SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Thèse de doctorat de Sorbonne Université

Sciences Cognitives & Neurosciences

Présentée par

Lucas BENJAMIN

Human sensitivity to statistical regularities at different scales in auditory sequences : evidence from electrophysiology and behavior in adults and neonates.

Soutenue le

20 Novembre 2023

Devant un jury composé de :

M.	Floris DE LANGE *Donders Institute, Nijmegen*	Rapporteur
Mme	Marcela PEÑA *Pontificia universidad católica de chile, Santiago*	Rapporteure
Mme	Lucia MELLONI *Max Planck for Empirical Aesthetics, Frankfurt*	Examinatrice
Mme	Maria CHAIT *University College London, London*	Examinatrice
M.	Valentin WYART *École Normale Supérieur, Paris*	Examinateur
Mme	Ghislaine DEHAENE-LAMBERTZ *CNRS, Neurospin, Gif-sur-Yvette*	Directrice de Thèse

NeuroSpin

Table des matières

Preface

Title :

Human sensitivity to statistical regularities at different scales in auditory sequences : evidence from electrophysiology and behavior in adults and neonates.

Titre :

Perception humaine des régularités statistiques à différentes échelles dans les séquences auditives: études comportementales et en électrophysiologie chez l'adulte et le nouveau-né.

Keywords :
Cognition, Development, Neuroscience, Experimental Psychology, Statistical learning, Network learning

Mots clés :
Cognition, Développement, Neurosciences, Psychologie expérimentale, Apprentissage statistique, Apprentissage de graphe

Abstract (4000 signs) :

How does a brain manage to extract information from the environment and discover relevant structures in it, all the more so when it is immature and with little experience, as a newborn baby's brain can be? This question of how a newborn discovers the structure of a sound sequence, the scales of the structures it can extract, and the brain mechanisms at its disposal to do so, motivated our work. We studied the learning abilities of sleeping newborns and compared their performance with that of awake adults. Based on a set of

results, we argue that a large number of sequential structures covering different scales of dependencies could be handled by a single brain computation based on highly automatic associative learning principles. The framework we propose here brings together results hitherto considered as separate competences: on the one hand "statistical learning", and on the other "network learning". To test the properties and limits of this framework, we present here a series of experiments on the different scales of sequential learning in a variety of populations, from comatose patients to sleeping neonates to awake adults. We used behavioral measures, modeling approaches and neuroimaging techniques such as electroencephalography (EEG) and magnetoencephalography.

First, using EEG to assess the learning abilities of non-responding populations, we studied sleeping neonates and patients suffering from disorders of consciousness in their abilities to perceive simple associations between consecutive items in a continuous speech sequence. This ability was hypothesized as a major mechanism enabling babies to find words in a continuous stream of speech. Indeed, awake adults, sleeping neonates and even some comatose patients showed comparable signs of learning these local dependencies, arguing for a highly automatic process that does not require explicit attention and is already functional from birth on. Furthermore, awake adults and sleeping neonates revealed a similar word-length limitations, demonstrating that despite the immaturity of the infant brain, the statistical learning capabilities and limitations in this context are very similar to those of a mature adult brain.

Secondly, we proposed a new paradigm, the sparse community paradigm, to simultaneously study statistical and network learnings within a single sequence. Adult behavioral responses showed that humans parsimoniously completed the representation of the network despite incomplete evidence. Modelling approaches enabled us to

hypothesize that local and high-order sequential learning might be supported by a single cognitive computation rather than by separated processes: statistical learning for local transitions and network learning for high order structure. The use of MEG on a similar paradigm revealed neural correlates compatible with this hypothesis, showing that local statistical learning and network learning could be supported by the same associative learning implementation.

Finally, we reviewed the literature on statistical and network learning and proposed a new framework unifying these different learning scales under a common principle. We show that most of the results previously reported in the literature and explained with a variety of models, might rely on a single associative model. This hypothesis provides a more parsimonious explanation of how humans statistically extract a structure and makes a number of new predictions that have yet to be tested.

