (ROIs)

Sensory-Motor regions: In Figure 5.3 we display the contours of primary sensory-motor regions, taken from the cytoarchitectonic SPM toolbox[59], projected on the anatomy of Subject 1 alongside the gray matter mask, the brain glass template contours were adapted to the T1 anatomy of the subject.

Notice that the primary regions are broad, since we considered any voxel with non zero probability to be part of the region, and cover both hemispheres.

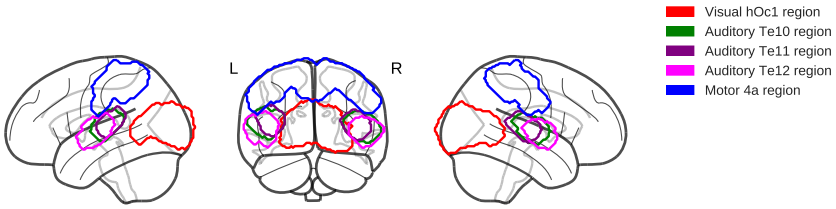


Figure 5.3: **Sensory-motor regions projected on *Subject 1* anatomy:** Contours are shown for the projected primary Visual, Auditory and Motor regions, alongside the subject extracted gray matter. The brain glass template contours were adapted to the T1 anatomy of the subject.

Language regions: Two sets of language regions were selected for the analysis. The first set of regions, shown in Figure 5.4, was selected to evaluate the quality of the language localizer contrasts from a study done by Mahowald and Fedorenko[118]. In this study activation parcels were derived from similar language localizer acquisitions in hundreds of subjects, covering the whole fronto-temporal language network.

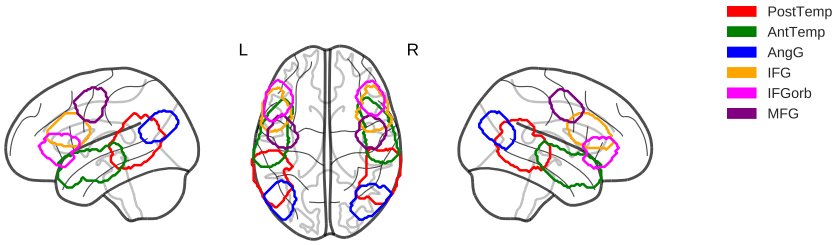


Figure 5.4: **Language localizer parcels projected on *Subject 1* anatomy:** Contours are shown for the projected language localizer parcels reported by Mahowald and Fedorenko. The brain glass template contours were adapted to the T1 anatomy of the subject.

The second set of regions is shown in Figure 5.5. Diverse regions, also covering the fronto-temporal language network, that have been directly linked to binding or constituency effects, from different sources, were selected to facilitate the analysis and interpretation of the results[4].

First we considered the left visual word form area (VWFA) that has been linked to binding of visual and verbal representations in both words and pseudowords, for early stages of language processing[41; 194; 50; 75; 205]. The VWFA was built as a 4 mm sphere centered at the x=-46, y=-61 and z=-10 in MNI space[50].

Second we considered the left hemispheric regions derived from neural activation clusters related to phrase constituency effects, observed in the experiment of Pallier et al.[154]. In this experiment two groups of clusters were found to respond differently to constituency manipulations in phrases and jabberwocky stimuli. Some regions responded only to semantic coherence from phrase stimuli, namely the anterior superior temporal sulcus (aSTS), the temporal pole (TP) and the temporo parietal junction (TPJ). Other regions responded also to syntactic coherence from the jabberwocky stimuli that

[4] Besides the ROIs finally considered, we explored peaks of pseudoword phonetic and morphological effects from various meta-analysis[195; 178]. The effects reported were numerous and spread across the whole fronto-temporal network. We verified that the ROIs covered most of the effects and opted to perform the analysis in a smaller set of bigger ROIs than what would be obtained from spheres centered at the reported effect peaks. It could be argued that we are missing specific effects, but since we will implement a searchlight selection procedure of voxels, any specific effects should be selected inside their containing ROIs for the decoding models

contained pseudowords to minimize semantic content, namely the posterior superior temporal sulcus (pSTS) and the inferior frontal gyrus pars triangularis and pars orbitalis (IFGtri and IFGorb)[154].

Finally we considered Broca's complex for its long standing link to binding operations in language[82]. We took the Broca 44 and 45 regions from the cytoarchitectonic SPM toolbox[59], which are broad due to non zero probability consideration of voxels in the probabilistic map.

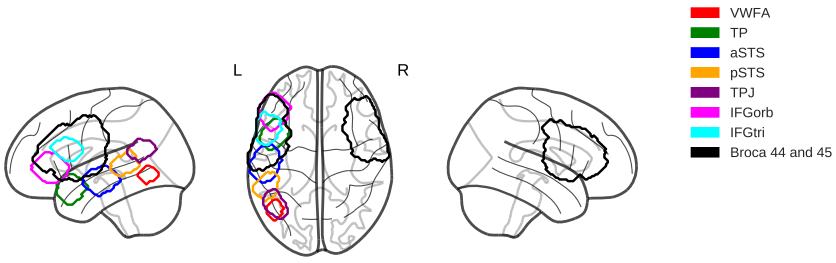


Figure 5.5: **Language regions of interest projected on *Subject 1* anatomy:** Contours are shown for the projected left hemispheric language regions of interest. We include the VWFA, the 6 regions reported by Pallier et al. related to constituency effects and the joint Broca 44 and 45 regions taken from the cytoarchitectonic SPM toolbox. The brain glass template contours were adapted to the T1 anatomy of the subject.

Sanity checks

To verify the integrity of the language localizer acquisitions, we compared the thresholded activations of the contrasts (word sequence over non words sequence), with the parcels of Mahowald and Fedorenko. These parcels represent probable activation derived from thresholded maps at $p < 0.001$ for hundreds of subjects, so not being able to cover them with our language localizer could signal problems with the acquisition and limit our interpretation of syllabic representations in the derived language network.

In the case of the pseudoword matching task runs, we validated the estimated activation maps in two ways. First we verified the statistical effect of the left vs right motor response contrast and checked that we could decode left and right response activation maps derived from the GLM estimation. Second we looked for expected retinotopic effects of the centered text in the visual modality, that implied a separation of the statistical effects of the first syllable position and second syllable position in the right and left hemispheres respectively.

Sensory-Motor Classification methods

Classification of motor responses: Motor classification was simply performed on the average beta maps of each session derived from the smoothed images GLM model, masked by the motor 4a region of the cytoarchitectonic maps. We standardized voxel activations to form the features used for training a nonlinear SVC classifier based on a radial basis functions kernel with default parameters from the Scikit-Learn[155] software. We employed the multiclass One Vs Rest (ovr) classification strategy, such that the decision function is

based on one classifier per condition, with a Leave One Out Cross Validation (LOOCV) procedure based on sessions.

Sensory and language classification methods

Classification models: In each Region of Interest (ROI), we trained three decoding models: the first identifying full bi-syllabic pseudowords (CVCV model), the second identifying only first position syllables (CV1 model) and the third identifying only second position syllables (CV2 model). Chance on the CV1 and CV2 models was 33.33% for the three conditions "fi", "gu" and "na", and 11.11% on the CVCV model for the nine pseudoword conditions. Moreover we trained one model per sensory modality, so we trained one model on the visual stimuli and one model on the auditory stimuli for each ROI, except for the visual and auditory regions. In all models we tested generalization to the opposite sensory modality.

Searchlight and voxel selection procedure for ROI analysis: The ROIs that we considered had thousands of voxels (features), which could impact negatively the performance of the classifiers, so we first decided to select promising voxels by running a Searchlight[61] classification procedure on a 5 millimeter radius spheres. We ran the searchlight on a selection of voxels in the gray matter mask of each subject constrained by additional statistical considerations. For all regions we considered only voxels on which a 3 mm sphere centered on them contained at least one statistical effect of syllable in first position, of syllable in second position or position interaction with a p-value < 0.001. For language regions we also constrained the 3 mm voxel sphere to contain at least one statistical effect of the language network contrast with a p-value < 0.001. For the searchlight classifiers we employed the average beta maps of each session from the non smoothed images GLM model. We ran in each sphere the three classifier models for each sensory modality dataset. The classifiers accuracy was assigned to the center voxel of each sphere, resulting in three accuracy maps. Then voxels from each map were ranked and the top "n" voxels of each map alongside a 3 mm sphere around them were taken as features for the final ROI classifier. The number "n" of top voxels to consider was cross validated in a parameter grid search of the ROI classifiers, taking values from 1 to 40 in sensory regions and from 1 to 20 in language regions.

Classification procedure: For all classification models, we employed the multiclass One Vs Rest (ovr) classification strategy, such that the decision function is based on one classifier per condition. A Leave One Out Cross Validation (LOOCV) procedure based on sessions was implemented for all trained classifiers, taking into account activation maps from the 720 trials of each sensory modality. We took into account only the voxels (features) selected by the previously explained searchlight preprocessing procedure in

an ROI. The trials in the training set were employed to standardize the beta activation values of all trials inside each feature. The standardized features were then passed to a NuSVC linear classifier, for which we performed a grid search for the best value of the "nu" parameter taken from 0.2, 0.5 and 0.8, alongside the number "n" of top voxels.

To estimate p-values for accuracy and other values taken from the classifier, we retrained a model 100 times with the same dataset but shuffled labels (shuffled models)[5]. From each classification model we extracted confusion matrices and the model coefficients for further analysis.

[5] Taking into account only a 100 permutations introduced a limited precision of 10^-2 in the estimation of p-values, such that 0.01 is the best threshold that can be tested. This had to be done to reduce the computational time that was, already for a 100 permutations, around 2 hours for each model per ROI in a parallelized setup on a machine with an 8 cores 3 Ghz AMD CPU

Structural tests of representations

Null distributions for interpretation of representations tests: We will test the superposition principle and the locality of representations by interpreting measurements taken from the confusion matrix and coefficient weights of the NuSVC linear classifiers. Since the classifier has particular biases, it is important that we define appropriately a null distribution, such that we can assess how extreme or significant are the obtained measurements for a given dataset and condition labels. As was done to evaluate the classifiers accuracy, we built the null distribution by repeating the measurement in the results of the 100 shuffled models, in which the same dataset was employed but condition labels were uniformly shuffled. For demonstration here we took the shuffled models of an example dataset corresponding to the selected voxels for the visual hOc1 region classifier of Subject 4. As an additional check, we trained a 100 NuSVC classifiers with a fake white noise dataset of same dimension as the example dataset (white noise models), such that an alternative null distribution was generated using the same 100 label shuffles of the shuffled models. In the following paragraphs we demonstrate that the null distribution given by the shuffled models is similar to that of the white noise models.

Reminder of the implications of the superposition principle: The superposition principle predicts that neural representations (voxel activations) should follow Equation 5.1. This means that the activation value of a pseudoword at a voxel is the sum of the activation value of a syllable bound to the first position and the activation value of the other syllable bound to the second position. Then we expected the representation patterns of pseudowords sharing syllables in the same position to be more similar to each other than completely unrelated pseudowords. Moreover we expected pseudowords sharing syllables in different positions to not be more similar to each other than to other unrelated pseudowords, since the neural activity of a syllable is meant to change after being bound to its position according to Smolenky's framework.

$$Activation = Syllable \times Position_1 + Syllable \times Position_2 + Noise \quad (5.1)$$

Testing superposition with confusion matrices: In Figure 5.6 we identify each of the cells of a classifier confusion matrix according to the relationship between the syllables of the true and predicted pseudowords. Besides the diagonal of the confusion matrix, that represents predicting the same original pseudoword label, encoding the accuracy of the condition, we have three more types of cells: when the pseudowords have a syllable in the same position (overlapping syllables); when there is a shared syllable but in a different position (shared syllables); and when there are no common syllables between pseudowords (different syllables).

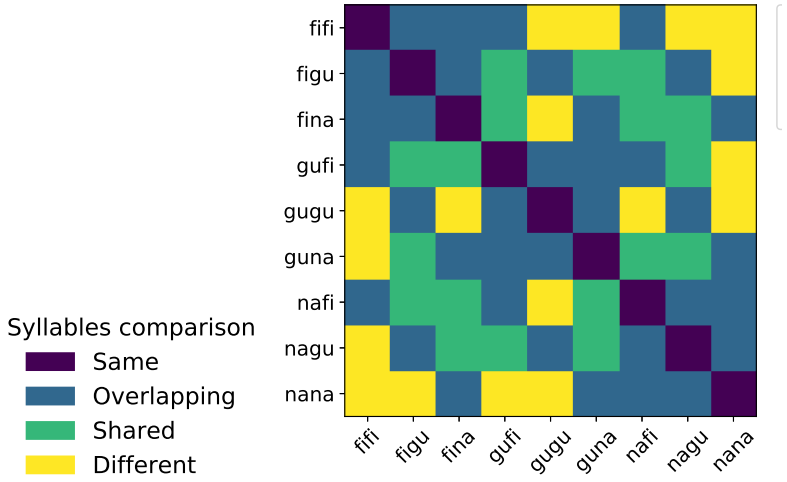


Figure 5.6: **Cell types in the confusion matrix of a pseudoword classifier:** The diagonal represents the same true and predicted pseudoword. The rest of the pairs correspond to pseudowords that have an overlapping syllable in one position (Overlapping), a common non overlapping syllable (Shared) or no common syllables (Different)

The representation similarity structure given by the linear terms in the superposition equation should be reflected in the confusion between conditions in a linear classifier, which means that we can compare the mean confusion of the different cell groups to provide evidence for or against superposed representations. The principle predicts that the mean confusion between conditions with overlapping syllables should be higher than between those sharing syllables with no overlap or with different syllables. In Figure 5.7 we show how the mean confusion of the different cell groups are related in the case of the null distributions. We provide evidence in favour of superposition if the mean confusion values of the cell groups in a tested model are located above the diagonal of both plots in the Figure 5.7.

Also we have to verify that mean confusion values of a tested model have a distance from the center reference higher than chance, which is given by the vector (0.11, 0.11, 0.11) that describes equal confusion for all pseudoword categories. To make this confirmation, we computed the empirical distribution of distances between the mean confusion vectors of the shuffled models and the chance vector, from which we calculate a p-value for the vector of a tested

model. We observe in the Figure 5.7 plots that the projected distribution of the cell groups mean confusion of the shuffled models is similar to that of white noise models, so we consider sensible to take the empirical distribution of shuffled models as reference to estimate p-values.

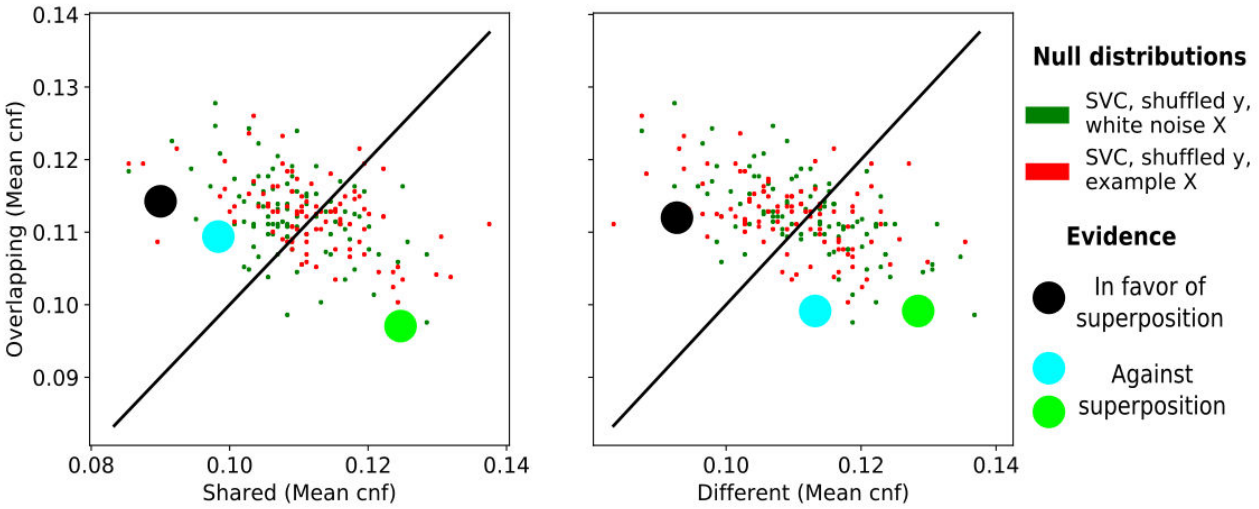


Figure 5.7: **Superposition test:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The red dots correspond to example shuffled models taken from the classifier of the visual hOc1 region of Subject 4. The green dots correspond to the white noise models using the same shuffled labels as the example shuffled models. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition.

Locality of syllable representations: We also tested if representations of syllables in different positions are partitioned (semilocal representations), which is expected for example in visual areas due to the hemispheric separation of syllable positions given by retinotopy. Smolensky's framework propose that completely distributed representations are more likely to be implemented due to their memory efficiency over local representations. Moreover if this was the case, our BOLD-fMRI voxel decoding should be more affected by the neural sparsity encoding considerations mentioned in the Introduction Section 4.1.

From ranking the feature (voxel) coefficients of the linear classifiers of the CV1 and CV2 models we can get an idea of the level of partition of position related information in a region. Based on the voxel selection procedure, we have that the voxels selected for both models are the same or at least one set of voxels is completely contained in the other. Thanks to this we can take the "N" best voxels subset of each model and then look at the proportion of shared voxels between the two sets. We expect an statistically extreme overlap of the best voxels subsets in case of distributed representations, while we expect less overlap than that given by chance in the case of semilocal representations. We can obtain the null distribution for the overlap of each N best voxels subset from the shuffled CV1 and CV2 models.

We show in Figure 5.8 several curve distributions to demonstrate how our argument operates in practice: a red distribution derived from an example CV1 and CV2 shuffle models taken from the visual hOc1 region of Subject 4; a green distribution derived from the corresponding white noise models.

and a theoretical blue curve distribution that illustrates the overlap in subsets taken from two ordinary lists, that share voxel indexes, uniformly permuted to create fake random voxel rankings, which reflects our intuition of the amount of overlap that can be achieved by random uniform permutations of rankings;

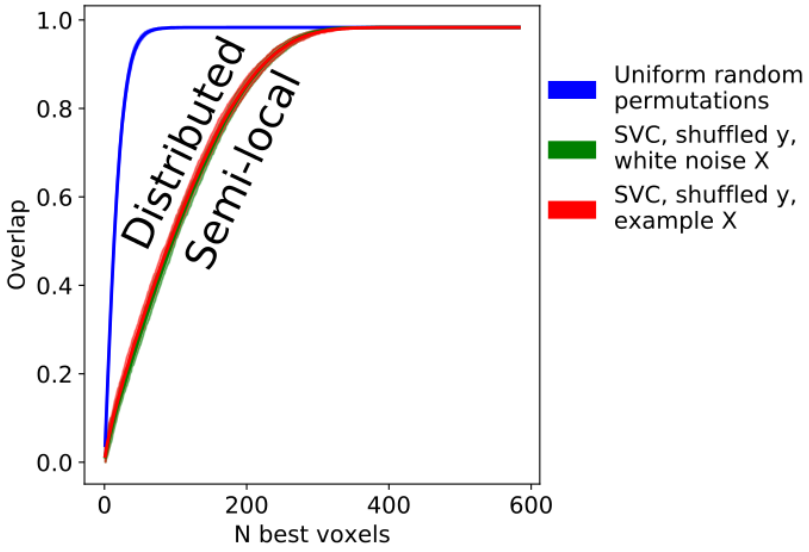


Figure 5.8: **Locality test:** Each curve in the plot represents the proportion of overlap between best voxel subsets from CV1 and CV2 models. The blue curve distribution represents repeated overlap comparison of uniform random permutations of an index list to create fake random rankings. The red curve distribution was derived from an example set of shuffled models taken from the visual hOc1 region of Subject 4. The green curve distribution was derived from shuffled models trained on Gaussian noise data with the same shuffled labels as the example set.

As we can appreciate from the green curve distribution, the amount of overlap introduced by an SVM model trained on white noise is quite different from the intuition given by a simple uniform random permutation of rankings, suggesting the need to estimate an empirical distribution from each SVM model. We also observe that the shuffled models null distribution behave similar to the white noise models null distribution, so it is sensible to use the empirical distribution of the shuffled labels to estimate p-values of the low deviation, towards semilocal representations, or high deviation, towards distributed representations. We will test the overlap deviation for each "N" best voxel subset of the target CV1 and CV2 models.

6 Experimental results

In this chapter we report the data analysis results. We comment on the successful pass of all required sanity checks and analyse the properties of pseudoword representations in the selected brain regions. In particular we will demonstrate evidence in favour of superposition in anterior brain regions and other interesting effects.

6.1 Behavioral performance

Four subjects (1 to 4), had a behavioral performance above 97% in both visual and auditory *CVCV Pseudowords presentations*, while *Subject 05*, that reported concentration span issues over all the acquisition, had a lower overall performance of 90%. Note that due to the experimental design structure, in which we only query few random samples, small score decrements can imply distraction over an important task segment. *Subjects 01 and 04* reported in the second auditory session that the volume was not high enough to be comfortable, although this did not reflect on their behavioral performance. So we consider all subjects data apt to neuroimaging interpretation, with caution over *Subject 05*. Behavioral performance details are provided in Table 6.1.

Subject	Visual (%)	Auditory (%)	Overall (%)
01	97.22	97.22	97.22
02	100.00	98.61	99.31
03	97.92	97.22	97.57
04	99.31	99.31	99.31
05	92.36	88.89	90.62

Table 6.1: **Behavioral performance on the *Pseudowords matching task*:** Performance correspond to correctly identifying if the pseudowords were the same or different, with no answer considered as incorrect. Visual and Auditory headers refer to the sensory modality of the task, where overall is the mean performance of both modalities.

6.2 Sanity checks

Language localizer activations: The contours of the language localizers' contrasts, thresholded at p-value < 10e-3, for both auditory and visual modalities are presented in Figure 6.1. We also show in Figure 6.2 the coverage of Mahowald et al. parcels[118] by the thresholded language localizers for all

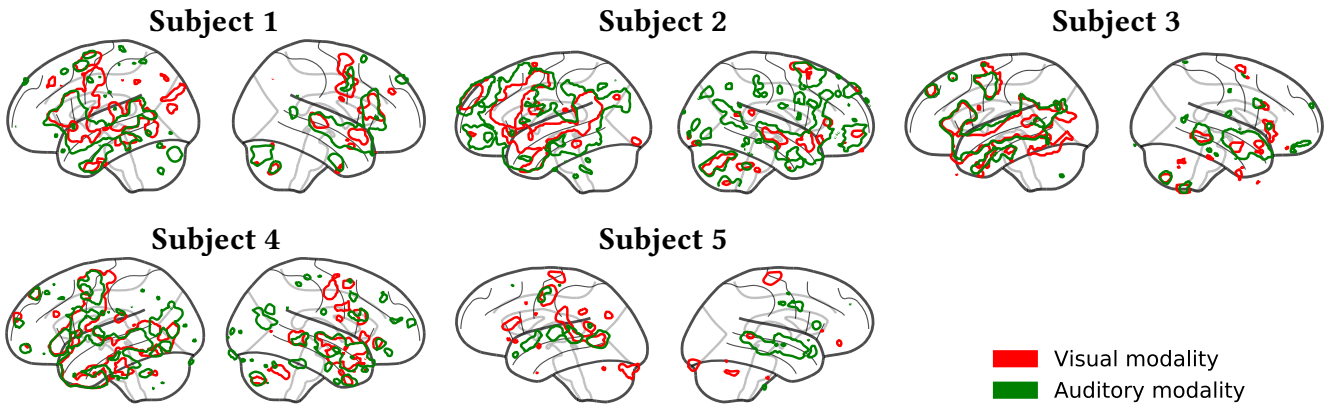


Figure 6.1: **Language localizers:** We show left and right hemispheric contours of the language localizer contrast of word sequences over control stimuli (consonant strings or scrambled recordings), thresholded at a p-value < 10e-3. Statistical images are projected in the anatomical space of each subject.

subjects. We observe a left lateralization of the detected language network with more than 40% coverage of all the language parcels, which covers the fronto-temporal language system that has been well depicted in previous imaging studies[118; 63; 51; 16]. There is variability between the modalities, that particularly disfavours activations of the visual one, in which the subjects can get distracted from perceiving and processing the stimuli more easily, than in the auditory case. This could be expected from the intrinsic variability of different experimental designs in language localizers as demonstrated by Mahowald et al.[118]. *Subjects 1 and 5* have a deficient coverage that will diminish our capacity to interpret syllabic representation effects along their cortex. In particular *Subject 5*, who reported concentration problems, have an extremely deficient coverage of the language network.

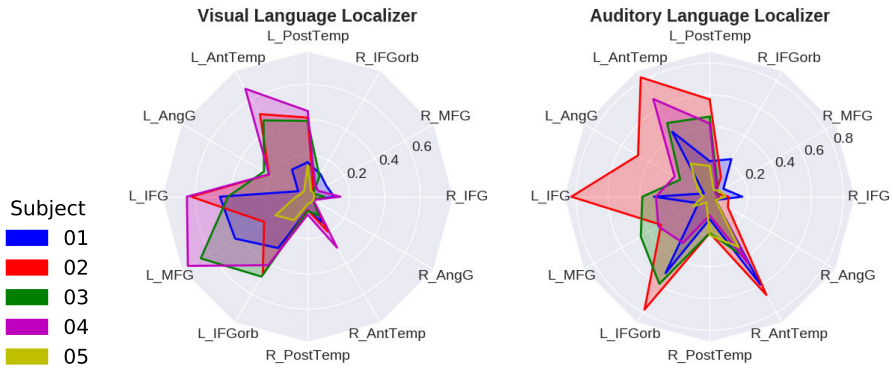


Figure 6.2: **Language localizer parcel coverage:** We show the parcel coverage of each language localizer for the 6 language parcels derived by Mahowald et al. in both hemispheres. Each subject is represented in a radial chart to emphasize the overall coverage of the language localizers of each subject. Also the left and right hemisphere parcels have been arranged symmetrically in the radial charts.

Motor activations: We verified the integrity of the activation maps of the *CVCV Pseudowords presentation* with statistical tests portraying the left and right hand button press contrast. Z score maps of the left over right button press contrast, for all subjects, are shown in Figure 6.3, confirming a good statistical separation of hand responses.

We also verified that we can employ a Support Vector Classifier (SVC) to

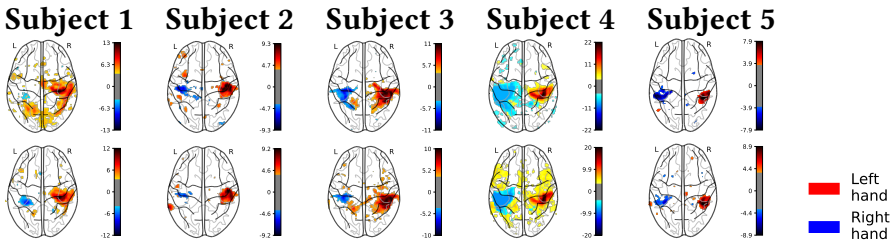


Figure 6.3: **Button press effects:** We show the left button press over right button press contrast Z scores from the auditory modality, thresholded at p < 10e-4, for all subjects. Statistical images correspond to the anatomical space of each subject.

distinguish left and right button press average activation maps derived from the *CVCV Pseudowords presentation* General Linear Model (GLM) runs. There were in total 32 maps for each condition corresponding to one map per run per session (8 runs in 4 sessions). As can be seen in Table 6.2, we achieve high classification scores of right and left button press events for all subjects. Moreover, the classification generalize across sensory modalities.

(Train, Test) Subject	(V, V) (%)	(A, A) (%)	(V, A) (%)	(A, V) (%)	(V-A, V-A) (%)
01	84.38***	93.50***	80.84***	76.94***	90.88***
02	95.38***	92.50***	84.03***	93.03***	95.31***
03	98.00***	99.00***	93.91***	98.75***	100.00***
04	97.38***	99.50***	97.50***	98.72***	100.00***
05	86.62***	77.62***	90.12***	74.28***	93.56***

Table 6.2: **Classification of left and right button press maps of *CVCV Pseudowords presentation*:** *"V"* corresponds to the Visual modality and *"A"* to the Auditory modality. *"V-A"* corresponds to pooling together both datasets for training and testing.
chance: 50%
** : p < 10e-2,*
*** : p < 10e-3,*
**** : p < 10e-4*
Bonferroni corrected for 25 similar tests performed

Visual activations: From the statistical tests performed in the GLM beta maps, of syllable position effects and position interaction, we observed that the statistical effects (any syllable difference) in left and right syllable position in the Visual hOc1 region corresponded to inversed hemispheric projections. In the experimental design we asked the subjects to fixate a centered green dot before stimuli presentation. The inversed hemispheric effect can be seen in Figure 6.4, where *Subjects 1 and 4* have the clearest retinotopic activations.

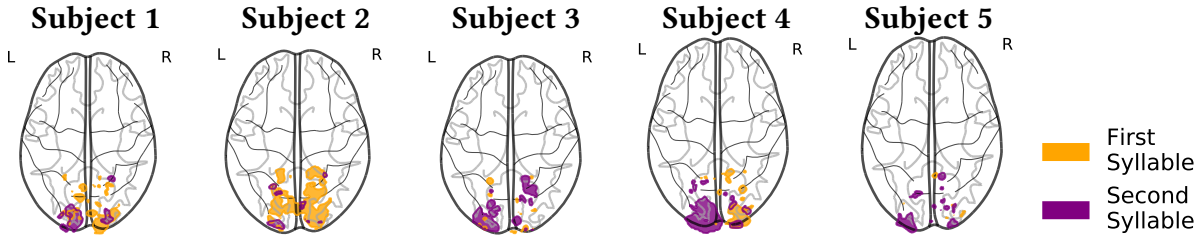


Figure 6.4: **Retinotopic effect:** We show first and second syllable position effects masked by the Visual hOc1 region, thresholded at a p-value < 0.005. Statistical images correspond to the anatomical space of each subject.

6.3 Superposed semi-local representations in Visual region (hOc1)

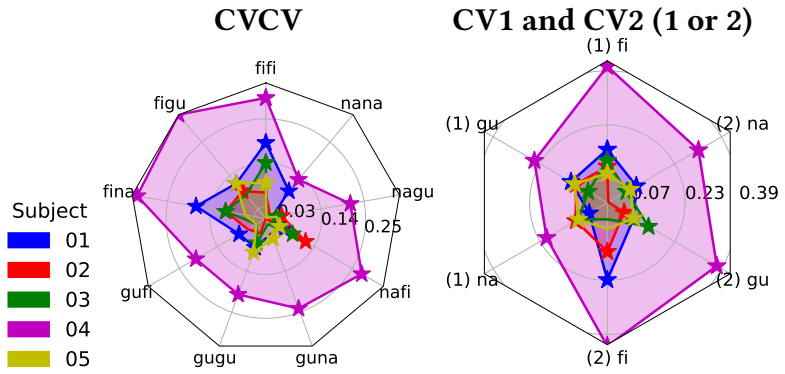


Figure 6.5: **Accuracy in Visual-h0c1:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

We obtained significant classification scores for almost all condition categories in all subjects, with subject 4 having an exemplary performance, distinguishing significantly all conditions in all classifiers. We show in Figure 6.5 accuracy scores from which chance baseline was substracted for each condition. All the classifiers were trained on the visual stimuli and were not able to generalize to auditory stimuli, as would be expected from primary visual areas. Significant scores are marked with a star in case of a p-value < 0.05. We observe, from the relative area of accuracy above chance, that we could decode syllables in each position and pseudowords best in Subjects 1 and 4. Moreover Subject 5, that reported problems with attention, had the worst classifier performance.

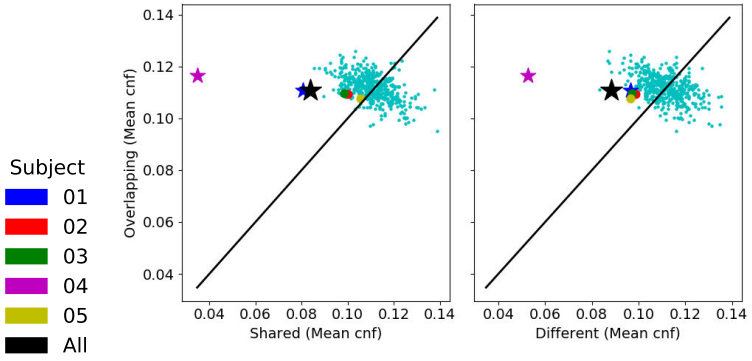


Figure 6.6: **Superposition test in Visual-h0c1:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. A star means significance with a p-value < 0.05. The pattern of all Subjects support superposition, where Subjects 1 and 4 and the group are significant.

We also observe in Figure 6.6, evidence in favour of superposed representations, as all subjects have a higher mean confusion values on pseudowords with position overlapping syllables. Subjects 1 and 4, that had the highest classification scores, as well as the group as a whole have a significant mean confusion vector, with significance given by a p-value < 0.05. We also observe in Figure 6.7 significant segments of semi-local

representations in all Subjects except Subject 5. The best segment belongs to Subject 4 that had the most accurate models. More details about decoding performance in this region can be verified in the Appendix section A.1.

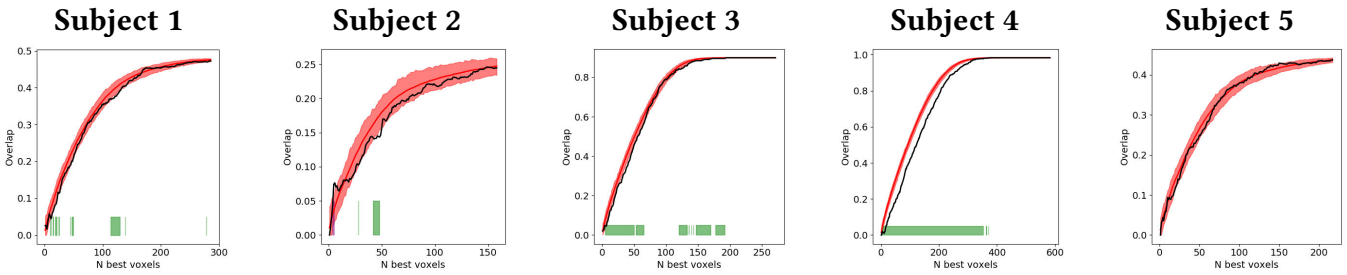


Figure 6.7: **Locality test in visual regions:** We show in black the overlap of the "N" best voxel sets given by the two syllable position classifiers. In red we show the overlap distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05 with respect to the shuffled distribution.

6.4 Superposed semi-local representations in anterior auditory regions (Te12)

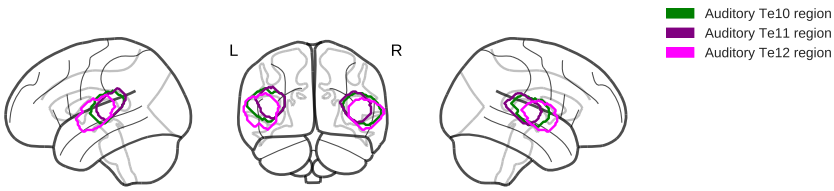


Figure 6.7: **Auditory regions projected on *Subject 1* anatomy:** Contours are shown for the projected auditory regions. Area Te12 extends from Te10 towards anterior regions while Te11 extends to posterior regions. The brain glass template contours were adapted to the T1 anatomy of the subject.

We trained separate models for the auditory hierarchy of regions, shown in Figure 6.4. We observed significant classification results in all regions for all Subjects. In all CVCV models there were 5 or less pseudowords with individual significant accuracy scores and 4 or less syllables with significant scores considering the CV1 and CV2 models together. Nonetheless the level of classification in auditory areas was far less than that obtained in visual areas and for any subject only five or less pseudowords had an individual significant accuracy score. In Figure 6.8 we show the high variability in accuracies in some conditions with respect to others in the CVCV models of all regions.

The more anterior auditory region Te12 shows evidence in favour of superposition at the group level, contrary to Te10 and Te11 that show no particular pattern. We show in Figure 6.9 the pattern change with respect to superposition from region Te10 to Te12. Moreover while region Te10

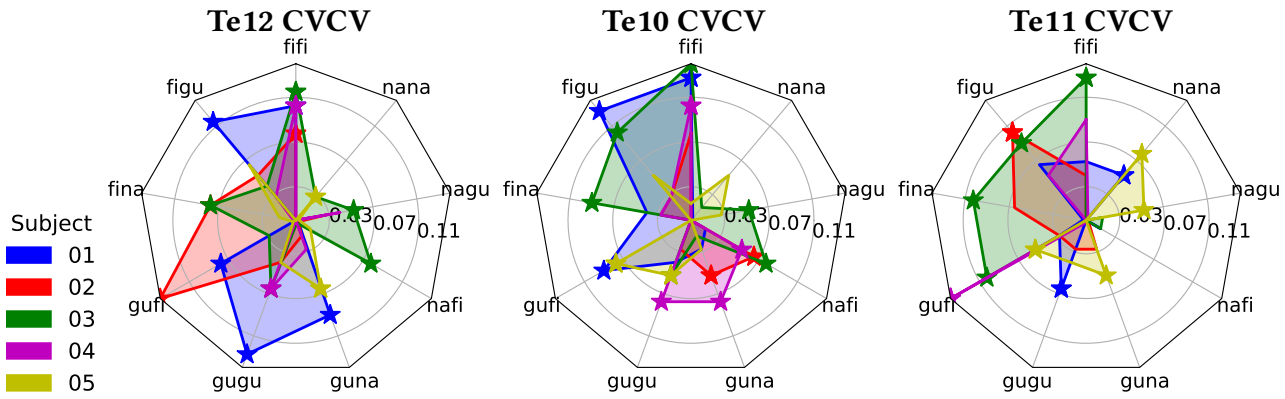


Figure 6.8: **CVCV accuracy in auditory regions:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for all pseudoword conditions. The accuracy points are denoted with stars whenever they are significant. Significance represents a p-value < 0.05 derived from the shuffled models. We present first the most anterior region and last the most posterior region.

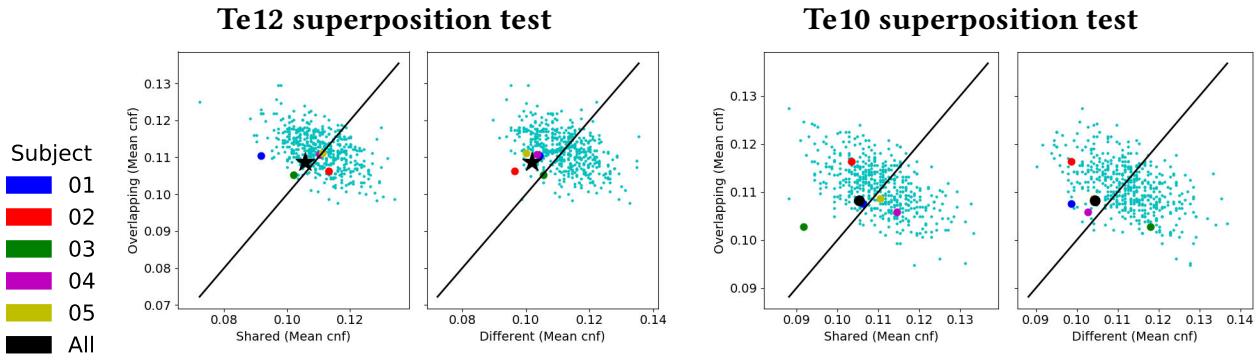


Figure 6.9: **Superposition test pattern change from Te10 to Te12:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

displays significant segments of distributed representations in Subject 5 and 1, region Te12 changes in behavior to display important segments of semi-local representations in all Subjects except Subject 4. We show in Figure 6.10 the locality test pattern changes in Subjects 1 and 5 from Te10 to Te12. Region Te11 shows no particular distinctions from Te10.

In summary more anterior auditory regions seem to encode semi-local superposed representations of syllables. More details on the locality test plots of regions Te10 and Te12 can be verified in the Appendix Figures A.36 and A.42 respectively. More details on the decoding models of the regions Te10, Te11 and Te12 can be checked in the Appendix sections A.12, A.13 and A.14 respectively.

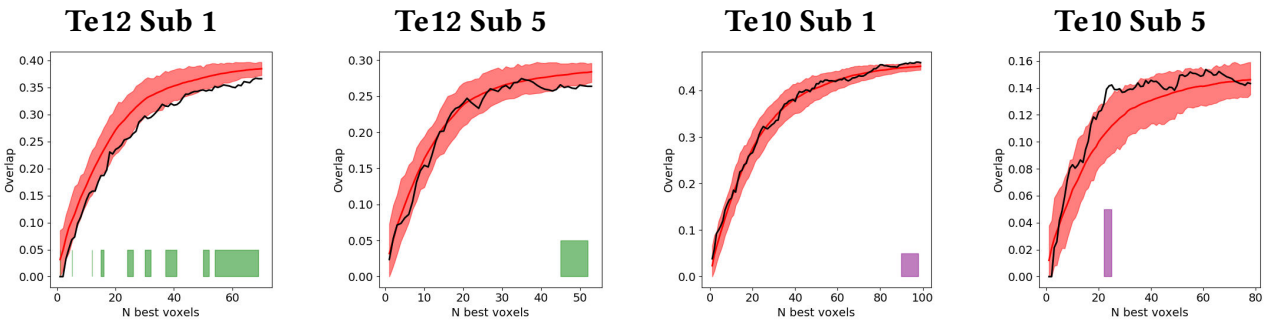


Figure 6.10: **Locality comparison between Te10 and Te12:** We show in black the overlap of the "N" best voxel sets given by the two syllable position classifiers. In red we show the overlap distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05 with respect to the shuffled distribution.

6.5 Superposed distributed representations in Broca's complex

We observed significant classification scores ($p < 0.05$) in all Subjects and sensory modalities for Broca 44 and Broca 45, except for the auditory dataset of Subject 4 in Broca 44. In all CVCV models there were 5 or less pseudowords with individual significant accuracy scores and 4 or less syllables with significant scores considering the CV1 and CV2 models together. No model generalized from the sensory modality in which they were trained to the other. Both Broca regions have some significant subject or group effect in favor of superposition in at least one sensory modality, while the rest of non significant patterns are coherent with superposition as well. In Figure 6.11 we show the significant patterns in favour of superposition for the visual modality in Broca 44 and the auditory modality in Broca 45.