Résumé (4000 caractères) :

Comment un cerveau parvient-il à extraire des informations de l'environnement et à y découvrir des structures pertinentes, a fortiori lorsqu'il est immature et peu expérimenté, comme peut l'être le cerveau d'un nouveau-né ? La question de savoir comment un nouveau-né découvre la structure de la parole, la complexité des structures qu'il peut extraire et les mécanismes cérébraux dont il dispose pour le faire, a motivé notre travail. Nous avons étudié les capacités d'apprentissage de nouveau-nés endormis et comparé leurs performances à celles d'adultes éveillés. Sur la base d'un ensemble de résultats, nous soutenons qu'un grand nombre de structures séquentielles couvrant différentes échelles de complexité pourraient être traitées par un seul calcul cérébral basé sur le principe de l’apprentissage associatif. Le cadre que nous proposons ici rassemble des résultats jusqu'à

présent considérés comme des compétences distinctes : d'une part "l'apprentissage statistique" et d'autre part "l'apprentissage de réseaux". Pour tester les propriétés et les limites de ce cadre, nous présentons une série d'expériences d'apprentissage séquentiel à différentes échelles et dans différentes populations : des patients comateux aux adultes éveillés en passant par les nouveau-nés endormis. Nous avons utilisé des mesures comportementales, des approches de modélisation et des techniques de neuro-imagerie telles que l'électroencéphalographie (EEG) et la magnétoencéphalographie.

Tout d'abord, en utilisant l'EEG pour évaluer l'apprentissage de populations non verbales, nous avons étudié l'aptitude de nouveau-nés endormis et de patients dans le coma à percevoir des associations simples entre des syllabes consécutives dans une séquence de parole continue. L'hypothèse a été faite que cette capacité est un mécanisme majeur permettant aux bébés d'extraire les mots d'un flot de parole continu. En effet, les adultes éveillés, les nouveau-nés endormis et même certains patients dans le coma montrent des signes comparables d'apprentissage de ces dépendances locales, ce qui plaide en faveur d'un processus hautement automatique ne nécessitant pas d'attention explicite, et qui serait fonctionnel dès la naissance. En outre, les adultes éveillés et les nouveau-nés endormis ont montré des limitations similaires quant à la longueur des mots possibles à extraire, ce qui démontre que malgré l'immaturité du cerveau du nourrisson, les propriétés et les limitations de l'apprentissage statistique dans ce contexte sont très similaires à celles d'un cerveau adulte mature.

Deuxièmement, nous avons proposé un nouveau dessin expérimental, le paradigme de 'communauté incomplète', pour étudier simultanément les apprentissages dits statistiques et ceux de graphe dans une unique séquence. Les réponses comportementales d'adultes ont montré que les humains complétaient leur

représentation des structures, malgré une présentation incomplète. Des approches de modélisation nous ont permis d'émettre l'hypothèse que ces différents ordres d'apprentissage des séquences relèveraient en fait d'un seul et même processus cognitif. L'exploration MEG sur ce paradigme a révélé des corrélats neuronaux compatibles avec cette hypothèse, montrant que l'apprentissage statistique local et l'apprentissage de graphes pouvaient être sous-tendus par un unique mécanisme d'apprentissage associatif.

Enfin, nous avons passé en revue la littérature sur l'apprentissage statistique et l'apprentissage de réseaux, et proposé un nouveau cadre unifiant ces différentes échelles d'apprentissage sous un principe commun. Nous avons montré que la plupart des résultats précédemment rapportés dans la littérature, avec une variété de modèles à considérer, peuvent correctement s'expliquer par un seul modèle d'apprentissage associatif. Cette hypothèse fournit une explication plus parcimonieuse des mécanismes permettant aux humains d'extraire une structure statistique et fait de nombreuses nouvelles prédictions.

Short abstract (1000 signs) :

This PhD investigates how our brain learns relevant structures in sequential inputs. We used behavioral measures, modeling approaches, and neuroimaging techniques (EEG, MEG) to study sequential learning and the associated neural correlates in comatose patients, sleeping neonates, and awake adults.

We first showed that sleeping neonates and comatose patients perceive local associations between consecutive items in speech sequences. Then, we introduced the sparse community paradigm to simultaneously investigate local and higher-scale statistical learning. Behavioral and MEG results suggest that both scales might

be supported by a single associative learning mechanism. Finally, we formalized a unified framework, integrating statistical and network learning under a common associative learning principle, offering a more parsimonious and biologically plausible explanation of the literature.

Résumé court (1000 caractères)

Ce doctorat étudie comment notre cerveau apprend la structure séquentielle de notre environnent. Nous avons utilisé des approches de modélisation du comportement et des méthodes de neuro-imagerie (EEG, MEG), pour étudier cet apprentissage et ses corrélats neuronaux, chez des patients comateux, des nouveau-nés endormis et des adultes éveillés.

Nous avons d'abord montré que des nouveau-nés endormis et des patients dans le coma perçoivent les associations locales entre syllabes consécutives d'un flux de parole. Nous avons ensuite introduit le paradigme de communauté incomplète afin d'étudier simultanément les apprentissages locaux et d'ordre supérieur. Les mesures comportementales et MEG suggèrent que ces deux types d'apprentissages sont soutenues par un unique mécanisme cognitif.

Enfin, nous avons formalisé un cadre unifiant les différentes échelles d'apprentissage sous un principe commun, offrant une explication plus simple et biologiquement plausible de la littérature existante.

To all the neonates and their families,

for their time and kind participation.

ACKNOWLEDGMENTS

Despite the traditional single authorship of a thesis manuscript, this work is the result of constant scientific interactions and strongly supportive environment from all colleagues, family and friends that I would like to acknowledge.

First of all, I would like to warmly thank my thesis supervisor, Ghislaine Dehaene-Lambertz. These four years under your supervision have been immensely instructive and you have taught me so much about cognition and babies' development, but also about how to do good science and how to systematically doubt about results. I cannot be grateful enough for all the time you've given me, often going from a "quick question" to a two-hours discussion on how amazing learners babies are. Thanks to your never-ending enthusiasm, you made this PhD as enjoyable as it has been interesting and I really hope that we'll continue to share ideas and work together.

I would also like to address a very special thanks to Ana Fló and Fosca Al Roumi. Your day-to-day help, your amazing ideas and your support clearly shaped this thesis and I wouldn't have been able to do this work without your help. I want to thank all the amazing colleagues and friends for their valuable impact on this project : Mathias Sablé-Meyer (MEG partner in crime), Lucia Melloni, Shruti Naik, Stanislas Dehaene, Lorenzo Ciccione, François Leroy, Marie Palu (thanks for the hours testing at the maternity), Séverine Becuwe, Channel Valera, Elyes Tabbane, Florent Meyniel, Christophe Pallier Gaëlle Mediouni, Veronique Joly-Testault, Laurence Labruna and all the others. It would also not have been possible without the kind collaboration of Orsay and Port-Royal

maternities and staff. I also want to highlight how thankful I am to Vanna Santoro: your tremendous help and effort to make all this runs smoothly is invaluable.

I would like to take this opportunity to warmly thank all the extraordinary researchers who led me to cognitive neuroscience. Maria Chait, you are the one who introduced me to the wonderful field of cognitive science, who convinced me to pursue in it and who has followed and supported me ever since, thank you very much for that! Robert Zatorre, Philippe Albouy and Benjamin Morillon, thank you for opening the doors of your Montreal lab to me, and for trusting a complete beginner to work with you on your fMRI project.

Enfin, je tiens à remercier toute ma famille et mes amis. Merci à Agnès pour ton soutien indéfectible mais aussi pour les heures passées à m'écouter parler de ces recherches (je suis désolé, ça risque de continuer en post-doc...), tu en sais sans doute autant que moi sur ce sujet maintenant. Merci à mes parents, mes grands-parents, à mon frère, à ma sœur et à toute ma famille. Se poser beaucoup de questions, éveiller notre curiosité et discuter de tout à table est probablement à l'origine de mon envie de faire de la recherche et de mon goût pour les sciences. Merci à Christophe, Isabelle, Clarisse et Quentin de m'accueillir dans votre famille depuis le début. Enfin, à Sieg, Pauline, Caro, Gab, Lisa, Margot, Mathias, Fosca et tous les autres, vous savez ce que vous représentez pour moi, merci pour votre soutien et merci pour tout !

Finally, experimental neuroscience and human psychology can only be carried out thanks to the kind collaboration of the subjects who agreed to take part in the study. I can't thank enough all the subjects and in particular the neonates and families who took a little of their precious time at the maternity to take part in our experiments with no other motivation than to 'help science'.

INTRODUCTION

Foreword

To open this dissertation, I have decided to provide a brief general introduction to language and its development, not because the primary focus of this dissertation is ultimately on language per se, but because this was our entry point and the reason why we have come to consider the questions I will develop later. Subsequently, our thinking moved away from language, but the roots of these questions, the experimental paradigms we used, and the general context are directly affiliated to the experimental psychology of language and its acquisition as developed by *J. Mehler, R. Aslin*, or *J. Saffran* and of course by *G. Dehaene-Lambertz*, my thesis supervisor.

Each chapter correspond to a study that is, or aim to be, published. At the beginning of each chapter, a novelty statement specifies if the content of the chapter is published, in preprint, or novel to the thesis. It has then been written with the intention of being understandable as standalone articles. This might cause repetitions especially in the introductions of the different chapters. In particular the last chapter, the general discussion overlap with the thesis introduction. Moreover, each article is presented with its supplemental information, complementary analysis and comments. These supplementary analyses, as well as all the appendices, are not essential to the overall understanding of the work but are presented here for the sake of completeness.