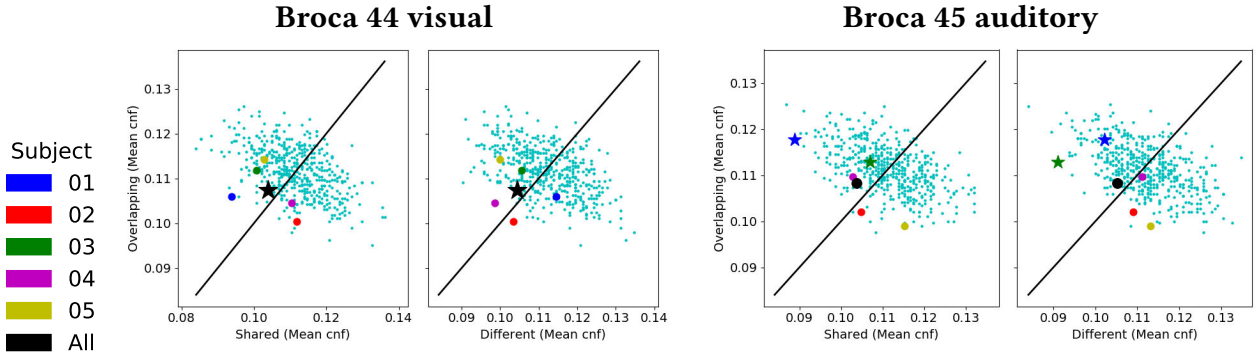


Figure 6.11: **Superposition tests in the Broca's complex:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

In the case of the locality test, in the visual modality, Subject 2 displays a segment of significant distributed representations in Broca 44 and in the

auditory modality Subjects 3 and 5 in Broca 44 and Subject 1 in Broca 45. We show the distributed representation segments of the subjects in Figure 6.12. Alongside the significant segments appreciated there are no subjects displaying strong patterns of semi-local representations, which lead us to interpret representations in the whole Broca complex as distributed. More details on the locality test plots of Broca 44 in visual and auditory modalities and Broca 45 in the visual and auditory modalities, can be verified in the Appendix Figures A.27, A.63, A.30 and A.66 respectively. More details on the decoding models can be checked in the Appendix sections A.9, A.21 A.10 and A.22 respectively.

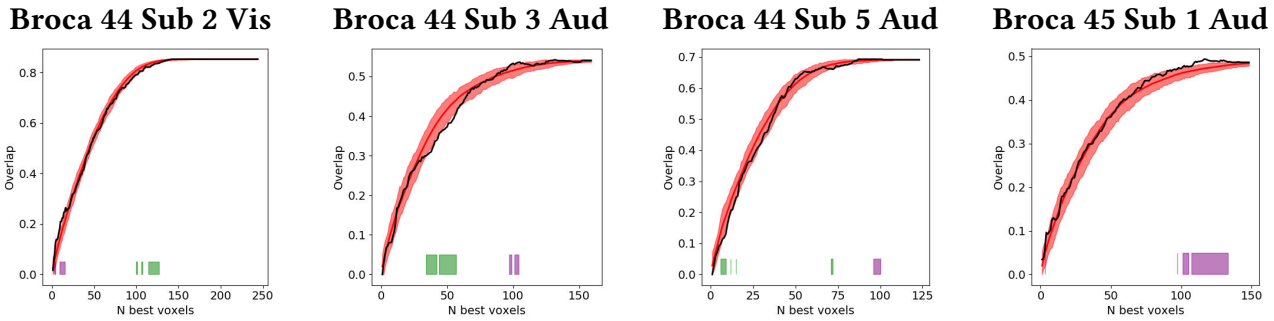


Figure 6.12: **Distributed representations in Broca's complex:** We show in black the overlap of the "N" best voxel sets given by the two syllable position classifiers. In red we show the overlap distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05 with respect to the shuffled distribution.

6.6 Weak evidence for non additive representations in the VWFA

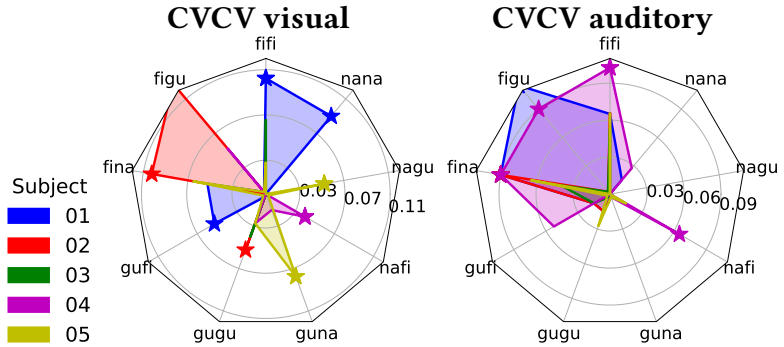


Figure 6.13: **Accuracy in VWFA:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. We show at the left the CVCV model of the visual modality and at the right the CVCV model of the auditory modality. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

We observed that few Subjects had significant accuracy scores in the CVCV model, with few significant pseudoword individual accuracies, as can be seen in Figure 6.13. There seems to be a bias in the models towards pseudowords containing the syllable "fi", which is particularly emphasized by the accuracy

score patterns of Subjects 1 and 4 in the auditory CVCV model. No model generalized from the sensory modality in which they were trained to the other.

Although not significant, patterns of the superposition test suggest evidence against superposition in this area, supporting instead a non additive model. We observed in Figure 6.14 that the mean confusion between pseudowords with overlapping syllables is less than that of pseudowords with shared syllables or different syllables for all Subjects. We did not find substantial patterns in the locality test to support semi-local or distributed representations. More details on the visual and auditory decoding models can be checked in the Appendix sections A.2 and A.11 respectively.

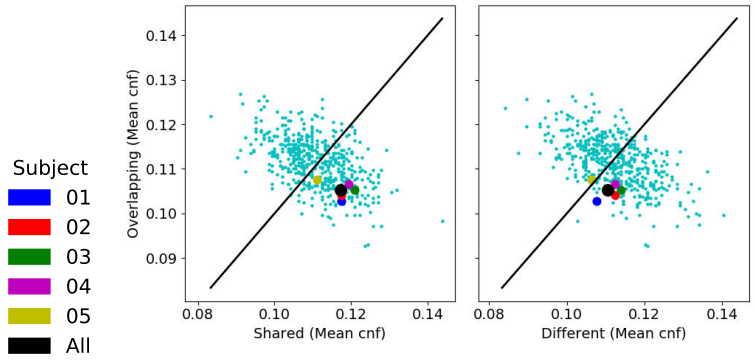


Figure 6.14: **Non additive representations in VWFA:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

6.7 Bimodal distribution of pseudoword accuracy scores

A recurrent pattern in the pseudoword accuracy scores of the CVCV models with significant overall accuracy was that some of the pseudowords (4 or more) would have a non significant accuracy very close to chance levels, while the significant ones seemed to have extremely better accuracy values. This motivated us to verify the distribution of pseudoword accuracy scores from the CVCV models of regions that demonstrated evidence in favor of superposition, namely the auditory region Te12 and Broca's complex. In Figure 6.15 we show that is possible a bimodal distribution describes the accuracy scores according to their significance. From the three regions there were in total 24 significant models (with p-value < 0.05) and from these models we considered separately the accuracies of pseudowords that were significant, shown in red, and those that were not, shown in blue.

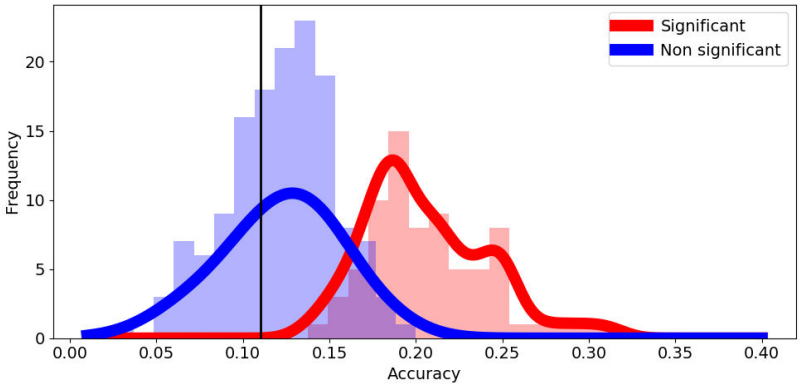


Figure 6.15: **Possible bimodal distribution of accuracy scores in superposition regions:** From the regions Te12, Broca 44 and Broca 45 we considered all Subject CVCV models for which the overall accuracy was significant with a p-value < 0.05. In total there 24 models were significant. From these models we considered separately the individual accuracies of pseudowords that were significant with a p-value < 0.05, in red, and those that were not, in blue. A bimodal distribution appears to describe the accuracy scores categories. The black line indicates the 0.11 chance level of classification.

6.8 Final remarks

So far we have not mentioned results related to the language constituency regions extracted by *Pallier et al.*, namely aSTS, TP, TPJ, pSTS, IFGorb and IFGtri. The reason can be verified in the decoding model details provided in Appendix A. All these regions have very low accuracy scores, with only few subjects showing significant accuracy scores in a few conditions, which adds difficulties to the interpretation of any patterns in the locality or superposition tests. Moreover their superposition tests are inconsistent, for example Subject 4 has a significant value against superposition in the visual modality of IFGorb, but every other Subject, although not significant, follow a pattern that would be congruent instead with superposition. We show this inconsistency in Figure 6.16. It seems that we were not able to decode well the bi-syllabic pseudoword representations in any region along the temporal lobes and we did not find any CVCV, CV1 or CV2 model that generalized their predictive power across sensory modalities.

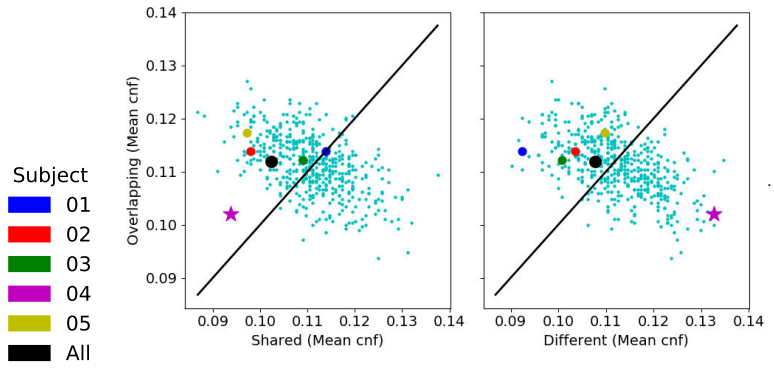


Figure 6.16: **Inconsistent evidence in IFGorb:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

To summarize, although we were not successful at decoding in several language regions, we found several effects of representations in anterior regions, namely the auditory Te12, Broca 44 and Broca 45 shown in Figure

6.8. All this regions provided evidence in favour of superposition and demonstrated support for different levels of locality in representations, where Te12 strongly supported semi-local representations while Broca's complex pointed at distributed representations. Moreover we found in these regions evidence for a bimodal distribution of the particular pseudoword accuracy scores. Finally we provided weak evidence for non additive representations in the VWFA and modulation of the region by auditory stimuli.

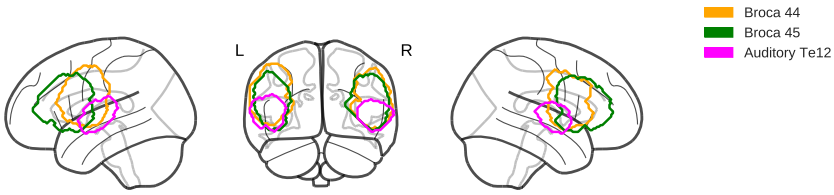


Figure 6.16: **Superposition regions projected on *Subject 1* anatomy:** Contours are shown for the projected superposition regions. The brain glass template contours were adapted to the T1 anatomy of the subject.

7 Discussion

In this chapter we interpret the results obtained from the analysis of sensory and language related representations of pseudowords. Then we comment on the decisions and limitations of the experimental design. Finally we provide our perspective for further experimentation to test Smolensky's superposition principle and other properties of tensor product representations.

7.1 Results interpretation

We expected, from well known retinotopic effects in the primary visual cortex [180], to see an hemispheric partition of left and right syllable position effects, such that left syllable effects would be emphasized in the right hemisphere and right syllable effects in the left hemisphere. This was the case, but we could also appreciate in the images that some subjects did not manage to completely follow the fixation instruction, since effects of both positions were present together in both hemispheres. It could have been useful to have eye tracking recordings to be able to account in visual areas for gaze position when evaluating representations, nonetheless this was not crucial for us since superposition of representation in visual areas is known and was just looked for as a quality check of the activation maps. The size of the text presented in the *Pseudowords matching task*, around five horizontal degrees, allowed us to induce enough spatial spread in the voxel activations, due to retinotopic mapping. Thanks to this we obtained high classification results across all pseudoword conditions, even while ignoring effects of gaze movement.

In the case of auditory representations, since we maximized the featural distance between consonants and between vowels[20], we expected to be able to decode syllables with higher accuracies than the ones we observed. Nonetheless other experiments in which syllable representations have been decoded employed fast sparse protocols that allowed presentation of syllables during a silence gap[62]. Our decision of not employing a different acquisition protocol across sensory modalities, to have comparable results in abstract representations, compromised our capacity to perform decoding in the auditory cortex.

Still our decoding results are coherent with evidence of hierarchical organization of the auditory and speech pathways. Previous evidence suggest that speech-specific responses to isolated syllables are only observed in later stages of processing[90; 112; 183], and we decode syllables better on the Te12 more anterior region of the primary auditory areas. Moreover we were able to demonstrate evidence in favour of superposition and semi-local representations in this region.

We also found evidence for superposition in the Broca's complex (Broca 44 and Broca 45), that had the best significant accuracy scores from the language related regions tested. It has been shown through a series of neuroimaging studies pooled in a meta-analysis[206], that Broca 44 is consistently engaged with syntactic binding operations, alongside the posterior superior temporal sulcus (pSTS) and the superior temporal gyrus (STG). In the same metaanalysis it is argued that Broca 44 is a pure syntactic processor, while pSTS and STG integrate syntactic and semantic information. The fact that we also find evidence for distributed representations in the Broca's complex turns it into the most promising region to further test Smolensky's tensor product representations.

In the rest of the language regions extracted from the study of Pallier *et al.*[154], for which semantic and syntactic coherence effects of constituency were demonstrated, we were not able to find any clear patterns to report and most accuracy scores were insignificant. In the case of the regions aSTS, TP and TPJ that were only sensitive to semantic coherence, not finding pseudoword representations could be expected. On the other hand pSTS, IFGorb and IFGtri were also sensitive to syntactic coherence, so we considered the possibility of finding pseudoword representations. The fact that we did not find any significant representations in these regions could be explained by the claims of the meta-analysis of Zaccarella *et al.*[206], in which IFG was not particularly linked to binding operations and pSTS was linked to the integration of syntactic and semantic information that we lack in pseudowords. Moreover Matchin *et al.*[126] demonstrate that pSTS, IFGtri and IFGorb might be related to top-down syntactic prediction instead of basic syntactic combination. Since we are presenting pseudowords in isolation we would not expect top-down syntactic predictions to take place.

The VWFA, linked to binding of visual and verbal representations in both words and pseudowords, for early stages of language processing[41; 194; 50; 75; 205], showed evidence against superposition or in favor of non additive models. This result goes in hand with the study of Glezer *et al.*[75] that argues against theories of sublexical representation in the VWFA. Moreover the fact that we found significant accuracy scores in the auditory modality supports previous evidence about speech modulation of the VWFA[205].

One important clarification to make regarding evidence against superposition is that such evidence do not necessarily immediately discards Smolensky's model of generalized tensor product representations, but only

its basic version, in which the final composition step of symbolic structures is given by addition. In the generalized version other operations are allowed to take place after construction of the final symbolic structure, like tensor contractions, exemplified by the memory efficient holographic reduced representations of Plate[158]. More analysis of the origin of the non additive pattern observed in the VWFA will be necessary to completely discard application of Smolensky's framework to its internal representations.

We observed an extreme variability between pseudowords with significant and non significant accuracy scores. We confirmed this variability by simply plotting the histograms of the separate distributions of significantly and non significantly classified conditions. We found an approximate bimodal distribution. This result is difficult to interpret without more detailed inspection of the decoding models. Other factors related to model training could influence this result, like the bias introduced for not doing a nested cross validation or the greedy voxel spheres selection approach implemented. Nonetheless we think this result can be explained by a lack of sparsity or variability in the spread of neural activations of the underlying neural unit patterns. Lacking variability in the spatial distribution of activations decreases the probability of finding substantial differences in the aggregated neural activity values of voxels. An example in which underlying neural patterns lead to aggregated activity in voxels that can not differentiate pseudoword conditions is shown in Figure 7.1.

A final unexpected result was the complete lack of generalization between sensory modalities in all classification models. This can be accommodated by two different interpretations. On one hand it is possible that noise in BOLD-fMRI measurements or non unique spatial assignment of neural vectors to neural units do not let us generalize across datasets. On the other hand it is also possible that there are no amodal abstract representations for simple stimuli like bi-syllabic pseudowords. We would require further tests of stability, outliers and to assess generalization across more datasets to confirm which is the case.

7.2 Limitations of the experimental design and methodology

With the objective of testing the superposition principle on syntactic operations of language, we opted for the simplest stimuli we could use as a first approach, namely two-syllabic pseudowords. Nonetheless even with this simple stimuli, due to the nature of BOLD imaging, our experiment suffered from several methodological limitations.

Following Devonshire *et al.*[53] guidelines to counteract possible non-linearities in the mapping from neural activity to the BOLD response, we designed a task to keep a pseudoword in memory to prolong its duration and tried to extend ISI as much as possible, 7 seconds, to still preserve a good sample size of the 9 stimuli conditions, 40 samples per condition per

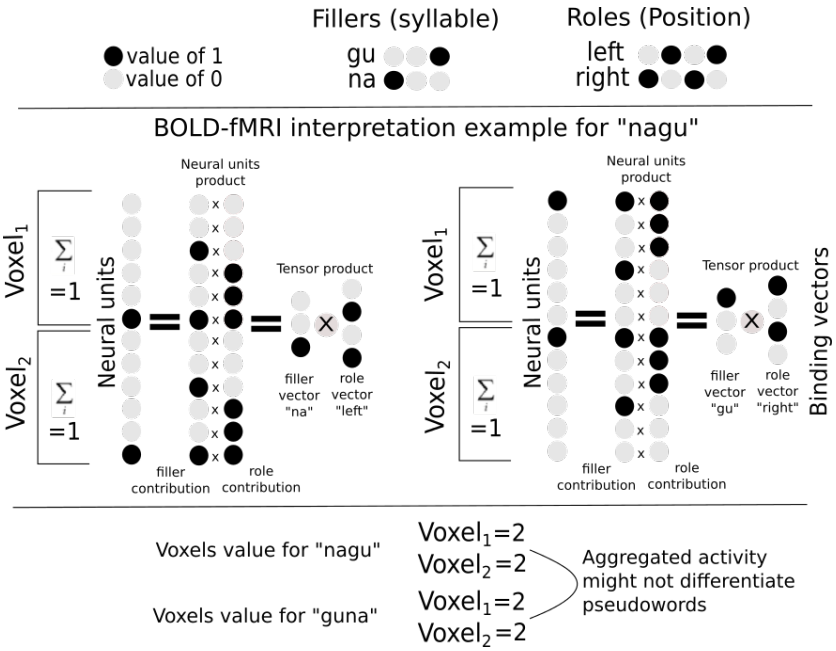


Figure 7.1: **Illutration of superposed tensor product representation in BOLD-fMRI:** We present the example neural vectors of the syllables "na" and "gu" bound to the left and right positions of a bi-syllabic pseudoword. We illustrate how the level of BOLD activity should reflect the aggregated activity of a segment of the neural units that form a representation. The superposition principle implies the sum of the vector values from each binding to give the final total activity in a voxel. The voxel values of the pseudoword "nagu" correspond to the plots of the neural vectors and those of the pseudoword "guna" were derived in a similar way. Due to the effect of aggregation, no voxel in the example permits differentiating the two pseudowords, even though the neural unit patterns are linearly independent between *Roles* and between *Fillers*.

session. Nonetheless this is far from the actual long stimuli durations of 40 s at which linearity was ensured[53]. We consider testing in the future the modulation of BOLD responses to the target stimuli instead of using heuristics to setup design parameters, which will be important to test this kind of model assumptions, that are sensitive to the underlying neural interpretation of the BOLD response.

Regarding the task, the fact that we did not probe every trial limited our capacity to assess attention modulation and outliers of the internal representations. Considering that we did not find a uniform increase in accuracy across the different pseudoword conditions, it could have been useful to assess if the variability in representations could be explained by correlates of attention.

In the case of the decoding methodology, there are several decisions that were made heuristically to save computational resources. For example we did not smooth the data to avoid inducing additional voxel correlations that would complicate interpretation of the feature coefficients of the classifier and to better exploit any extreme effects in particular voxels, but we could have explored the effect of different smoothing kernel sizes. Also for the searchlight voxel selection procedure we fixed the radius of the spheres to 5 mm, which means a 2 voxel radius for our acquisition parameter to search for local effects, and passed the complete spheres to the classifier. Instead we could have also determined empirically how this radius affects classifiers performance. Moreover the fact that we performed a grid search without a nested cross validation could have introduced a small positive bias in the

classification results[33]. To improve classification accuracy and compensate the high number of features in the classifiers, a 100 or more, we decided to ask the same subjects to come for several sessions to increase our sample size, such that we would have at least 80 samples per condition per sensory modality, but with respect to the number of features in classifiers this remains a very small sample size.

7.3 Future perspective

In this work we selected the simplest stimuli possible as a first approach to test the superposition principle in syntactic operations of language, but it will be interesting to go further and test superposition in more complex syntactic stimuli like pseudoword lists and jabberwocky phrases. Nonetheless this would increase the challenges faced when working with BOLD-fMRI by introducing additional variables in the experimental design like stimuli duration, length in terms of number of words and rate of word presentation.

All this additional experimental factors have been shown to induce nonlinear BOLD responses. Saturation from long phrase reading and nonlinear modulation from word presentation rate have been demonstrated[164]. Nonlinear effects of presentation rate have been shown to be similar in words and pseudowords and spatially heterogeneous across brain regions[132]. Also nonlinear effects of stimuli duration have been shown to be spatially heterogeneous[17]. If we expect representations of multiple words to be completely distributed we also have to be careful about the rate of presentation due to possible neural adaptation effects[103]. It will be necessary to study in detail the optimal setup of the mentioned experimental parameters, to diminish or correct the nonlinearities that can affect evidence for additive linear models of composition like the superposition principle.

In our experiment we only found evidence for superposition in a small set of regions located close to each other, namely Broca's complex and the anterior primary auditory region Te12. Considering that there is spatial heterogeneity of BOLD activation patterns across the brain, the best path of action would be to focus future acquisition of images in specific brain regions. Focusing on acquiring only a sub-volume of the cortex can facilitate improving spatial and temporal resolution of the BOLD signal. Moreover the uneven classification of individual pseudowords conditions, that we interpret as lack of variability in the spatial distribution of neural activations at the 3T 1.5 mm isometric resolution analysed, suggest to attempt similar and new experiments at higher imaging resolutions. For example high resolution laminar imaging with boundary based surface registration has been shown to reveal internal visual representations discernible with the bare eyes[111]. In addition, focusing on specific regions facilitates the design of functional localizer paradigms to better segment target regions for analysis and reduction of the amount of voxels (features) provided to decoding models.

Also exploring sub-volumes of anterior brain regions in future experiments suggest to rely more on speech than reading. Since we also found evidence for superposition in auditory regions linked to later stages of speech processing (Te12), it will be interesting to study in detail how the properties of auditory representations change from non additive to superposed and from semi-local to completely distributed. Lack of consideration of the problems introduced by the fMRI acoustic noise greatly diminished the performance of our classifiers. Future experiments should carefully pilot the effect of fast sparse protocols on the study of the properties of representations like the superposition principle, since they add their own constraints to the experimental design[157]. Studying in detail superposition and hierarchical processing of individual pseudowords in auditory regions with laminar fMRI, might be a good first step before continuing the analysis to the Broca's complex with pseudoword lists and jabberwocky.

Regarding our findings in the VWFA, we consider running future tests on this region to confirm in more detail the non additive nature of its representations will be interesting. For future experiments, considering the small size of the VWFA, we recommend also designing a localizer task to delineate with more certainty its location in individual subjects.

Besides further testing the superposition principle, it will also be important to better assess the stability of representations, which we considered was a weakness in our work. We had significant classification scores, but these were still quite low, only around 20%, to evaluate individual representations. We were not able for example, to determine if the bimodal distribution of accuracy scores could be explained by outliers or attention modulation effects. Since neural firing thresholds are known to alter according to arousal state[129], it will be important to include in future tasks processing confirmation of individual stimuli and assessment of attention modulation.

In conclusion, we think we have provided enough evidence for the superposition principle in anterior brain regions to motivate further experimentation based on Smolensky's tensor product representations. We expect to have illustrated well the great challenges behind testing experimentally even the simplest assumptions of this theoretical model. Considering the contrast between the maturity of theoretical models and the lack of empirical tests of their most basic assumptions, we hope to incentivize more work in the experimental direction.

Part III

The neural dynamics of binding in language with the Neural Blackboard Architecture

8 Language binding effects in neuroimaging and the Neural Blackboard Architecture

In this chapter we present some language neuroimaging studies of binding that we consider important and interesting to attempt reproduction with simulation of the Neural Blackboard Architecture (NBA). We also introduce the application of the NBA to syntactic representations in phrases.

8.1 Some language neuroimaging studies of binding

Most linguistic theories assume a constituency property that allows to combine and replace smaller phrases in larger phrases. Since solving variable binding requires an explanation of how to implement links between bits of information - like words and word types - to create basic data structures, like phrases in language, it is likely to also explain how to create links between such basic structures.

Behavioral evidence for constituents in phrases has been around for a while[15; 3], with more recent studies demonstrating the reuse of recently heard syntactic structures through syntactic priming experimental paradigms[19; 23]. But only recently we have started to characterize the detailed neural correlates of constituency and word binding with diverse brain-imaging techniques [141; 64; 25; 54; 12; 153; 11; 117].

We selected The ECoG analysis of Nelson *et al.*[141] as the first study to compare to our model. It is one of the only two studies so far demonstrating spatially specific and temporally detailed neural dynamics of phrase processing, made possible by analyses of intracranial neurophysiological data taken from epileptic patients. Moreover it is the first one to characterize the specific patterns of phrase-structure formation, possibly revealing the first neural signatures of variable binding related operations. Nelson *et al.* refer to them as "merge" operations that combine syntactic objects (word types and phrase types). In the study words were presented sequentially to patients in a screen to be read under a Rapid Serial Visual Presentation paradigm. The task was to keep a phrase of up to 10 words in memory to compare it just after with a

probe sentence composed of 2 to 5 words. We will show that simulation of the NBA portion responsible for variable binding, while only tuned for correct operation, generates strikingly similar temporal patterns of neural activity when aggregating the binding operations corresponding to complete phrase processing, assuming the phrase grammar and bottom-up parsing scheme employed by Nelson *et al.* in their analyses.

As a second study, we selected an fMRI experiment [153] to portray the capacity of the model to capture results from multiple neuroimaging spatio-temporal scales. In this experiment, trials with lists of 12 words obtained by concatenating phrases of a given length, were presented to healthy subjects. Conditions were formed from all combinations of m by n that give 12, satisfying the form n phrases of m words, like 2 phrases of 6 words. Besides normal words, the design also included pseudoword conditions that maintained morphological markers and closed-class (function) words. This allowed the authors to demonstrate a clear separation of syntactic and semantic binding neural activation patterns in language related regions, which is interesting to us, since syntactic specific patterns are the closest to the abstract considerations of binding of our model, assuming the same phrase grammar and parsing scheme employed for comparison with the ECoG results. The authors found a sub-linear pattern of neural activation as the number of constituents increase, which could not be explained by a simple "accumulation" model motivated by measurements of sequence learning tasks in awake macaque monkeys. The Neural Blackboard Architecture predicts this sub-linear effect from the circuit recruitment process required by the number of binding operations, alongside expected patterns of hemodynamic peak onset differences from delay activity considerations.

8.2 The Neural Blackboard Architecture (NBA) applied to language

The details presented in this section are a literal reminder of those already developed in subsections 1.3 and 1.3 of Chapter 1, so in case that chapter was consulted recently we recommend skipping to the next chapter 9. What we present here are only the key aspects of the Neural Blackboard Architecture that must be understood to follow details of the circuit implementation presented in the following chapters. To understand more details about the properties of neural representations in the NBA please consult section 1.3 of Chapter 1.

There are several previous instantiations of sub-circuits of the NBA with varying degrees of biological plausibility, the latest relying mostly on Wilson Cowan population dynamics[52]. Some of the previous simulations attempted to address diverse aspects of language processing, such as ambiguity[67] and learning control from syntactic stimuli[188]. Other simulations addressed circuit implementation issues like how to develop a connectivity matrix with

randomly connected networks[189] and how to implement a central pattern generator sub-circuit for sequential activation [191]

In the following paragraphs we summarize the main abstract mechanisms and assumptions behind the NBA to implement binding operations. A complete illustration of the blackboard architecture is provided in Figure 8.1. For a deeper review we recommend reading a recent paper with a circuit design and examples that focus on sentence processing[48], as well as the original framework proposal introducing abstract combinatorial structures[187].

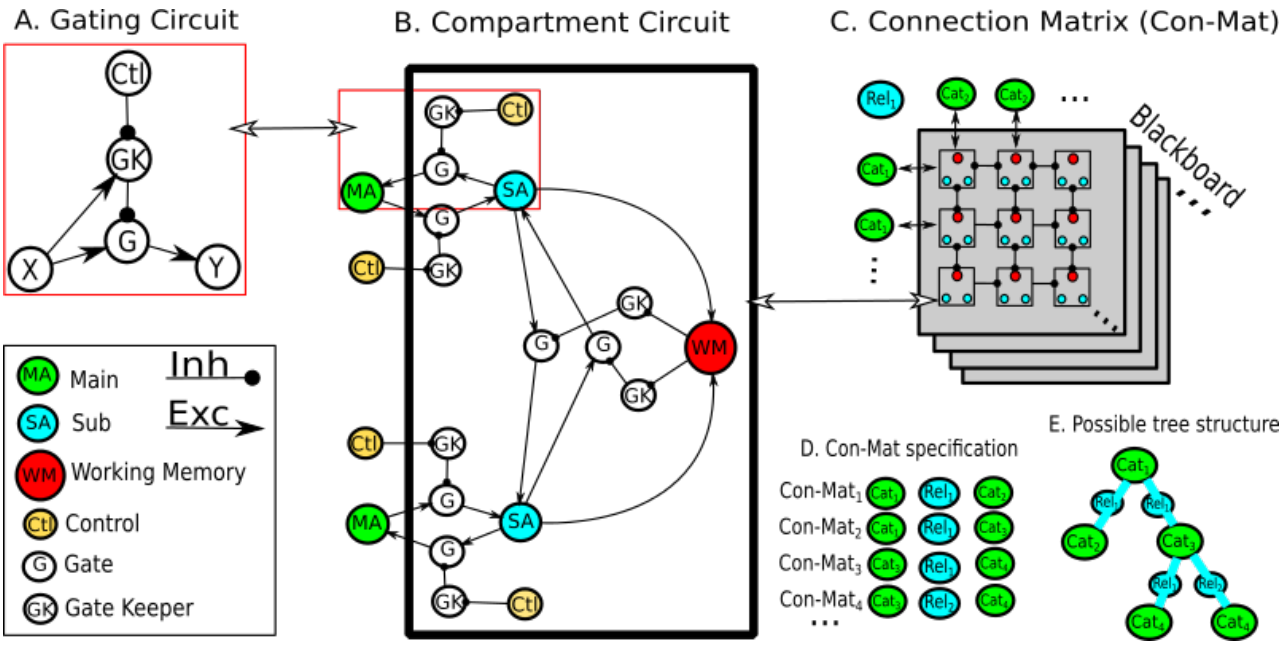


Figure 8.1: **The Neural Blackboard architecture: A.** Gating circuit that allows the implementation of conditional neural activity transfer between Neural assemblies X and Y through a gate assembly. The gate keeper assembly (GK) is activated by the X assembly and then inhibits the gate assembly (G). To let information flow through the gate assembly, a control assembly (Ctl) must therefore inhibit the gate keeper assembly. **B.** Architecture of a single compartment circuit of a connection matrix. Six gating circuits are arranged in a way that makes conditional bidirectional neural activity flow between two main assemblies possible. Control assemblies regulate the direction of information flow and allow the activation of sub assemblies. The two sub assemblies excitep the working memory assembly which, once activated, encode the binding of the main assemblies and allow activation to flow between them if the controls allow it too. **C.** Each connection matrix contain n by m compartment circuits that encode the same relationship type between the same pair of assembly categories. There are m available assemblies for one category and n available assemblies for the complementary category and only one cell circuit can activate its working memory assembly to link two particular assemblies due to mutual row and column inhibition of cells in the connection matrix. The size of the connection matrix effectively represents memory limitations. A blackboard is composed of an arbitrary number of connection matrices that encode different relationship types for a pair of assembly categories. **D.** A blackboard is composed of multiple connection matrices, where each of them is defined by two node categories and a relationship type between them. **E.** Example of a possible tree structure that can be represented based on the specified connection matrices.

Nodes in Figures 8.1.A and 8.1.B represent neural assemblies that can be interpreted as linked spiking neural populations. The most basic component of the NBA is a "Gating Circuit" illustrated in Figure 8.1.A. The main idea is that neural activity would flow from the assembly X to the assembly Y, but is blocked by the Gate Keeper (GK) assembly, which itself is excitepd by assembly X. So to allow directional activity flow from X to Y, a Control (Ctl) assembly has to inhibit the GK assembly. Notice that it is trivial to extend the gating circuit for bidirectional control of activity flow as illustrated in Figure 8.1.B. Introducing bidirectional conditional control signals is what gives the NBA the possibility of implementing separately queries like 'what follows X?' or 'what follows Y?'.

Another basic component of the NBA is a proposal for working memory (WM). Persistent neural activity in response to stimuli is considered to be the neural process underlying active (working) memory, and its implementation is hypothesized to be based on excitatory reverberation[199]. Based on this, the NBA considers a Delay Activity[45] mechanism as a biologically plausible implementation of WM. It consists on a neural assembly, that after being excited beyond a certain threshold, achieved by the co-activation of input populations, will maintain a constant amount of activation for a short period of time. By maintaining its activity, WM acts as a short lived bidirectional link between two assemblies. This mechanism can be considered as the creation of an implicit pointer from one assembly to the other, such that future reactivation of one assembly can be driven from the other to perform query operations. This conforms a "Memory Circuit" as depicted in Figure 8.1.B.

Two bidirectional "Gating Circuits" connected by a "Memory Circuit" form a "Compartment Circuit" capable of implementing variable binding and query operations. The key point of this circuit is that Main assemblies (MA), representing grounded concepts or instances of variables types, activate Sub assemblies (SA) if a control signal driven by another mechanism allows it. Then co-activation of SAs is what realizes a temporary binding of MAs by activating WM. So one "Compartment Circuit" models specifically the neural activity of a variable binding operation. It is operated by a mechanism that drives control signals simultaneously in multiple "Compartment Circuits" to instantiate binary tree like data structures on which query/unbinding operations can be performed later.

As might be evident by now, applying the NBA to syntactic processing in language consists of two simple assumptions. First, equating the parsing mechanism to the control mechanism that coordinate binding events of words and word types and phrase types. Second, determining the number of compartment circuits necessary to instantiate a complete syntactic structure and the content of MA nodes from a grammar theory. In this work we will only employ a phrase grammar and bottom-up parsing scheme following theoretical assumptions of selected neuroimaging experiments. Nonetheless, a promising feature of the NBA is that it has the flexibility to test any arbitrary

parsing mechanism incorporating top-down considerations and an important variety of alternative theories of grammar based on binary trees. For example dependency grammars that assume multiple direct word bindings instead of the hierarchical phrase bindings modelled in this work have been employed in previous simulations[188].

To understand how a sentence is processed in the NBA, let us consider first the simplest case of binding two words, like "Sad student", belonging to grammatical categories instantiated in the MAs of one "Compartment Circuit", such that one MA is an "Adjective" corresponding to "sad" and the other one is a "Noun" corresponding to "student". The MAs activate with timings corresponding to word presentation, so we are assuming that words were recognized to motivate their corresponding instantiated grammatical categories before we attempt to link them. Then an assumed parsing mechanism determines that a link operating on "Adjective" and "Noun" types is necessary in the blackboard, driving activity in several "Compartment Circuits" from which only one, that we consider as the recruited "Compartment Circuit", completes co-activation of SAs to drive WM and realize binding between the word types.

In the case of a complete phrase, like "Fat sad student", if we are assuming the instantiation of phrase types that form a hierarchical tree theorized by a phrase grammar, then the time at which the binding of the instantiated grammatical categories of "sad student" takes place would be the time at which a "Noun Phrase" is activated and bound to the "Adjective" corresponding to "Ten".

Finally, a "Connection Matrix", portrayed in Figure 8.1.C, allows the implementation of a complete "Blackboard". It contains variable type relations learned by the "Blackboard" as sets of mutually inhibitory "Compartment Circuits" that enable the selection of the "Compartment Circuits" requested by the control mechanism. We portray the "Blackboard" as a regular grid for illustrative purposes, although there is already a proof of concept implementation with randomly connected networks[189]. Nonetheless in this work we will ignore the "Connection Matrix" dynamics by considering the "Compartment Circuits" as individual isolated circuits, since we lack information to form hypothesis about the size of the Blackboard, total number of Connection matrices and other important parameters. Simplifying our simulation by ignoring the "Connection Matrix" dynamics should only affect substantially predictions on language processing variables unrelated to binding, like memory constraints, which we do not explore in this work.

To implement a general syntactic control mechanism, although challenging, should be feasible, as suggested by the Feed-forward artificial neural networks employed in previous NBA simulations [188] and recent state of the art feed-forward network architectures that have shown top performance for diverse language parsing tasks [6]. Moreover a more recent proposed extension of the NBA, that imitates the motor circuit of the marine mollusc

Tritonia diomedea, shows how to generate patterns for sequential activation control[191]. Nonetheless we considered that simulating the higher level mechanisms of control is a task out of the scope of this work, since we focus specifically on reproducing the neural signatures of variable binding operations.

9 Simulation setup of the Neural Blackboard Architecture

In this chapter we present the architectural decisions of the simulation, how we determined the diverse parameters of the Compartment Circuit of the Neural Blackboard Architecture (NBA) and the experiments performed to tune the circuit for correct binding operation.

9.1 NBA simulation

Previous simulations of the NBA approximate the mean activity of neural assemblies with Wilson Cowan dynamics [67]. Nonetheless, as explained in Chapter 2 Section 2.2, direct simulations of leaky-integrate-and-fire (LIF) neurons [150] have different transient behaviour than the dynamics described by the Wilson Cowan equations.Since we are interested in modelling the transient dynamics of variable binding in order to compare the simulation with real temporally detailed patterns of intracortical neural measurements like ECoG, we feel the need to model spiking neuron dynamics is important.

The decision to use AdEx, rather than LIF neurons has two motivations: first, adaptation is ubiquitous and its inclusion has a substantial impact on the dynamical range allowed within the constraints of the blackboard architecture. Second, it has been shown that 2D models, like AdEx, can already predict correctly 96% of the spikes of detailed conductance models[27]. Also, this model reproduces many known electro-physiological features, as can be appreciated in the spike-frequency adaptation review of Benda *et al.* [13; 14]. Our approach is consistent with a trend towards simpler, geometrically motivated 2D models that preserve the essence of more complex biophysically motivated models [97].

AdEx is now available in MIIND. To our knowledge this is the first time that the AdEx model will be employed to approximate the neural dynamics of a circuit of this magnitude reproducing cognitive function.

In the case of Delay Activity (DA) populations like Working Memory (WM), we decided as a first approach to model such a mechanism phenomenologically. We plan to address the different alternatives to model persistent cortical activity with interacting neural populations in future work. As suggested by de Kamps[45] not only models of recurrent excitation but also recurrent inhibition

can account for this phenomena. In the current simulation, a constant firing rate for DAs is kicked off by a specified level of input, resulting in activation that is sustained for a predetermined period of time. Contrary to previous simulations [186], we do not consider Sub-Assemblies (SAs) as DA populations. We find that SAs can show rich and interesting dynamics just by fulfilling their function of mediating activation for WM.

We model Main-Assemblies (MAs) as receiving input from DA populations, representing word types in some cases, and WM populations representing phrase types in other cases. We do this to satisfy the assumptions of a phrase grammar that requires representation of deep tree hierarchical structures, so that we can separate the notion of a phrase resulting from previous word type bindings stored in WM, from the recruitment of MAs representing word grammatical category instantiations that take place during sentence processing. Note that for other grammar types, like dependency grammars considered in previous NBA simulations[186], to consider words as nodes in their syntactic representations, we only need to model word types for the MAs of the necessary compartment circuits.

9.2 Compartment circuit parameters

The compartment circuit contains two different types of neural populations. Artificial neural populations following a boxcar event model, shown in Figure 9.1.B and biological neural populations following LIF or AdEx neural models. We took LIF parameters from Omurtag *et al.* (2000) [150] and AdEx parameters from Brette and Gertsner [26].

As a first step we wanted to only explore the general behavior of the circuit of neural populations following well studied sets of parameters. Nonetheless it is clear that studying the neural dynamics of specific brain regions might require adapting the parameters of the neural models to local measurements. Each neural population is either excitatory or inhibitory; this means that a population that is excitatory (inhibitory) on one population is excitatory (inhibitory) on others as well, respecting Dale's law.

The dynamics of most populations are given by the PDTs and ultimately determined by the underlying model of spiking neurons. These neural populations comprise a pair of Main Assemblies (MA), a pair of Sub Assemblies (SA), six Gate Assemblies (G) and six Gate Keeper Assemblies (GK).