Excerpt from Marietta Ren's mural in Neurospin galleria © M.Ren

Representation of the type of experiments conducted in Neurospin and which were carried out during this thesis: syllabic sequences perception in infants and adults using EEG and MEG recordings.

Introduction

Human language is probably one of the most fascinating of our cognitive abilities. It allows, by the simple act of putting words together, to transmit thoughts, feelings, complex images, lived experiences, future desires, from one mind to another. This incredibly powerful mechanism is at the heart of human cultural transmission and inter-individual learning. Indeed, most of what we learn in childhood and throughout our lives is explicitly taught, through oral transmission or written explanations. Even though cumulative cultural development is not a uniqueness of human beings, the efficiency and versatility of our communication system might have amplified this phenomenon. Learning language to communicate with our fellow human beings is thus crucial in the first years of life. The nature vs nurture question of language in humans has then been a center of interest for centuries for philosophers and scientists. Without going through a full history of this debate about the innateness of the language capacity in humans, let's briefly explore some of the questions and arguments that have been made to support one or the other idea. For example, Plato considered that humans must be born with innate knowledge given the great accomplishment one can have despite a short existence. Learning everything from scratch would necessitate more input than what human can be exposed to during its short life. This idea was later named 'Plato's Problem' by N.Chomsky, invoking a similar argument from *poverty of stimulus* as evidence of innate constraints for language acquisition in humans ***(Chomsky, 1986)***. For him, human language acquisition cannot be only a matter of imitation. At the opposite, empiricists like Locke or Skinner argued against innate knowledge. Locke re-used the metaphor of a *blank slate* or

tabula rasa to describe the mind of newborn humans. He argues that everything -including language- is learned through our sensory experience of the world ***(Locke, 1690)***.

One must admit that this debate is still unsolved, but recent technical achievements have opened new perspectives for exploring this issue. The first one is the recent tremendous progress in the ability of large linguistic models to mimic human language. Going against Chomsky's hypothesis, without having any particular "innate" constraints but just after intensive training, these models produce a language that is now difficult to differentiate from a human being. Nevertheless, the poverty of the stimulus argument probably cannot apply here, because the amount of data used to train today's large linguistic models is out of all proportion to the amount of data needed for human newborns to acquire language in their first years of life.

The second is the development of non-invasive neuroimaging, which enables the cognitive abilities of preverbal infants to be finely tested and dissected. For a long time, assessing the cognitive abilities of preverbal infants remained a puzzle, as researchers could not ask for introspective judgements or easily test learning in newborns. They have shown great ingenuity in measuring gaze time, heart rate or sucking behavior as indicators of their cognitive abilities. But recent advances in developmental neuroimaging capabilities with fMRI, fNIRS or EEG have considerably improved our ability to accurately test and measure the early abilities of infants, and even sleepy newborns just a few days old. These studies showed that despite immaturity, the baby brain was able of many remarkable cognitive capacities and variety of activations (including frontal activity) that were previously described as acquired later in life, conflicting with purely constructivist theories like the stages of development proposed by J.Piaget for example ***(Piaget, 1964)***. Conducting such experiments is still complex, and the difficulty of collecting and processing the data often makes the results of the analysis more delicate to interpret. Nonetheless, I strongly

believe that the developmental approach is essential to better understand human cognition and to study the respective biological and environmental constraints on the human mind.

Picture of a sleeping neonate wearing a 128-electrodes EEG system to monitor brain electrical activity.

Another approach to explore the roots of human language is to explore its appearance in our lineage. Paleoanthropologists tried to find clues for language capacities along the hominid's lineage. As cognitive capacities cannot be tested, they had to rely on very sparse evidences such as anatomical properties and DNA. Some explored asymmetries between the lobes, suggesting language-like areas on the left hemisphere in *H. habilis* ***(Hublin and Seytre, 2011; Picq et al., 2008)***. DNA sequencing of *H. neanderthalensis* revealed that he possessed the same variant of FOXP2 gene as modern humans ***(Krause et al., 2007)*** known to be essential for language ***(Lai et al., 2001; MacDermot et al., 2005)***. None of these bring conclusive evidence and overall, little is known about the language capacities of other hominids.

After this brief exploration of the variety of approaches to investigate the *language problem*, we will now restrain our interest and only focus on a developmental cognitive approach, exploring studies on language acquisition in humans.

Language development

The question of language acquisition is a complex problem to tackle. How come newborns can learn their native language and flexibly use it in just a few years ? Why does infant have such incredible language learning capacities that decreases in adulthood ? Learning a language is a multicomponent problem and many cognitive capacities are necessary for a successful learning. Here is a non-exhaustive list of cognitive capacities that have been considered as pre-requisite for language learning and investigated in a developmental approach. Indeed, before learning a complex grammar structure, infants must :

First, to correctly parse the auditory scene, it requires to differentiate between different speakers. Already at birth, neonates can differentiate between voices and even prefer to listen to their mother voice as compared to other voices ***(DeCasper and Fifer, 1980; Mehler et al., 1978; Querleu et al., 1984).*** This learning most probably occurs before birth as fetuses are already sensitive to speech stimuli ***(DeCasper et al., 1994; Lecanuet et al., 1992)***.

It also requires being sensitive to all relevant acoustical cues that carry information while ignoring meaningless acoustical variations in the auditory signal. While at birth neonates already prefer listening human and non-human primates' vocalizations compared to artificial sound ***(Vouloumanos et al., 2010)***, they later narrow this preference to human speech only and even to their native language ***(Kuhl, 2004; Kuhl et al., 1992; Mehler et al., 1988; Vouloumanos et al., 2010; Vouloumanos and Werker, 2007, 2004; Werker et al., 2012)***. Moreover, babies are sensible to subtle auditory contrasts between two acoustically close

syllables from a nonnative language that adults cannot differentiate anymore ***(Werker et al., 2012, 1981; Werker and Tees, 1984)***. These contrasts are only conserved in adulthood if they are meaningful and used in the native language.

Prosody (the way we speak) is another essential information to grasp as it contains useful information about the meaning. At birth, infants are already sensitive to the prosody of their mother tongue enabling them to distinguish their lative language from another language ***(Mehler et al., 1988)***

Infants also quickly learn to rely on auditory-visual integration for speech perception just like the adult by mapping syllable sounds to particular lips movements ***(Rosenblum et al., 1997)***.

Finally, building a vocabulary and learning word meaning is also an essential feature of language development ***(Bergelson and Swingley, 2012; Fenson et al., 1994)***. The mechanisms used by infants to discover and extract words out of a continuous speech signal has been much debated and is known as the bootstrapping problem. Prosody, statistical regularities between syllables, repetitions, position in the sequence… many proposals have been made to explore how babies could learn which group of syllables form words and which correspond to consecutive syllables of two adjacent words in a sentence. Among those, one is particularly driving my attention because it seems a very general mechanism that could have preceded language and be recycled in language acquisition but not limited to it : ***statistical learning***

Statistical Learning

Statistical Learning in cognition refers to the ability of organisms to learn dependencies between sensory inputs. This is a crucial capacity of the human mind to be able to build a mental model of the world around us and make predictions about possible future events. To do so, we constantly search for structure and statistical dependencies in the

uninterrupted flow of sensory input that we face. A very simple example would be that, through experience, we have learned to associate cloudy skies with high probability of rain in the near future. Many researchers explored the range of statistical dependencies that humans and other animals can acquire using different paradigms and modalities. It ranges for Pavlovian classical conditioning, statistical dependencies between adjacent elements, analysis of visual scenes, sequence learning, temporal expectations or even network learning.

Here, we will mainly limit ourselves to the human ability to extract statistical regularities from auditory sequences, even though most of the descriptions and experiments we will present here have been, or could be, translated into the visual domain.

Learning local statistics

Description

Learning local regularities in sequences has been reliably demonstrated in many species and population with a great variety of paradigms and approaches. Classical conditioning ***(Pavlov 1860)*** has been an early demonstration that co-occuring events could be associated, changing the behavioral response to the environment in non-human animals. In humans, the frequent association of syllables have been proposed as a precursor for speech intelligibility ***(Harris, 1955)***.

Later on, many researchers used artificial grammar formalism to systematically test human ***(Reber, 1967)*** and non-human animal capacities in learning such regularities. It can range from pure associative learning (A implies B) to more complex streams derived out of arbitrarily long artificial grammar. Sensitivity to local dependencies, following such grammars, has then been observed in humans ***(Maheu et al., 2019; Milne et al., 2018)***, but also non-human primates ***(Milne et al., 2018)***, rodents ***(Toro and Trobalón, 2005)*** and bees ***(Avarguès-Weber et al., 2020; Nicholls and Hempel de Ibarra, 2014).*** Different songbird species also learn

association of syllables from their parents, demonstrating a learning of the local dependencies present in their parents' vocal productions ***(James et al., 2020; Menyhart et al., 2015; Santolin et al., 2016; Takahasi et al., 2010)***.