Nonetheless there are a few other populations for which we simplified the simulation to the phenomenological level with an imitation of Delay Activity, which means that, after transient stimulation, a population retains its activation above a certain threshold for a given period of time. For instance, the biophysical mechanisms of WM are still not understood completely, but its characterization as Delay Activity is relatively uncontroversial. We modelled in this way, Control assemblies (Ctl), Working memory assemblies (WM), Event Input Assemblies (Inp) and a Baseline Assembly (B) that drives baseline

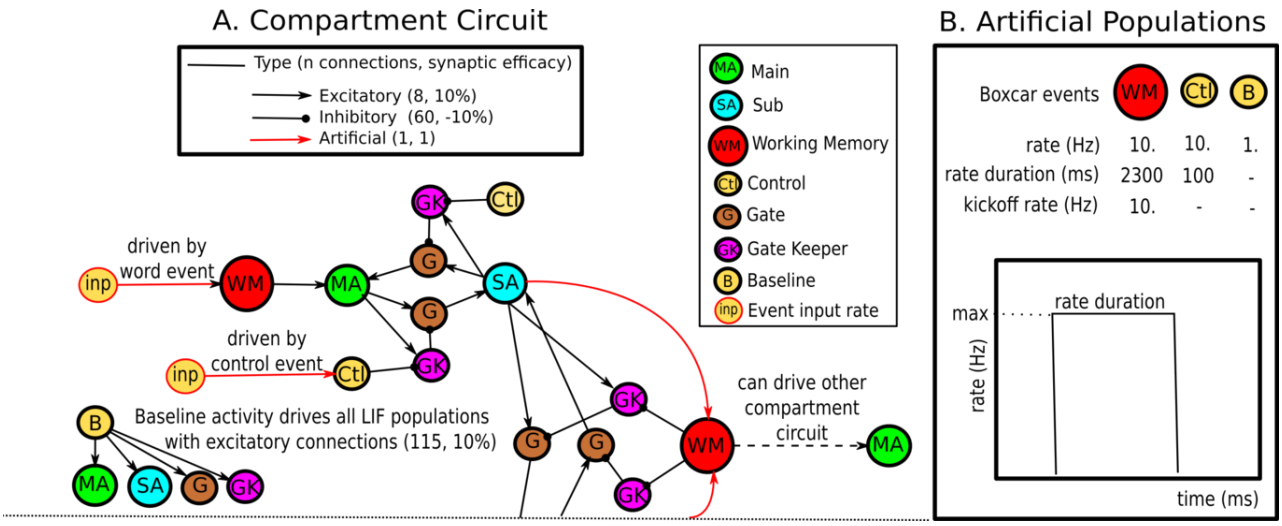


Figure 9.1:**Compartment circuit example:** A. Details of the Compartment Circuit implementation. Only half of the circuit is shown since the design is symmetric. The baseline (B) and Event input (Inp) populations are part of the simulation and not of the original abstract circuit proposal. B. The behavior of the artificial neural populations and their selected parameters is shown

neural activity of all completely simulated neural populations. A complete diagram of the compartment circuit with example parameter values for LIF populations is given in Figure 9.1.

We use a boxcar event model for persistent activity. This model requires specification of the starting point of events, the persistent firing rate of the population and the duration of the persistent activity. In the case of the Delay Activity of WM we also have to provide a kickoff input rate threshold that automatically triggers the boxcar event instead of providing a start time point. The duration of persistent activity was pragmatically set up long enough for the neural dynamics to reach steady state and allow the formation of all required bindings between phrase types and word types. Finally the persistent activity rate and kickoff rate threshold were arbitrarily selected from possible parameter range values as a result of simulations of the circuit dynamics that will become clear in the following section.

Selecting firing rates to tune the compartment circuit is a complex task given the contrast between the extremely simplified circuit and real neural networks that contain multiple types of neurons with diverging behavior across cortical layers [202]. Wohrer et al [202] show, from measurements in rat cortex, that the actual firing rate distributions of neural networks do not differ much between resting state and evoked activity. The small difference would come from very few neurons that manage to drive up the mean firing

rate in recordings while most neurons in the population are almost silent, some with rates as low as 0.1 Hz [102], whose activity might not even be picked up by most recording devices. Although theoretical analysis of the distribution of firing rates in randomly recurrently connected networks of LIF neurons near the fluctuation-driven regime suggests considering mean firing rates around 6.4 Hz [167]. Based on the review of Wohrer *et al.* [202], particularly on the firing rate in motor areas of behaving macaques, we decided to kickstart biological neural populations activity up to a conservative baseline firing rate of 1 Hz and study the neural dynamics of circuit input firing rates of up to 10Hz.

There are two parameters governing transmission of neural activity between neural populations. First, the synaptic efficacy of connections, which was setup to be uniform across the circuit under the lack of appropriate hypothesis to tinker it in a detailed manner. According to London [116], current understanding of synapses is limited and contextual measurements and parametrization of efficacy might be more appropriate than fixing individual connection parameters. For example recent evidence [28] shows that synaptic efficacy might be modulated by attention processes. In the study of Briggs [28] neurons of the thalamus were stimulated while measuring evoked responses from corresponding monosynaptically connected neurons in primary visual cortex. With this procedure the authors showed that, the percentage of shocks that evoke a postsynaptic response, the average efficacy, ranged from 28% to 36% depending on the type of neurons considered and the attention state. Considering the possible efficacy variability in cortex, we decided to verify, through simulations of a sub-circuit, the sensitivity of the circuit temporal dynamics to low (10%) and high (30%) values of synaptic efficacy, where percentages are taken with respect to the difference between equilibrium and threshold potential, for both LIF and AdEx populations.

The second parameter governing transmission of neural activity was the number of connections between a pair of neural populations. Unlike synaptic efficacy, the number of connections were determined from a series of simulation experiments. First the number of connections from baseline persistent activity was set such that, during rest, the circuit steady state activity would stabilize around 1 Hz. The number of baseline connections necessary is a function of input firing rate, synaptic efficacy and neural model, such that a lower synaptic efficacy required a higher number of connections. Then the number of connections coming from excitatory populations was determined such that bidirectional gating circuits would have a stable steady state firing rate when both Gs allow neural activity to be transmitted. Finally the number of connections coming from inhibitory nodes were setup high enough to block neural activity flow in a gating circuit, which means that GKs driven by MAs would be able to completely inhibit activity in Gs. Our simple approach to neural rate transmission ignores many intricacies like activity regimes that might allow rich internal computations. [151]. Also connections distribution

might have an impact in spike based communication [144]. Still we decided to keep connections between populations as simple and homogeneous as possible for a first approach.

9.3 Simulation experiments performed

Since it is possible to tune the circuit to reproduce a wide range of firing rate absolute values under which circuit dynamics are similar and stable, we simply aimed at picking reasonable parameter values such that the circuit would maintain overall modest firing rate values with respect to the literature of neural measurements. To setup parameters and compare in detail the compartment circuit dynamics for LIF and AdEx neural populations, four simulation experiments were performed taking different sub-circuits into account. A diagram of each sub-circuit is shown in Figure 9.2.

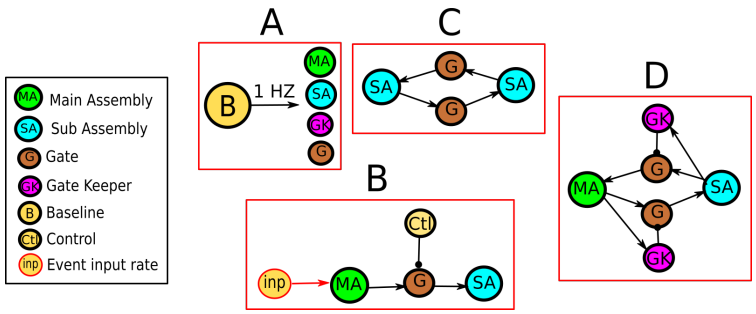


Figure 9.2: **Sub-circuit simulation topologies:** For better visualization baseline activity nodes are excluded from the topologies. A. Single neural population driven by baseline activity. This topology reminds of the fact that all MA, SA, G and GK populations are driven initially in the same way by a persistent baseline fixed rate. B. Chain of populations where activity is temporally interrupted by a control node. C. Excitatory loop between SAs when Working Memory is activated. D. Excitatory loop broken thanks to GKs inhibition.

The first simulation simply consists of the activity of one neural population driven by a fixed activity rate of 1 Hz. We used this simulation to explore the necessary number of baseline connections to drive baseline activity in the circuit to approximately 1 Hz. The second simulation allowed us to explore how neural activity flows through a chain of neural populations being regulated by a control mechanism. The third simulation explores how neural activity is enhanced by a closed loop between a MA and SA, since it will be the case in the memory sub-circuit that activity is allowed to flow bidirectionally once the WM delay activity is unleashed. Finally the fourth simulation consists on adding GKs to the closed loop sub-circuit of the second simulation to explore how many inhibitory connections are necessary to keep activity from flowing in the circuit unless the control mechanism allows it.

After determining reasonable parameter values, we simulated the complete circuit, shown in Figure 9.1, for both LIF and AdEX neural populations. Then we compared the resulting neural patterns of the MA, SA, G and GK neural populations to binding and constituency effects available in the neuroimaging literature.

We simulated the binding activity related to the processing of complete phrases, by assuming a syntactic tree structure given by a phrase grammar

and the order of control events given by a bottom up parsing scheme. As a first simplified approximation to the NBA dynamics, we instantiated the required compartment circuits independently to represent the complete assumed tree structure and temporally align their neural signals according to input onsets. Like this we obtained entire phrase neural time series, by summing activity across similar node categories of the multiple independent compartment circuits instantiated. We used this procedure to simulate the neural activity of simple phrases, corresponding to increasing size right branching tree structures, to be compared with two different neuroimaging signals.

First, we showed similarities between the activity of simple phrases and ECoG time series patterns of binding revealed by Nelson et al[141]. We naively compared the firing rates of our simulation directly to the patterns observed in ECoG recordings, considering the correlation that exist between the high gamma power of local field potential signals and firing rates[163; 119]. Nonetheless a quantitative comparison would require a more careful consideration, employing recent models tuned to electro-physiological measurements that offer a way to translate neural activity to local field potentials[127; 81].

Second, we concatenated simple phrases to reproduce the stimuli of Pallier *et al.* (2011)[153]. Then we convolved the stimuli neural time series with the Glover Hemodynamic Response Function[77]. This allowed us to make a qualitative comparison with the hemodynamic constituency effects depicted by Pallier *et al.* (2011)[153].

Since the quantitative level of neural activity can be easily tuned for a wide range of parameter values with similar behavior, when comparing the circuit neural dynamics with the neuroimaging literature, we only focused on the qualitative neural temporal patterns observed.

10 Simulation outcomes

In this chapter we present the outcome of the circuit tuning experiments, the phrase syntactic processing patterns of the simulator after tuning and how we reproduce diverse evidence from BOLD-fMRI and ECoG neuroimaging experiments.

10.1 Sub-circuit simulations

Experiment 1: Simple neural population

In the first experiment we explored the steady state rate and temporal behavior of the different neural models with different synaptic efficacies. As indicated in the circuit topology of Figure 10.1, neural populations were driven by a persistent 1 Hz input rate. We show the steady state rate as a function of the number of baseline connections in the top plots of each neural model in Figures 10.1 and 10.2. In the bottom plots we display the respective firing rate dynamics for different number of connections.

In the case of a LIF population, by manipulating the number of connections, we can tune to any value the steady state rate. For all synaptic efficacy values, the firing rate increases smoothly until achieving the steady state at approximately 200 ms. The AdEx population has a different temporal behavior. An immediate transient peak of activity on initial stimulation is driven down by adaptation, achieving a steady state at approximately 600 ms. The adaptation effect, on a 30% synaptic efficacy, limits the range of values that the steady state rate can take by manipulating the number of connections.

As explained in the Methods section 9.2, binding takes place in the Compartment Circuit when the kickoff input rate threshold of the Working Memory (WM) population is reached. The total input rate of WM depends on the sum of the firing rate of both Sub-Assemblies in the Compartment Circuit, which themselves are driven by separate input events. Since steady state rate values are limited in the AdEx model with high synaptic efficacy, operation of the circuit would be more constrained with non simultaneous input events, than in the low synaptic efficacy case.

Because we wanted to explore the behavior of the Compartment circuit for all possible timings of input events, we decided to restrict all remaining

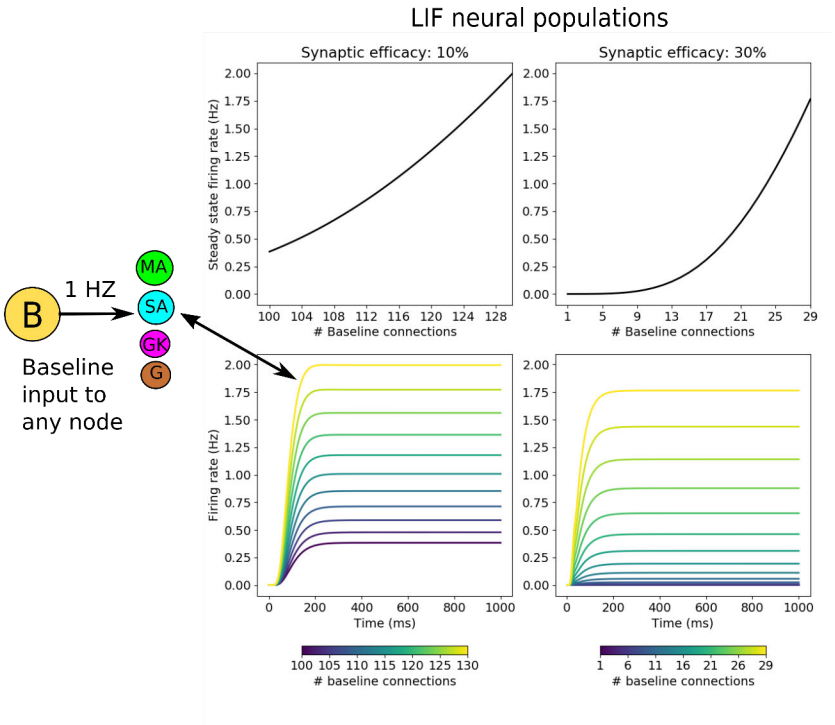


Figure 10.1: **LIF Baseline neural dynamics:** The plots at the top show how the steady state rate of a neural population relates to the number of baseline connections for a baseline input of 1Hz. The plots at the bottom show the temporal dynamics for different number of baseline connections.

simulations to a 10% synaptic efficacy. We also fixed the number of baseline connections to 115 and 1646, for LIF and AdEx populations respectively, since these values best approximated the desired 1Hz steady state firing rate under a 10% synaptic efficacy.

Experiment 2: Neural activity flow and control release

For the second experiment we wanted to understand how firing rate, in the Sub-assemblies of the Compartment Circuit, would vary with the timing of the onset of input and control events. To accomplish this we employed the sub-circuit topology presented in Figure 10.3. In this topology the Gate (G) population is permanently inhibited by a Control (Ctl) population with persistent activity, such that the Sub-Assembly (SA) can not be driven by the Main-Assembly (MA) until a control event, that inhibits the Control population, takes place. For this experiment, the number of excitatory connections was fixed to 9 for LIF populations and 20 for AdEx populations. The effect of modifying the number of excitatory connections will be explored in Experiment 3 in Results section 10.1.

We considered two possible persistent rates for the input event, 10 Hz or 20 Hz for the LIF model and 20 Hz or 30 Hz for the AdEx model. We needed higher input rates for the AdEx model since adaptation induces smaller steady state rates with respect to the LIF model. There are three possible extreme

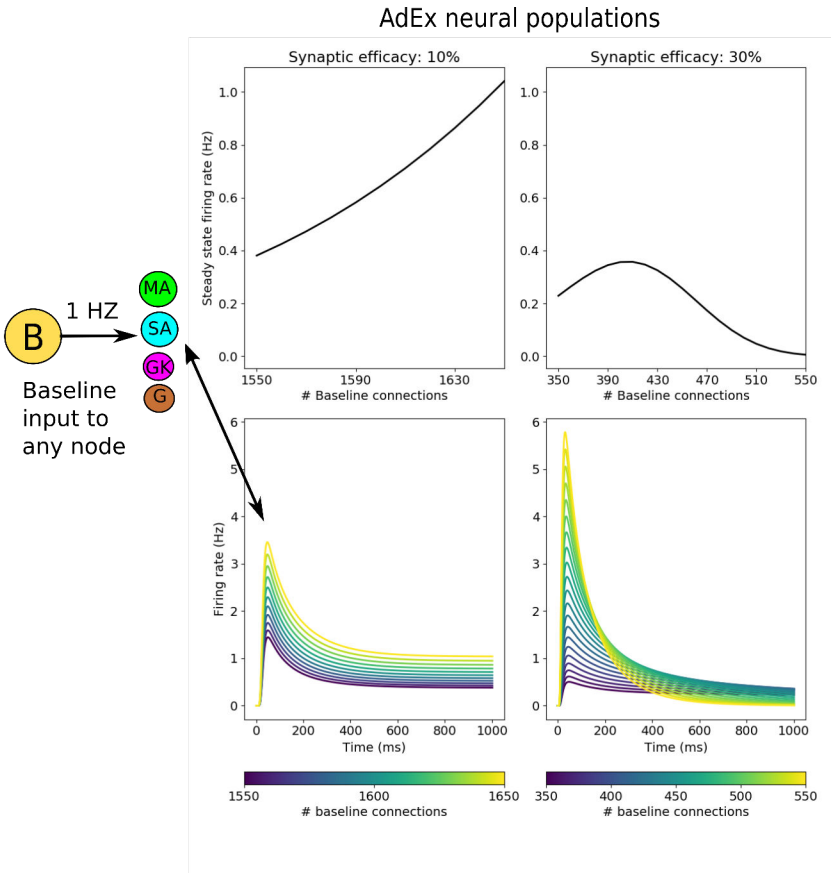


Figure 10.2: **AdEx Baseline neural dynamics:** The plots at the top show how the steady state rate of a neural population relates to the number of baseline connections for a baseline input of 1Hz. The plots at the bottom show the temporal dynamics for different number of baseline connections.

cases of timing between the input and control events; When the input event takes place at 0 ms and the control event at 1000 ms (Input First); When both events start at 1000 ms (Simultaneous); And when the control event starts at 0 ms followed by the input event at 1000 ms (Control First). These timing of events are extreme cases because 1000 ms is enough time for the neural populations to achieve a steady state rate after any event initiated at 0 ms. Any other timing in which populations have still not achieved a steady state before the arrival of the second event would produce neural dynamics with patterns in between the extreme cases. For language stimuli, timing cases can be interpreted as different types of parsing mechanisms, where Control First corresponds to a predictive (top-down) one and Simultaneous and Input First to a reactive (bottom-up) one. We show in Figure 10.3 the firing rate time series of the Sub-Assembly (SA) for all possible event timing cases and input firing rates.

First we observe that the input rate do not change the relative behavior of the timing cases but only increase the steady state rate and transient fluctuations. We see that the timing cases do not modify the final steady state rate, which

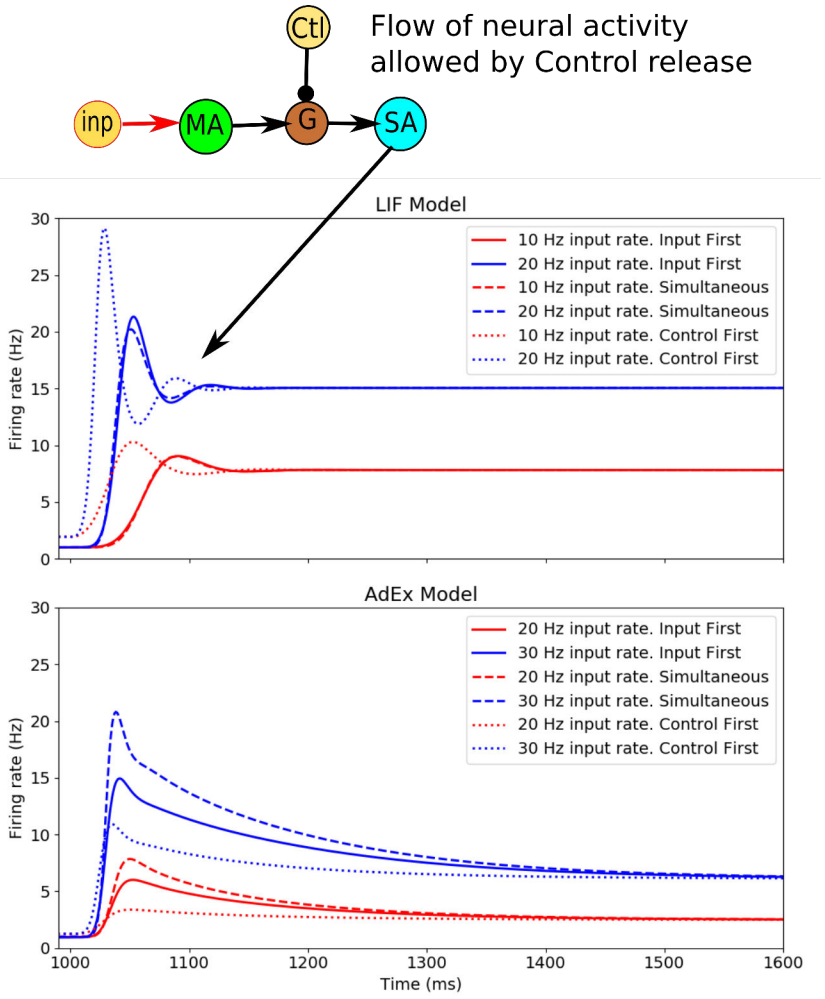


Figure 10.3: **Neural dynamics of input and control events:** We plot the temporal dynamics of the Sub Assembly population corresponding to the sub-circuit topology shown. 9 and 20 excitatory connections are assumed for the LIF and AdEx models respectively. We show the time series after 1000 ms, time at which all neural populations have achieved a steady state rate from their initial events at time 0. For each neural model two constant input rates are simulated for the input events, 10 Hz and 20 Hz for LIF, and 20 Hz and 30 Hz for AdEx. There are three possible extreme cases of timing between the input and control events; When the input event takes place at 0 ms and the control event at 1000 ms (Input First); When both events start at 1000 ms (Simultaneous); And when the control event starts at 0 ms followed by the input event at 1000 ms (Control First).

only depends on the input rate, but influence the maximum rate of the transient activity fluctuations. In the case of AdEx, the speed at which the steady state is approximated is also affected by the timing cases, for example the Simultaneous case takes approximately 400 ms more than the Control First case, to achieve the steady state, for a 30 Hz input rate. The steady state rate is in most cases and neural models the lowest firing rate, with some short transient exceptions. Moreover the timing cases have different relative behaviors depending on the neural model, as can be seen from the Control First case that has the lowest transient rates for AdEx but the highest ones for LIF.

Successful binding in the Compartment Circuit depends on the sum of activity of two SAs, that reaches the kickoff threshold rate of the Working Memory (WM) population. Assuming activity of SAs is driven by two separate input events, like two words to be bound presented 200 ms apart, the timing of the two input events and the timing cases of their respective control events

will determine the possible range of values for the WM kickoff threshold. We can also think the other way around and say that the range of values of the WM kickoff threshold constrain the possible timing of all events.

An example scenario, illustrated in Figure 10.3 for a LIF population with 20 Hz input, would be that the onset of input events correspond to the onset of word presentation, 200 ms apart, where the timing of the first SA input event follows the Input First case and the timing of the second SA input event follows the Control First case. In that scenario any WM kickoff threshold between 16 Hz and approximately 44 Hz would be reached by the sum of the 15 Hz steady state rate of the first SA and the firing rate of the second SA achieving a transient maximum of approximately 29 Hz.

Since we wanted to consider all possible event timings when studying the Compartment Circuit dynamics, we took from this experiment the cases with the highest transient rates for each neural model, to later analyse the circuit parameter space. We see in Figure 10.3 that the Control First case has the highest transient rate for the LIF model, while the Simultaneous case has the highest transient rate for the AdEx model.

Experiment 3: Circuit operation according to the parameter space

In a third experiment, we studied the parameter space of the input rate, the number of excitatory connections and the WM kickoff activation threshold, to understand the operational, event timing related, constraints of the Compartment Circuit when attempting to instantiate binding under different regions of the parameter space. As shown in Figure 10.4, to explore the circuit behavior, we have to consider the Sub-Assembly (SA) temporal dynamics presented in Results section 10.1 and a sub-circuit topology representing an excitatory loop between two SAs.

As shown in the Compartment Circuit diagram of Figure 9.1 of Methods section 9.2, once the Working Memory (WM) Delay activity is unleashed, both Gate Keepers (GKs) are inhibited, creating an excitatory loop between the Sub-Assemblies (SAs). Beyond a certain number of excitatory connections, there is the possibility of runaway activity in the excitatory loop, which motivates a constraint in the parameter space of the Compartment Circuit. The excitatory loop activity considered is only driven by the 1 Hz baseline input rate, as would be the case in the circuit once the input events stop driving activity in Main-Assemblies (MAs) and as a consequence in SAs. In Figure 10.4 we plot the space of excitatory connections up to 11 connections and 21 connections for LIF and AdEx respectively, values at which we observed runaway activity in the excitatory loop.

Alongside the excitatory loop baseline steady state rate curve of the SA, we also plot the input driven maximum transient firing rate and steady state rate of an SA, according to the different events' timing behavior presented in

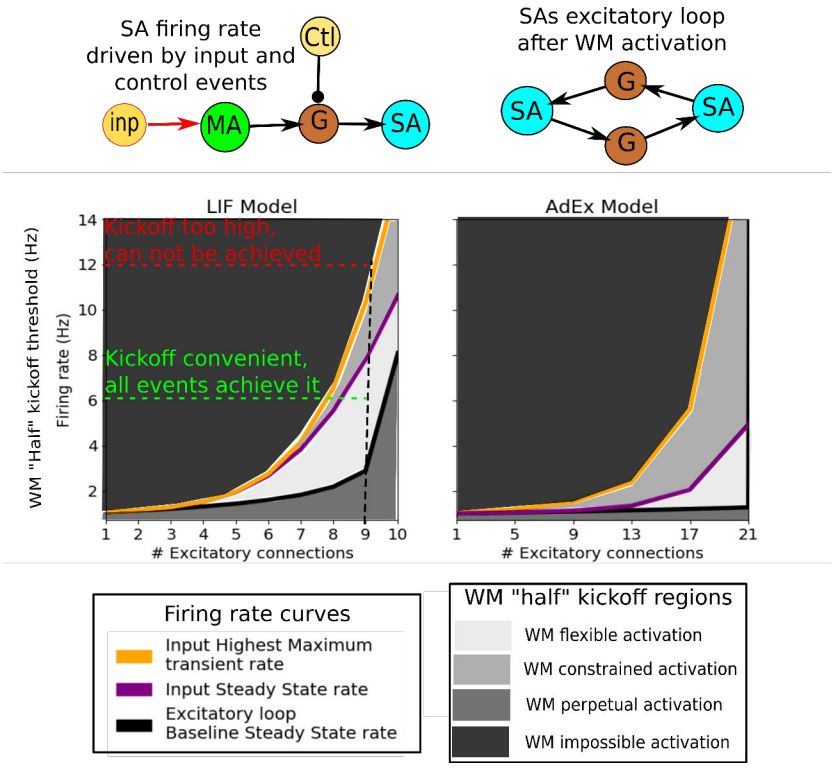


Figure 10.4: **Excitatory loop and WM activation parameter regions:** At the top the two sub-circuit topologies from which SA firing rate curves are derived. Rate curves consist on firing rate as a function of the number of excitatory connections for a given input rate of 10 Hz and 25 Hz for the LIF and AdEx models respectively. From the chained neural population topology we consider the highest maximum transient rate and the steady state rate. From the excitatory loop topology we consider the steady state rate driven only by baseline activity. We color the regions between the curves to indicate the different WM activation cases determined by the value of the WM "half" kickoff threshold rate. The four parameter regions refer to the possible combination of input and control events that would allow binding to take place if the WM "half" kickoff threshold falls in the region: The perpetual activation region implies that WM will get permanently reactivated; The flexible activation region implies that all events cases can produce binding; The constrained activation region implies that only some combination of events' timings can permit binding; Finally the impossible region implies that no binding can take place for the given WM kickoff rate.

Results section 10.1. The firing rate curves correspond to an input of 10 Hz and 25 Hz for LIF and AdEx populations respectively. All the firing rate curves correspond to the activity of only one SA, so whenever we represent the WM kickoff rate threshold in Figure 10.4, we refer to the "Half" kickoff threshold. For example the convenient "Half" kickoff rate threshold of 6 Hz, marked with a green line in the LIF Model plot, implies a total WM kickoff rate threshold of 12 Hz.

From the relationship between the three firing rate curves, we can establish four parameter regions with different implications for the behavior of the Compartment Circuit: First, below the excitatory loop baseline steady state rate, we have a parameter region for which WM would be continuously reactivated. The initial activation of WM leads to the excitatory loop steady state rate, so if the kickoff threshold is below it, WM will be reactivated perpetually. We call this the WM perpetual activation region; Second, in the area between the loop steady state and the input steady state curves, all input and control event timing cases will lead to activation of WM, which can be explained by the steady state rate being the lowest transient rate. We call this the WM flexible activation region; Third, in the region between the input driven maximum transient rate and the steady state rate curves, activation of WM will not take place for some timings of input and control events. The higher the WM kickoff threshold in this region, less input and control event

timing cases can activate WM. We call this the WM constrained activation region; Finally, above the input driven highest maximum transient rate, it is clear that activation of WM can not be achieved under any circumstance, which is why we denote it as the WM impossible activation region.

To understand the constrained activation region, it helps to take a look back at Figure 10.3 of Results section 10.1. Consider the AdEx model with a 30 Hz input rate. We can see that a WM kickoff rate of 14 Hz would be reached by adding the steady state of one SA and the transient rate of any events' timing case for the second SA. If we raise the WM kickoff rate to 20 Hz then we would need the events driving the second SA activity to follow the Input First or Simultaneous timing cases, while raising it further to 25 Hz would leave the Simultaneous case as the only option.

We still do not know the parameter variability allowed by the cortex to implement the circuit, so we consider the proportion between the constrained and flexible activation parameter regions as a indicator of the difficulty to operate the Compartment Circuit under the different neural models. Based on this, we observe in Figure 10.4 that the AdEx model is more likely to induce constraints in the timing of input and control events to perform the bindings necessary to represent complete structures in cortex. To allow the most flexible behavior exploration of the Compartment Circuit, when simulating language processing, we decided to select parameters in the flexible activation region. We selected a combination of 10 Hz and 20 Hz input rates, 8 and 20 excitatory connections and 10 Hz and 9 Hz WM kickoff rates for LIF and AdEx populations respectively.

Experiment 4: Inhibition of undesired activity spill

In the fourth experiment, we tuned the amount of inhibitory connections between Gate Keepers (GKs) and Gates (Gs) to avoid undesired spill of neural activity from the Main Assemblies (MAs) to the Sub-Assemblies (SAs). We wanted any spill to be practically insignificant for any number of excitatory connections and arbitrary input activity fluctuations to which the AdEx model is sensitive. We decided to study this with the sub-circuit topology of Figure 10.5.

We plot, in Figure 10.5, the maximum transient firing rate of the SA as a function of the number of inhibitory connections for a varied number of excitatory connections. If the amount of inhibitory connections is not enough, transient activity of the SA will be increased beyond its baseline activity, denoted with a black line. We determined how many inhibitory connections are necessary by looking at the amount of inhibitory connections at which the maximum firing rate becomes practically insensitive to the number of excitatory connections. It is clear from the plots that, after a certain number of inhibitory connections, unidirectional activity will be allowed only by controlled inhibition of the GKs. From these experiment observations, we

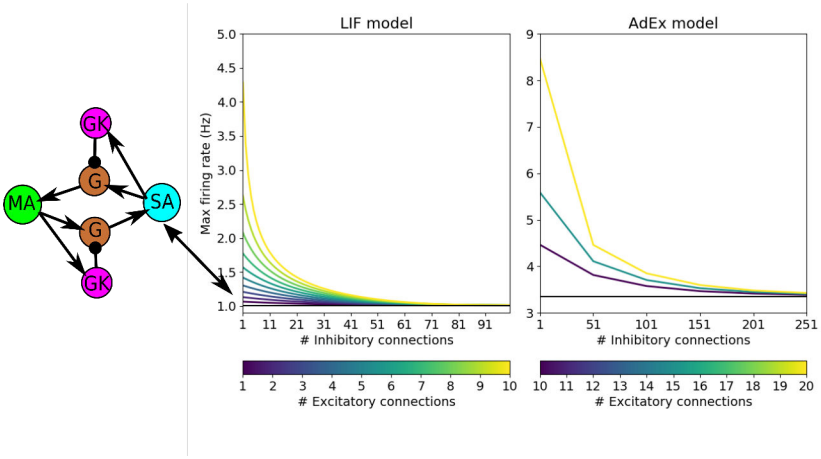


Figure 10.5: **Inhibition to avoid excitatory loop:** The sub-circuit topology at the right depicts the inhibition of Gates (Gs) by the Gate Keepers (GKs) that are driven themselves by the Main and Sub Assemblies (MA and SA) to avoid an excitatory loop between them. Activity in the sub-circuit is driven only by a 1 Hz baseline rate. Each curve in the plots represent how the maximum transient rate of SA for a given number of excitatory connections varies as we increase the number of inhibitory connections. We present one plot for each neural model (LIF and AdEx). The maximum firing rate is employed instead of the steady state rate to observe sensitivity to transient rate fluctuations.

decided to set the number of inhibitory connections to 70 and 250 for LIF and ADEX populations respectively.

10.2 Complete compartment circuit simulations

After selecting a set of parameters in line with the previous experiments, we analysed the behavior of the complete compartment circuit simulation. The dynamics of the compartment circuit can be summarized by a combination of the input events that drive activity in Main-Assemblies (MAs) and the control events that inhibit Gate Keepers (GKs) such that activity can flow from MAs to Gates (Gs) and from the latter to Sub-Assemblies (SAs). In Table 10.1 we present a summary of the parameters taken for LIF and AdEx simulations and in Figure 10.6 we present the temporal dynamics of the compartment circuit for a complete and incomplete binding.

Parameter	**LIF**	**AdEx**
baseline connections	115	1646
excitatory connections	8	20
inhibitory connections	70	250
Input rate (Hz)	10	20
WM/Ctl rate (Hz)	10	20

Table 10.1: **Complete simulation parameters**

First, we show the baseline dynamics of the circuit when no event takes place in part A of figure 10.6. In this case all neural populations are only receiving an input baseline rate of 1 Hz. So the different populations just reflect with their firing rate the architecture of the circuit. Gs show a low rate of activation due to GKs inhibition, while GKs show the highest rate driven by MA and baseline activity. MAs show an activation close to the approximated 1 Hz baseline as well as SAs that have been isolated in the circuit thanks to

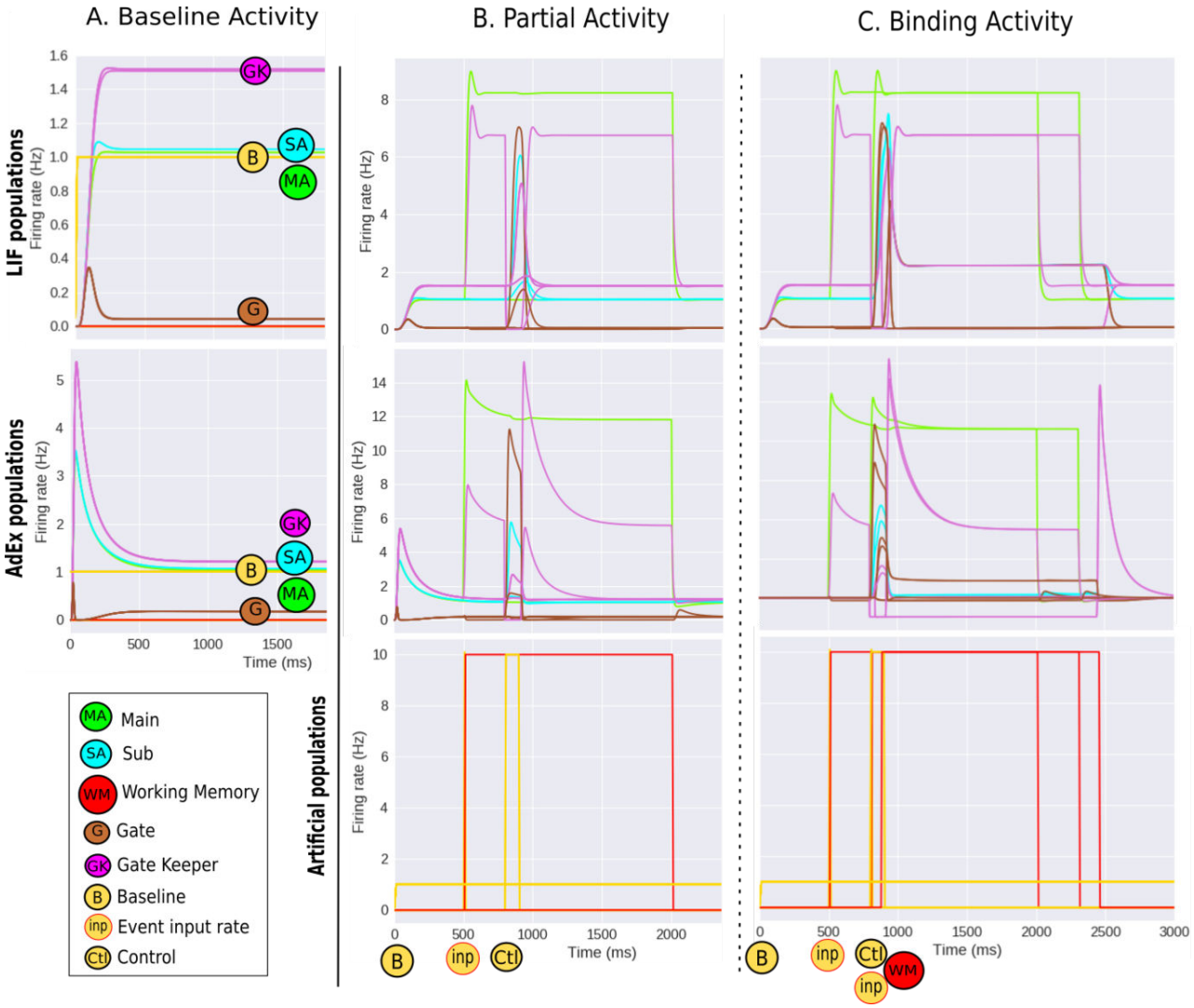


Figure 10.6: **Profiles of neural activity: A.** Neural activation driven only by baseline input. **B.** Neural activation of the circuit when only one MA is activated by a word event or WM at 500 ms. Shows the neural activity related to an erroneous control signal at 800 ms. It is possible to see that the steady state of neural activity is resilient to a slip of control, going to the appropriate levels of neural activity once the control activity is over. **C.** Neural activity of the Compartment Circuit for a successful binding. The second MA gets activated at 800 ms alongside the controls. Since both MAs are active, the SAs manage to activate WM to instantiate the binding of the MAs. Two interesting dynamics arise from the binding: The first is that a spike of activity in SAs, GKs and Gs takes place due to the sudden inhibitory activity of WM on the GKs; The second is that the memory circuit internally raises its baseline activity due to the excitatory loop formed.

GKs inhibition.

Second, we show the activity of the circuit for an incomplete binding in part B of Figure 10.6. This means that only one MA is driven by an input event,

after which a Control (Ctl) event allows activity flow from both MAs to SAs, even though there is no binding to be done. Due to stimulation of the MA, the GK firing rate raises to stop activity to flow to the SA until the control event takes place to inhibit the GK. As only one SA is driven by input, the total rate contribution to the WM population do not achieve the WM kickoff threshold rate necessary to perform a binding. Both neural models display a transient spike of neural activity in the SA, G and memory sub-circuit GKs during the time window the control permits activity to flow to the SA. In the case of the AdEx dynamics, shown in Figure 2.8, there is the possibility of an activity rebound after inhibition, in which neurons will respond more vigorously than if they would not have been inhibited, reflected in the GKs after control stops.

Third, we show the circuit dynamics of a successful binding in part C of Figure 10.6. When both MAs are driven by an input event and a control event takes place. In this case the added activity of the SAs reaches the WM kickoff threshold and kickstarts the Delay activity of WM. Then activity in the SAs and Gs of the memory sub-circuit raise to a new baseline due to the excitatory loop created by WM inhibition of GKs, which also generates an initial transient spike of activity in SAs. A similar behavior to this one, simulating sentence parsing, was also reported by previous work with the NBA[67]. Finally, after the WM Delay activity stops, the LIF model activity goes back to baseline, but the AdEx model exhibits a final transient rise of firing rate in the GKs of the memory sub-circuit, similar to that of the GKs affected by control inhibition release.

10.3 Simulation of complete phrase processing

With the neural dynamics of several isolated Compartment Circuits, simulated independently of each other, we approximated the binding of complete phrases. As explained in the Introduction section 8.2, we simplified the simulation of the Blackboard by ignoring mutually inhibitory Compartment Circuits dynamics determined by a Connection Matrix. The right branching hierarchical structure that corresponds to an example phrase of 4 words, determined by a phrase grammar, is shown in Part A of Figure 10.7. In this example only three Compartment Circuits are necessary to realize all the bindings that would correspond to the phrase processing, and the exact input event onsets were taken from the LIF simulation. The onset of input events driving Main Assemblies that represent word grammatical categories were matched to word presentation onsets spaced 600 ms apart from each other. In the case of phrasal nodes, we assumed that their input event onset corresponds to the previous realization of a binding, determined by the moment at which their respective Working Memory population was activated. In this way, phrasal nodes can be represented by activity in the Main-Assemblies of a Compartment Circuit and be bound to other word grammatical categories or phrasal nodes.