In humans, computing local dependencies in auditory streams was proposed as a major mechanism to structure the input, available from an early age since ***(Saffran et al., 1996a)*** showed that 8-month-old infants could use transition probabilities - P(B|A) - between syllables to extract words from a monotonous stream with no other available cues – see also ***(Lew-Williams et al., 2011; Saffran et al., 1999; Saffran and Kirkham, 2018)*** . Indeed, they presented infants with a continuous stream of syllables composed of a random concatenation of four tri-syllabic pseudowords (we note A B C the syllables with an index i for the pseudowords they belong to : $A_iB_iC_i$). Thus, syllables within a pseudoword were fully predictable (for example /ra/ was always followed by /fi/ and so P(/fi/|/ra/) = 1 : $P(B_1|A_1) = 1$) while between words the transition probability dropped to 1/3 (the last syllable of a word can be followed by the first syllable of one of the three other words $P(A_2|C_1) = 1/3$) see *fig 0.1*. Using time looking paradigm, they showed that 8-months old infants were more familiar to Words ($A_iB_iC_i$) compared to so-called Part-Words (triplets straddling a word boundary $B_iC_iA_j$) demonstrating that using transitions probabilities between syllables, infants have learnt to extract and memorize the pseudowords from the sequence. They subsequently showed that those extracted words could be used by infants as labels ***(Hay et al., 2011)***, supporting the role of this mechanism in vocabulary construction. Since then, the sensitivity of humans to local dependencies has been robustly demonstrated in the auditory, visual and tactile domains ***(Conway and Christiansen, 2005; Fiser and Aslin, 2002)***, without the focus of attention ***(Batterink and Choi, 2021; Batterink and Paller, 2017; Benjamin et al., 2021)*** ,to some extent in sleeping adults ***(Batterink and Zhang, 2022)***, neonates ***(Benjamin et al., 2022b; Bulf et al., 2011; Fló et al., 2022a)*** or even comatose patients ***(Xu et al., 2022)***. Here again, this does not seem to be a human uniqueness as non-human

primates ***(Conway and Christiansen, 2001; Hauser et al., 2001)***, Dogs ***(Boros et al., 2021)*** and to some extent songbird ***(Takahasi et al., 2010)*** exhibit similar behavior.

Similar effects have also been observed with backward transition probabilities ***(Pelucchi et al., 2009a; Tummeltshammer et al., 2017)***, suggesting bi-directionnality in the computation of local relations between elements in streams (see *fig 0.3*).

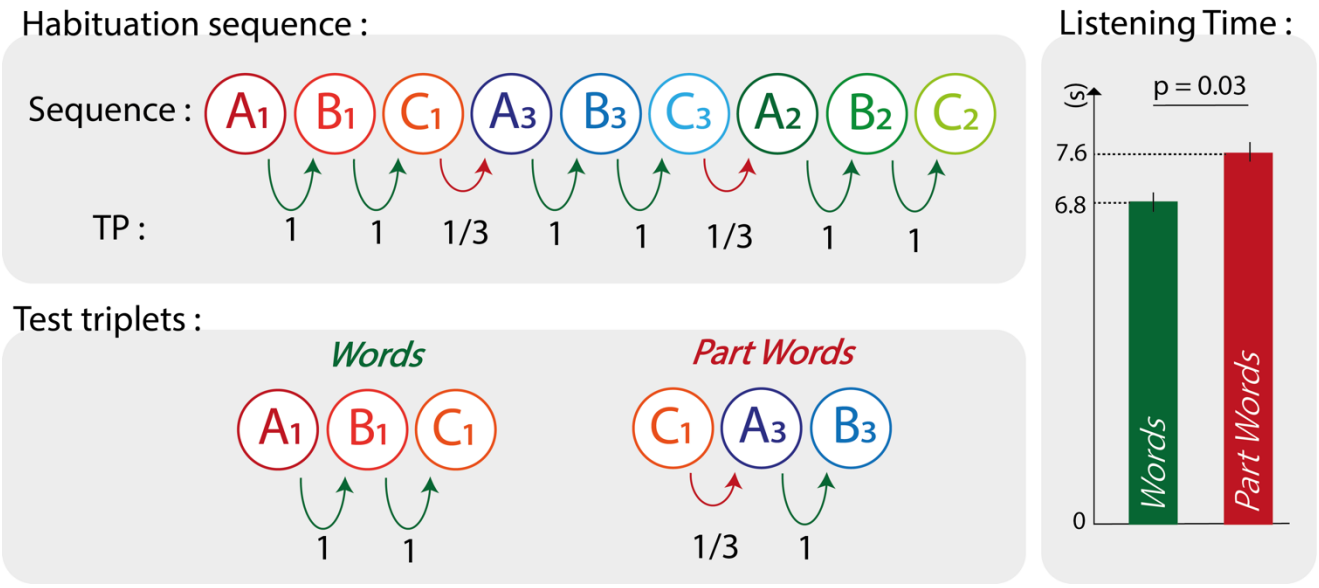


Figure 0.1 : Description of the Saffran paradigm and results from Saffran et al 1996*. 8 months old infants were exposed to a habituation sequence composed of four trisyllabic words pseudo randomly concatenated. Each syllable only belongs to a single word and so the local transition probabilities (TP) between syllables belonging to the same words are higher than between words (1 vs 1/3). To asses learning, infants were presented after 2 min of habituation with repetitions of test triplets that were either Words ($A_iB_iC_i$) or PartWords ($C_iA_jB_j$); Infants significantly listened longer at PartWords (red bar) than Words (green bar) showing a correct learning of transition probability and segmentation of the sequence into word-like units.*