We needed to prolong the Main-Assemblies and Working Memory activity

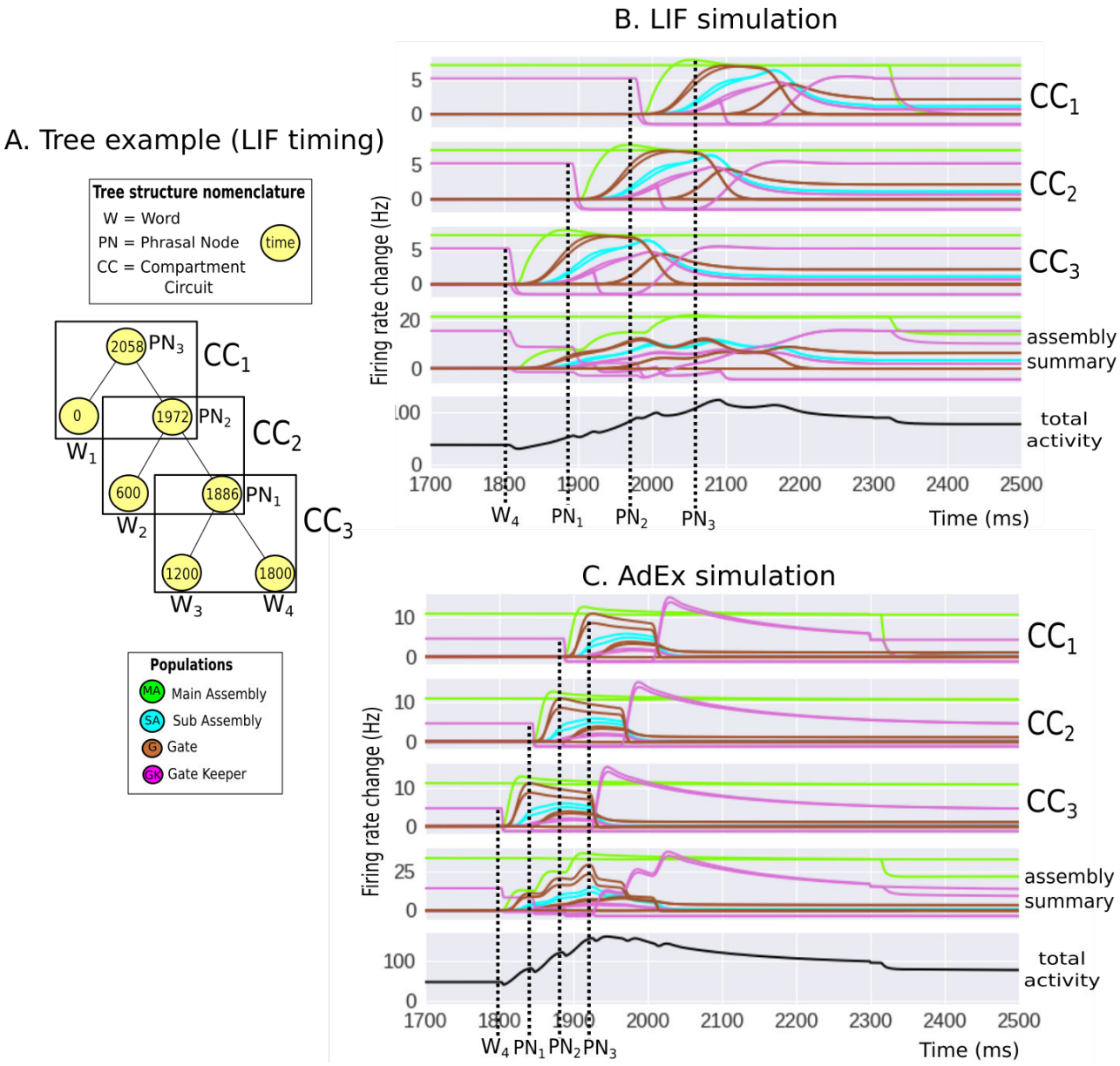


Figure 10.7: **Sentence processing example:** A. Tree structure hypothesized for a given 4 words phrase. It is shown how compartment circuits correspond to sections of the tree structure and how the nodes corresponding to grammatical categories of words processed or phrase nodes are instantiated in time under a bottom-up parsing approach. B. Blackboard time series that correspond to the simulated processing of the considered tree structure and time of activation of the nodes. The separate activity of the LIF populations of each compartment circuit are shown separately, followed by their summary and total activity. C. Same as B but for AdEx populations.

long enough to instantiate all the necessary bindings, so in this example we assumed WM and input events to last 2300 ms for all simulations. As indicated

in the second phrasal node (PN2) of the tree example, if input events were active for less than 1972 ms then activation of the first word MA would cease before the accompanying phrasal node MA comes into play to realize the last binding. In the Compartment Circuit simulation presented in Figure 10.6, there was a difference in timing of WM activation between the LIF and the AdEx neural models, that was not easy to see in the plots. The Working Memory population became active 86 ms after all input and control events take place in the LIF simulation, while in the AdEx simulation this only took 42 ms. This time difference originates in the faster initial transient response of the AdEx dynamics in contrast to the LIF dynamics, that can be seen in Figures 10.2 and 10.1 respectively. By contrasting the LIF and AdEx complete phrase simulations in Figure 10.7 we can better appreciate how this difference adds up to accelerate phrase processing in the AdEx model.

To later compare the phrase processing simulation with neuroimaging patterns, we first substracted baseline activity from the time series of each neural population in each Compartment Circuit. Then we summed the aligned time series of the same neural population category belonging to different Compartment Circuits. Finally, to obtain total neural activity of phrase processing, we summed activity from all the non phenomenological neural populations and the Working Memory population, such that they would all be equally weighted under the absence of a more detailed hypothesis about the neural population sizes and their spatial distribution in the cortex.

10.4 Qualitative reproduction of ECoG patterns

As presented in the Introductory section 8.1, the ECoG analysis of Nelson *et al.*[141] is the first to characterize the specific temporal patterns of phrase-structure formation from intracranial neurophysiological data, possibly revealing the first neural signatures of binding operations. Nelson *et al.* demonstrate two patterns that are of particular interest to our simulations: first, the average temporal dynamics of processing increasing size right branching phrases. Second, the average neural dynamics for hypothesized number of pending binding operations, during phrase processing, under a bottom-up parsing approach. In Figure 10.8 we show the aggregated neural activity predicted by our LIF and AdEx simulations, alongside the temporal dynamics of phrase processing presented by Nelson *et al.*, from the mean high gamma power of the intracortical recordings.

As can be seen in the top plots, our simulations suggest the existence of four qualitatively different segments of neural dynamics: first, as words are presented to the circuit, input events drive activity in Main-Assemblies (MAs) corresponding to the grammatical categories of words. The activity of all the MAs accumulate but still do not change the activity of other neural populations on the Compartment Circuits, since for parsing a right branching tree under a bottom up parsing scheme, control events that allow bindings, do

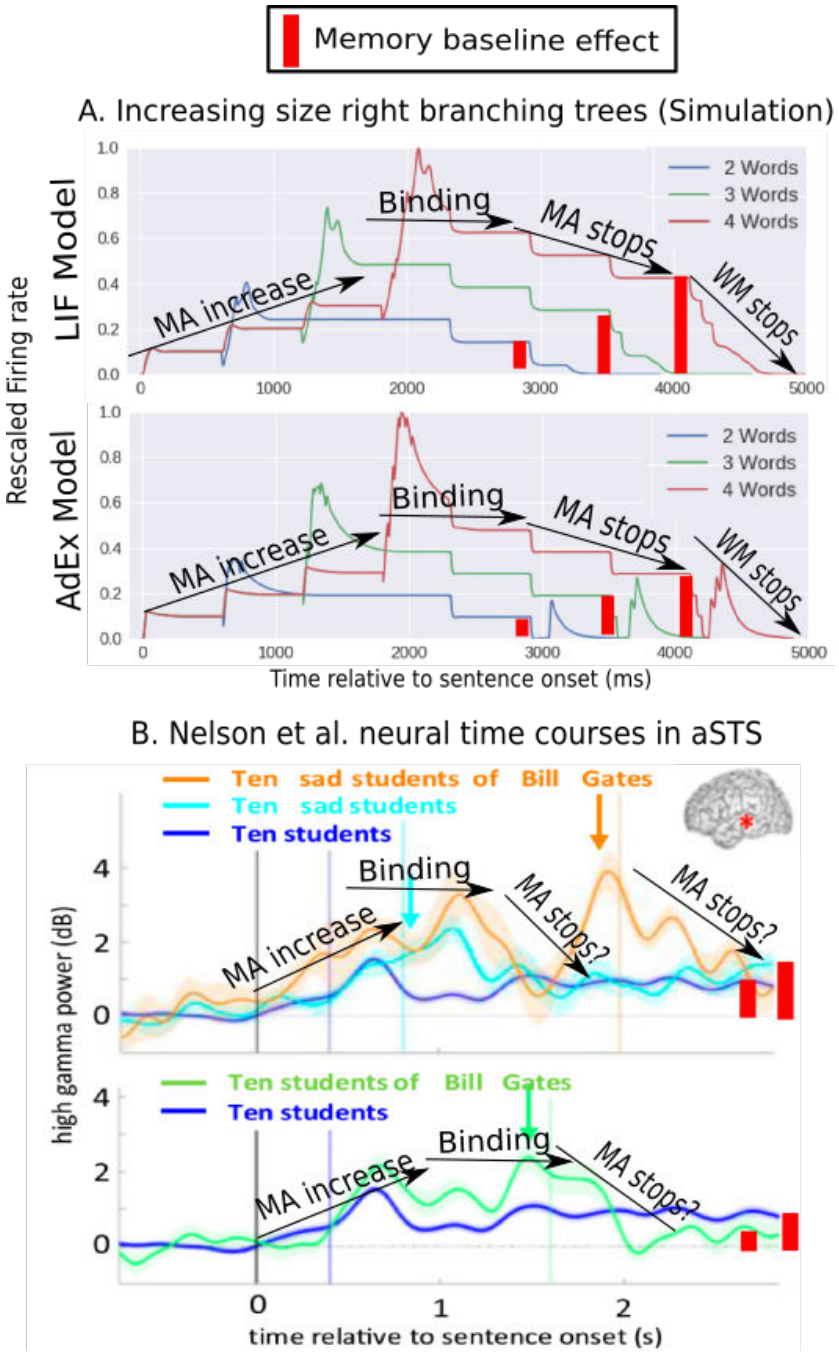


Figure 10.8: **Simulation comparison with intracortical (EcoG) recordings:** In the top plots we show phrase processing for the LIF and AdEx simulations. We denote with arrows the four segments of neural dynamics identified in the simulations; The Main Assemblies (MAs) activity increase the segment; The accumulated binding operations segment; The Main Assemblies (MAs) activity release segment; And the Working Memory (WM) release segment. We denote with red bars the magnitude of Working Memory activity in the circuit that depends on phrase length and remains at the end of phrase processing. In the bottom plots we identify, in Figures modified from Nelson *et al.*, the segments of intracortical recordings that resemble the simulation and denote with red bars the possible Working Memory related activity that remains at the end of phrase processing.

not occur until the last word is presented. The second segment correspond to the succession of bindings that take place after the last word of the phrase is processed. The neural activity allowed by the control events creates a transient rise in activity that stabilizes with the accumulated Delay activity of the Working Memory populations and the still ongoing input activity. The

third segment is characterized by the gradual drop of input related activity. And the fourth segment corresponds to the final drop of Working Memory activity, such that all the neural populations return to their baseline steady state rate.

We see in the bottom plots of Figure 10.8, modified from the Figures in Nelson *et al.*, that we can qualitatively identify the three initial segments predicted by the simulation in the high gamma power time series. We observe an initial increase in neural activity, for which a later onset and higher magnitude of the peak appear to depend on phrase length, as would be explained by the first segment of the simulation based on an increase of activity in Main Assemblies (MAs). The following transient fluctuations of the ECoG time series could be identified with the binding related segment and the final activity drop with the release of MA activity. In the simulation, because we deactivate MAs on discrete time steps, we observe plateaus of MA activity, while the ECoG time series suggest a more abrupt drop after bindings have taken place, which complicates distinguishing the neural fluctuations related to the binding operations, from those related to the MAs activity release. In the longer 6 words phrase "Ten sad students of Bill Gates", there is a middle sentence high transient fluctuation that is not expected from a bottom up parsing scheme.

We indicate with red bars, that the activity drop of the ECoG time series stops at a higher level than the initial baseline, which is compatible with the hypothesized ongoing Working Memory (WM) activity of the simulation. The AdEx model distinguishes itself from the LIF model, during WM inactivation, by predicting a final burst of activity due to the inhibition release of the Gate Keepers in the memory circuit. Nonetheless, due to the task of the ECoG experiment, which requires retaining in memory the phrase for later comparison with another phrase, we should not be able to observe the final drop of WM activity predicted by the simulation, as is the case.

In Figure 10.9 we show, in the top plots, the simulation time series aligned on the last word onset, to demonstrate the neural activity fluctuations linked to the number of accumulated and executed binding operations, which Nelson *et al.* refer to as the number of nodes closing. In the bottom plots we show modified Figures from Nelson *et al.*, where the effect is demonstrated in the case of middle sentence operations and sentence end operations.

10.5 Qualitative reproduction of BOLD-fMRI patterns

As explained in the Introductory section 8.1, we also reproduced patterns from an experimental design employed to show constituency effects with BOLD-fMRI[153]. Stimuli, presented to a subject in a trial, consisted of a list of phrases with the same number of words (constituents), such that in total 12 words would be presented. All phrases correspond to right branching trees according to the phrase grammar considered by the authors. The conditions

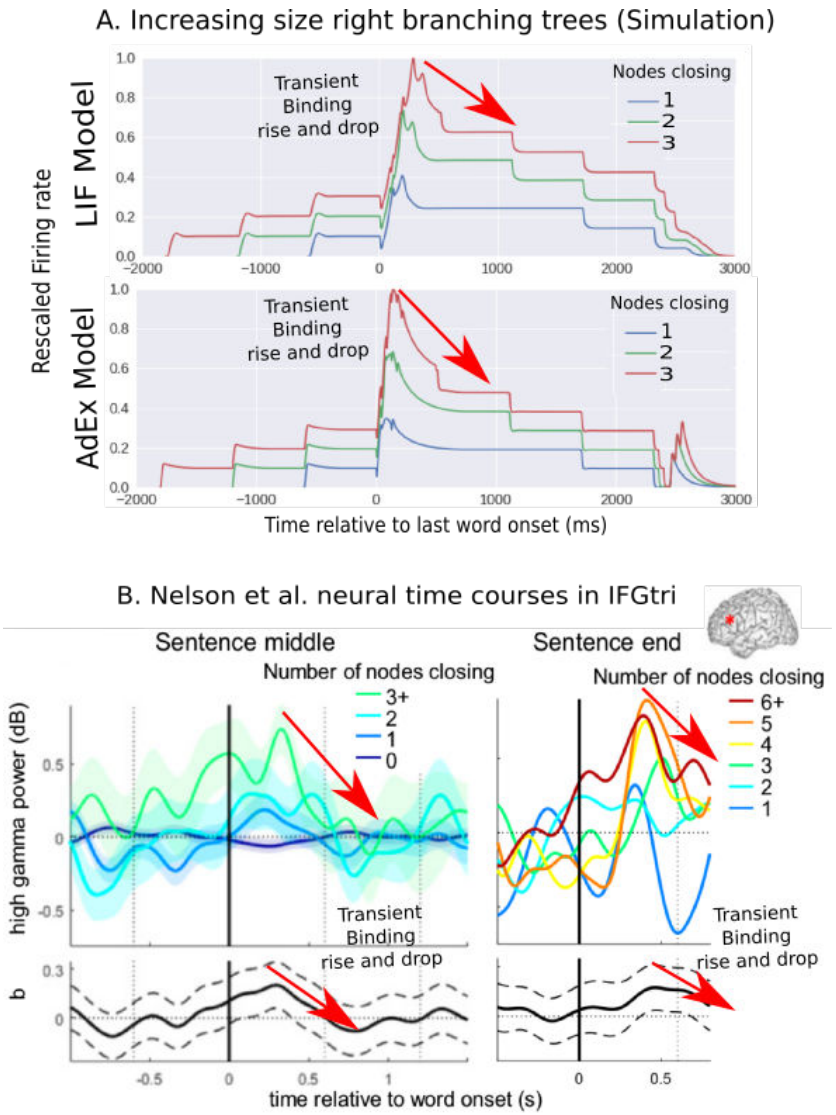


Figure 10.9: **Effect of number of executed binding operations:** In the top plots we show the phrase processing time series of the LIF and AdEx simulations, aligned on the onset of the last word. We denote with arrows the segment of transient rise and drop of neural activity hypothesized to be linked to the number of executed pending binding operations, which we refer to as number of nodes closing in the plots, following terminology from Nelson *et al.* In the bottom plots we show, in Figures modified from Nelson *et al.*, the intracortical recordings effect of executed pending binding operations at the middle and end of phrases.

were one list of 12 unconnected words (c01), 6 phrases of 2 words (c02), 4 phrases of 3 words (c03), 3 phrases of 4 words (c04), 2 phrases of 6 words (c06) and 1 phrase of 12 words (c12).

Besides normal words, the design also included pseudoword conditions that maintained morphological markers and closed-class (function) words. We will compare our simulation with the pseudoword effects of Pallier et al, since they provide syntactic specific patterns that can be interpreted closer to the abstract binding operations of our simulation. Moreover we continue to assume the same phrase grammar and bottom-up parsing scheme employed for comparison with the intracortical recordings of Nelson *et al.* To simulate the Pallier *et al.* stimuli, we added the repeated neural time series of each of the right branching trees in a condition. So, for example, to simulate the 4

phrases of 3 words condition (c03), we aligned and summed, based on word onsets, the neural activity of 4 simulations of a 3 words phrase.

In the standard analysis of BOLD-fMRI time series, events are modelled as a constant stepwise function that reflects the duration of the stimuli, called a boxcar model. The boxcar model events are then convolved by an Hemodynamic Response Function (HRF), for which we considered the HRF proposed by Glover[77], available in the python open source package Nistats [1]. The convolved events are then used in a general linear model (GLM) to obtain a peak estimate of hemodynamic responses for the different conditions, as was the done in the Pallier *et al.* study.

[1] https://github.com/nistats/nistats

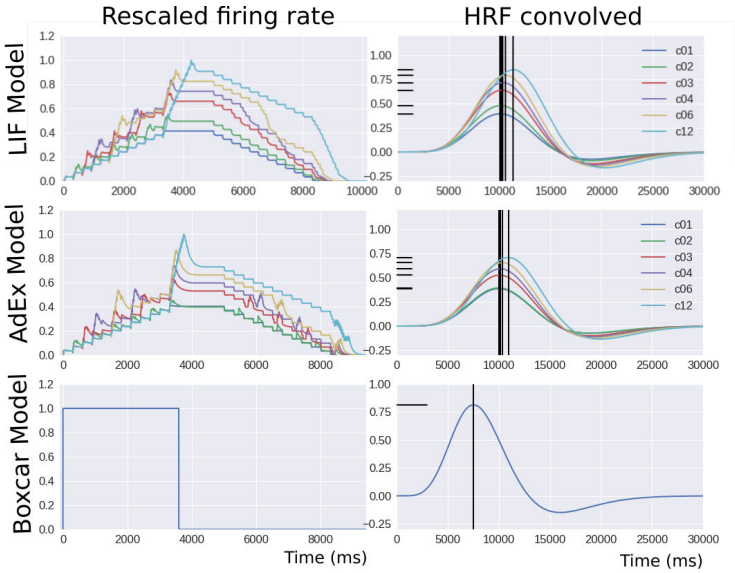


Figure 10.10: **Hemodynamic interpretation of the simulation:** At the top and middle plots we show the rescaled time series of the LIF and AdEx simulations respectively, alongside the HRF convolved time series. At the bottom we show a boxcar event of 3600 ms and its convolution, as was employed by Pallier *et al.* to estimate the amplitude of responses for the different conditions from the BOLD-fMRI time series. we considered the HRF proposed by Glover, available in the open source python package Nistats.

We generated a prediction of hemodynamic responses from our simulations by rescaling the conditions' time series by the maximum firing rate of all conditions and then convolving them with the HRF. We present the predicted hemodynamic responses in the top and middle plots of Figure 10.10. Since in the Pallier *et al.* study, 12 words are presented every 300 ms, we considered the last word onset of 3600 ms as the duration of the stimuli for a traditional boxcar event model, shown in the bottom plots, to compare it with our models. We mark the HRF peak and its onset with black lines on all the HRF convolved time series.

We observe that the neural time series would predict in all cases a peak onset displaced many seconds with respect to the traditional boxcar event that only represents the duration of the stimuli. Looking at the time series, this would be expected, since the HRF peak onset depends on the center of mass of the accumulated neural activity, which continues several seconds after the last word onset in our simulations. The peak onset in the LIF and AdEx models follow a super-linear increase with respect to the number of constituents, at

odds with with sub-linear patterns reported by Pallier *et al.* Also the LIF neural model introduces an slightly longer onset delay with respect to the AdEx neural model, due to its slower activation of Working Memory populations.

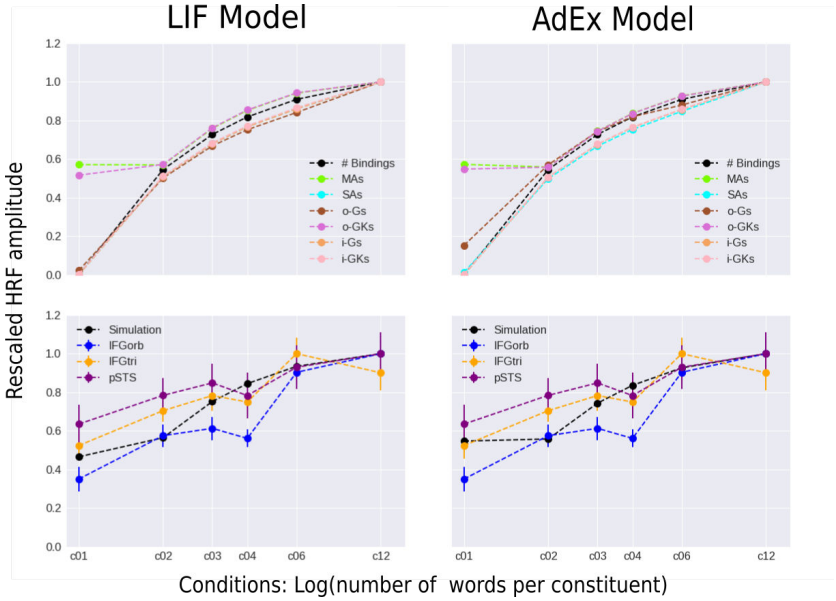


Figure 10.11: **Hemodynamic peak magnitudes comparison with BOLD-fMRI experiment:** The top plots show the number of bindings executed for each condition alongside the rescaled Hemodynamic Response Function (HRF) amplitudes of each of the Compartment Circuit neural populations. We demonstrate that the hemodynamic pattern of the neural populations in the simulation follow closely the number of bindings executed. In the bottom plots we contrast the pattern of the total neural activity in the simulation alongside the sub-linear patterns reported by Pallier *et al.* in the pSTS, IFGorb and IFGtri brain regions.

In the case of the HRF peak amplitudes, we show in Figure 10.11 that both LIF and AdEx models predict a sub-linear pattern of peak amplitudes as a function of the number of constituents. We demonstrate in the top plots that the HRF magnitudes of added neural activity in all neural populations of the Compartment Circuit follow the pattern given by the number of executed bindings in a condition. It is unlikely then, that the sub-linear pattern appreciated in the HRF amplitudes would be qualitatively changed by manipulating other parameters of the circuit, like the duration of the input to Main-Assemblies and Working Memory that could modify qualitatively the peak onsets pattern.

Pallier *et al.* reported constituent sub-linear responses in the language areas TP, aSTS, pSTS, TPJ, IFGorb and IFGtri, but only the regions pSTS, IFGorb and IFGtri showed a similar response pattern when minimizing the semantic content of phrases with pseudowords. Since our simulation puts aside semantic considerations, we consider this type of experimental manipulation to be a better reflection of the binding activity modelled in the Compartment Circuit. In the bottom plots of Figure 10.11, we show the similarity between the HRF magnitude pattern of the total neural activity in the simulation models with what is reported by Pallier *et al.* in the pSTS, IFGorb and IFGtri brain regions.

11 Discussion

In this chapter we discuss results obtained from the Neural Blackboard Architecture simulation and comment on future perspectives of the framework for further experimental work.

11.1 The neural models and circuit architecture

Regarding the neural model parameter values, we considered those from Omurtag[150] and Brette *et al.*[26] for a first approximation of the neural dynamics. We left for future work consideration of values based on electrophysiological recordings from specific brain regions. For example, there are different adaptation constants along the cortex, that could change the AdEx model dynamics. Since we have compared the simulation with neural activity in specific brain regions like aSTS, pSTS, IFGtri and IFGorb, it would be reasonable to fit the simulations to their specific biological reality.

In the case of the Compartment Circuit assumptions, we made many simplifications that should be revised in future work. We approximated baseline dynamics with a low constant input rate instead of considering the natural oscillatory activity of the cortex, homeostatic mechanisms in cortical circuits[182] and balanced networks[203]. Also we adopted homogeneous synaptic connections instead of testing different synaptic distributions that could have an impact in the neural dynamics. Moreover, if we allowed random connectivity to shape the Compartment Circuit architecture our capacity to control its dynamics with the number of connections would be restrained.

The explicit simulation of Delay Activity in Working Memory was left out of the current work due to its flexible and still debated implementation[43; 68]. Studying it could reveal important neurobiological limitations on the way we assess the relative proportion of neural activity between Main-Assemblies and Working Memory. Also it could provide a more limited set of hypothesis about the spatio-temporal memory limitations of the Neural Blackboard Architecture, to be contrasted with neuroimaging and psycholinguistic evidence.

Out of two options, we took the decision to allow the existence of excitatory loops after Working Memory activation, although this permits the possibility of unstable runaway neural activity. Neural activity related to these excitatory

loops was regulated in the Working Memory sub-circuit by careful tuning of the number of excitatory connections and the number of inhibitory connections that would close the loops. The second option was to introduce in the Working Memory sub-circuit a bidirectional control mechanism similar to that employed to regulate communication between Main-Assemblies (MAs) and Sub-Assemblies (SAs). Nonetheless the second approach implies additional complexity in the number of nodes, connections and events that we have to consider for the circuit operation. Since we do not really know what is closer to the biological reality of the cortex, we decided to show how the less complex architecture that includes excitatory loops could be made stable, but consideration of a more complex architecture would also be possible.

To our knowledge, this is also the first time complex neural models like AdEx are simulated alongside LIF for variable binding and language function related circuits. In contrast to previous simulations [189; 188; 67; 191], we employed population density techniques implemented in the MIIND software[49], that allowed us to approximate the transient fluctuations of the different binding related events. Thanks to this, we found that the circuit implementation and neural dynamics interpretation can depend on the underlying neural model in non trivial ways. For example we observed that in a LIF model there was a non-consequential trade-off between synaptic efficacy and number of excitatory connections to control the steady state rates of the circuit. On the other hand the AdEx model was very sensitive to changes in synaptic efficacy due to adaptation effects, to the point of making us unable to control the magnitude of the steady state rate of the circuit for high synaptic efficacy values. If the physical reality of the cortex was closer to an AdEx model with high synaptic efficacies we would then need to restrict our hypothesis about the circuit operation with input and control events to a subset of the possibilities explored in our simulation. Adaptation in the AdEx model also had an important effect in the case of lower synaptic efficacies, making coordination of input and control events more restricted in a larger portion of the circuit parameter space. Since we have to take into account the possibility of random variation of those parameters in the cortex, this effect can be crucial to understand limits and constraints of language processing in different brain regions.

Another important distinction observed between the AdEx and LIF model was how dynamics after inhibition are qualitatively different under the influence of adaptation. While in the LIF circuit, neural activity on a population would smoothly recover back to its steady state after inhibition stops, that of an AdEx circuit would show a renovated burst of activity due to adaptation decreasing during the inhibition period. The effect might be strong enough to suggest it as a predictive marker for certain events in the circuit, like the release of Working Memory activity.

Moreover, characterizing the Working Memory activation parameter regions was important to understand the reliability of the circuit if exposed to noisy input rates, arbitrary timing coordination of events, control mistakes

or anticipatory control signals. Although for a bottom-up parsing approach, we can safely assume control events to take place after input events, this might not be the case for other parsing strategies like top-down, that could be implemented with anticipatory control events. Since some parameter regions restrict the timing of input and control events, we might get insights into the possible set of parsing mechanisms directly from the anatomical structure of the cortex that constrain the parameter boundaries.

Finally, the question of how a Compartment Circuit and the Neural Blackboard Architecture could be formed during brain development and modified by learning is still work in progress, partially tackled in a previous study[189]. Demonstrating how neural mechanisms approximated by the architecture can be implemented with biological realistic Hebbian or STDP rules alongside random connectivity constraints, during development and learning, would be an important avenue of future research.

11.2 Circuit implications of the linguistic hypothesis

A strength of the current simulation is its flexibility to predict the neural activity of diverse grammar theories and parsing schemes, which we only explore partially in this work. We could in principle, without circuit modification, predict the binding activity for any structure that can be represented by a binary tree. This is the case of the phrase grammar of the minimalist program of Chomsky[38], that represent phrases as binary trees, and also the case of other theories like dependency grammars[142] that represent grammatical relations between words. Nonetheless in the case of dependency grammars, as they do not require a hierarchical representation, we would not need to assume that the Working Memory of an executed binding drives the Main-Assembly of another Compartment Circuit.

Because we only modelled a bottom-up parsing scheme, we considered activation of the Main-Assemblies corresponding to phrase nodes only after the binding that produces the corresponding phrase took place. For example, for the phrase "the black cat" we would create an input event for the phrase node of "black cat" after "black" and "cat" have been bound. If we consider instead a pure top-down parsing scheme, that implies prediction of future bindings, or the generalized left corner parsing scheme proposed by Hale[83], there would be three additional mechanistic options to consider: First, we could start input events for Main-Assemblies representing the phrase nodes before their corresponding bindings and only start the control event after the bindings have been confirmed; Second, we could start the control events beforehand, which is an option explored in the simulation, and still make input events follow the corresponding bindings; Third, we could go ahead and perform bindings ahead of time, that would need to be deactivated by an error signal provided by the parsing mechanism. This last option would allow to simulate the possibility of multiple parallel phrase representations, from

which only one survives at the end.

A simplification was made regarding the Compartment Circuit selection mechanism in the Neural Blackboard Architecture. We did not model the dynamic inhibition of competing Compartment Circuits belonging to the same Connection Matrix. To do it we would require an hypothesis about the size of the Neural Blackboard, governed by memory limitations and the total number of possible grammatical category combinations given by a grammar. Forming such an hypothesis was out of the scope of this work, so we opted to assume the simplest selection mechanism possible based on uniform random selection, which is how we justify simply recruiting Compartment Circuits as needed. Nonetheless we are only able to ignore the inhibitory activity of competing Compartment Circuits in complete Connection Matrices because we are not planning to explore the effects of memory limits under time compressed sentence processing scenarios or memory tasks. Otherwise important deviations in background neural activity due to depletion of available Compartment Circuits and additional inhibitory activity would become a crucial factor for the simulation. We plan to explore this in future work, to try to reproduce temporal bottleneck effects shown by Vagharchakian *et al.* on hemodynamic responses, based on a BOLD-fMRI experiment with an experimental design containing compressed speech and reading conditions[185].

With respect to the parsing mechanism, we only model its interface with the Compartment Circuit that implements binding, through the assumed control signals. We considered that understanding how a parsing algorithm is learned and implemented by the cortex, such that it can provide the respective control signals, was a separate research question. Previous work has shown the feasibility to implement a parsing mechanism with neural networks in connection to the Neural Blackboard Architecture[188], for a limited set of possible syntactic structures.

As can be inferred from this discussion, there is already great potential for exploration of linguistic hypothesis with the current simulation developed, but there are also many open questions left for future development. We believe that taking into account more experimental evidence from psycholinguistics and neuroimaging studies is necessary to guide future refinements of the circuit architecture and simulation.

11.3 Qualitative reproduction of neuroimaging evidence

Comparison of our simulation with neuroimaging measurements revealed striking qualitative similarities, even though the circuit was only tuned for its correct operation, with respect to binding execution. We aggregated the time series of the simulation in the simplest way possible, uniformly, under the lack of more precise hypothesis about the spatial distribution of the Neural Blackboard Architecture in the cortex. Although we interpreted reports of

the high gamma power of Local Field Potentials and hemodynamic responses separately, there is the potential to integrate all these different measurements as coherent quantitative evidence thanks to recent efforts on modelling their relationship[92].

High gamma power has been shown to be correlated to the firing rate time series of spiking neurons from in-vivo recordings[163], and we decided to make a direct qualitative comparison between the firing rate of the simulation and high gamma power time series. Nonetheless a future quantitative comparison would require a more precise mapping from the simulation firing rates to local field potentials, as has been done recently[127; 81].

An important discrepancy between the simulation and the high gamma power time series, was that the simulation segments of neural activity identified with binding and Main-Assemblies transient activity drop were not as clearly separable in the intracortical recordings. Moreover the data seem to suggest an immediate Main-Assemblies transient drop after a binding event takes place, instead of the paced inactivation assumed during the simulation. This would suggest the addition to the Compartment Circuit of a feedback mechanism from the Working Memory populations to the Main-Assemblies to knock out their unnecessary activity once binding has been established. It would be an efficient strategy from an energetic point of view at the cost of extra complexity in the circuit architecture.

We also observed a middle phrase activity drop in the intracortical time series of the longest phrase, which was not coherent with a bottom-up parsing hypothesis. In the phrase "Ten sad students of Bill Gates" the activity drop took place after "Ten sad students", and was compensated immediately after to bind the remaining phrase "of Bill Gates". Two possibilities arise from this observation: The obvious first one is to consider an alternative parsing mechanism combining a bottom-up and top-down approach, a generalized left corner parsing scheme, to explain the fluctuation; The second one is difficulties of the Compartment Circuit to sustain local activity in Main-Assemblies for prolonged periods of time, such that they need to be reactivated if a binding has still not taken place. If this was the case, we could also explain the previously explained apparent immediate Main-assemblies activity drop after binding as a side effect of an imminent deactivation that was going to take place independently of binding.

To approximate hemodynamic responses, we resorted to a naive approximation that has to be interpreted with caution since the relationship between neural activity, cerebral blood flow and blood oxygenation can be non-linear under certain circumstances[70; 31] and better represented by the balloon model than the gamma function considered in this work[197]. A more precise translation from firing rates to an hemodynamic response would allow a quantitative fit of simulation parameters and to test linguistic hypothesis. At the moment we show that the simulation could be adapted to other hemodynamic peak onset patterns and that it naturally reproduces

magnitude patterns, although we do not attempt to tune the simulation to reproduce the relative differences between conditions.

Regarding the hemodynamic peak onsets, our first observation was that persistent neural activity in Main-Assemblies and Working Memory can substantially delay the onset of the hemodynamic response, with respect to that given by a traditional boxcar model event. Such a large delay demonstrates the importance of modelling neural dynamics to avoid an event model misspecification. It has been reported that parametric estimation of gamma based models, used for General Linear Model estimation to analyse BOLD-fMRI experiments, quickly deteriorates as model misspecification increases [115]. To realize a future quantitative comparison between the generated simulation time series and hemodynamic measurements, we would need to fit a new linear model for each simulation hypothesis to the available BOLD time series, looking for the best fit.

The super-linear increase pattern of peak onset we observed was not coherent with sub-linear patterns reported by Pallier *et al.* Nonetheless the peak onset of our simulation depends on the input events and Working Memory durations, that were arbitrarily set to a constant duration. The Neural Blackboard Architecture does not provide a particular hypothesis on the timing of the deactivation of Main-Assemblies and Working Memory, which is why durations were simply set to a pragmatic constant that secured binding of the last phrasal node with the first word of the longest phrase. Comparison of our simulation with intracortical recordings in results section 10.4 suggested a quicker drop of the Main-Assemblies activity after binding operations were executed, instead of the current choice of persistent activity for a constant amount of time after binding. Modifying the simulation to drop activity in Main-Assemblies after binding, would permit emulating sub-linear patterns of peak onset as necessary to reproduce the hemodynamic measurements.

Regarding the hemodynamic peak amplitudes, future quantitative comparison of the levels of neural activity between the word list condition (c01) and rest of the conditions in which binding takes place, could give insights into the relative proportion of Main-Assemblies activity and the rest of populations in the circuit. At the moment, the simulation's initial slope of hemodynamic peak amplitude increase was lower than that reported by Pallier et al, which can be interpreted as an underestimation of the binding related populations contribution to the total neural activity. Pallier *et al.* initially hypothesized a linear pattern of peak amplitudes instead of the sub-linear one observed. Their initial hypothesis was based on a simple "accumulation" model where each new word presented would add a constant amount of neural activity until a binding was not possible, leading to a sudden drop of activity back to baseline. After their findings, the authors revised their hypothesis to propose instead a model that assigns a logarithmic increase of activity to each new word presented. Nonetheless our simulation suggest another explanation for the sub-linear pattern as a direct reflection of the number of binding

operations executed during phrase processing. It turns out that the type of stimuli employed by the authors consisted exclusively of right branching trees and that their concatenation lead to a sub-linear increase of number of binding operations, which is why our simulation is at a first sight coherent with the logarithmic word activity addition model. Then our simulation suggest the possibility that assigning a logarithmic increase of activity to the next word presented in a phrase is an artefact of the experimental design, due to missing consideration of other syntactic tree structures for phrases containing the same number of words.

11.4 Future perspective

Even though the current simulation can still be improved in many ways, we would like to emphasize with this work the quick progress in the development of biologically plausible models of cognition. New computational methods like population density techniques have made it tractable to approximate, at a circuit scale, point neural models as complex as the adaptive exponential. With an additional modelling effort at the level of the neural populations, we could close the gap that has delayed physical mechanistic testing of computational linguistic hypothesis with direct neuroimaging measurements. Taking into account cytoarchitectonic details, tailored to different brain regions, would allow to study the spatial distribution in the cortex of the Neural Blackboard Architecture and other circuit alternatives. Modelling these details would allow better physical reproduction of temporally and spatially detailed signals, like Local Field Potentials (LFP)[127; 81] and hemodynamics (BOLD)[31]. Moreover, it would also be possible to integrate the evidence from multiple spatio-temporal scales in a coherent way, such as has been done in the literature, taking as example recent work linking LFP and BOLD signals[92].

We selected two experiments that we considered best characterized key neuroimaging evidence of binding in phrase processing. Moreover we think these experiments, coming from different spatio-temporal scales and experimental designs, demonstrate the potential of our simulations to integrate varied experimental paradigms. Many other experiments could inform different parameters and circuit assumptions from the ones explored in this work. For example we could look at processing speed and memory constraints of the Neural Blackboard architecture with the BOLD-fMRI manipulation of Vagharchakian *et al.*[185] based on compressed speech and reading conditions. Creating a database of such neuroimaging experiments alongside psycholinguistic behavioral evidence would create the opportunity to incrementally and systematically test linguistic computational hypothesis and their brain implementation.

As we commented in Discussion section 11.2, the current implementation of the Compartment Circuit allows us to test any grammar theory providing binary tree representations, combined with any parsing scheme that

determines the timing of input and control events in the circuit. Although we focused on one parsing scheme and grammar, it is evident that we can explore all other alternatives in the future. This means that as we refine the boundary of the circuit parameters and operating assumptions, we can obtain for any corpus the neural activity of all its phrases for all the different linguistic hypothesis available. From such artificial dataset we could motivate experimental designs and tests that would be optimal to explore the linguistic hypothesis space. For example controlling for diverse variables like phrase length or number of syllables, we could estimate the likelihood of a phrase grammar versus a dependency grammar theory, by comparing a set of phrases that maximize neural activity differences between the theories with respect to a set of control phrases.

We think that the proposed framework could lead to quick progress in our understanding of language function if accompanied by the most recent neuroimaging techniques. We would imagine a setting in which intracortical recordings can be systematically positioned with information coming from quick and reliable fMRI language localizer paradigms[118]. From a language localizer and anatomical scans, it would be possible to take advantage of 3d printing techniques, already tested in non-human primates[37], to make frames perfectly adapted to the skull of patients, with electrodes precisely positioned at the peaks of hemodynamic effects. Moreover recent advances in laminar fMRI[111] are an exciting possible addition for the tuning of models approximating cortical columns with cytoarchitectonic constraints, which we propose to extend our simulations.

In conclusion we hope to have demonstrated that we are close to producing biologically realistic mechanistic neural models of cognitive function. In particular to provide new ways of testing linguistic hypothesis integrating evidence from varied neuroimaging techniques with different spatio-temporal scales. With this work we expect to inspire further efforts in this direction.

Part IV

Concluding remarks

12 Final remarks

12.1 Summary of findings

In the experimental part of this work, we identified the superposition principle to be one of the crucial assumptions of Smolensky's basic tensor product representations. To test the superposition principle we created and fMRI dataset from which we could extract spatial representations of bi-syllabic pseudowords in visual and auditory sensory modalities.

The decoding analysis in sensory brain regions revealed the highest accuracy scores and reproduced known effects like the superposed semi-local representations induced by retinotopy. In the case of auditory regions we found weak evidence in favour of local superposed representations in anterior areas higher in the auditory processing hierarchy. Decoding on language related regions only revealed significant classification in Broca's complex (44 and 45), for which we could provide evidence in favour of superposition and more distributed representations. Finding superposed representations in Broca is interesting, since this region has been shown in a meta-analysis of fMRI studies[206] to be consistently engaged with syntactic binding manipulations. We were also able to provide evidence against superposition or in favour of non additive models in the visual word form area (VWFA), which is coherent with previous evidence of whole word representations in that region[75].

There were also other findings not directly related to the superposition principle. We verified that it was possible to decode auditory representations form the VWFA, providing additional evidence to the literature body claiming that this region can be modulated by speech as well as reading[205]. Moreover we were surprised by a global lack of generalization from decoding models trained in one sensory modality to the other, which can be either interpreted as a lack of sensitivity due to variability of the representations signal or as the absence of amodal representations for simple bi-syllabic pseudowords. Finally we observed in most regions with significant classification scores, except Visual, extreme variability in the accuracy scores of individual items, such that few had particularly high scores while most remained closer to chance level. We demonstrate this effect with an approximate bimodal distribution of the accuracy scores and we think this pattern could be explained by lack of sparsity and low variability in the spatial distribution of values of the neural

vectors underlying the neural representations.

In the modelling part of this work, we created a new implementation of the Neural Blackboard Architecture (NBA) based on population density techniques, that allowed us to make temporal high resolution predictions of neural dynamics linked to the binding process. Our simulations were based on the dynamics of spiking point model neurons: leaky-integrate-and-fire (LIF) and adaptive-exponential-integrate-and-fire (AdEx) neurons. Contrasting LIF and AdEx models allowed us to demonstrate that, although they are not importantly differentiated by average dynamics, their parametrization have strong implications for the timing and control of phrase processing events.

We also showed that an NBA implementation, only implementing the binding mechanism and tuned to operational constraints, qualitatively reproduces the neural activity patterns of at least two neuroimaging experiments involving linguistic binding at different spatio-temporal scales. We qualitatively reproduced three out of four predicted temporal segments of the neural dynamics of sentence comprehension revealed by intracortical recordings (ECoG)[141]. Moreover our simulation provides a similar drop of neural activity related to the moment at which a binding operation takes place, by activating the working memory mechanism, and an increasing activity baseline that depend on the number of bindings performed. We also reproduce qualitatively sub-linear patterns of hemodynamic responses caused by phrase constituency manipulations[153]. Our simulation provides an alternative hypothesis to explain the sub-linear pattern, based on the number of binding operations executed during phrase processing. Alongside these results, we illustrate the flexibility of the NBA to represent arbitrary binary tree structures and parsing schemes, which makes it a promising tool for linguistic hypothesis exploration and future refined quantitative and integrated accounts of multi-scale neuroimaging measurements.

12.2 Global perspectives

In this work we parallely explored two modelling approaches to the binding problem. We selected these approaches for how powerful they are to handle several aspects of language modelling: like answering Jackendoff's challenges[98], being able to represent multiple levels of hierarchical language processing and flexibly implement multiple linguistic hypothesis. Alongside being quite powerful, both approaches appear to be importantly distinct in their underlying assumptions, as we explained in Chapter 1 Section 1.4. In Table 12.1 we provide a reminder of the comparison.