Most of those results have been described by the authors as evidence for the computation of transition probabilities - P(B|A) – between consecutive elements in a sequence.

Learning statistics or chunks ?

While the Word vs Part-word effect is consensual, the mechanism at stake is debated in this literature. Indeed, do we have a statistical approach and compute transition probabilities between elements of the sequence ***(Endress and Johnson, 2021; Endress and Mehler, 2009; Fiser and Aslin, 2005; Saffran and Wilson, 2003)*** or a chunking approach where we recognize repeating triplets that we store into memory ***(French et al., 2011; Isbilen et al., 2020; Perruchet, 2019, 2019; Perruchet and Poulin-Charronnat, 2012; Perruchet and Vinter, 1998; Slone and Johnson, 2018)***. In order to disentangle between the two mechanisms, authors have proposed a new condition of phantom words ***(Endress and Mehler, 2009; Polyanskaya, 2022; Slone and Johnson, 2018)*** (triplets that never occur together but in which local transition probabilities are correct, for example in the sequence are embedded /ra/ /ti/ /fu/ and /bo/ /ti/ /ma/, the corresponding phantom word would be /ra/ /ti/ /ma/). Indeed, high familiarity with phantom words would support the statistical learning mechanism only, while rejection of those would argue more for a chunking mechanism. Unfortunately, both results have been found in the literature, leaving the problem unsolved ***(Endress and Mehler, 2009; Perruchet and Poulin-Charronnat, 2012; Slone and Johnson, 2018).***

Impact of priors

Statistical learning is not isolated from previous belief and priors on how the words are organized. For example, isolated words improve subsequent segmentation as babies knows what to look for ***(Lew-Williams et al., 2011)***. More than that, isolated words of a certain length before exposure can implies priors on the word length to look for in the sequence

and inhibit participants to perform the task if the priors on word length are not in accordance with the embedded words of the sequence ***(Lew-Williams and Saffran, 2012)***. Moreover, when studying adults with expertise in their oral native language, priors on the transition probability distribution or set of rules of these languages can affect performances ***(Elazar et al., 2022a; Onnis and Thiessen, 2013a; Potter et al., 2017; Siegelman et al., 2018; Wang and Saffran, 2014)***.

Learning non-adjacent dependencies

Computation of relations between elements is not limited to adjacent elements but can be extended to non-adjacent syllables - P(C|XA) - that could account for non-adjacent dependencies in language ***(Buiatti et al., 2009; Newport and Aslin, 2004; Pena, 2002)***. In such paradigms, participants are exposed to triplets (AXC) were the first element (A) predicts the last one (C) but the middle one (X) is variable. Although harder than adjacent transition probabilities, this kind of dependencies is learnable by humans but also non-human primates ***(Newport et al., 2004; Sonnweber et al., 2015)***. It has also been shown that the number of possible elements in the middle (X) position influence the capacity to learn the A-C relation ***(Gomez et al 2002)*** with more variability increasing our ability to learn the A-C dependency (see *fig 0.3*).

Learning high order structure

Several studies used a network approach to investigate how humans encode visual sequential information ***(Garvert et al., 2017; Kahn et al., 2018; Karuza et al., 2016, 2017, 2019; Lynn et al., 2020b; Lynn and Bassett, 2020; Mark et al., 2020; Ren et al., 2022; Schapiro et al., 2013, 2016; Stiso et al., 2022)*** but much less in the auditory modality ***(Benjamin et al., 2023a)*** despite the sophisticated auditory sequence processing abilities observed in humans compared to other primates ***(Dehaene et al., 2015)*** and better statistical learning capacities in the auditory

domain compared to other modalities ***(Conway and Christiansen, 2006)***. Using fMRI while participants had to navigate within networks, Garvert et al ***(2017)*** showed that the brain activity from hippocampus and entorhinal cortex correlates with *Communicability.* Communicability is a metric of Node distance in networks, more complex than just local properties (like the adjacency of nodes) but taking into account the full structure of the network to estimate how easy it is to navigate from one node to another in the network.