Although there seem to be many differences between the modelling approaches, we think that the computational operations supporting bindings, binding and unbinding are the truly fundamental differences between them. The other differences are linked to implementational issues that most likely will converge as we better understand the structural and functional properties

Aspect	Smolensky's TPR	NBA
About modelling:		
Neural simulation	Artificial NN	Spiking NN
Temporal dynamics	Not included	Included
Representation	Neural unit vectors	Neural assemblies
Parallel repr model	Memory slot roles?	Separate neural assemblies
Representation properties:		
Declaration	Explicit	Implicit
Spatial stability	Static (temporally stable)	Dynamic (temporally unstable)
Locality	Distributed or local	Local
Operation implementation:		
Composition of bindings	Superposition (addition)	Compartment recruitment
Binding	Tensor product	Working memory assembly activation
Unbinding	Inner product	Reactivation of bound neural assemblies

Table 12.1: **Modelling approach comparison:** We present all binding related aspects studied in this work about Smolensky's tensor product representations and the Neural Blackboard Architecture.

of cortical circuits. There are several aspects related to implementation details that could and should be reconciled to properly compare the approaches in future work. Two in particular that we thought about are the basic representational units assumed by the models and the inclusion of temporal dynamics in Smolensky's framework.

Neurons are still considered by most models as simple compartment units although their superior information processing power has been known for some time[106]. In this regard both approaches might require a reinterpretation of their implementation. In the case of the Neural Blackboard Architecture (NBA), its fundamental mechanisms are based on the idea of a gating circuit and a short term memory device. Although these mechanisms have been interpreted at the level of a circuit of neural assemblies and reverberating activity, an alternative implementation at the cellular level for gating[109] and synaptic short term memory[138] have been demonstrated in the literature. This means that it could be possible to reimplement the functionality of complete Compartment Circuits of a Blackboard with few neurons to bring its implementation at the neural unit level. In the case of Smolensky's framework the mapping of cellular activity to theoretical values of the neural units is not clear and several alternatives based on the computational complexity of a single neuron should be considered in the future.

The NBA provide a clear temporal depiction of the control and memory mechanisms necessary to implement binding. On the other hand Smolensky presents in the *Harmonic mind*[172], the implementation of tensor products abstractly as matrix multiplication, were the matrix coefficients are interpreted

as synaptic weights in a layered network. The problem with this interpretation is the infinite possible concrete network configurations that are equivalent and that the dynamics of computation in real networks is ignored. Recent work from Smolensky on a dynamic optimization scheme to instantiate input representational vectors[173] could help bridge the gap on temporal predictions to compare it with the NBA.

For future experimental designs, besides the intricacies that can be introduced by the particular neuroimaging modalities employed, there are three aspects that should be emphasized. First, that future work should still focus on simple syntactic structures, namely syllable combinations to form pseudowords, short pseudoword lists and short jabberwocky phrases. It seems clear from recent meta-analysis[206] that limiting the semantic content of stimuli importantly reduce the number of brain regions involved in its processing. For example only Broca 44 is constantly involved in purely syntactic operations while the the posterior superior temporal sulcus (pSTS) and the superior temporal gyrus (STG) seem to be involved with syntactic and semantic integration[206]. Second, based on the previous point and our finding of superposition in Broca's complex and auditory regions, future experiments should focus specifically on anterior brain regions, including auditory areas higher in the processing hierarchy and Broca's complex. Focusing in specific regions would also facilitate targeted acquisition with different neuroimaging modalities that have less spatial coverage but high temporal resolution like intracortical recordings (ECoG). Third, the stage at which abstract representations arise in the brain, which we were not able to demonstrate with simple bi-syllabic stimuli, should be explored in more detail. The process of formation of abstract representations could be linked to audiovisual integration and related cognitive phenomena in language like the McGurk effect[179].

Regarding the available linguistic computational hypothesis, it is important to seriously consider simultaneously grammar alternatives like Phrase structure grammars and Dependency grammars and parsing schemes like bottom-up, top-down and generalized-left corner parsing. Both modelling approaches have the capacity to interpret the diversity of linguistic hypothesis and their test is intrinsically related to the hypothesis considered. For very simple stimuli like bi-syllabic stimuli, this do not seem to be a crucial issue, but processing of jabberwocky phrases is already subject to highly divergent linguistic theories and we can not avoid assuming one or another when matching model predictions with neuroimaging measurements. Incrementally testing the different linguistic hypothesis alongside the modelling approaches would be an important complementary extension to the efforts of this work.

Finally we would like to emphasize the recent advances in neuroimaging techniques that will provide even richer evidence to future experimental efforts, like laminar fMRI[111] and increasingly available intracortical (ECoG) recordings. Also there are diverse theoretical and computational advances in

the simulation of neural cortical columns that allow to reproduce complex neural signals like Local Field Potentials (LFP)[127; 81], hemodynamics (BOLD)[31] and their relationship[92]. Perhaps soon it will be possible to produce precise mechanistic predictions of neural signals out of linguistic computational hypothesis and that future work should aim to arrive to such reality.

12.3 Conclusion

Neuroscientific models of language have matured, while empirical tests of their assumptions have been left behind. New neuroimaging techniques and the recent possibility to simulate some of their corresponding neural signals should not be ignored and lead to a new wave of experiments capable of mechanistically testing linguistic computational hypothesis. We think we have been able to give a glance at the value of the approaches considered, namely Smolensky's tensor product representations and the Neural Blackboard Architecture, and the challenges we face to test them empirically. In this work we have covered just a small segment of the path leading to understanding of variable binding in symbolic structures and hope to motivate more work in this direction.

13 Other contributions during the PhD

Other experimental work not included in the manuscript

In addition to the experimental work presented, there are several other projects that were conducted or are currently in progress, but were not included in this manuscript.

During my master thesis I performed an empirical investigation, using fMRI, of the brain regions involved in representing the syntax of mathematical formula, manipulating their complexity and using structural repetition priming, which was finished during the first months of the PhD, but I decided then to focus completely on the binding problem for the PhD and this manuscript. Also as part of the tests of the superposition principle, we had the idea to also run a two-digit numbers version of the bi-syllabic pseudowords experiment. Nonetheless decoding of the number conditions was not sensitive enough, so I decided to concentrate on language stimuli for the rest of PhD and this manuscript.

There is currently work in progress on the analysis of an ECoG dataset of a phrase and word list reading task, to better understand the timing of events related to the temporal segments and neural assemblies predicted by the Neural Blackboard Architecture. We are employing supervised learning techniques to characterize time segments linked to different grammatical features, possibly connected to the dynamics of some NBA neural assemblies. We are considering grammatical features from alternative grammar theories, a phrase grammar and a dependency grammar, such that we can also explore tests to empirically evaluate the likelihood of these theories. We are also exploring the application of unsupervised learning techniques, based on time series alignment with dynamic time warping, to extract clusters of electrodes with similar neural signatures related to binding dynamics.

Contributions to a study in Pediatric neuro-oncology

As a side project, I carried a substantial contribution to the statistical analysis and methodological development of a clinical study where we investigated the relationships between the changes in different cognitive scores and radiation dose distribution in 30 children treated for a posterior fossa tumor. We showed two cases for which there was a relationship between the radiation dose in

specific brain areas and particular cognitive decline. From my participation I was recognized as third author of the published study[55]

Children treated for posterior fossa tumor with cranial radiation therapy often suffer from cognitive impairments. Radiotherapy might specifically impact brain regions implicated in different cognitive functions. Therefore, identifying regional effects of radiotherapy on cognitive functions may help to propose specific rehabilitation interventions adapted to the risk of cognitive impairment.

Open source software development and assistance to open science initiatives

I contributed as well code to several Python open source libraries linked to machine learning and statistical analysis in neuroimaging: Nilearn, Nipype and Pypreprocess. In the process I became one of the main contributors of the Nistats library that offers an alternative for complete statistical analysis of BOLD-fMRI datasets. This experience lead me to also get involved with open science data standards initiatives linked to open sharing of raw BOLD-fMRI datasets (BIDS) and open sharing of statistical results (NIDM), due to which I participated in several coding sprints in Paris and Stanford.

A Appendix. Superposition experiment ROIs decoding and tests

A.1 Visual-h0c1 (Visual dataset)

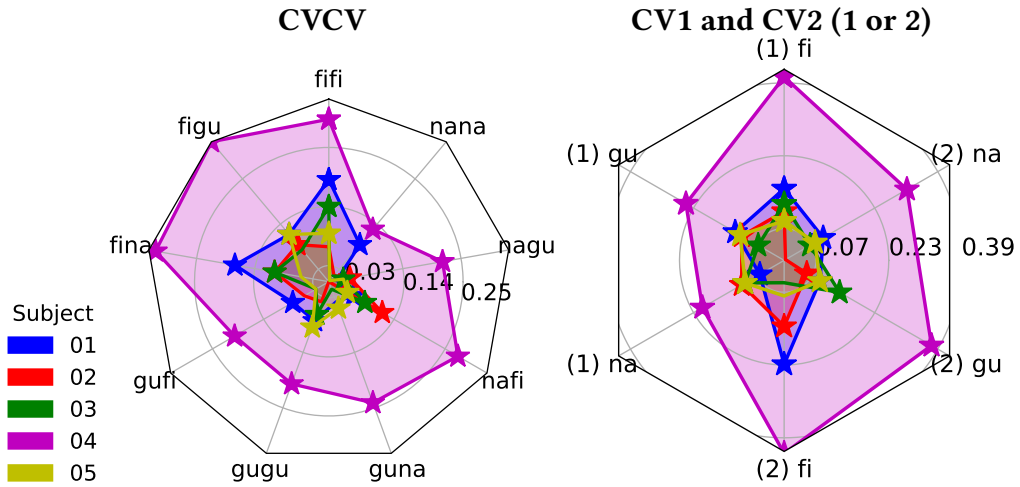


Figure A.1: **Accuracy in Visual-h0c1:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.30**	0.23**	0.29**	0.19*	0.19*	0.15	0.16*	0.12	0.20**	0.20**
02	0.17	0.20*	0.21*	0.16	0.15	0.09	0.23**	0.15*	0.12	0.17**
03	0.25*	0.19	0.21*	0.14	0.17**	0.12	0.19**	0.14*	0.11	0.17**
04	0.41**	0.45**	0.44**	0.31**	0.31**	0.35**	0.39**	0.33**	0.24**	0.36**
05	0.20*	0.23**	0.16	0.14	0.20**	0.16*	0.15*	0.14	0.11	0.17**

Table A.1: **Accuracy Visual-h0c1 CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.49**	0.45**	0.39*	0.44**	0.56**	0.42**	0.43**	0.47**
02	0.44**	0.44**	0.44**	0.44**	0.47**	0.39*	0.33	0.40**
03	0.45**	0.40*	0.43**	0.43**	0.38	0.47**	0.40*	0.42**
04	0.73**	0.58**	0.54**	0.62**	0.75**	0.70**	0.64**	0.70**
05	0.42*	0.44**	0.43**	0.43**	0.41	0.42**	0.41**	0.41**

Table A.2: **Accuracy Visual-h0c1 CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

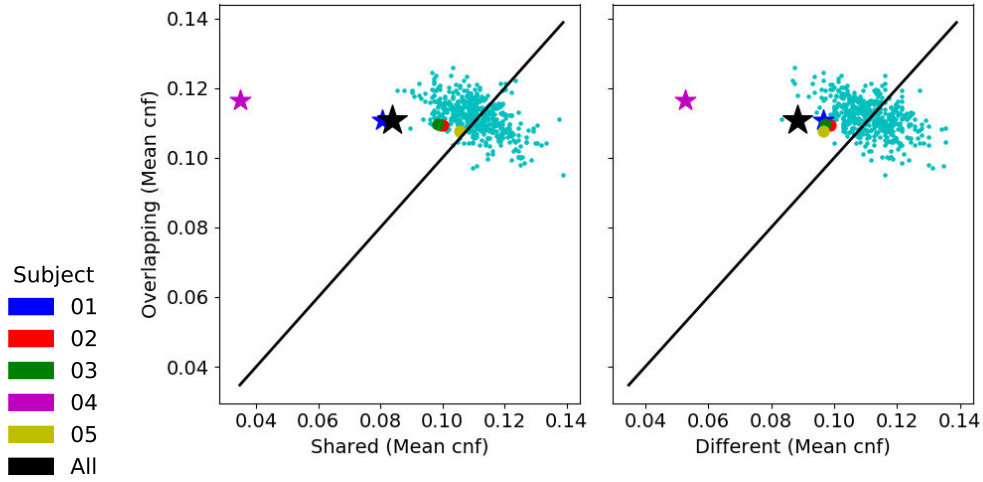


Figure A.2: **Superposition test in Visual-h0c1:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

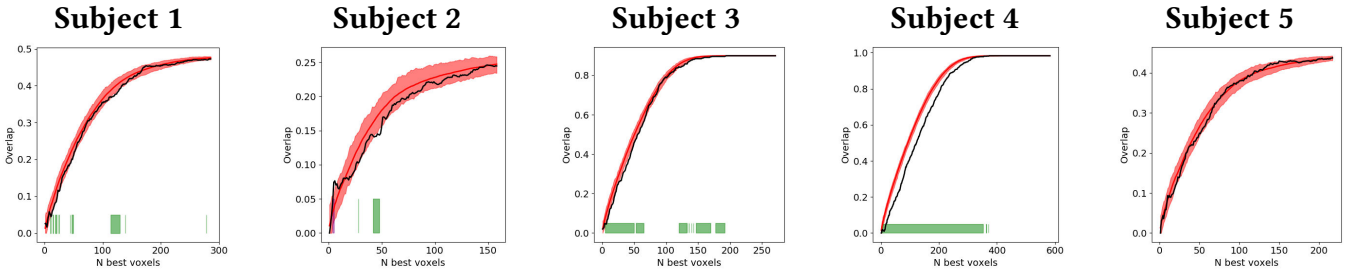


Figure A.3: **Locality test in Visual-h0c1:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.2 VWFA (Visual dataset)

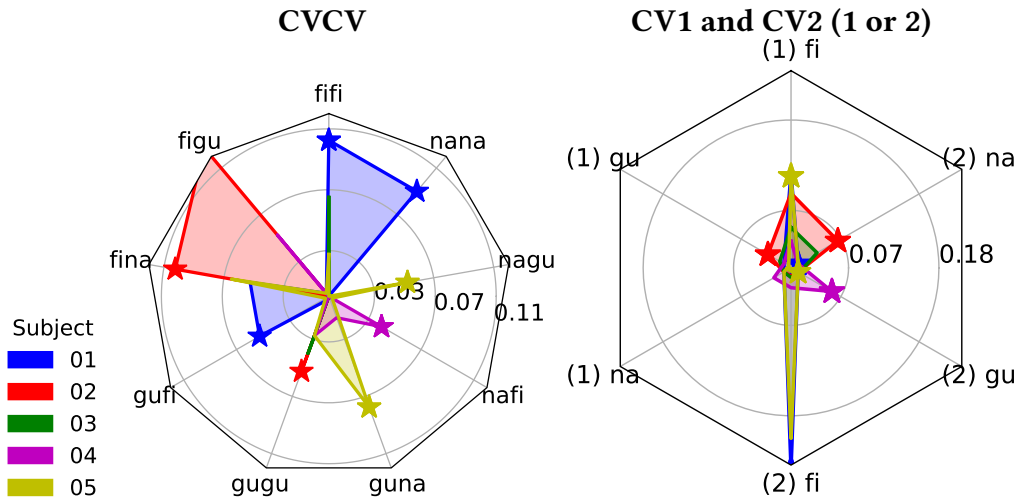


Figure A.4: **Accuracy in VWFA:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.21*	0.11	0.16	0.16*	0.11	0.07	0.11	0.10	0.20**	0.14*
02	0.10	0.24**	0.21*	0.07	0.16*	0.09	0.06	0.07	0.10	0.12
03	0.17	0.10	0.11	0.07	0.15	0.10	0.09	0.11	0.07	0.11
04	0.07	0.16	0.10	0.10	0.14	0.12	0.15*	0.10	0.04	0.11
05	0.14	0.11	0.17	0.11	0.14	0.19*	0.11	0.16*	0.07	0.13*

Table A.3: **Accuracy VWFA CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.45	0.31	0.31	0.35	0.57**	0.28	0.30*	0.38**
02	0.42	0.36*	0.33	0.37*	0.33	0.29	0.40**	0.34
03	0.38	0.35	0.31	0.34	0.34	0.34	0.37	0.35
04	0.36	0.30	0.35	0.34	0.35	0.39*	0.32	0.35
05	0.44**	0.33	0.26	0.35	0.54	0.30*	0.22	0.35

Table A.4: **Accuracy VWFA CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

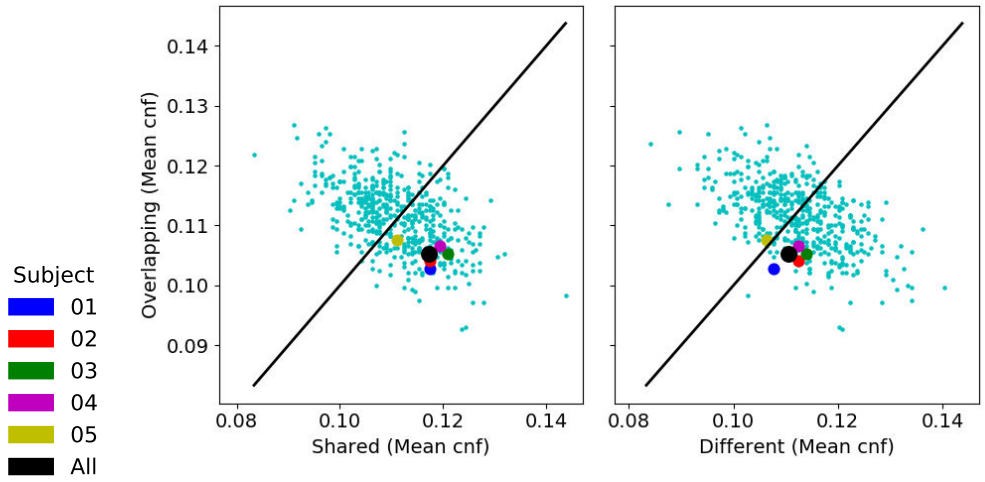


Figure A.5: **Superposition test in VWFA:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

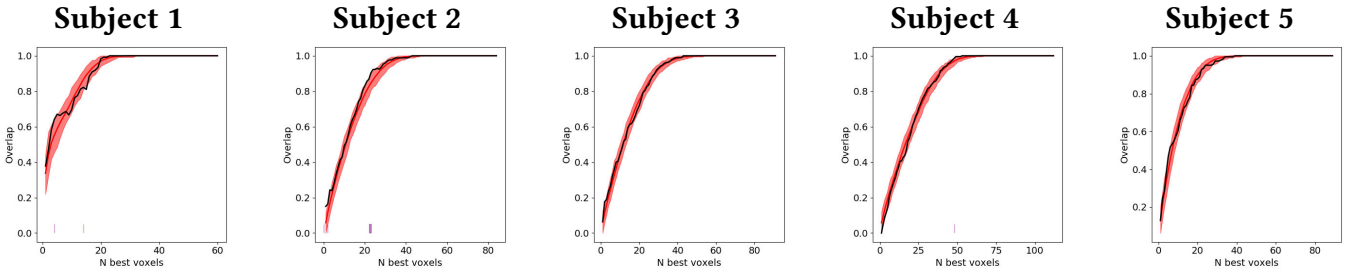


Figure A.6: **Locality test in VWFA:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.3 TP (Visual dataset)

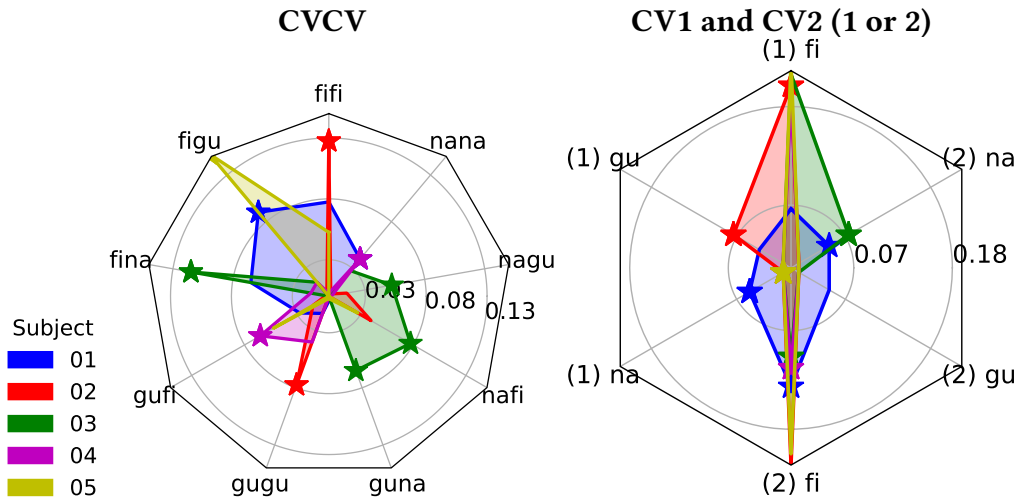


Figure A.7: **Accuracy in TP:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.19	0.20*	0.17	0.14	0.12	0.11	0.11	0.10	0.15*	0.14**
02	0.24**	0.11	0.11	0.12	0.19*	0.09	0.15	0.12	0.11	0.14*
03	0.10	0.12	0.23**	0.06	0.10	0.17*	0.19*	0.16*	0.14	0.14**
04	0.11	0.12	0.12	0.17*	0.15	0.10	0.09	0.11	0.15*	0.13
05	0.16	0.26**	0.10	0.16	0.11	0.11	0.14	0.07	0.09	0.13*

Table A.5: **Accuracy TP CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.40	0.37	0.38*	0.38*	0.46**	0.38	0.38*	0.41**
02	0.53**	0.40**	0.31	0.42**	0.55	0.30	0.25	0.37*
03	0.55	0.25	0.28	0.36	0.43*	0.32	0.40**	0.38**
04	0.53	0.31	0.30*	0.38**	0.44*	0.33	0.32	0.36
05	0.56	0.27	0.33**	0.38**	0.54	0.30	0.27	0.37*

Table A.6: **Accuracy TP CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

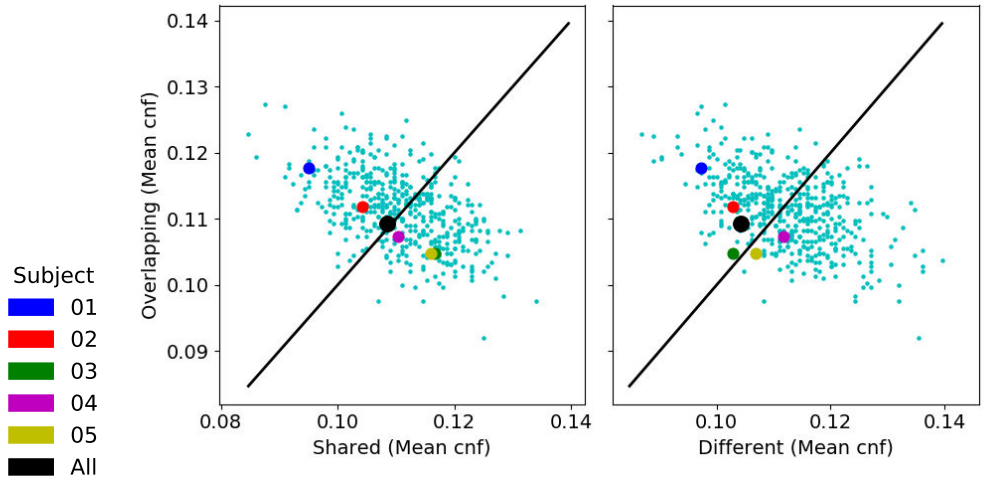


Figure A.8: **Superposition test in TP:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

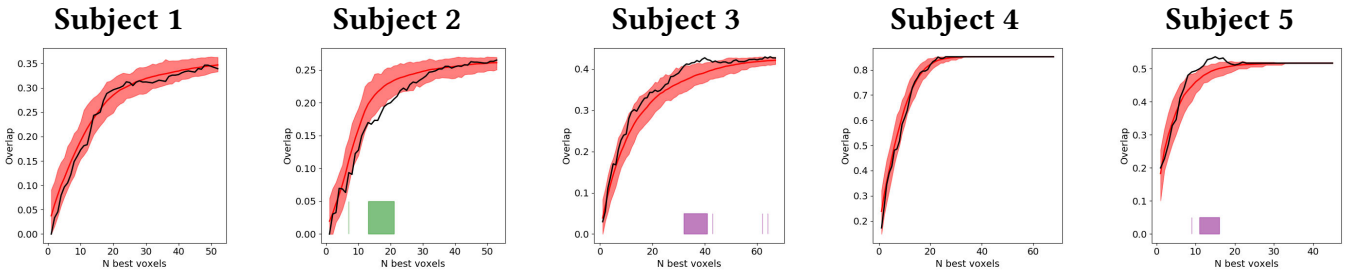


Figure A.9: **Locality test in TP:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.4 TPJ (Visual dataset)

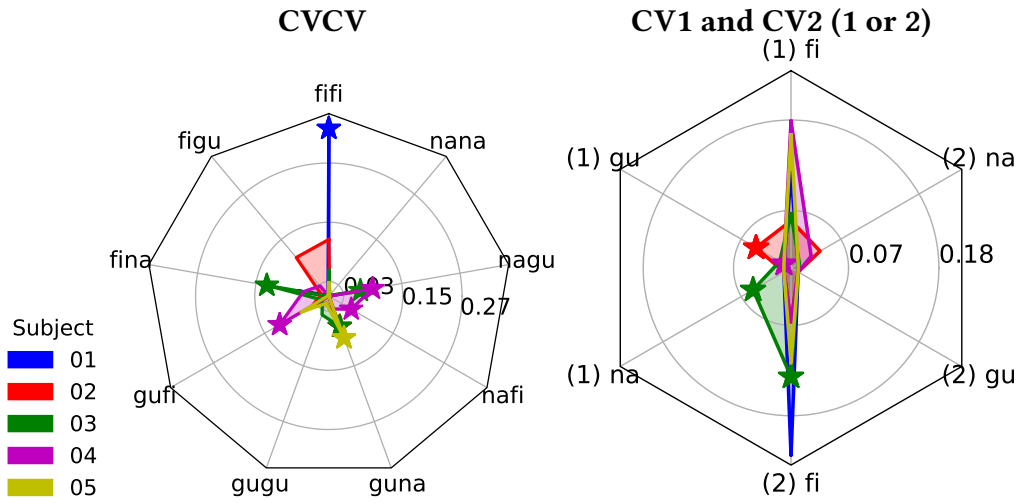


Figure A.10: **Accuracy in TPJ:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.45**	0.11	0.17	0.06	0.09	0.11	0.00	0.00	0.06	0.12
02	0.23	0.21	0.12	0.15	0.12	0.17**	0.03	0.05	0.11	0.13*
03	0.16	0.09	0.24**	0.12	0.15	0.17*	0.07	0.17*	0.11	0.14**
04	0.06	0.14	0.16	0.23*	0.06	0.14	0.16*	0.20*	0.10	0.14*
05	0.14	0.11	0.11	0.17	0.12	0.20*	0.05	0.11	0.11	0.13*

Table A.7: **Accuracy TPJ CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.45	0.26	0.30	0.34	0.56	0.16	0.28	0.33
02	0.39	0.38**	0.33	0.37*	0.40	0.31	0.37	0.36
03	0.40	0.35	0.38*	0.38*	0.46**	0.32	0.34	0.37*
04	0.51	0.32*	0.19	0.34	0.40	0.31	0.36	0.36
05	0.49	0.31	0.28	0.36	0.45	0.26	0.30	0.33

Table A.8: **Accuracy TPJ CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

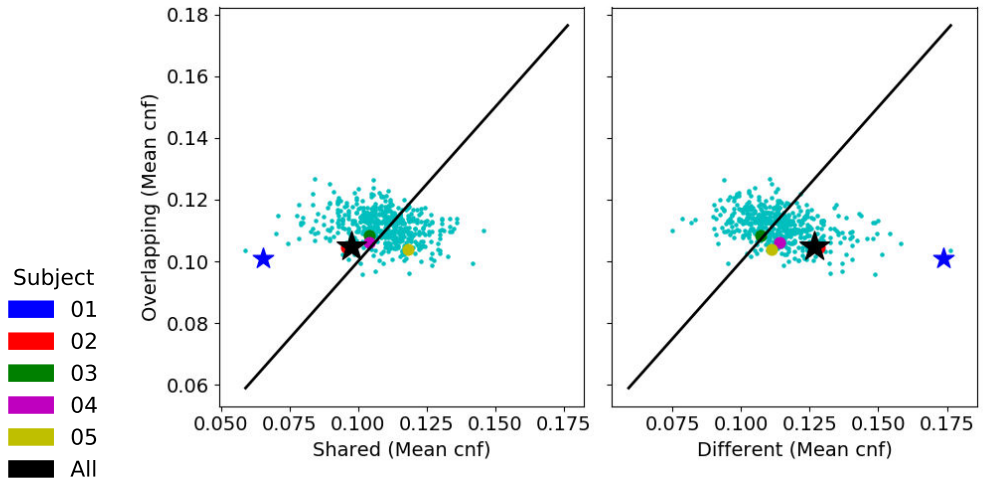


Figure A.11: **Superposition test in TPJ:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

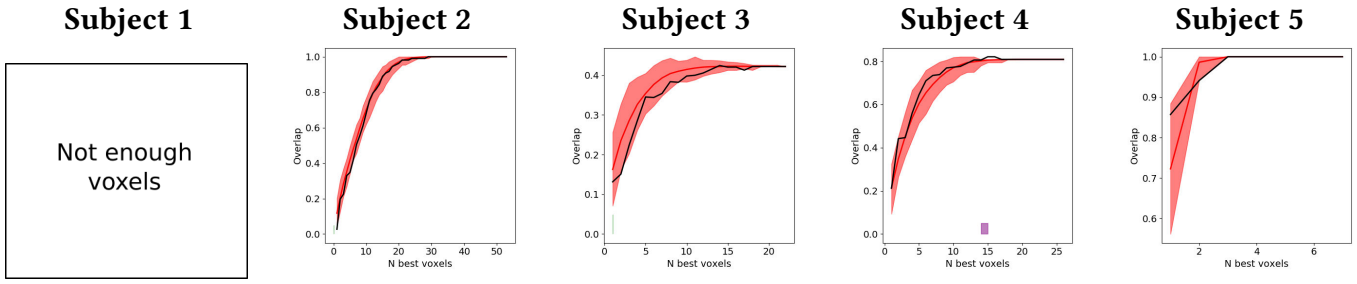


Figure A.12: **Locality test in TPJ:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.5 aSTS (Visual dataset)

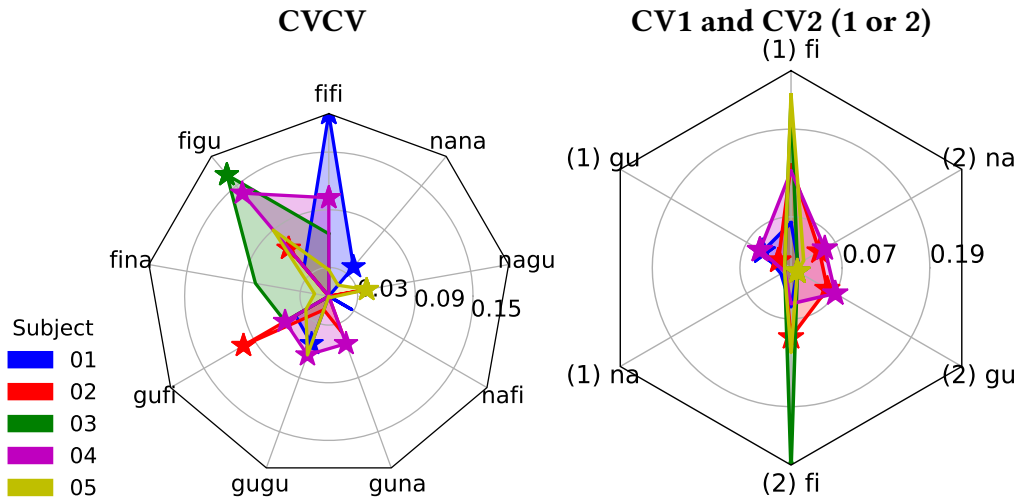


Figure A.13: **Accuracy in aSTS:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.30**	0.15	0.10	0.15	0.16*	0.10	0.14	0.05	0.15*	0.14**
02	0.10	0.17*	0.10	0.21*	0.12	0.15	0.09	0.15	0.10	0.13*
03	0.17	0.28**	0.19	0.16	0.10	0.11	0.09	0.06	0.06	0.14
04	0.21*	0.25**	0.09	0.16*	0.17*	0.16*	0.05	0.11	0.10	0.15**
05	0.14	0.20	0.12	0.14	0.17	0.05	0.11	0.15*	0.12	0.13*

Table A.9: **Accuracy aSTS CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.39	0.38*	0.35	0.37	0.38	0.33	0.34	0.35
02	0.47	0.35**	0.33	0.38**	0.42*	0.39*	0.38**	0.40**
03	0.56	0.28	0.28	0.38*	0.61*	0.33*	0.20	0.38*
04	0.46	0.38**	0.29	0.38**	0.38	0.40*	0.38*	0.39**
05	0.57	0.25	0.25	0.35	0.45	0.33*	0.35	0.38*

Table A.10: **Accuracy aSTS CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

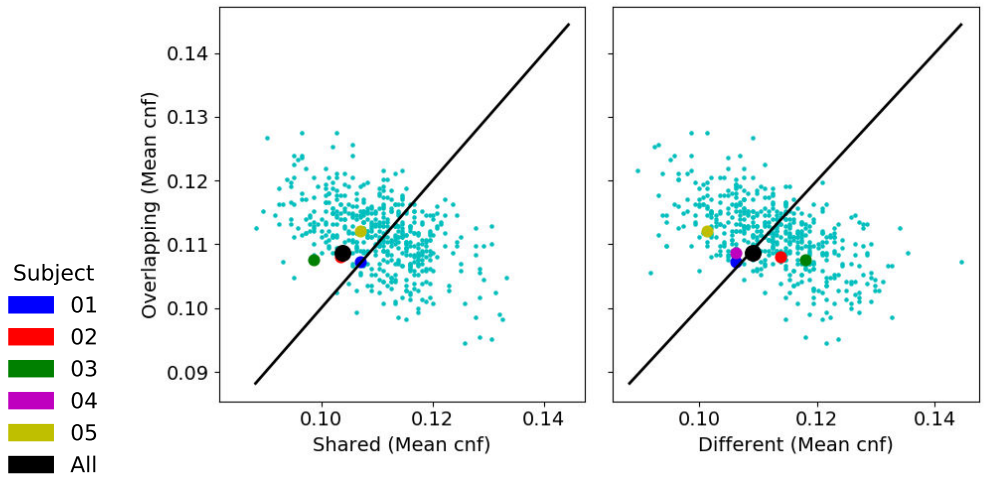


Figure A.14: **Superposition test in aSTS:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

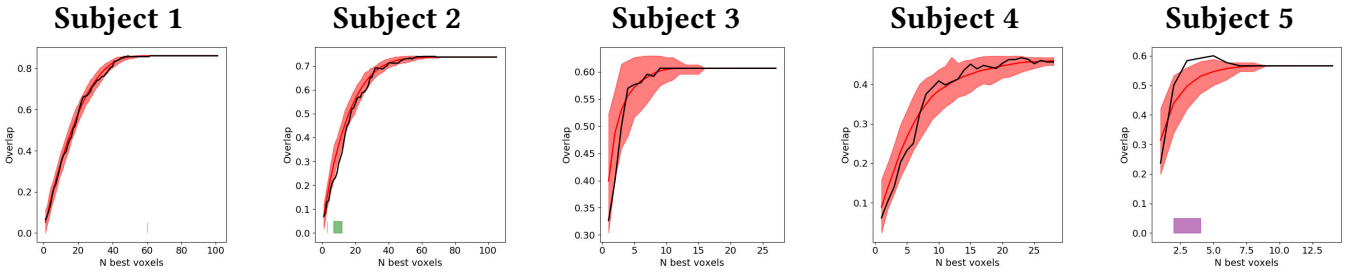


Figure A.15: **Locality test in aSTS:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.6 pSTS (Visual dataset)

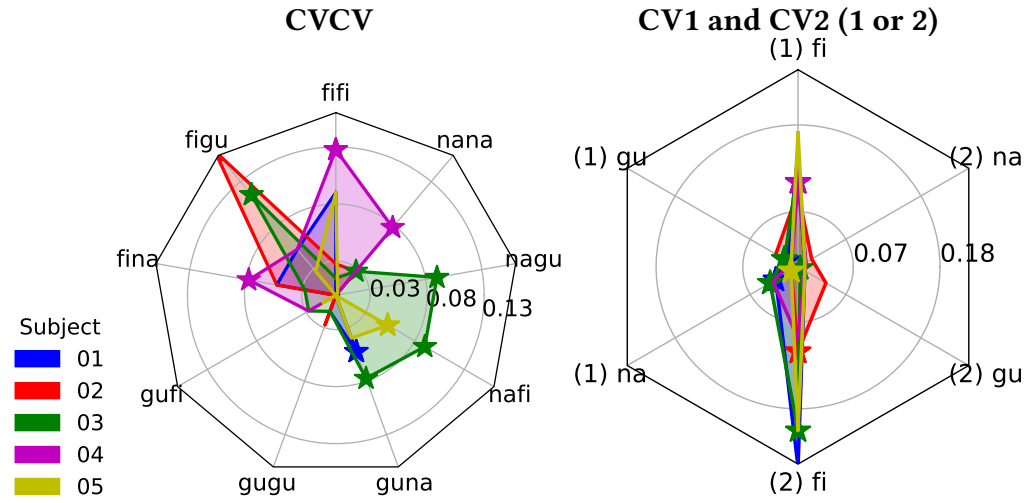


Figure A.16: **Accuracy in pSTS:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.20	0.16	0.16	0.07	0.12	0.16*	0.10	0.11	0.09	0.13*
02	0.14	0.28**	0.16	0.10	0.14	0.07	0.10	0.09	0.14	0.13*
03	0.12	0.23*	0.14	0.14	0.12	0.19**	0.20**	0.20**	0.14*	0.16**
04	0.24*	0.16	0.19*	0.14	0.09	0.11	0.11	0.05	0.19*	0.14**
05	0.20	0.14	0.06	0.12	0.06	0.15	0.16**	0.06	0.11	0.12

Table A.11: **Accuracy pSTS CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.47	0.35*	0.36**	0.39**	0.58	0.28	0.23	0.36*
02	0.42	0.36	0.33	0.37	0.44**	0.37	0.35	0.39**
03	0.45	0.35**	0.37**	0.39**	0.54*	0.33*	0.22	0.36*
04	0.44*	0.32	0.37	0.37*	0.42	0.30	0.34	0.35
05	0.50	0.29	0.28*	0.36	0.54	0.31	0.23	0.36

Table A.12: **Accuracy pSTS CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

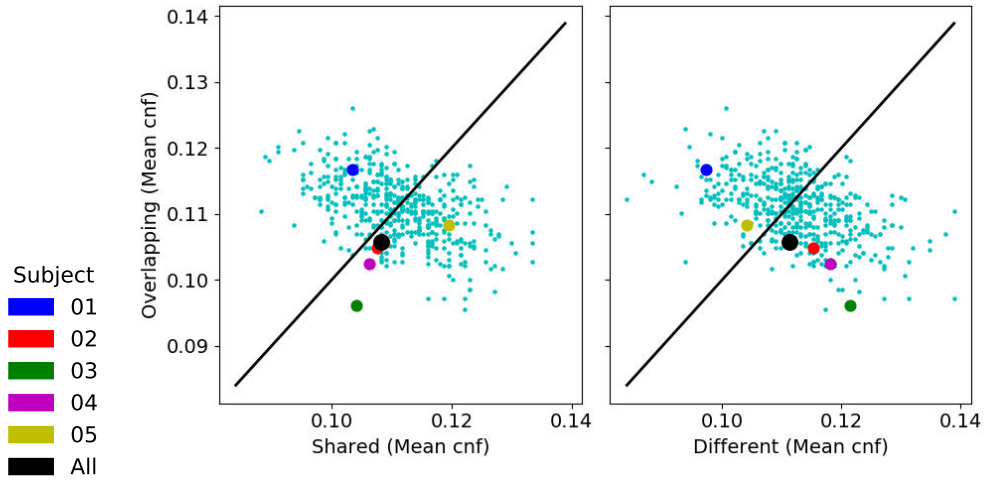


Figure A.17: **Superposition test in pSTS:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

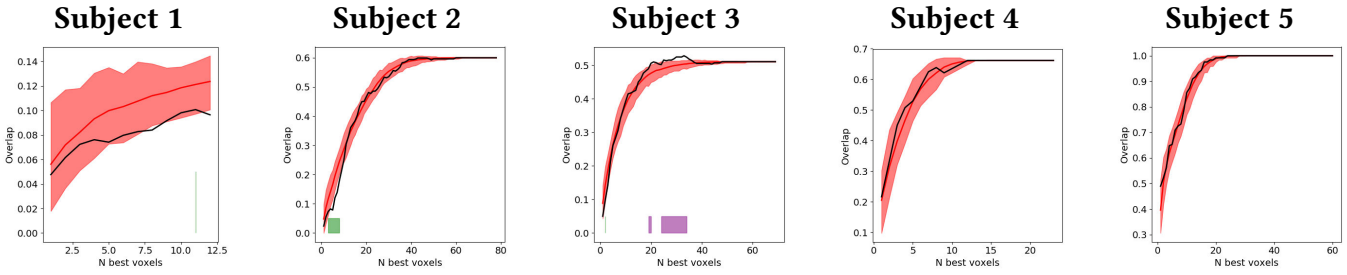


Figure A.18: **Locality test in pSTS:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.7 IFGorb (Visual dataset)

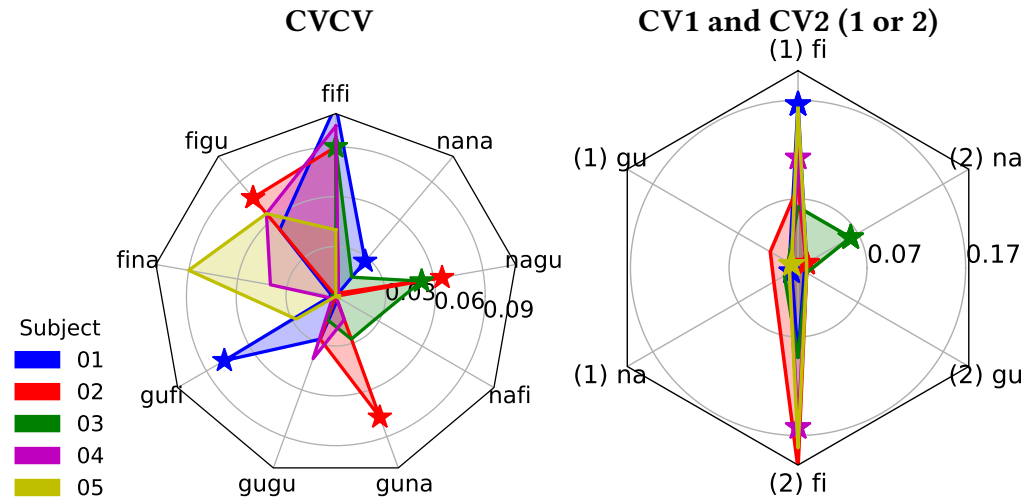


Figure A.19: **Accuracy in IFGorb:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.23**	0.16	0.11	0.19**	0.14	0.11	0.07	0.04	0.14*	0.13**
02	0.20	0.19*	0.09	0.09	0.14	0.19**	0.10	0.17**	0.11	0.14**
03	0.20*	0.07	0.11	0.11	0.12	0.14	0.14	0.16*	0.12	0.13*
04	0.21	0.17	0.15	0.11	0.15	0.12	0.10	0.11	0.11	0.14*
05	0.15	0.17	0.20	0.14	0.05	0.05	0.11	0.11	0.06	0.12

Table A.13: **Accuracy IFGorb CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.50*	0.31	0.34*	0.38**	0.42	0.34	0.30	0.35
02	0.41	0.36	0.36	0.38*	0.53	0.28	0.30*	0.37*
03	0.39	0.34	0.35	0.36	0.42	0.30	0.39*	0.37*
04	0.44*	0.34	0.31	0.36	0.49*	0.32	0.30	0.37*
05	0.49	0.34*	0.32	0.38*	0.51	0.27	0.27	0.35

Table A.14: **Accuracy IFGorb CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

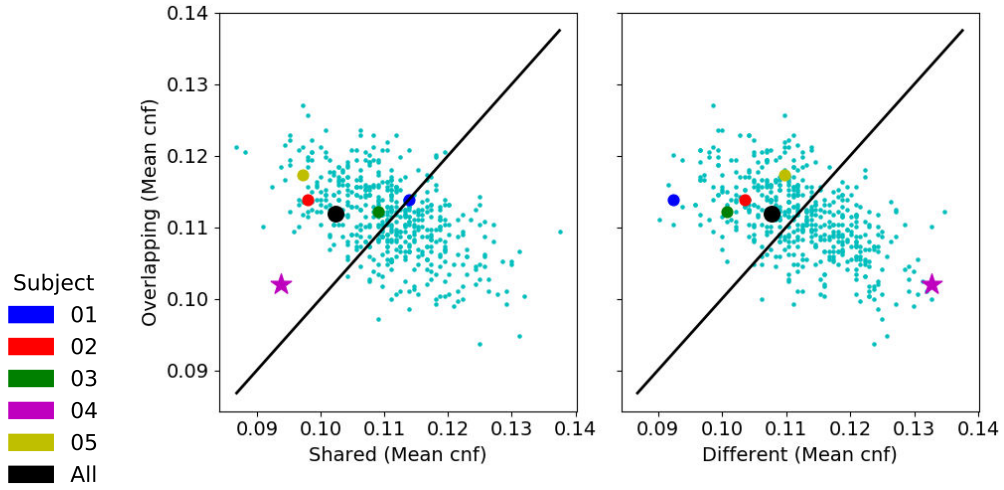


Figure A.20: **Superposition test in IFGorb:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

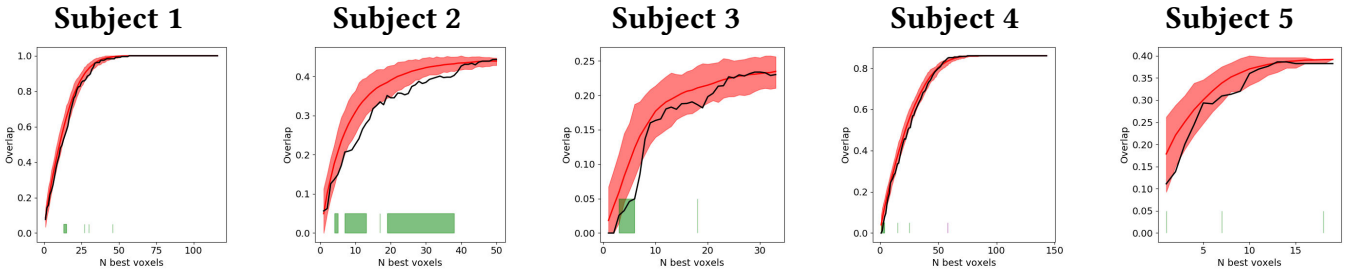


Figure A.21: **Locality test in IFGorb:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.8 IFGtri (Visual dataset)

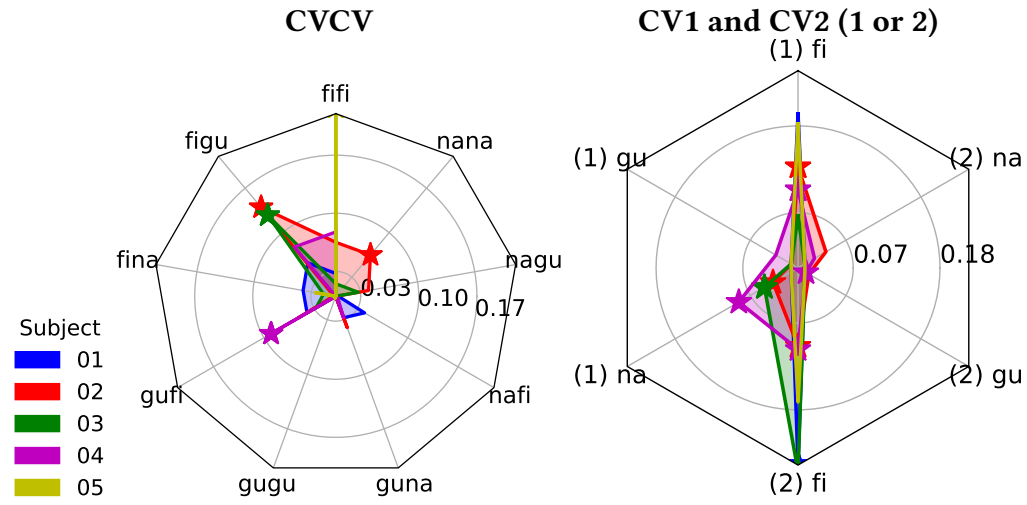


Figure A.22: **Accuracy in IFGtri:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.14	0.16	0.15	0.15	0.06	0.14	0.15	0.11	0.06	0.12
02	0.17	0.25**	0.11	0.16	0.04	0.15	0.10	0.15	0.17**	0.15**
03	0.12	0.24*	0.12	0.14	0.04	0.10	0.09	0.14	0.12	0.12
04	0.19	0.19	0.09	0.20*	0.10	0.14	0.11	0.07	0.07	0.13
05	0.34*	0.10	0.14	0.07	0.11	0.07	0.10	0.09	0.01	0.12

Table A.15: **Accuracy IFGtri CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.53	0.32	0.26	0.37*	0.58**	0.27	0.28	0.38*
02	0.46**	0.32	0.37*	0.38**	0.43*	0.35	0.37	0.38*
03	0.40	0.31	0.38*	0.36	0.59*	0.31	0.20	0.36
04	0.43*	0.36	0.42**	0.40**	0.43*	0.34*	0.35	0.38**
05	0.51	0.23	0.24	0.33	0.50	0.26	0.25	0.34

Table A.16: **Accuracy IFGtri CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

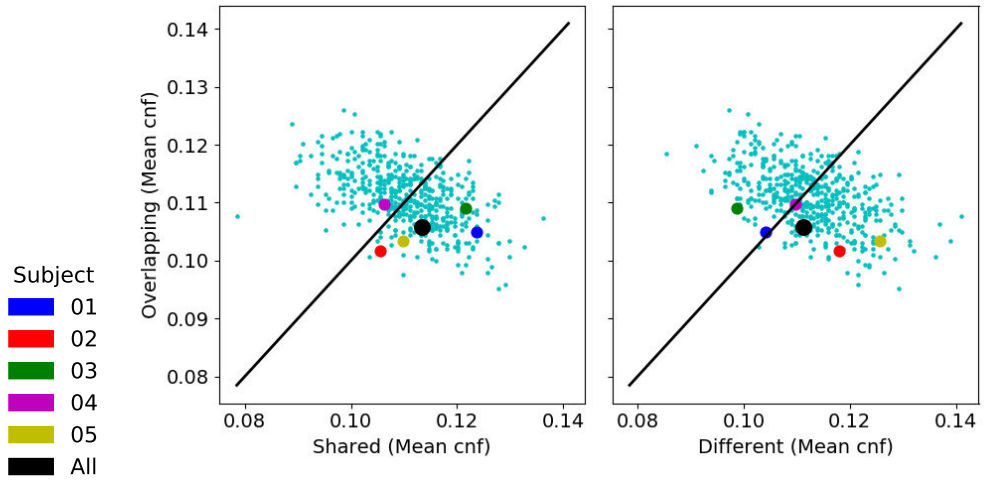


Figure A.23: **Superposition test in IFGtri:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

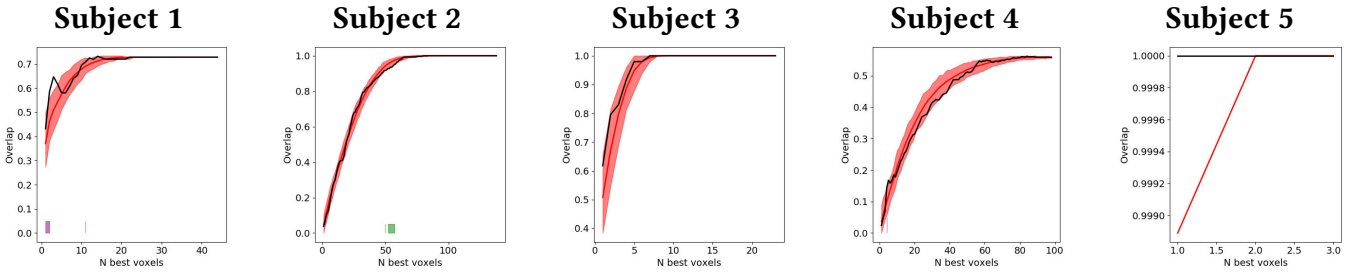


Figure A.24: **Locality test in IFGtri:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.9 Broca-44 (Visual dataset)

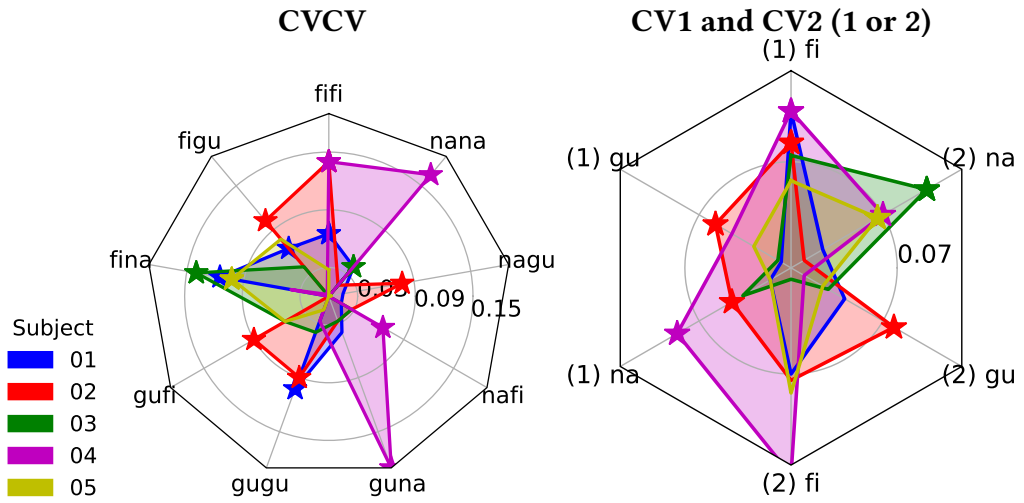


Figure A.25: **Accuracy in Broca-44:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.17*	0.17*	0.23**	0.10	0.21**	0.15	0.12	0.12	0.15	0.16**
02	0.25**	0.21**	0.06	0.20*	0.20**	0.14	0.14	0.19*	0.12	0.17**
03	0.07	0.15	0.25**	0.16	0.15	0.14	0.14	0.05	0.15*	0.14**
04	0.25**	0.11	0.15	0.03	0.14	0.30**	0.17*	0.05	0.28**	0.16**
05	0.14	0.19	0.21**	0.16	0.12	0.10	0.10	0.11	0.10	0.14*

Table A.17: **Accuracy Broca-44 CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.43	0.34	0.35	0.37*	0.40	0.37	0.35	0.38*
02	0.41*	0.39*	0.38*	0.39**	0.40	0.41**	0.32	0.38*
03	0.40	0.32	0.37	0.36*	0.34	0.36	0.43**	0.38*
04	0.43*	0.38	0.42**	0.41**	0.47**	0.33	0.40**	0.40**
05	0.39	0.36	0.35	0.36	0.41	0.35	0.40**	0.39*

Table A.18: **Accuracy Broca-44 CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

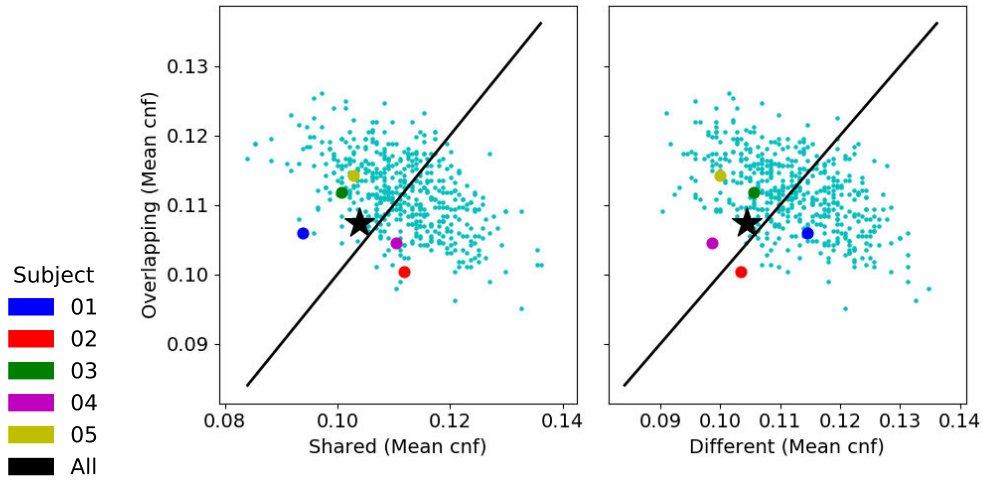


Figure A.26: **Superposition test in Broca-44:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

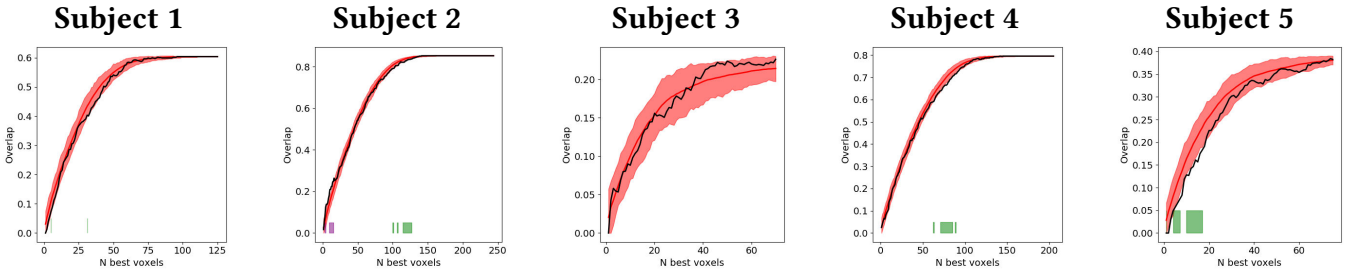


Figure A.27: **Locality test in Broca-44:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.10 Broca-45 (Visual dataset)

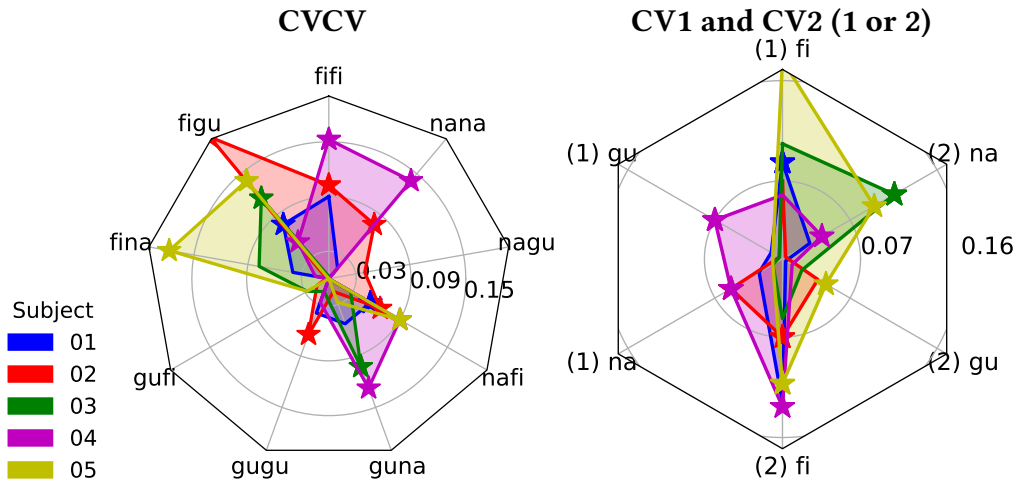


Figure A.28: **Accuracy in Broca-45:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.20	0.19*	0.15	0.10	0.15	0.16	0.16*	0.10	0.12	0.15**
02	0.21*	0.31**	0.11	0.12	0.17*	0.12	0.17*	0.15	0.19**	0.17**
03	0.11	0.23**	0.19	0.14	0.12	0.21**	0.14	0.06	0.06	0.14**
04	0.26**	0.16*	0.12	0.06	0.14	0.24**	0.20**	0.07	0.25**	0.17**
05	0.07	0.25**	0.29**	0.14	0.09	0.14	0.20**	0.10	0.07	0.15**

Table A.19: **Accuracy Broca-45 CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.42*	0.34	0.35	0.37	0.46	0.33	0.36	0.38**
02	0.40	0.33	0.38*	0.37*	0.40*	0.37	0.33	0.37*
03	0.43	0.33	0.31	0.36	0.38	0.35	0.45**	0.39**
04	0.39	0.40*	0.38*	0.39*	0.46*	0.33	0.37*	0.39*
05	0.51	0.28	0.25	0.34	0.44*	0.38**	0.42**	0.41**

Table A.20: **Accuracy Broca-45 CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

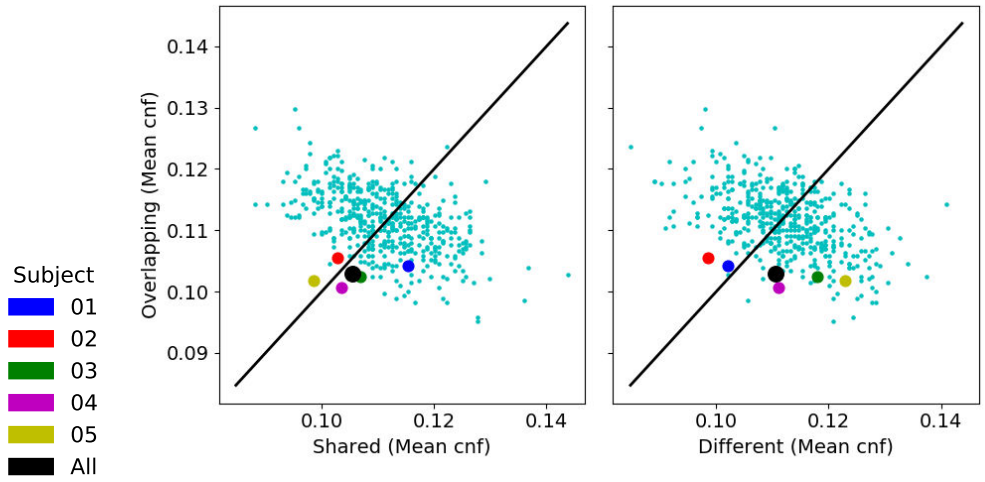


Figure A.29: **Superposition test in Broca-45:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

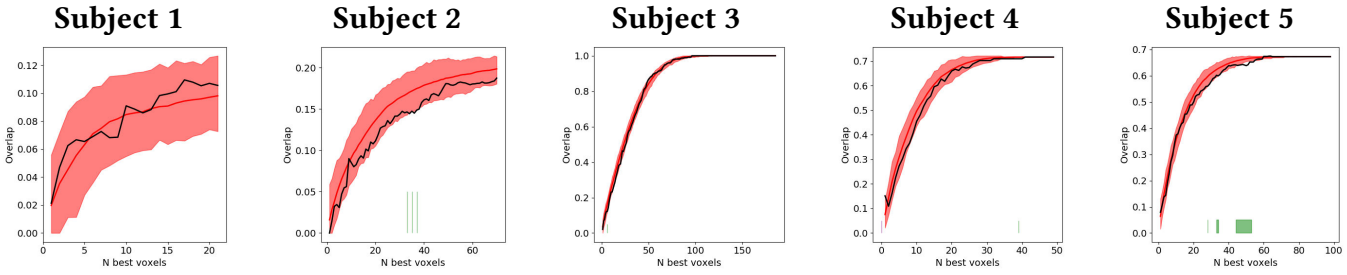


Figure A.30: **Locality test in Broca-45:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.11 VWFA (Auditory dataset)

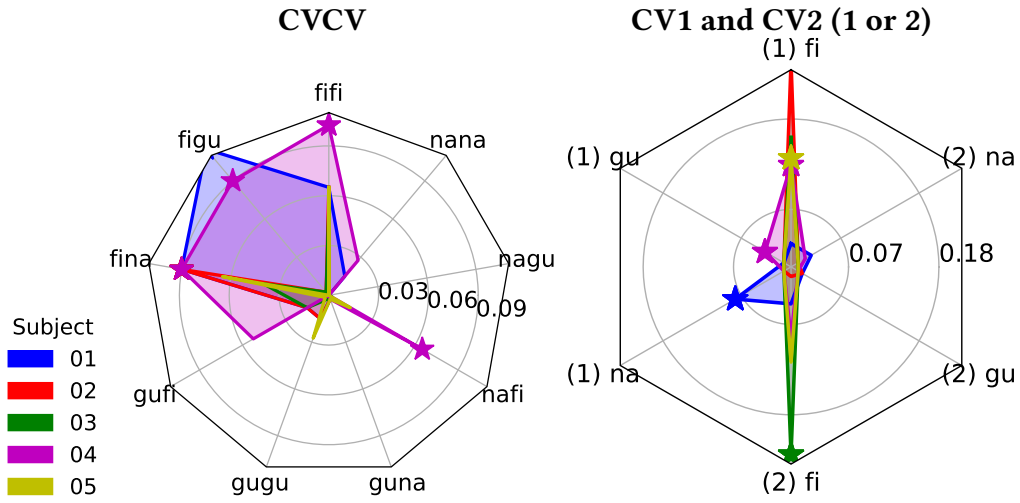


Figure A.31: **Accuracy in VWFA:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.17	0.23**	0.20*	0.12	0.12	0.11	0.11	0.10	0.12	0.14*
02	0.17	0.11	0.20*	0.12	0.12	0.11	0.05	0.07	0.09	0.12
03	0.16	0.11	0.15	0.12	0.11	0.07	0.11	0.05	0.09	0.11
04	0.21*	0.20*	0.20*	0.16	0.10	0.11	0.17**	0.09	0.14	0.15**
05	0.17	0.09	0.17	0.11	0.14	0.11	0.12	0.10	0.10	0.12

Table A.21: **Accuracy VWFA CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.36	0.33	0.41*	0.37	0.38	0.35	0.36	0.36
02	0.58*	0.25	0.24	0.36	0.34	0.35	0.30	0.33
03	0.49	0.30	0.24	0.34	0.56*	0.22	0.23	0.34
04	0.45*	0.37*	0.28	0.37*	0.42	0.27	0.35	0.34
05	0.46*	0.31	0.29	0.36	0.45	0.26	0.27	0.33

Table A.22: **Accuracy VWFA CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

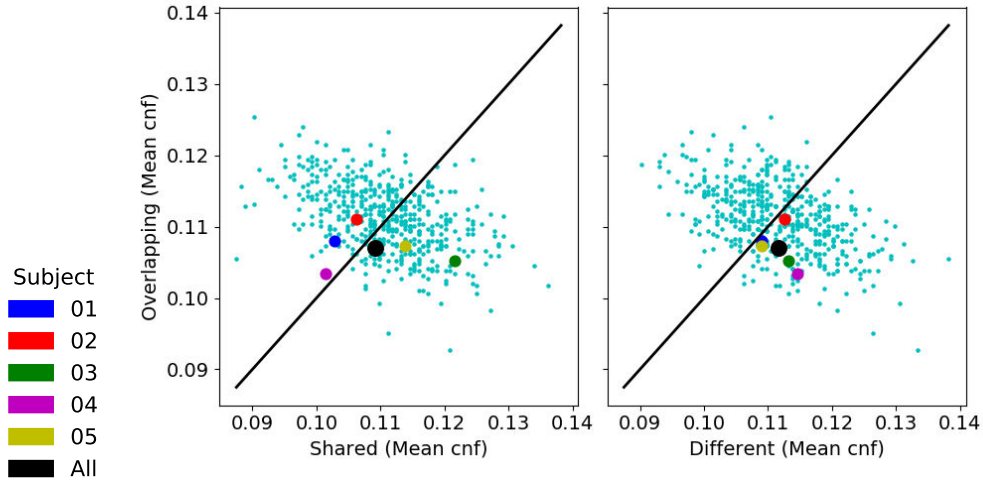


Figure A.32: **Superposition test in VWFA:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

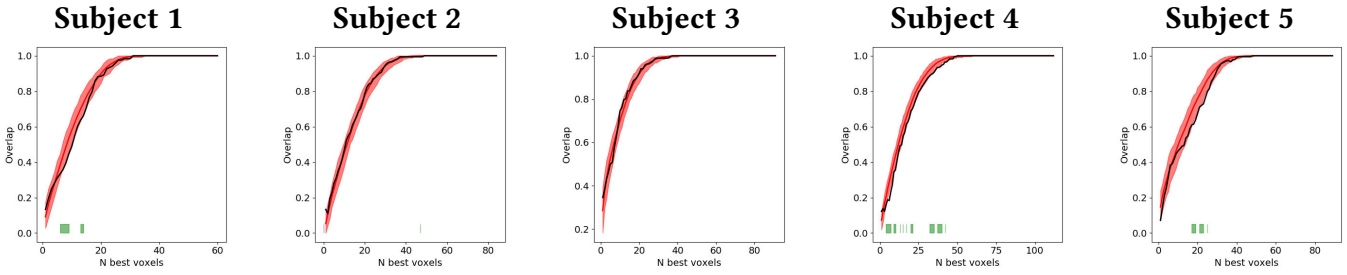


Figure A.33: **Locality test in VWFA:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.12 Auditory-Te10 (Auditory dataset)

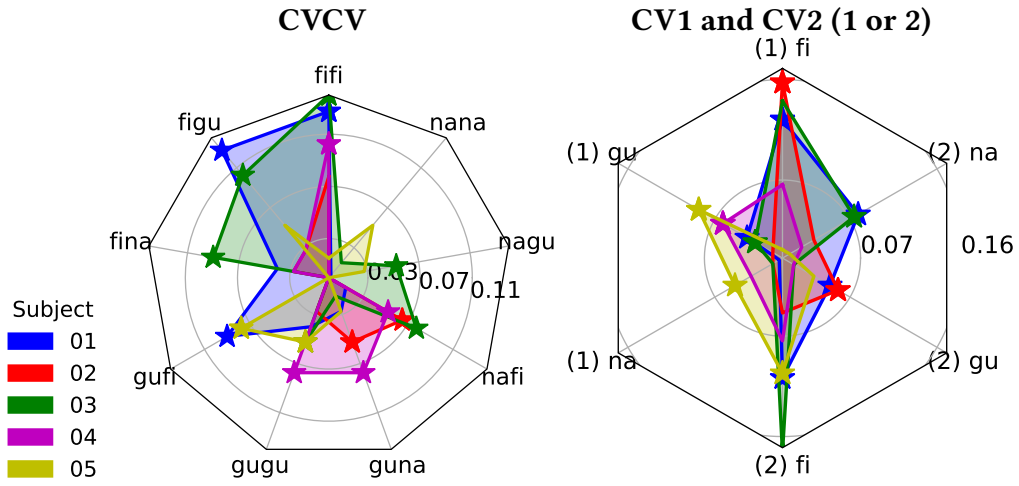


Figure A.34: **Accuracy in Auditory-Te10:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.24**	0.24**	0.15	0.20*	0.15	0.14	0.12	0.09	0.11	0.16**
02	0.19	0.14	0.14	0.10	0.14	0.16*	0.17**	0.09	0.05	0.13*
03	0.25**	0.21*	0.20**	0.10	0.16*	0.12	0.19**	0.16*	0.12	0.17**
04	0.21**	0.14	0.14	0.07	0.19**	0.19**	0.16*	0.09	0.09	0.14**
05	0.12	0.16	0.09	0.19*	0.16*	0.14	0.06	0.14	0.16	0.14*

Table A.23: **Accuracy Auditory-Te10 CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.45*	0.37*	0.33	0.38**	0.44*	0.38*	0.41*	0.41**
02	0.49**	0.29	0.32	0.37*	0.38	0.39*	0.36	0.38*
03	0.47	0.36*	0.33	0.39*	0.50**	0.34	0.40**	0.42**
04	0.40	0.39**	0.36	0.38*	0.40	0.34	0.35	0.37
05	0.34	0.42*	0.38*	0.38*	0.43*	0.36	0.34	0.38**

Table A.24: **Accuracy Auditory-Te10 CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

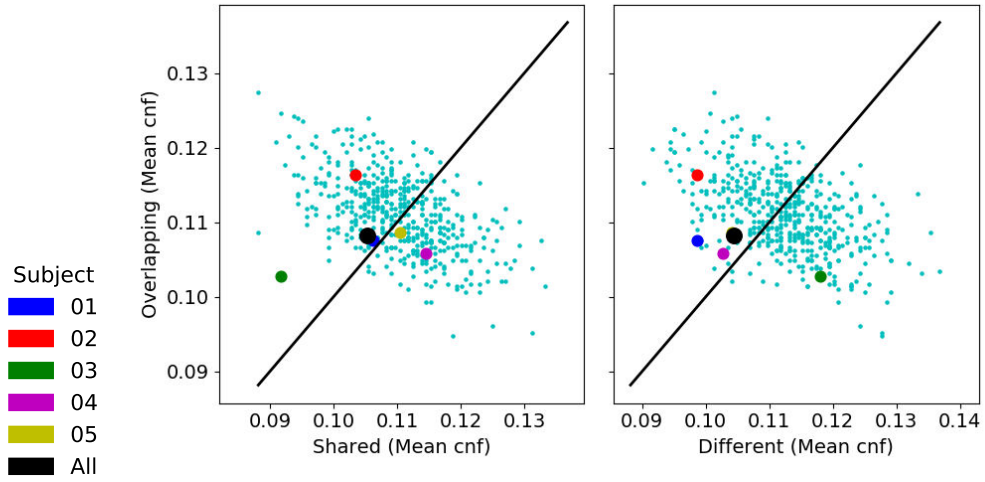


Figure A.35: **Superposition test in Auditory-Te10:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

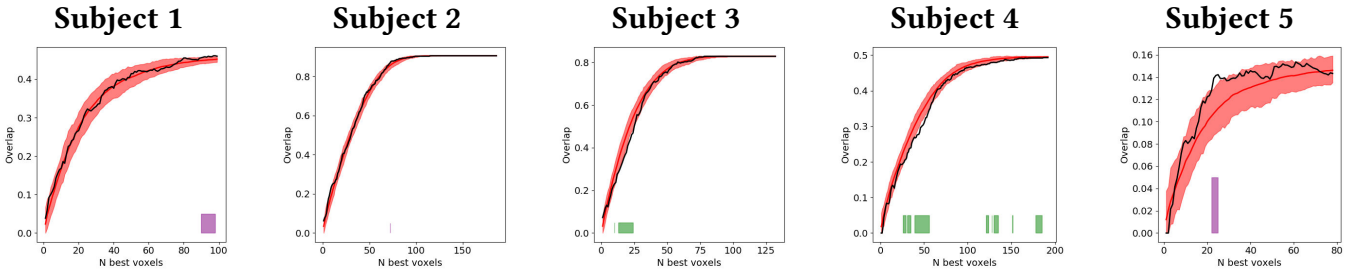


Figure A.36: **Locality test in Auditory-Te10:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.13 Auditory-Te11 (Auditory dataset)

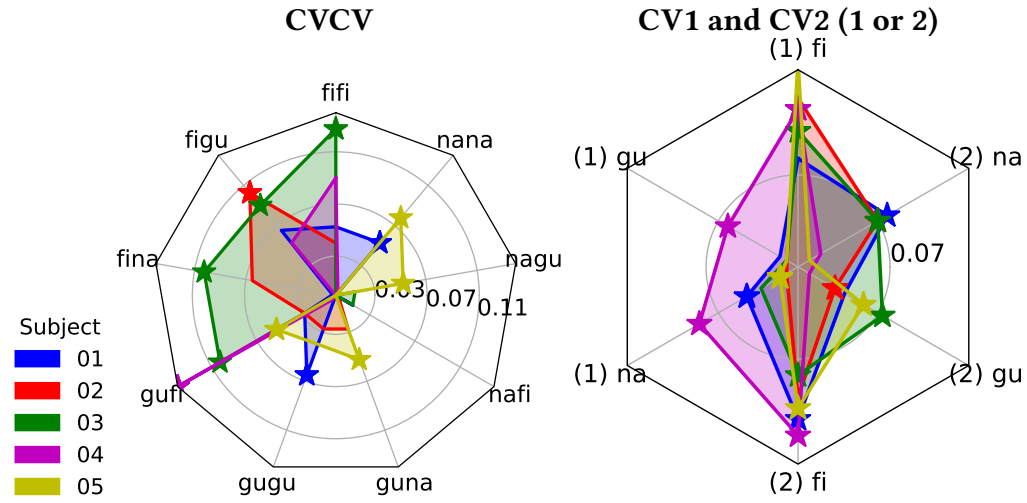


Figure A.37: **Accuracy in Auditory-Te11:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.16	0.17	0.11	0.14	0.17*	0.09	0.12	0.10	0.16*	0.14*
02	0.15	0.21*	0.17	0.14	0.14	0.14	0.11	0.11	0.09	0.14*
03	0.24**	0.20*	0.21**	0.21**	0.10	0.10	0.12	0.12	0.09	0.16**
04	0.20	0.16	0.10	0.25**	0.11	0.10	0.09	0.10	0.11	0.14*
05	0.09	0.10	0.06	0.16*	0.15	0.16*	0.09	0.16*	0.19*	0.13*

Table A.25: **Accuracy Auditory-Te11 CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.41	0.35	0.38*	0.38*	0.45*	0.37	0.41*	0.41**
02	0.46	0.32	0.32	0.37*	0.43	0.36*	0.40**	0.40**
03	0.43**	0.33	0.36	0.37*	0.41*	0.40*	0.40**	0.41**
04	0.45*	0.39*	0.42**	0.42**	0.46*	0.30	0.35	0.37*
05	0.49	0.28	0.35*	0.37*	0.44*	0.39**	0.33	0.38**

Table A.26: **Accuracy Auditory-Te11 CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

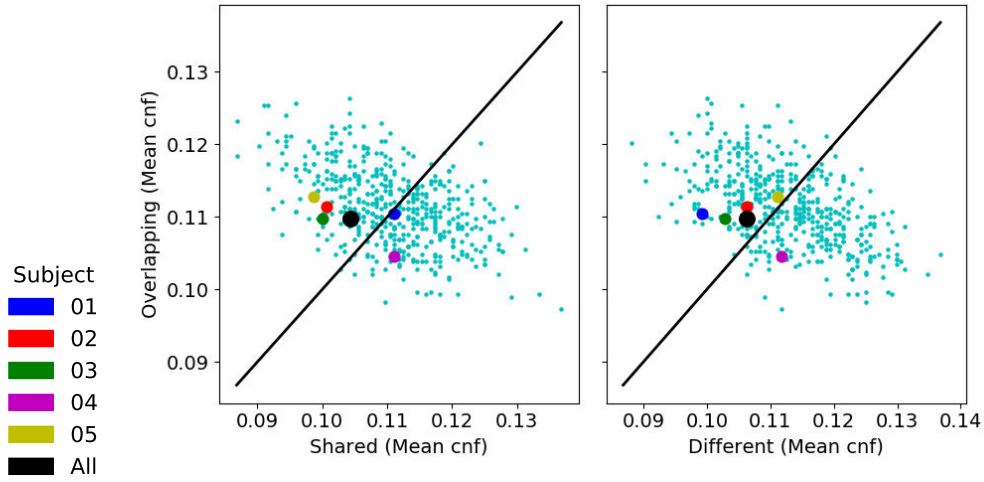


Figure A.38: **Superposition test in Auditory-Te11:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

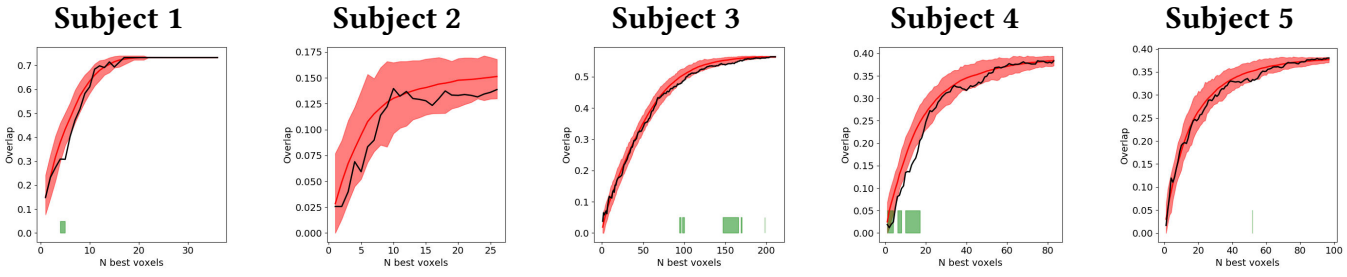


Figure A.39: **Locality test in Auditory-Te11:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.14 Auditory-Te12 (Auditory dataset)

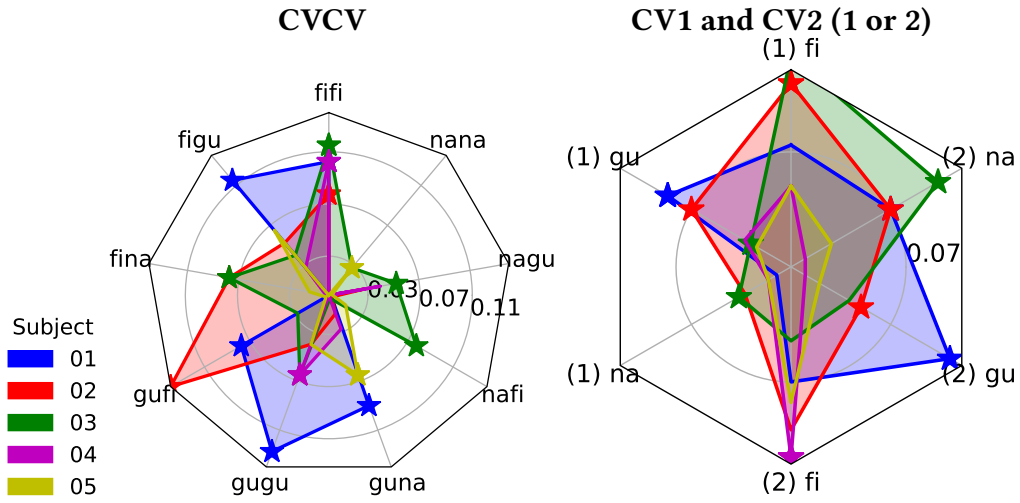


Figure A.40: **Accuracy in Auditory-Te12:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.21*	0.23**	0.10	0.19*	0.24**	0.20**	0.11	0.12	0.10	0.17**
02	0.19*	0.16	0.19*	0.25**	0.15	0.12	0.11	0.14	0.09	0.16**
03	0.23**	0.15	0.19*	0.14	0.17*	0.11	0.19**	0.16*	0.14	0.16**
04	0.21**	0.14	0.11	0.05	0.17*	0.14	0.10	0.15	0.09	0.13*
05	0.09	0.17	0.12	0.11	0.15	0.17*	0.12	0.11	0.14*	0.13*

Table A.27: **Accuracy Auditory-Te12 CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.40	0.42**	0.33	0.38*	0.40	0.44**	0.40*	0.41**
02	0.44**	0.40**	0.36	0.40**	0.43	0.38**	0.40**	0.40**
03	0.46*	0.36*	0.37**	0.39**	0.38	0.37	0.43**	0.39**
04	0.38	0.36	0.35	0.36*	0.45**	0.31	0.30	0.35
05	0.38	0.35	0.35	0.36	0.41	0.35	0.36	0.37

Table A.28: **Accuracy Auditory-Te12 CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

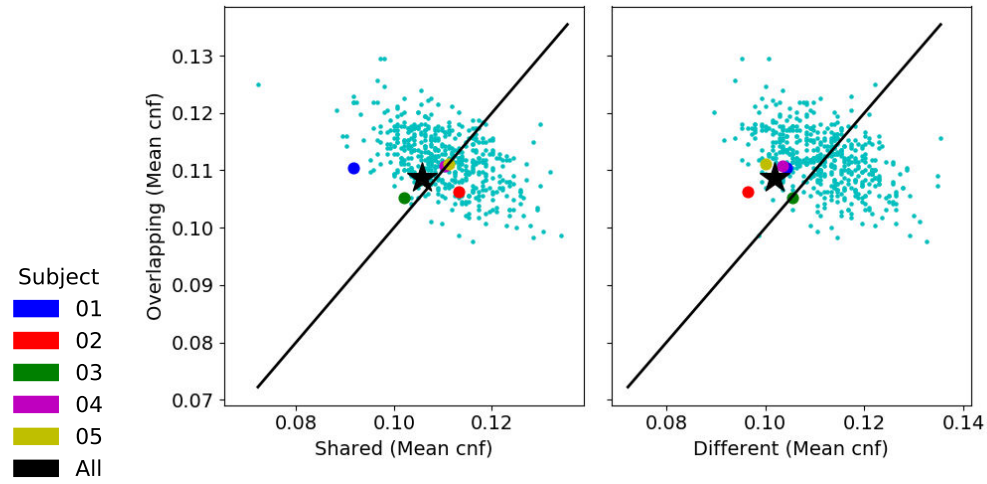


Figure A.41: **Superposition test in Auditory-Te12:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

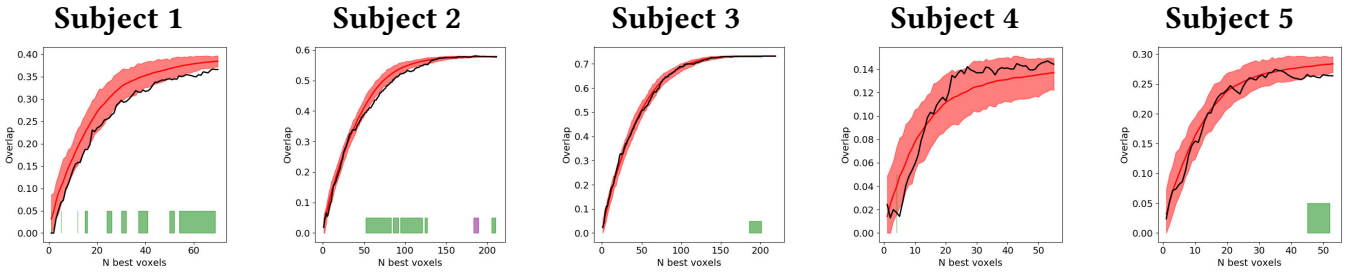


Figure A.42: **Locality test in Auditory-Te12:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.15 TP (Auditory dataset)

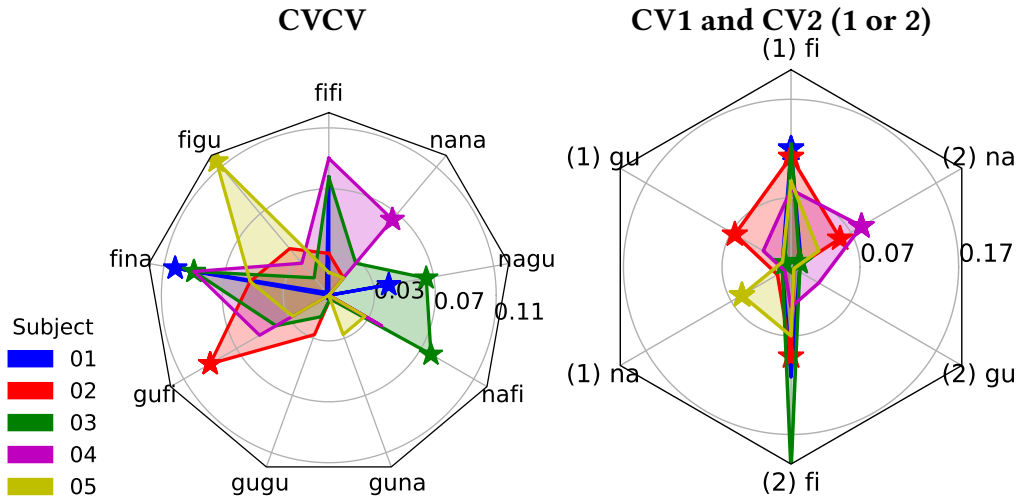


Figure A.43: **Accuracy in TP:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.19	0.11	0.21*	0.05	0.11	0.11	0.06	0.15*	0.07	0.12
02	0.14	0.15	0.16	0.20**	0.14	0.11	0.14	0.10	0.12	0.14*
03	0.19	0.12	0.20**	0.15	0.12	0.11	0.19*	0.17**	0.14	0.16**
04	0.20	0.14	0.20	0.16	0.09	0.05	0.15	0.07	0.17**	0.14*
05	0.12	0.23**	0.16	0.14	0.10	0.14	0.14	0.10	0.12	0.14*

Table A.29: **Accuracy TP CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.45*	0.31	0.31	0.36	0.44	0.32	0.34	0.37**
02	0.44**	0.40**	0.35	0.39*	0.42*	0.33	0.39*	0.38*
03	0.45	0.34*	0.28	0.36	0.54	0.29	0.29*	0.37*
04	0.41	0.36	0.34	0.37	0.37	0.36	0.41**	0.38*
05	0.42	0.33	0.39**	0.38*	0.40	0.33	0.36	0.37*

Table A.30: **Accuracy TP CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

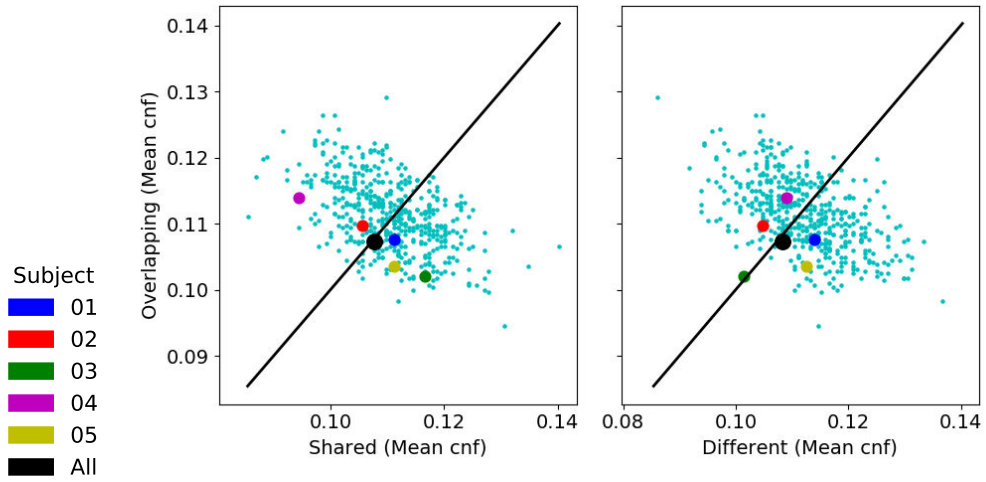


Figure A.44: **Superposition test in TP:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

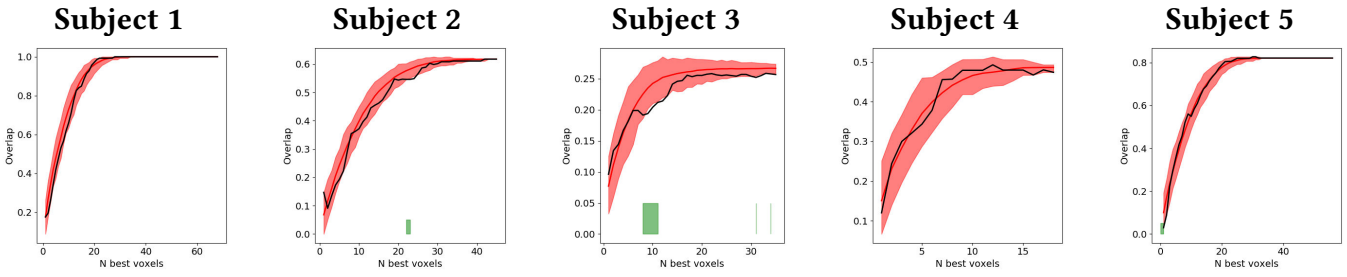


Figure A.45: **Locality test in TP:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.16 TPJ (Auditory dataset)

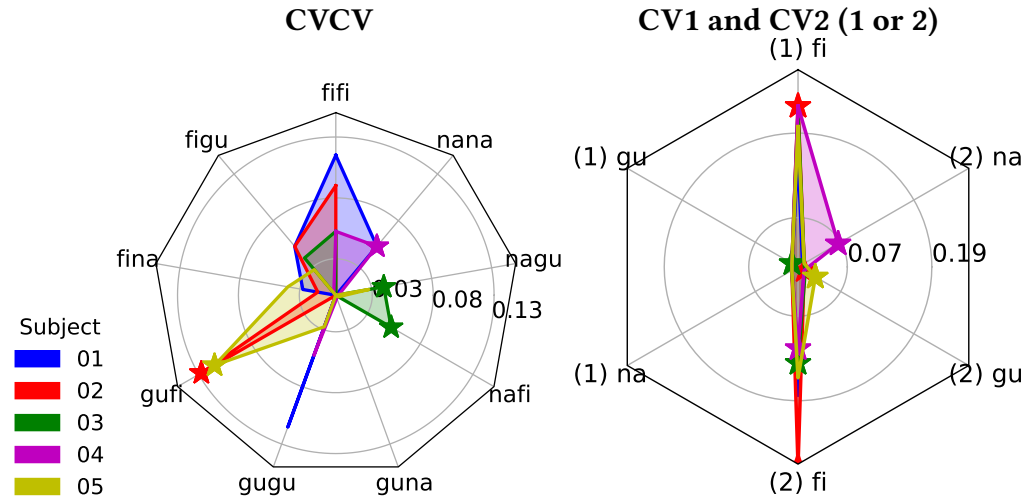


Figure A.46: **Accuracy in TPJ:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.23	0.16	0.14	0.00	0.23	0.07	0.00	0.05	0.16	0.12
02	0.20	0.16	0.12	0.24**	0.10	0.07	0.06	0.09	0.05	0.12
03	0.16	0.15	0.09	0.10	0.10	0.09	0.16*	0.15*	0.06	0.12
04	0.16	0.11	0.10	0.04	0.16	0.11	0.07	0.11	0.16*	0.12
05	0.09	0.14	0.15	0.23*	0.14	0.07	0.11	0.14	0.06	0.12*

Table A.31: **Accuracy TPJ CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.49	0.30	0.22	0.34	0.51	0.15	0.34	0.33
02	0.56*	0.30	0.26	0.37*	0.61**	0.33*	0.25	0.40**
03	0.55	0.30*	0.28	0.38**	0.47**	0.34	0.33	0.38**
04	0.56	0.24	0.25	0.35	0.45*	0.33	0.40*	0.39**
05	0.53	0.25	0.23	0.34	0.49	0.36*	0.30	0.38**

Table A.32: **Accuracy TPJ CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

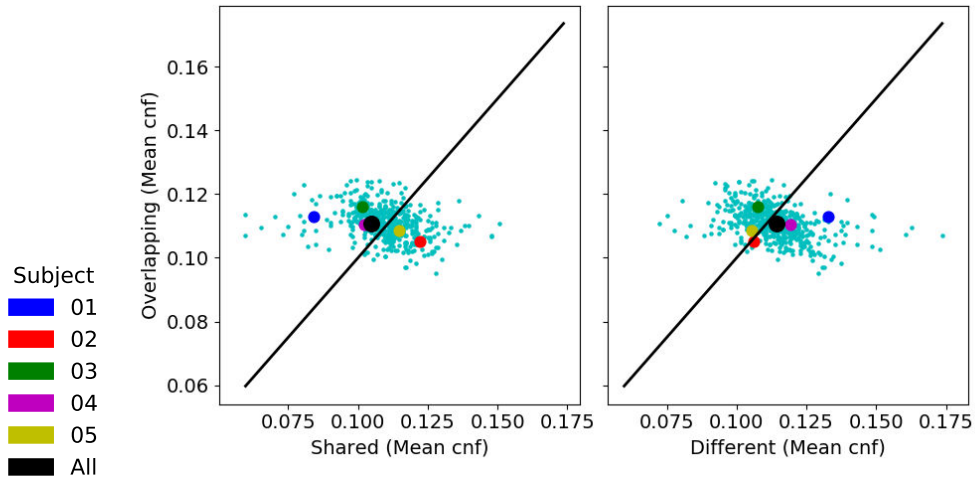


Figure A.47: **Superposition test in TPJ:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

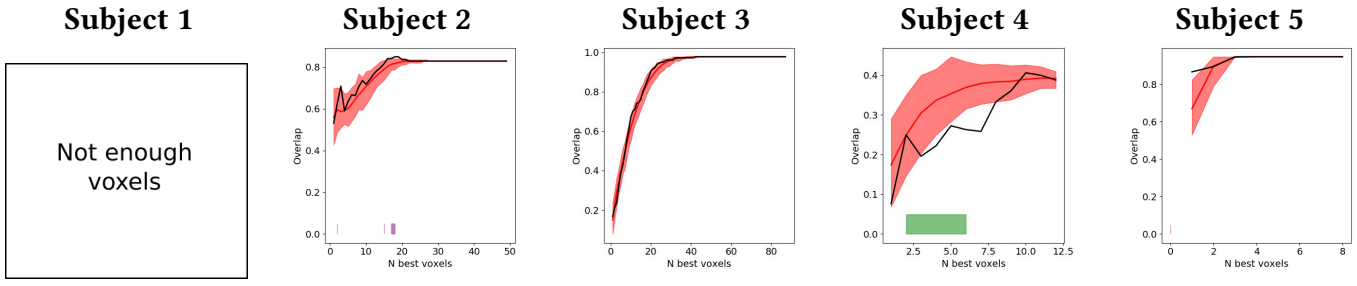


Figure A.48: **Locality test in TPJ:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.17 aSTS (Auditory dataset)

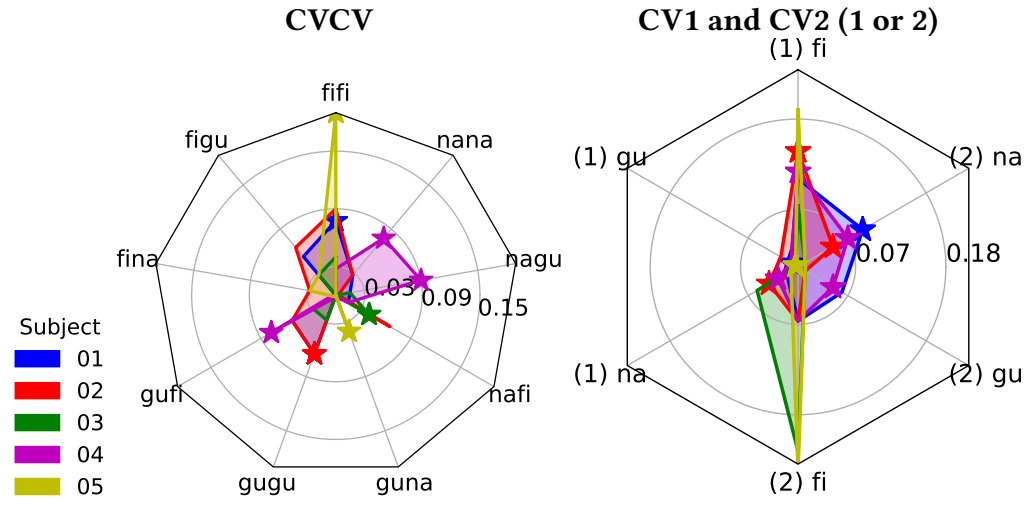


Figure A.49: **Accuracy in aSTS:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.19*	0.16	0.07	0.16	0.17*	0.10	0.12	0.12	0.14	0.14*
02	0.20	0.17	0.14	0.16	0.17*	0.09	0.17	0.11	0.14	0.15**
03	0.15	0.14	0.10	0.14	0.14	0.07	0.15*	0.12	0.07	0.12
04	0.14	0.07	0.11	0.19*	0.11	0.11	0.12	0.20**	0.19*	0.14*
05	0.30*	0.14	0.14	0.11	0.10	0.15*	0.07	0.10	0.06	0.13

Table A.33: **Accuracy aSTS CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.44	0.33*	0.35	0.37	0.40	0.39	0.42**	0.40**
02	0.47**	0.35	0.37*	0.40**	0.40	0.31	0.38*	0.36
03	0.40	0.33	0.39	0.38*	0.55	0.30	0.28	0.38*
04	0.45*	0.31	0.36**	0.37**	0.40	0.38*	0.40*	0.39**
05	0.52	0.33*	0.22	0.36	0.58	0.27	0.22	0.36

Table A.34: **Accuracy aSTS CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

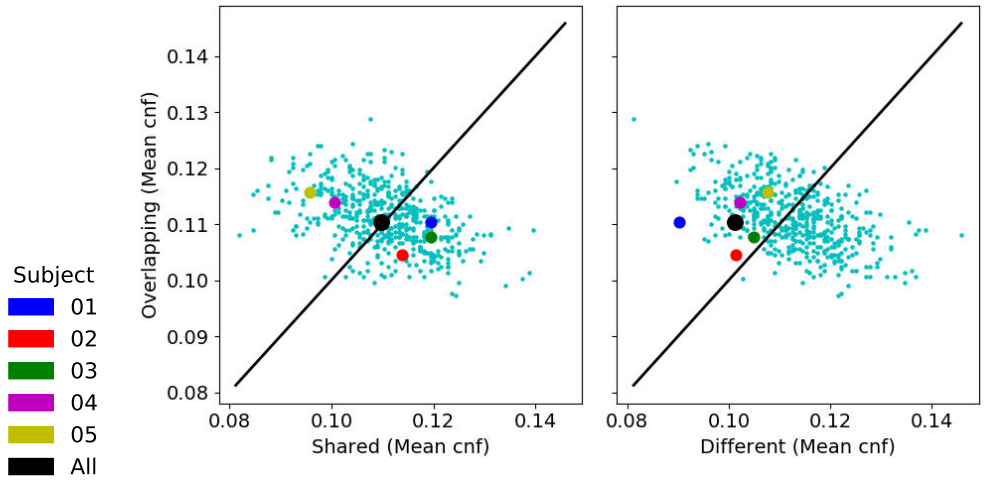


Figure A.50: **Superposition test in aSTS:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

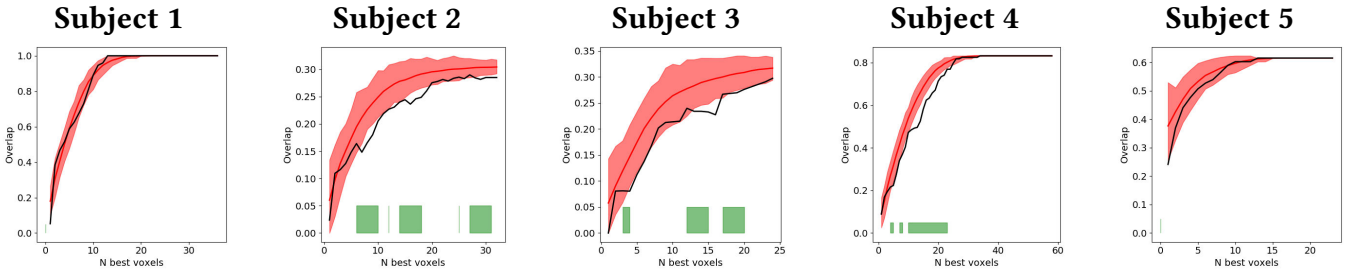


Figure A.51: **Locality test in aSTS:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.18 pSTS (Auditory dataset)

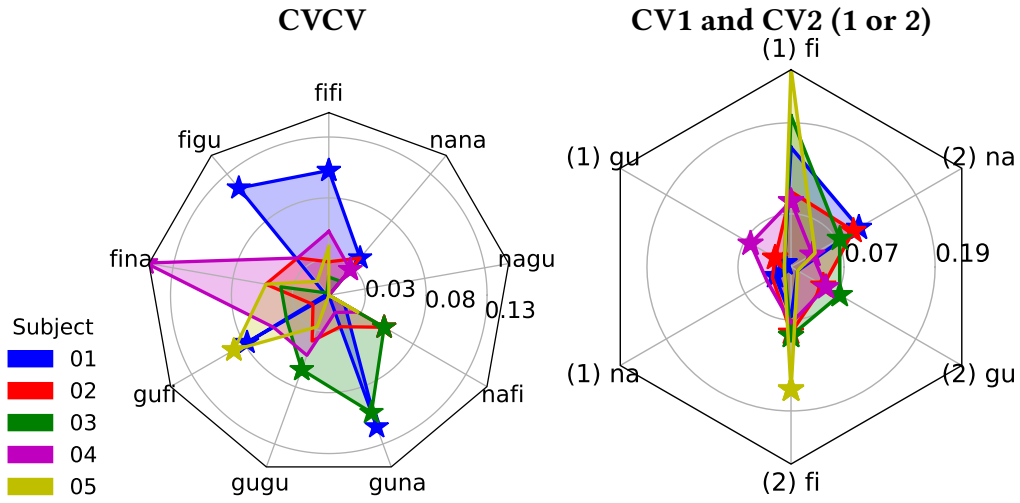


Figure A.52: **Accuracy in pSTS:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.21*	0.23*	0.11	0.19*	0.05	0.23**	0.12	0.06	0.15*	0.15**
02	0.14	0.15	0.16	0.12	0.15	0.14	0.16*	0.09	0.15	0.14*
03	0.12	0.11	0.15	0.15	0.17*	0.21**	0.16*	0.05	0.12	0.14**
04	0.16	0.15	0.26**	0.16	0.16	0.12	0.14	0.10	0.14*	0.16**
05	0.15	0.12	0.16	0.20*	0.14	0.10	0.14	0.10	0.10	0.13*

Table A.35: **Accuracy pSTS CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.49	0.33*	0.35**	0.39**	0.40	0.28	0.43**	0.37*
02	0.43	0.35*	0.36	0.38*	0.42*	0.38*	0.42*	0.41**
03	0.53	0.28	0.28	0.36*	0.42**	0.40*	0.40**	0.41**
04	0.42*	0.39*	0.37	0.39**	0.40	0.38**	0.36*	0.38**
05	0.59	0.29	0.22	0.37*	0.49**	0.26	0.37	0.37*

Table A.36: **Accuracy pSTS CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

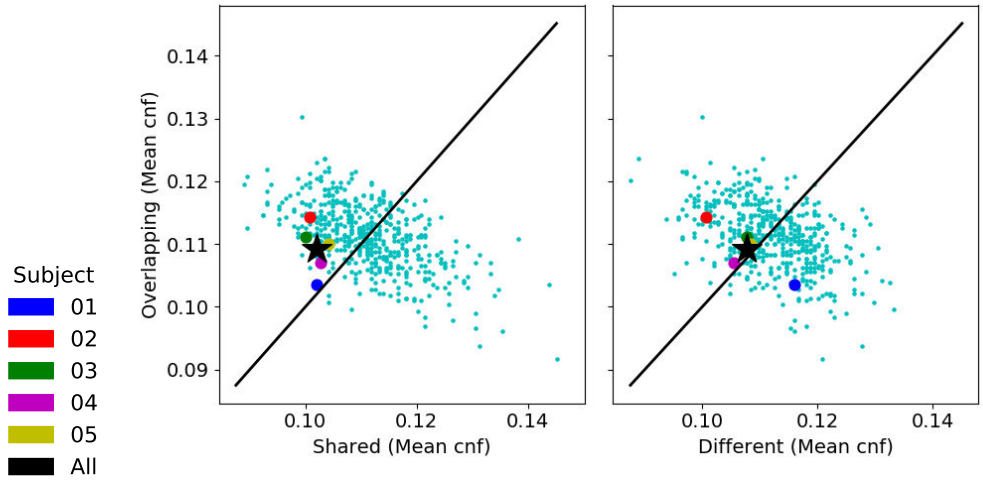


Figure A.53: **Superposition test in pSTS:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

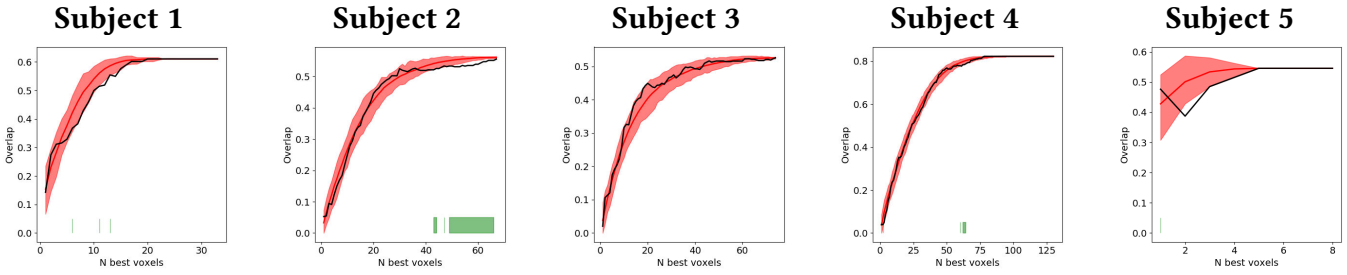


Figure A.54: **Locality test in pSTS:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.19 IFGorb (Auditory dataset)

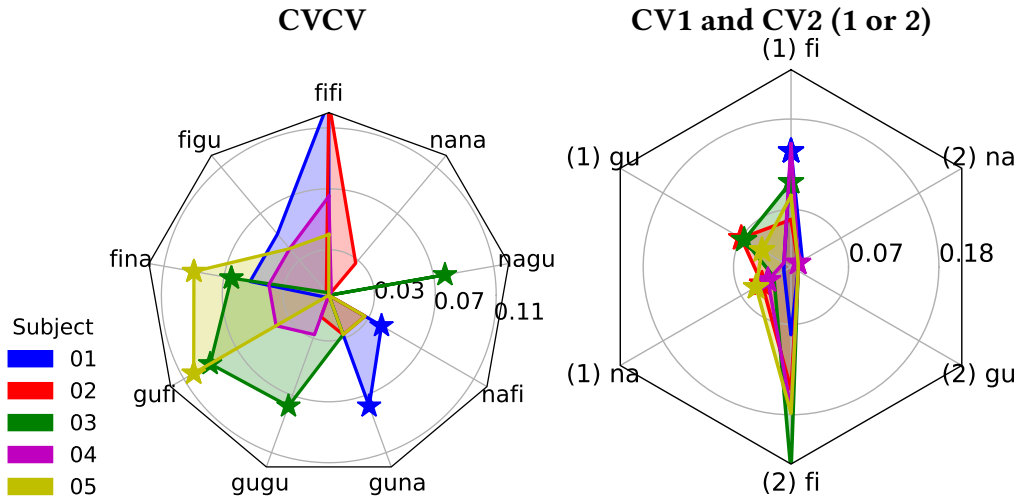


Figure A.55: **Accuracy in IFGorb:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.24*	0.16	0.16	0.11	0.09	0.19**	0.15*	0.06	0.06	0.14*
02	0.24*	0.11	0.10	0.07	0.12	0.14	0.14	0.09	0.14	0.13
03	0.11	0.11	0.17*	0.20*	0.19**	0.14	0.09	0.19*	0.10	0.14**
04	0.17	0.15	0.15	0.15	0.14	0.10	0.06	0.05	0.11	0.12
05	0.15	0.15	0.20*	0.21*	0.09	0.14	0.14	0.10	0.09	0.14**

Table A.37: **Accuracy IFGorb CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.47*	0.30	0.31	0.36	0.41	0.32	0.35	0.36
02	0.39	0.40**	0.37*	0.39**	0.48	0.33	0.33	0.38*
03	0.43*	0.40**	0.35	0.39**	0.58**	0.24	0.29	0.37
04	0.48	0.26	0.36*	0.37*	0.50	0.31	0.31*	0.37*
05	0.42	0.37*	0.38*	0.39**	0.51	0.26	0.31	0.36

Table A.38: **Accuracy IFGorb CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

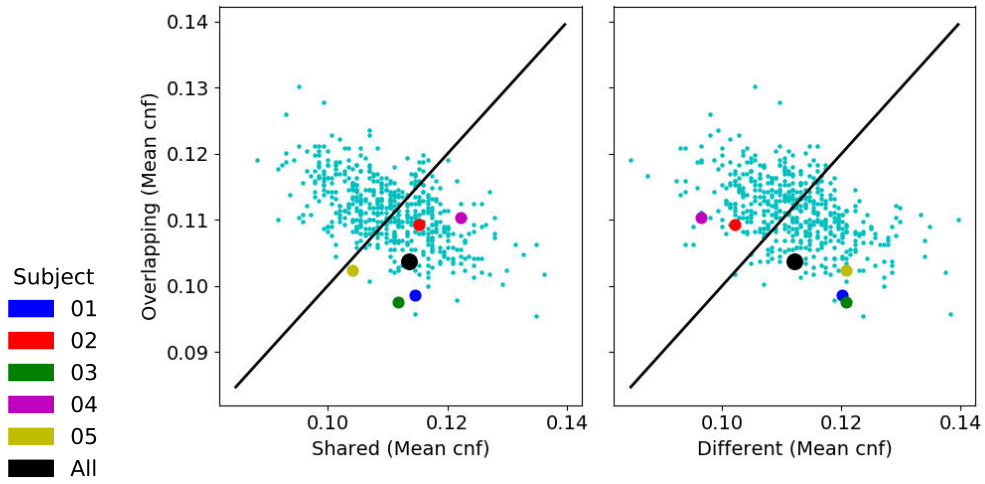


Figure A.56: **Superposition test in IFGorb:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

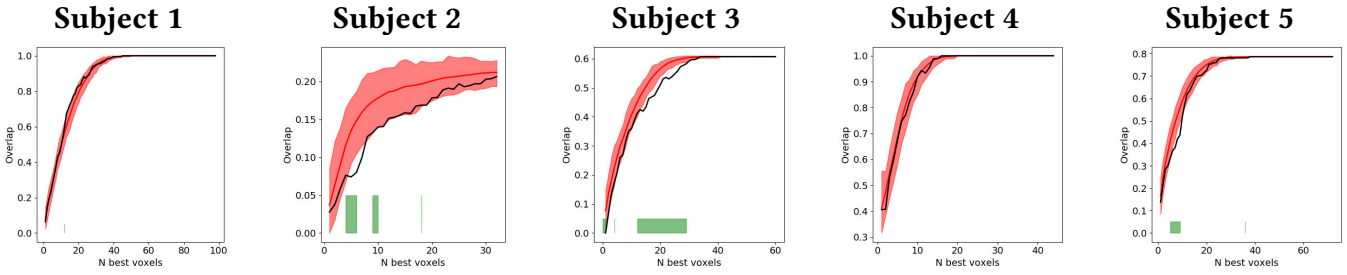


Figure A.57: **Locality test in IFGorb:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.20 IFGtri (Auditory dataset)

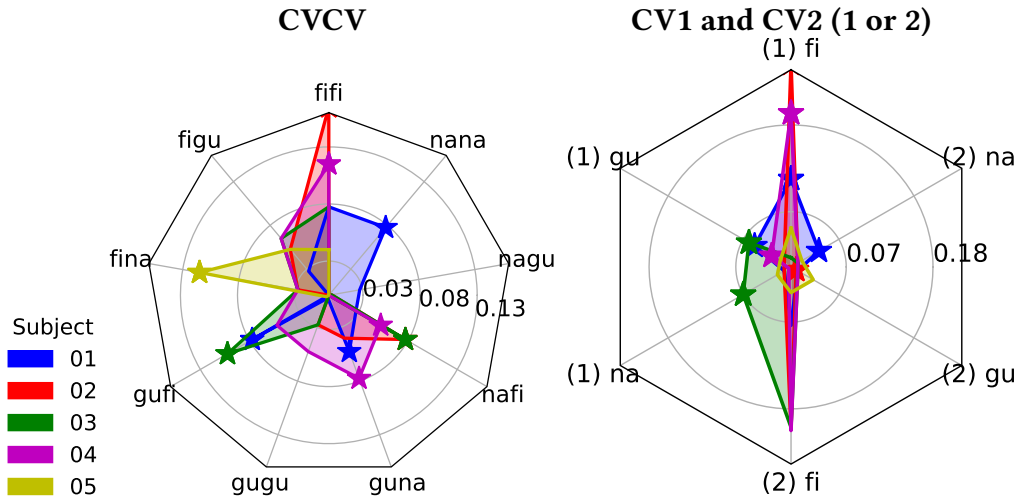


Figure A.58: **Accuracy in IFGtri:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.19	0.14	0.09	0.19*	0.11	0.16*	0.14	0.14	0.19*	0.15**
02	0.28**	0.16	0.14	0.10	0.14	0.15	0.19**	0.11	0.09	0.15**
03	0.19	0.17	0.14	0.21*	0.14	0.09	0.19**	0.11	0.05	0.14*
04	0.23*	0.17	0.14	0.16	0.16	0.19**	0.16*	0.07	0.09	0.15**
05	0.15	0.16	0.23*	0.11	0.09	0.10	0.10	0.05	0.09	0.12

Table A.39: **Accuracy IFGtri CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.44*	0.38*	0.33	0.38**	0.40	0.31	0.37**	0.36*
02	0.59*	0.26	0.28	0.38*	0.53	0.34*	0.27	0.38*
03	0.34	0.39*	0.40*	0.38	0.53	0.31	0.28	0.38**
04	0.53**	0.36*	0.33	0.41**	0.54	0.30	0.26	0.37
05	0.38	0.35	0.35	0.36	0.36	0.36	0.35	0.36

Table A.40: **Accuracy IFGtri CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

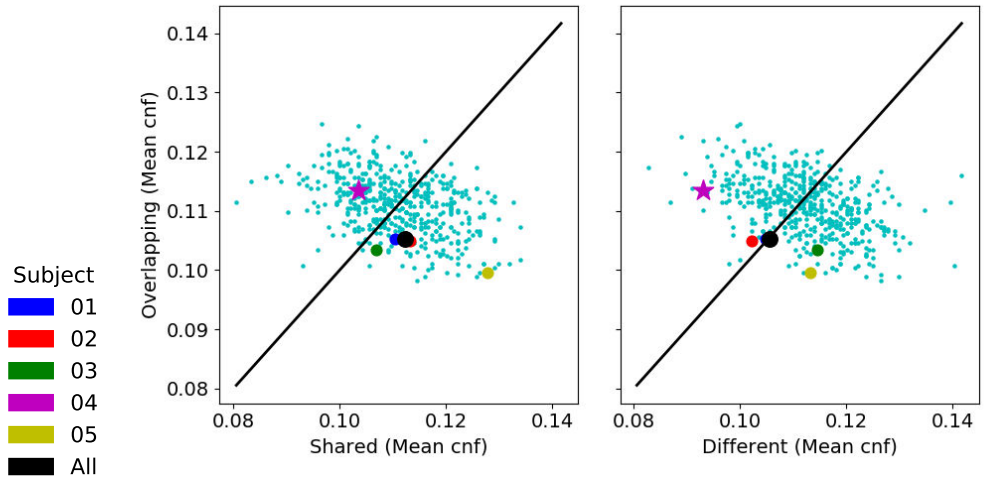


Figure A.59: **Superposition test in IFGtri:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

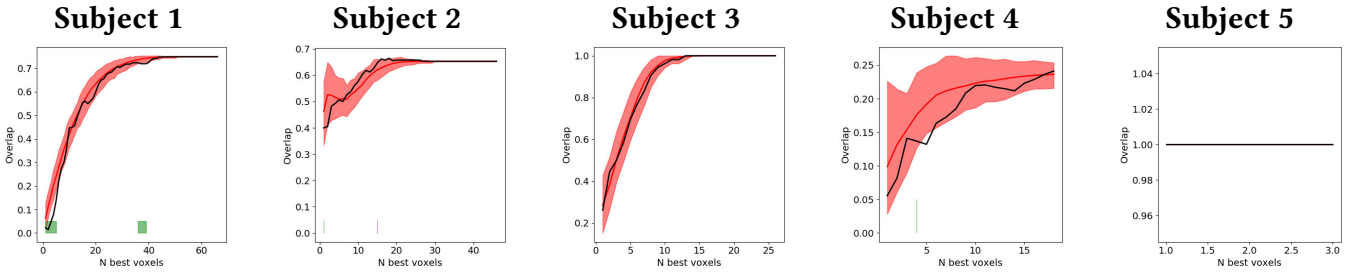


Figure A.60: **Locality test in IFGtri:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.21 Broca-44 (Auditory dataset)

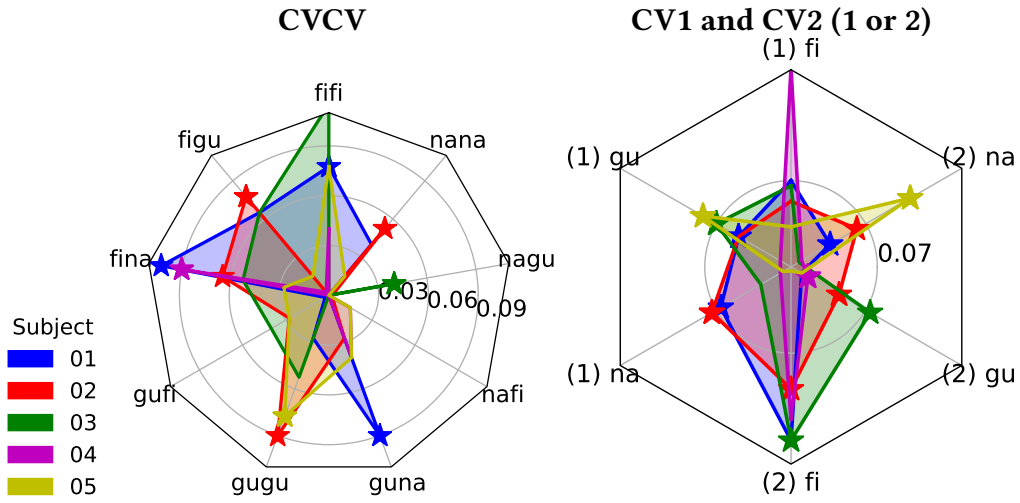


Figure A.61: **Accuracy in Broca-44:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.19*	0.17	0.21*	0.11	0.14	0.20*	0.07	0.10	0.15	0.15**
02	0.10	0.19*	0.17*	0.14	0.20**	0.14	0.11	0.11	0.16*	0.15**
03	0.23*	0.17	0.16	0.15	0.16	0.07	0.09	0.15*	0.09	0.14*
04	0.15	0.11	0.20*	0.10	0.09	0.15	0.12	0.10	0.10	0.12
05	0.19	0.12	0.14	0.14	0.19*	0.15	0.12	0.10	0.12	0.14**

Table A.41: **Accuracy Broca-44 CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.40	0.38*	0.40*	0.39*	0.47	0.30	0.37*	0.38**
02	0.38	0.38	0.40*	0.39**	0.43**	0.38*	0.39*	0.40*
03	0.40	0.40*	0.36	0.38**	0.47*	0.40**	0.34	0.40*
04	0.49	0.30	0.28	0.36	0.45	0.35*	0.33	0.38**
05	0.36	0.41*	0.34	0.37*	0.33	0.32	0.44**	0.36

Table A.42: **Accuracy Broca-44 CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

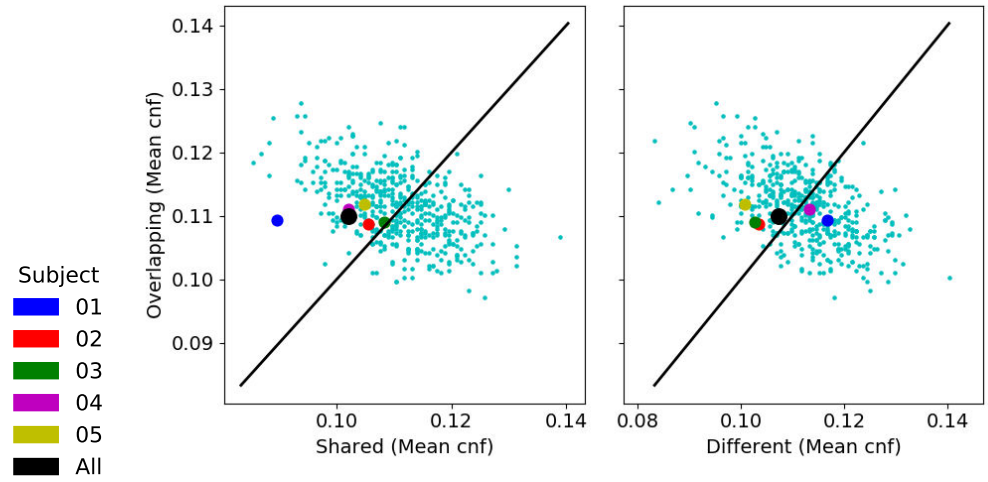


Figure A.62: **Superposition test in Broca-44:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

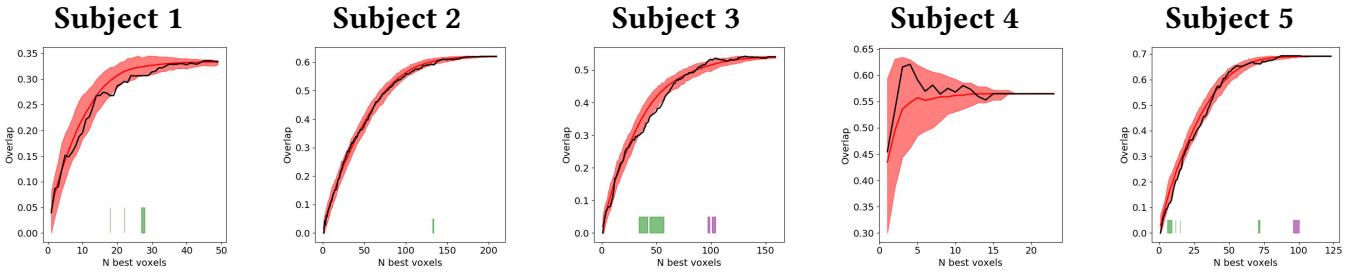


Figure A.63: **Locality test in Broca-44:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

A.22 Broca-45 (Auditory dataset)

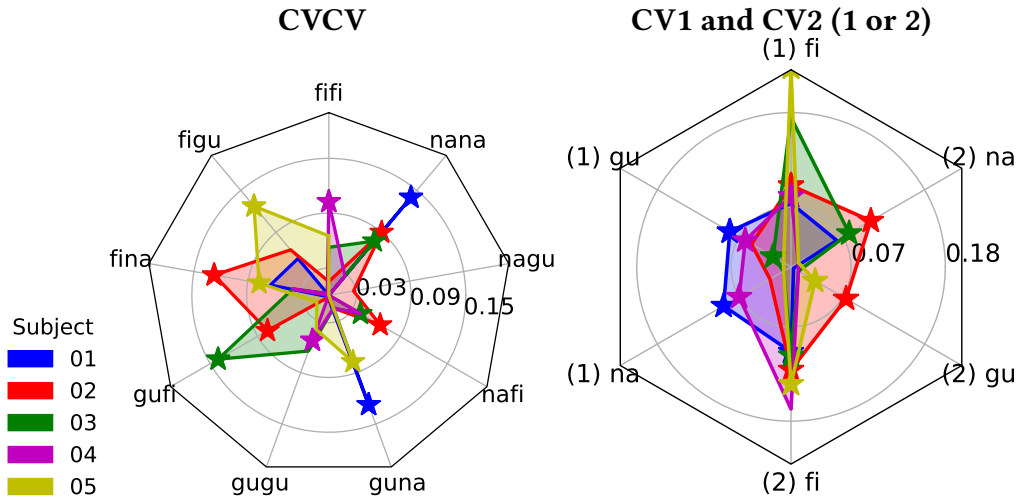


Figure A.64: **Accuracy in Broca-45:** Chance baseline has been substracted from all accuracy scores. Chance is 11.11% for the CVCV model. Chance is 33.33% for the CV1 and CV2 models. We show at the left the CVCV accuracy and at the right the CV1 and CV2 accuracy together. (1) denotes the CV1 model and (2) denotes the CV2 model. The accuracy score points are denoted with stars whenever they are significant with p-value < 0.05

Condition Subject	fifi	figu	fina	gufi	gugu	guna	nafi	nagu	nana	Mean
01	0.09	0.16	0.17	0.11	0.11	0.24**	0.09	0.10	0.25**	0.15**
02	0.12	0.17	0.24**	0.19*	0.11	0.12	0.17*	0.14	0.20*	0.16**
03	0.16	0.11	0.15	0.25**	0.17	0.12	0.15*	0.06	0.19*	0.15**
04	0.21*	0.11	0.15	0.06	0.16*	0.12	0.15	0.09	0.14	0.13*
05	0.17	0.24**	0.19*	0.12	0.15	0.19**	0.10	0.10	0.06	0.15**

Table A.43: **Accuracy Broca-45 CVCV:** * p-value < 0.05, ** p-value < 0.01.

Subject	(1) fi	(1) gu	(1) na	(1) Mean	(2) fi	(2) gu	(2) na	(2) Mean
01	0.40	0.41*	0.42*	0.41**	0.43*	0.33	0.39	0.38**
02	0.42*	0.38	0.36	0.39*	0.45*	0.40**	0.44**	0.43**
03	0.50	0.35*	0.26	0.37*	0.43*	0.33	0.41*	0.39*
04	0.41*	0.39*	0.40**	0.40**	0.50	0.31	0.23	0.35
05	0.56**	0.28	0.23	0.36	0.47*	0.36*	0.33	0.39*

Table A.44: **Accuracy Broca-45 CV1 and CV2:** * p-value < 0.05, ** p-value < 0.01.

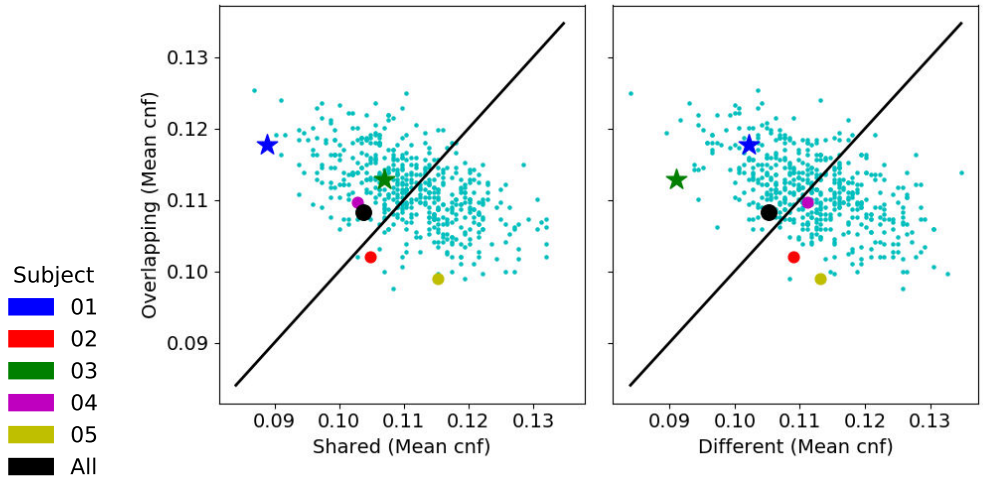


Figure A.65: **Superposition test in Broca-45:** We present the relationship between the mean confusion of cell groups representing overlapping syllables, shared syllables and different syllables. The smaller cyan dots correspond to the shuffled models of all subjects. All other dots correspond to subjects. If the mean confusion values of a tested model is reflected as a dot above the black line in both plots, then we have evidence for superposition. A star means significance with a p-value < 0.05

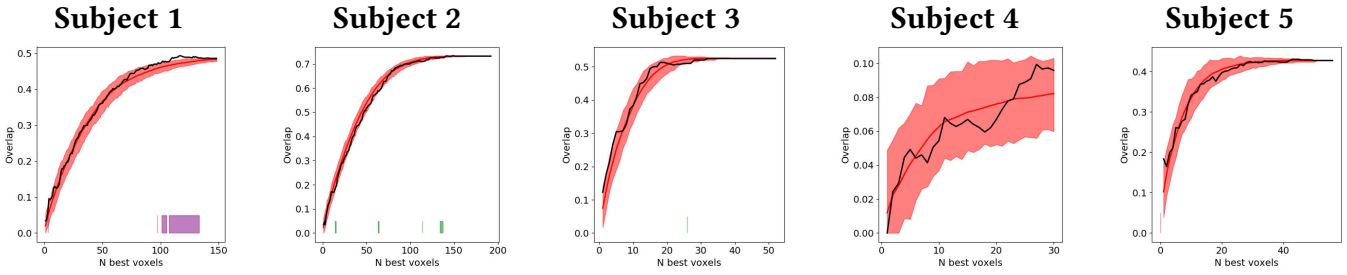


Figure A.66: **Locality test in Broca-45:** We show in black the overlap of the N best voxels subsets of the CV1 and CV2 models. In red we show the overlap null distribution given by the shuffled models. In green we denote segments of significantly inferior overlap with a p-value < 0.05. In magenta we denote segments of significantly higher overlap with a p-value < 0.05

Bibliography

[1] H. Abdulrahman and R. N. Henson. Effect of trial-to-trial variability on optimal event-related fmri design: Implications for beta-series correlation and multi-voxel pattern analysis. *NeuroImage*, 125:756–766, 2016.

[2] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8, 2014.

[3] K. Abrams and T. G. Bever. Syntactic structure modifies attention during speech perception and recognition. *The Quarterly journal of experimental psychology*, 21(3):280–290, 1969.

[4] J. Allman, F. Miezin, and E. McGuinness. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual review of neuroscience*, 8(1):407–430, 1985. URL `http://www.annualreviews.org/doi/pdf/10.1146/annurev.ne.08.030185.002203`.

[5] B. M. Ances, E. Zarahn, J. H. Greenberg, and J. A. Detre. Coupling of neural activation to blood flow in the somatosensory cortex of rats is time-intensity separable, but not linear. *Journal of Cerebral Blood Flow & Metabolism*, 20(6):921–930, 2000.

[6] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*, 2016.

[7] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[8] O. Arthurs and S. Boniface. What aspect of the fmri bold signal best reflects the underlying electrophysiology in human somatosensory cortex? *Clinical Neurophysiology*, 114(7):1203–1209, 2003.

[9] O. Arthurs, E. Williams, T. Carpenter, J. Pickard, and S. Boniface. Linear coupling between functional magnetic resonance imaging and evoked

potential amplitude in human somatosensory cortex. *Neuroscience*, 101 (4):803–806, 2000.

[10] C. I. Baker, M. Behrmann, and C. R. Olson. Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature neuroscience*, 5(11):1210, 2002.

[11] M. Bastiaansen, L. Magyari, and P. Hagoort. Syntactic unification operations are reflected in oscillatory dynamics during on-line sentence comprehension. *Journal of cognitive neuroscience*, 22(7):1333–1347, 2010.

[12] D. K. Bemis and L. Pylkkänen. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex*, 23(8):1859–1873, 2012.

[13] J. Benda and A. V. M. Herz. A Universal Model for Spike-Frequency Adaptation. *Neural Computation*, 15(11):2523–2564, nov 2003. DOI: 10.1162/089976603322385063.

[14] J. Benda and J. Tabak. Spike-Frequency Adaptation. In *Encyclopedia of Computational Neuroscience*, pages 1–12. Springer New York, 2014.

[15] T. G. Bever, J. Lackner, and R. Kirk. The underlying structures of sentences are the primary units of immediate speech processing. *Attention, Perception, & Psychophysics*, 5(4):225–234, 1969.

[16] J. R. Binder, J. A. Frost, T. A. Hammeke, R. W. Cox, S. M. Rao, and T. Prieto. Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, 17(1):353–362, 1997.

[17] R. M. Birn, Z. S. Saad, and P. A. Bandettini. Spatial heterogeneity of the nonlinear dynamics in the fmri bold response. *Neuroimage*, 14(4): 817–826, 2001.

[18] I. Bocancia. *A psycholinguistically motivated neural model of sentence comprehension*. PhD thesis, Universitat Ulm, 2014.

[19] K. Bock, G. S. Dell, F. Chang, and K. H. Onishi. Persistent structural priming from language comprehension to language production. *Cognition*, 104(3):437–458, 2007.

[20] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327, 2013.

[21] J. S. Bowers. On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological review*, 116(1):220, 2009. URL `http://psycnet.apa.org/psycinfo/2009-00258-008`.

[22] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, 16(13):4207–4221, 1996.

[23] H. P. Branigan, M. J. Pickering, and A. A. Cleland. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25, 2000.

[24] M. Breakspear, L. M. Williams, and C. J. Stam. A novel method for the topographic analysis of neural activity reveals formation and dissolution of 'dynamic cell assemblies'. *Journal of computational neuroscience*, 16 (1):49–68, 2004.

[25] J. R. Brennan, E. P. Stabler, S. E. Van Wagenen, W.-M. Luh, and J. T. Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94, 2016.

[26] R. Brette. Adaptive Exponential Integrate-and-Fire Model as an Effective Description of Neuronal Activity. *Journal of Neurophysiology*, 94(5): 3637–3642, nov 2005. DOI: 10.1152/jn.00686.2005.

[27] R. Brette and W. Gerstner. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of neurophysiology*, 94(5):3637–3642, 2005.

[28] F. Briggs, G. R. Mangun, and W. M. Usrey. Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. *Nature*, 499 (7459):476–480, jun 2013. DOI: 10.1038/nature12276. URL http://dx.doi.org/10.1038/nature12276.

[29] G. Brinker, C. Bock, E. Busch, H. Krep, K.-A. Hossmann, and M. Hoehn-Berlage. Simultaneous recording of evoked potentials and t2*-weighted mr images during somatosensory stimulation of rat. *Magnetic resonance in medicine*, 41:469–473, 1999.

[30] R. L. Buckner. Event-related fmri and the hemodynamic response. *Human brain mapping*, 6(5-6):373–377, 1998.

[31] R. B. Buxton, K. Uludağ, D. J. Dubowitz, and T. T. Liu. Modeling the hemodynamic response to brain activation. *NeuroImage*, 23:S220–S233, jan 2004. DOI: 10.1016/j.neuroimage.2004.07.013.

[32] M. M. B. Cardoso, Y. B. Sirotin, B. Lima, E. Glushenkova, and A. Das. The neuroimaging signal is a linear sum of neurally distinct stimulus- and task-related components. *Nature Neuroscience*, 15(9):1298–1306, July 2012. ISSN 1097-6256, 1546-1726. DOI: 10.1038/nn.3170. URL http://www.nature.com/doifinder/10.1038/nn.3170.

[33] G. C. Cawley and N. L. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010.

[34] L. Chaâri, J.-C. Pesquet, A. Benazza-Benyahia, and P. Ciuciu. A wavelet-based regularized reconstruction algorithm for sense parallel mri with applications to neuroimaging. *Medical image analysis*, 15(2):185–201, 2011.

[35] D. J. Chalmers. Syntactic Transformations on Distributed Representations. In *Connectionist Natural Language Processing*, pages 46–55. Springer Netherlands, 1992.

[36] W. Chen and S. Ogawa. 10 principles of bold functional mri. *Red*, 10:1, 1996.

[37] X. Chen, J. K. Possel, C. Wacongne, A. F. van Ham, P. C. Klink, and P. R. Roelfsema. 3d printing and modelling of customized implants and surgical guides for non-human primates. *Journal of Neuroscience Methods*, 286:38–55, 2017.

[38] N. Chomsky. *The Minimalist Program*. The MIT Press, dec 2014. DOI: 10.7551/mitpress/9780262527347.001.0001.

[39] M. H. Christiansen and N. Chater. Connectionist Natural Language Processing: The State of the Art. *Cognitive Science*, 23(4):417–437, oct 1999.

[40] J. M. Cisler, K. Bush, and J. S. Steele. A comparison of statistical methods for detecting context-modulated functional connectivity in fmri. *Neuroimage*, 84:1042–1052, 2014.

[41] L. Cohen, S. Dehaene, L. Naccache, S. Lehéricy, G. Dehaene-Lambertz, M.-A. Hénaff, and F. Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000.

[42] D. Collins, A. Zijdenbos, W. Baaré, and A. Evans. Animal+ insect: improved cortical structure segmentation. In *Information processing in medical imaging*, pages 210–223. Springer, 1999.

[43] A. Compte, N. Brunel, P. S. Goldman-Rakic, and X.-J. Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923, 2000.

[44] M. de Kamps. A simple and stable numerical solution for the population density equation. *Neural computation*, 15(9):2129–2146, 2003.

[45] M. de Kamps. A model for delay activity without recurrent excitation. *Artificial Neural Networks: Biological Inspirations–ICANN 2005*, pages 229–234, 2005.

[46] M. de Kamps. A Generic Approach to Solving Jump Diffusion Equations with Applications to Neural Populations. *arXiv preprint arXiv:1309.1654*, 2013.

[47] M. de Kamps. A generic approach to solving jump diffusion equations with applications to neural populations. *arXiv preprint arXiv:1309.1654*, 2013.

[48] M. De Kamps and F. Van der Velde. Combinatorial structures and processing in neural blackboard architectures. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo 2015)*, volume 1583. CEUR, 2016.

[49] M. de Kamps, V. Baier, J. Drever, M. Dietz, L. Mösenlechner, and F. van der Velde. The state of MIIND. *Neural Networks*, 21(8):1164–1181, oct 2008. DOI: 10.1016/j.neunet.2008.07.006. URL http://dx.doi.org/10.1016/j.neunet.2008.07.006.

[50] S. Dehaene and L. Cohen. The unique role of the visual word form area in reading. *Trends in cognitive sciences*, 15(6):254–262, 2011.

[51] S. Dehaene, F. Pegado, L. W. Braga, P. Ventura, G. Nunes Filho, A. Jobert, G. Dehaene-Lambertz, R. Kolinsky, J. Morais, and L. Cohen. How learning to read changes the cortical networks for vision and language. *science*, 330(6009):1359–1364, 2010.

[52] A. Destexhe and T. J. Sejnowski. The Wilson–Cowan model 36 years later. *Biological Cybernetics*, 101(1):1–2, jul 2009. DOI: 10.1007/s00422-009-0328-3.

[53] I. M. Devonshire, N. G. Papadakis, M. Port, J. Berwick, A. J. Kennerley, J. E. W. Mayhew, and P. G. Overton. Neurovascular coupling is brain region-dependent. *NeuroImage*, 59(3):1997–2006, Feb. 2012. ISSN 1053-8119. DOI: 10.1016/j.neuroimage.2011.09.050. URL http://www.sciencedirect.com/science/article/pii/S1053811911011153.

[54] N. Ding, L. Melloni, H. Zhang, X. Tian, and D. Poeppel. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158, 2016.

[55] E. Doger de Speville, C. Robert, M. Perez-Guevara, A. Grigis, S. Bolle, C. Pinaud, C. Dufour, A. Beaudré, V. Kieffer, A. Longaud, et al. Relationships between regional radiation doses and cognitive decline in children treated with cranio-spinal irradiation for posterior fossa tumors. *Frontiers in Oncology*, 7:166, 2017.

[56] P. F. Dominey, T. Inui, and M. Hoen. Neural network processing of natural language: II. Towards a unified model of corticostriatal

function in learning sentence comprehension and non-linguistic sequencing. *Brain and Language*, 109(2-3):80–92, may 2009. DOI: 10.1016/j.bandl.2008.08.002.

[57] D. M. Eagleman. TIMELINE: Visual illusions and neurobiology. *Nature Reviews Neuroscience*, 2(12):920–926, dec 2001. DOI: 10.1038/35104092.

[58] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, may 2017. DOI: 10.1016/j.neuroimage.2016.10.001.

[59] S. B. Eickhoff, K. E. Stephan, H. Mohlberg, C. Grefkes, G. R. Fink, K. Amunts, and K. Zilles. A new spm toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, 25 (4):1325–1335, 2005.

[60] D. Ellis. Time-domain scrambling of audio signals in matlab, 2010.

[61] J. A. Etzel, J. M. Zacks, and T. S. Braver. Searchlight analysis: promise, pitfalls, and potential. *Neuroimage*, 78:261–269, 2013.

[62] S. Evans and M. H. Davis. Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. *Cerebral cortex*, 25(12):4772–4788, 2015.

[63] E. Fedorenko, P.-J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, and N. Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104 (2):1177–1194, 2010.

[64] E. Fedorenko, T. L. Scott, P. Brunner, W. G. Coon, B. Pritchett, G. Schalk, and N. Kanwisher. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41): E6256–E6262, 2016.

[65] J. Feldman. The neural binding problem(s). *Cognitive Neurodynamics*, 7 (1):1–11, sep 2012. DOI: 10.1007/s11571-012-9219-8.

[66] B. Fontaine, D. F. Goodman, V. Benichoux, and R. Brette. Brian hears: online auditory processing using vectorization over channels. *frontiers in Neuroinformatics*, 5, 2011.

[67] V. D. V. Frank. Linking population dynamics and high-level cognition: Ambiguity resolution in a neural sentence processing model. *Frontiers in Neuroinformatics*, 8, 2014. DOI: 10.3389/conf.fninf.2014.18.00055.

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

[69] A. D. Friederici. Evolution of the neural language network. *Psychonomic bulletin & review*, 24(1):41–47, 2017.

[70] K. Friston, A. Mechelli, R. Turner, and C. Price. Nonlinear Responses in fMRI: The Balloon Model Volterra Kernels, and Other Hemodynamics. *NeuroImage*, 12(4):466–477, oct 2000. DOI: 10.1006/nimg.2000.0630.

[71] K. J. Friston, A. P. Holmes, J. Poline, P. Grasby, S. Williams, R. S. Frackowiak, and R. Turner. Analysis of fmri time-series revisited. *Neuroimage*, 2(1):45–53, 1995.

[72] M. Garagnani, G. Lucchese, R. Tomasello, T. Wennekers, and F. Pulvermüller. A Spiking Neurocomputational Model of High-Frequency Oscillatory Brain Responses to Words and Pseudowords. *Frontiers in Computational Neuroscience*, 10, jan 2017. DOI: 10.3389/fncom.2016.00145.

[73] C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*. Springer-Verlag, 1994.

[74] M.-O. Gewaltig and M. Diesmann. Nest (neural simulation tool). *Scholarpedia*, 2(4):1430, 2007.

[75] L. S. Glezer, X. Jiang, and M. Riesenhuber. Evidence for highly selective neuronal tuning to whole words in the "visual word form area". *Neuron*, 62(2):199–204, 2009.

[76] G. H. Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9:416–429, 1999.

[77] G. H. Glover. Deconvolution of Impulse Response in Event-Related BOLD fMRI1. *NeuroImage*, 9(4):416–429, apr 1999. DOI: 10.1006/nimg.1998.0419.

[78] K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. S. Ghosh. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 5, 2011.

[79] D. N. Greve and B. Fischl. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, 48(1):63–72, 2009.

[80] U. Guclu and M. A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014, jul 2015. DOI: 10.1523/jneurosci.5023-14.2015.

[81] E. Hagen, D. Dahmen, M. Stavrinou, H. Lindén, T. Tetzlaff, S. van Albada, S. Grün, M. Diesmann, and G. T. Einevoll. Hybrid scheme for modeling

local field potentials from point-neuron networks. *BMC Neuroscience*, 16(Suppl 1):P67, 2015. DOI: 10.1186/1471-2202-16-s1-p67.

[82] P. Hagoort. On broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9):416–423, 2005.

[83] J. T. Hale. *Automaton theories of human sentence comprehension*. CSLI Publications, 2014.

[84] G. S. Halford, W. H. Wilson, and S. Phillips. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, 21(6):803–831, 1998.

[85] D. A. Handwerker, J. M. Ollinger, and M. D'Esposito. Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651, 2004.

[86] K. D. Harris, J. Csicsvari, H. Hirase, G. Dragoi, and G. Buzsaki. Organization of cell assemblies in the hippocampus. *Nature*, 424(6948): 552, 2003.

[87] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2 edition, 2009.

[88] J. V. Haxby, A. C. Connolly, and J. S. Guntupalli. Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37:435–456, 2014.

[89] D. J. Heeger, A. C. Huk, W. S. Geisler, and D. G. Albrecht. Spikes versus bold: what does neuroimaging tell us about neuronal activity? *Nature neuroscience*, 3(7):631–633, 2000.

[90] A. Heinrich, R. P. Carlyon, M. H. Davis, and I. S. Johnsrude. Illusory vowels resulting from perceptual continuity: a functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, 20(10): 1737–1752, 2008.

[91] R. Henson and W. Penny. Anovas and spm. *Wellcome Department of Imaging Neuroscience, London, UK*, 2003.

[92] D. Hermes, M. Nguyen, and J. Winawer. Neuronal synchrony and the relation between the blood-oxygen-level dependent response and the local field potential. *PLoS biology*, 15(7):e2001461, 2017.

[93] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[94] C. R. Huyck. A psycholinguistic model of natural language parsing implemented in simulated neurons. *Cognitive Neurodynamics*, 3(4): 317–330, mar 2009. DOI: 10.1007/s11571-009-9080-6.

[95] C. R. Huyck and P. J. Passmore. A review of cell assemblies. *Biological Cybernetics*, 107(3):263–288, apr 2013.

[96] R. Iyer, V. Menon, M. Buice, C. Koch, and S. Mihalas. The influence of synaptic weight distribution on neuronal population dynamics. *PLoS computational biology*, 9(10):e1003248, 2013.

[97] E. M. Izhikevich. *Dynamical systems in neuroscience*. MIT press, 2007.

[98] R. Jackendoff. *Foundations of Language*. Oxford University Press, jan 2002. DOI: 10.1093/acprof:oso/9780198270126.001.0001.

[99] R. Jackendoff. Combinatoriality. In *Foundations of Language*, pages 38–67. Oxford University Press, jan 2002. DOI: 10.1093/acprof:oso/9780198270126.003.0003.

[100] B. H. Jansen and V. G. Rit. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological cybernetics*, 73(4):357–366, 1995.

[101] M. Jones, I. M. Devonshire, J. Berwick, C. Martin, P. Redgrave, and J. Mayhew. Altered neurovascular coupling during information-processing states. *European Journal of Neuroscience*, 27(10):2758–2772, 2008.

[102] J. N. D. Kerr, D. Greenberg, and F. Helmchen. From The Cover: Imaging input and output of neocortical networks in vivo. *Proceedings of the National Academy of Sciences*, 102(39):14063–14068, sep 2005. DOI: 10.1073/pnas.0506029102. URL `http://dx.doi.org/10.1073/pnas.0506029102`.

[103] T. Kim, K. Masamoto, M. Fukuda, A. Vazquez, and S.-G. Kim. Frequency-dependent neural activity, cbf, and bold fmri to somatosensory stimuli in isoflurane-anesthetized rats. *Neuroimage*, 52(1):224–233, 2010.

[104] A. Klein and J. Hirsch. Mindboggle: a scatterbrained approach to automate brain labeling. *NeuroImage*, 24(2):261–280, 2005.

[105] B. W. Knight. Dynamics of encoding in a population of neurons. *The Journal of general physiology*, 59(6):734–766, 1972.

[106] C. Koch and I. Segev. The role of single neurons in information processing. *Nature neuroscience*, 3(11), 2000.

[107] F. Krause and O. Lindemann. Expyriment: A python library for cognitive and neuroscientific experiments. *Behavior Research Methods*, 46(2):416–428, 2014.

[108] Y. M. Lai and M. de Kamps. Population density equations for stochastic processes with memory kernels. *Physical Review E*, 95(6):062125, 2017.

[109] M. Larkum. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in neurosciences*, 36 (3):141–151, 2013.

[110] M. Lauritzen. Relationship of spikes, synaptic activity, and local changes of cerebral blood flow. *Journal of Cerebral Blood Flow & Metabolism*, 21 (12):1367–1383, 2001.

[111] S. J. Lawrence, E. Formisano, L. Muckli, and F. P. de Lange. Laminar fmri: applications for cognitive neuroscience. *Neuroimage*, 2017.

[112] E. Liebenthal, J. R. Binder, S. M. Spitzer, E. T. Possing, and D. A. Medler. Neural substrates of phonemic perception. *Cerebral cortex*, 15(10):1621–1631, 2005.

[113] T. Lindeberg. Normative theory of visual receptive fields. *arXiv preprint arXiv:1701.06333*, 2017.

[114] M. A. Lindquist. The statistical analysis of fmri data. *Statistical Science*, pages 439–464, 2008.

[115] M. A. Lindquist, J. M. Loh, L. Y. Atlas, and T. D. Wager. Modeling the hemodynamic response function in fmri: efficiency, bias and mismodeling. *Neuroimage*, 45(1):S187–S198, 2009.

[116] M. London, A. Schreibman, M. Häusser, M. E. Larkum, and I. Segev. The information efficacy of a synapse. *Nat. Neurosci.*, 5(4):332–340, mar 2002. DOI: 10.1038/nn826. URL `http://dx.doi.org/10.1038/nn826`.

[117] O. Longe, B. Randall, E. A. Stamatakis, and L. K. Tyler. Grammatical categories in the brain: The role of morphological structure. *Cerebral Cortex*, 17(8):1812–1820, 2006.

[118] K. Mahowald and E. Fedorenko. Reliable individual-level neural markers of high-level language processing: a necessary precursor for relating neural variability to behavioral and genetic variability. *Neuroimage*, 139:74–93, 2016.

[119] J. R. Manning, J. Jacobs, I. Fried, and M. J. Kahana. Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. *Journal of Neuroscience*, 29(43):13613–13620, oct 2009. DOI: 10.1523/jneurosci.2041-09.2009.

[120] A. H. M. Marcus Gary and T. Dean. The atoms of neural computation. *Science*, 346(6209):551–552, 2014.

[121] H. Markert, A. Knoblauch, and G. Palm. Modelling of syntactical processing in the cortex. *Biosystems*, 89(1-3):300–315, may 2007. DOI: 10.1016/j.biosystems.2006.04.027.

[122] D. Marr. *Vision: A Computational Investigation Into*. WH Freeman, 1982.

[123] A. E. Martin and L. A. A. Doumas. A mechanism for the cortical computation of hierarchical linguistic structure. *PLOS Biology*, 15(3): e2000663, mar 2017. DOI: 10.1371/journal.pbio.2000663.

[124] J. Martindale, J. Mayhew, J. Berwick, M. Jones, C. Martin, D. Johnston, P. Redgrave, and Y. Zheng. The hemodynamic impulse response to a single neural event. *Journal of Cerebral Blood Flow & Metabolism*, 23(5): 546–555, 2003.

[125] K. Masamoto, M. Fukuda, A. Vazquez, and S.-G. Kim. Dose-dependent effect of isoflurane on neurovascular coupling in rat cerebral cortex. *European Journal of Neuroscience*, 30(2):242–250, 2009.

[126] W. Matchin, C. Hammerly, and E. Lau. The role of the ifg and psts in syntactic prediction: Evidence from a parametric study of hierarchical structure in fmri. *cortex*, 88:106–123, 2017.

[127] A. Mazzoni, H. Lindén, H. Cuntz, A. Lansner, S. Panzeri, and G. T. Einevoll. Computing the Local Field Potential (LFP) from Integrate-and-Fire Network Models. *PLOS Computational Biology*, 11(12):e1004584, dec 2015. DOI: 10.1371/journal.pcbi.1004584.

[128] J. L. McClelland and A. H. Kawamoto. Mechanisms of sentence processing: Assigning roles to constituents of sentences. *Parallel distributed processing*, 2:318–362, 1986.

[129] D. A. McCormick and T. Bal. Sleep and arousal: thalamocortical mechanisms. *Annual review of neuroscience*, 20(1):185–215, 1997.

[130] W. McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. SciPy Austin, TX, 2010.

[131] D. B. McMahon and C. R. Olson. Linearly additive shape and color signals in monkey inferotemporal cortex. *Journal of neurophysiology*, 101(4):1867–1875, 2009.

[132] A. Mechelli, K. J. Friston, and C. J. Price. The effects of presentation rate during word and pseudoword reading: a comparison of pet and fmri. 2006.

[133] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 343(6174):1006–1010, jan 2014. DOI: 10.1126/science.1245994.

[134] R. Miikkulainen. Natural language processing with subsymbolic neural networks. *Neural network perspectives on cognition and adaptive robotics*, pages 120–139, 1997.

[135] M. Minsky. A framework for representing knowledge. in the psychology of computer vision, ed. ph winton, 211–277, 1975.

[136] M. Misaki, Y. Kim, P. A. Bandettini, and N. Kriegeskorte. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, 53:103–118, 2010.

[137] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

[138] G. Mongillo, O. Barak, and M. Tsodyks. Synaptic theory of working memory. *Science*, 319(5869):1543–1546, 2008.

[139] F. Moradi and R. B. Buxton. Adaptation of cerebral oxygen metabolism and blood flow and modulation of neurovascular coupling with prolonged stimulation in human visual cortex. *NeuroImage*, 82:182–189, Nov. 2013. ISSN 1053-8119. DOI: 10.1016/j.neuroimage.2013.05.110. URL http://www.sciencedirect.com/science/article/pii/S1053811913006137.

[140] J. A. Mumford, B. O. Turner, F. G. Ashby, and R. A. Poldrack. Deconvolving bold activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, 59(3):2636–2643, 2012.

[141] M. J. Nelson, I. E. Karoui, K. Giber, X. Yang, L. Cohen, H. Koopman, S. S. Cash, L. Naccache, J. T. Hale, C. Pallier, and S. Dehaene. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678, apr 2017. DOI: 10.1073/pnas.1701590114.

[142] J. Nivre. Dependency grammar and dependency parsing. *MSI report*, 5133(1959):1–32, 2005.

[143] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

[144] J. nosuke Teramae, Y. Tsubo, and T. Fukai. Optimal spike-based communication in excitable networks with strong-sparse and weak-dense links. *Sci. Rep.*, 2, jul 2012. DOI: 10.1038/srep00485. URL http://dx.doi.org/10.1038/srep00485.

[145] D. Q. Nykamp and D. Tranchina. A population density approach that facilitates large-scale modeling of neural networks: Analysis and an

application to orientation tuning. *Journal of computational neuroscience*, 8(1):19–50, 2000.

[146] S. Ogawa, D. Tank, R. Menon, J. Ellermann, S. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89:5951–5955, 1992.

[147] R. C. Oldfield. The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, 9(1):97–113, 1971.

[148] T. E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3), 2007.

[149] B. A. Olshausen and others. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. URL `http://www.cs.ubc.ca/~little/cpsc425/olshausen_field_nature_1996.pdf`.

[150] A. Omurtag, B. W. Knight, and L. Sirovich. On the simulation of large populations of neurons. *Journal of computational neuroscience*, 8(1): 51–63, 2000.

[151] S. Ostojic. Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons. *Nature Neuroscience*, 17(4):594–600, feb 2014. DOI: 10.1038/nn.3658. URL `http://dx.doi.org/10.1038/nn.3658`.

[152] W. Ou, I. Nissilä, H. Radhakrishnan, D. A. Boas, M. S. Hämäläinen, and M. A. Franceschini. Study of neurovascular coupling in humans via simultaneous magnetoencephalography and diffuse optical imaging acquisition. *NeuroImage*, 46(3):624–632, 2009.

[153] C. Pallier, A.-D. Devauchelle, and S. Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, jan 2011. DOI: 10.1073/pnas.1018711108.

[154] C. Pallier, A.-D. Devauchelle, and S. Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 2011.

[155] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12 (Oct):2825–2830, 2011.

[156] F. Pérez and B. E. Granger. Ipython: a system for interactive scientific computing. *Computing in Science & Engineering*, 9(3), 2007.

[157] T. K. Perrachione and S. S. Ghosh. Optimized design and analysis of sparse-sampling fmri experiments. *Frontiers in neuroscience*, 7, 2013.

[158] T. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3):623–641, may 1995. DOI: 10.1109/72.377968.

[159] T. C. Potjans and M. Diesmann. The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model. *Cerebral cortex*, 24(3):785–806, 2012.

[160] A. Pouget and T. J. Sejnowski. Spatial transformations in the parietal cortex using basis functions. *Journal of cognitive neuroscience*, 9(2): 222–237, 1997.

[161] F. Pulvermüller. Brain embodiment of syntax and grammar: Discrete combinatorial mechanisms spelt out in neuronal circuits. *Brain and Language*, 112(3):167–179, mar 2010. DOI: 10.1016/j.bandl.2009.08.002.

[162] F. Pulvermüller and A. Knoblauch. Discrete combinatorial circuits emerging in neural networks: A mechanism for rules of grammar in the human brain? *Neural Networks*, 22(2):161–172, mar 2009. DOI: 10.1016/j.neunet.2009.01.009.

[163] S. Ray and J. H. R. Maunsell. Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biology*, 9(4): e1000610, apr 2011. DOI: 10.1371/journal.pbio.1000610.

[164] G. Rees, A. Howseman, O. Josephs, C. D. Frith, K. J. Friston, R. S. Frackowiak, and R. Turner. Characterizing the relationship between bold contrast and regional cerebral blood flow measurements by varying the stimulus presentation rate. *Neuroimage*, 6(4):270–278, 1997.

[165] C. Reverberi, K. Görgen, and J.-D. Haynes. Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex*, 22(6): 1237–1246, 2012.

[166] J. L. G. Rosa and A. B. da Silva. Thematic role assignment through a biologically plausible symbolic-connectionist hybrid system. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 2, pages 1457–1462. IEEE, 2004.

[167] A. Roxin, N. Brunel, D. Hansel, G. Mongillo, and C. van Vreeswijk. On the Distribution of Firing Rates in Networks of Cortical Neurons. *Journal of Neuroscience*, 31(45):16217–16226, nov 2011. DOI: 10.1523/jneurosci.1677-11.2011. URL `http://dx.doi.org/10.1523/jneurosci.1677-11.2011`.

[168] D. E. Rumelhart and J. L. McClelland. Learning the past tenses of english verbs: Implicit rules or parallel distributed processing. *Mechanisms of language acquisition*, pages 195–248, 1987.

[169] D. E. Rumelhart, G. E. Hinton, J. L. McClelland, et al. A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1:45–76, 1986.

[170] L. Shastri and V. Ajjanagadde. From simple associations to systematic reasoning: A connectionist representation of rules variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16(03):417, sep 1993. DOI: 10.1017/s0140525x00030910.

[171] J. C. W. Siero, D. Hermes, H. Hoogduin, P. R. Luijten, N. F. Ramsey, and N. Petridou. BOLD matches neuronal activity at the mm scale: A combined 7T fMRI and ECoG study in human sensorimotor cortex. *NeuroImage*, 101:177–184, Nov. 2014. ISSN 1053-8119. DOI: 10.1016/j.neuroimage.2014.07.002. URL http://www.sciencedirect.com/science/article/pii/S1053811914005746.

[172] P. Smolensky and G. Legendre. *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1.* MIT Press, 2006.

[173] P. Smolensky, M. Goldrick, and D. Mathis. Optimization and Quantization in Gradient Symbol Systems: A Framework for Integrating the Continuous and the Discrete in Cognition. *Cognitive Science*, 38(6): 1102–1138, jun 2013. DOI: 10.1111/cogs.12047.

[174] D. A. Soltysik, K. K. Peck, K. D. White, B. Crosson, and R. W. Briggs. Comparison of hemodynamic response nonlinearity across primary cortical areas. *Neuroimage*, 22(3):1117–1127, 2004.

[175] R. B. Stein. Some models of neuronal variability. *Biophysical journal*, 7 (1):37–68, 1967.

[176] M. Stimberg, D. F. Goodman, V. Benichoux, and R. Brette. Equation-oriented specification of neural models for simulations. *Frontiers in Neuroinformatics*, 8, 2014.

[177] S. C. Strother. Evaluating fmri preprocessing pipelines. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):27–41, 2006.

[178] J. Taylor, K. Rastle, and M. H. Davis. Can cognitive models explain brain activation during word and pseudoword reading? a meta-analysis of 36 neuroimaging studies., 2013.

[179] K. Tiippana. What is the mcgurk effect? *Frontiers in psychology*, 5, 2014.

[180] R. B. Tootell, N. K. Hadjikhani, J. D. Mendola, S. Marrett, and A. M. Dale. From retinotopy to recognition: fmri in human visual cortex. *Trends in cognitive sciences*, 2(5):174–183, 1998.

[181] H. Toyoda, K. Kashikura, T. Okada, S. Nakashita, M. Honda, Y. Yonekura, H. Kawaguchi, A. Maki, and N. Sadato. Source of nonlinearity of the BOLD response revealed by simultaneous fMRI and NIRS. *NeuroImage*, 39(3):997–1013, Feb. 2008. ISSN 1053-8119. DOI: 10.1016/j.neuroimage.2007.09.053. URL http://www.sciencedirect.com/science/article/pii/S1053811907008932.

[182] G. Turrigiano. Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Annual review of neuroscience*, 34:89–103, 2011.

[183] S. Uppenkamp, I. S. Johnsrude, D. Norris, W. Marslen-Wilson, and R. D. Patterson. Locating the initial stages of speech–sound processing in human temporal cortex. *Neuroimage*, 31(3):1284–1296, 2006.

[184] M. Ureshi, T. Matsuura, and I. Kanno. Stimulus frequency dependence of the linear relationship between local cerebral blood flow and field potential evoked by activation of rat somatosensory cortex. *Neuroscience research*, 48(2):147–153, 2004.

[185] L. Vagharchakian, G. Dehaene-Lambertz, C. Pallier, and S. Dehaene. A Temporal Bottleneck in the Language Comprehension Network. *Journal of Neuroscience*, 32(26):9089–9102, jun 2012. DOI: 10.1523/jneurosci.5685-11.2012.

[186] F. van de Velde and d. M. Kamps. Ambiguity resolution in a Neural Blackboard Architecture for sentence structure. *???*, 2015.

[187] F. van der Velde and M. de Kamps. Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, 29 (01), feb 2006.

[188] F. van der Velde and M. de Kamps. Learning of control in a neural architecture of grounded language processing. *Cognitive Systems Research*, 11(1):93–107, mar 2010. DOI: 10.1016/j.cogsys.2008.08.007.

[189] F. van der Velde and M. de Kamps. Development of a connection matrix for productive grounded cognition. In *2011 IEEE International Conference on Development and Learning (ICDL)*. Institute of Electrical & Electronics Engineers (IEEE), aug 2011. DOI: 10.1109/devlrn.2011.6037343. URL http://dx.doi.org/10.1109/devlrn.2011.6037343.

[190] F. van der Velde and M. de Kamps. The necessity of connection structures in neural models of variable binding. *Cognitive Neurodynamics*, 9(4):359–370, feb 2015. DOI: 10.1007/s11571-015-9331-7.

[191] D. van Dijk and F. van der Velde. A central pattern generator for controlling sequential activation in a neural architecture for sentence processing. *Neurocomputing*, 170:128–140, dec 2015. DOI: 10.1016/j.neucom.2014.12.113.

[192] G. Varoquaux and B. Thirion. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3(1):28, 2014.

[193] A. L. Vazquez and D. C. Noll. Nonlinear aspects of the bold response in functional mri. *Neuroimage*, 7(2):108–118, 1998.

[194] M. Vigneau, G. Jobard, B. Mazoyer, and N. Tzourio-Mazoyer. Word and non-word reading: what role for the visual word form area? *Neuroimage*, 27(3):694–705, 2005.

[195] M. Vigneau, V. Beaucousin, P.-Y. Herve, H. Duffau, F. Crivello, O. Houde, B. Mazoyer, and N. Tzourio-Mazoyer. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, 30(4):1414–1432, 2006.

[196] C. von der Malsburg. The Correlation Theory of Brain Function. In *Models of Neural Networks*, pages 95–119. Springer New York, 1994.

[197] L. Waldorp. Robust and Unbiased Variance of GLM Coefficients for Misspecified Autocorrelation and Hemodynamic Response Models in fMRI. *International Journal of Biomedical Imaging*, 2009:1–11, 2009. DOI: 10.1155/2009/723912.

[198] S. v. d. Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[199] X.-J. Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, 24(8):455–463, 2001.

[200] C. Wendelken and L. Shastri. Multiple instantiation and rule mediation in SHRUTI. *Connection Science*, 16(3):211–217, sep 2004. DOI: 10.1080/09540090412331311932.

[201] H. R. Wilson and J. D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.

[202] A. Wohrer, M. D. Humphries, and C. K. Machens. Population-wide distributions of neural activity during perceptual decision-making. *Progress in Neurobiology*, 103:156–193, apr 2013. DOI:

10.1016/j.pneurobio.2012.09.004. URL http://dx.doi.org/10.1016/j.pneurobio.2012.09.004.

[203] F. Wolf, R. Engelken, M. Puelma-Touzel, J. D. F. Weidinger, and A. Neef. Dynamical models of cortical circuits. *Current Opinion in Neurobiology*, 25:228–236, apr 2014. DOI: 10.1016/j.conb.2014.01.017. URL http://dx.doi.org/10.1016/j.conb.2014.01.017.

[204] B. Yeşilyurt, K. Uğurbil, and K. Uludağ. Dynamics and nonlinearities of the bold response at very short stimulus durations. *Magnetic resonance imaging*, 26(7):853–862, 2008.

[205] Y. N. Yoncheva, J. D. Zevin, U. Maurer, and B. D. McCandliss. Auditory selective attention to speech modulates activity in the visual word form area. *Cerebral Cortex*, 20(3):622–632, 2009.

[206] E. Zaccarella, M. Schell, and A. D. Friederici. Reviewing the functional basis of the syntactic merge mechanism for language: A coordinate-based activation likelihood estimation meta-analysis. *Neuroscience & Biobehavioral Reviews*, 80:646–656, 2017.