Shapiro and colleagues ***(Schapiro et al., 2016, 2013)*** tested human adults with a very particular network consisting of three communities (i.e. sets of nodes densely connected with each other and poorly connected with the rest of the graph - ***(Newman, 2003))*** where transitions between all elements were equiprobable (each node had the same degree) - See *fig 0.2* -. This community structure is an extreme version of the communities and clustering properties often found in real-life networks, whether social, biological or phonological ***(Girvan and Newman, 2002; Karuza et al., 2016; Siew, 2013)***. The authors observed that subjects discriminated transitions switching between communities from those staying within communities (*fig 0.2* - see also ***(Benjamin et al., 2022a; Lynn et al., 2020a; Ren et al., 2022; Stiso et al., 2022))***. One particularity of this design is that all nodes of the networks have exactly four neighbors either in the same community (the nodes deep inside the communities are linked with the four other nodes of the same community only), or three within the same community and one belonging to another community for the nodes at the border. Then, a sequence derived from such network does not show any pattern in its local transition probabilities between elements. Crucially, transitions that stay inside a community or that cross between two communities are associated with the same transition probability of ¼. Since local properties (TP) were not informative, the behavioral response observed when the sequence switch between community revealed participants' sensitivity to higher-order properties not explained by local or intermediate probabilistic models. The author speculated that humans could be sensitive to cosine similarity and use

this information to distinguish transitions within and between communities. Recently, Lynn and colleagues proposed a possible computation to account for such high order dependencies that takes into account the probability of memory errors in the computation of the transition probabilities (TP) between the elements of the stream.

It is worth noting that, if here we present those as statistical learning studies, they are most of the time discussed outside of the statistical learning literature. Indeed, network

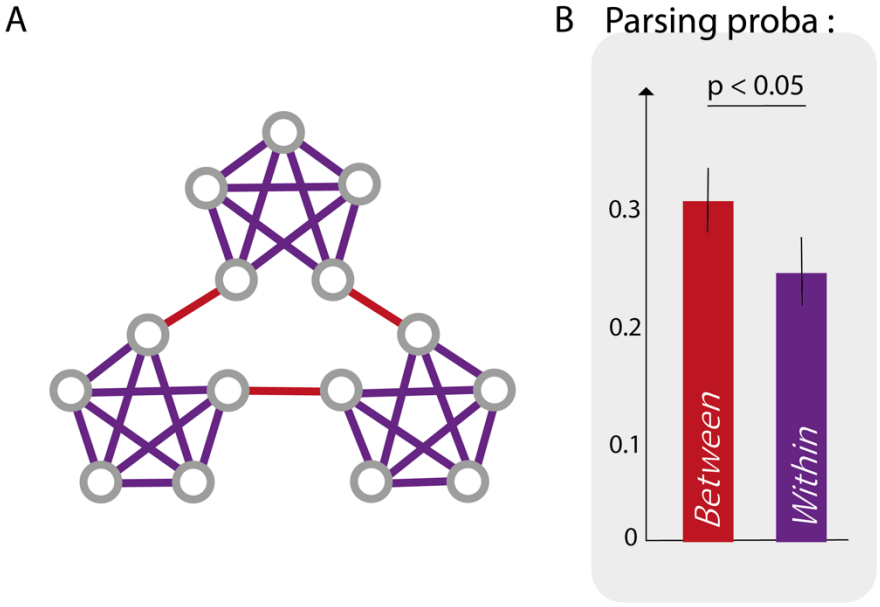


Figure 0.2: Description of the community paradigm proposed by Schapiro and colleagues (2016). *A : representation of the community structure, each node of the network represents a stimulus and each edge a possible transition between two stimuli. The twelve stimuli are organized into three communities where all nodes are connected to all the other nodes of the community (except the two nodes at the edge of the community). Only one transition is possible from one community to another. Each node exactly has four neighbors so that local transition probabilities between elements are always ¼.* ***B*** *When exposed to a sequence derived of this network and asked to press a button when they feel a natural break in the sequence, subjects pressed more when switching communities (red bar) compared to staying within a community (purple bar). As both Within and Between community transitions have the same local transition probability of 1/4 , this shows as sensitivity to the high order structure of the sequence, beyond local properties.*

learning (ie. learning topological properties of network structures such as clustering properties or centrality for example) is often presented as a distinct cognitive capacity, relying on the creation of maps, and navigation into abstract spaces. We have decided to present it as statistical learning (learning statistical relationships between elements of a sequence) of high order structures, because in this thesis, we will push the idea that learning networks rely on similar cognitive mechanisms than local statistical learning.

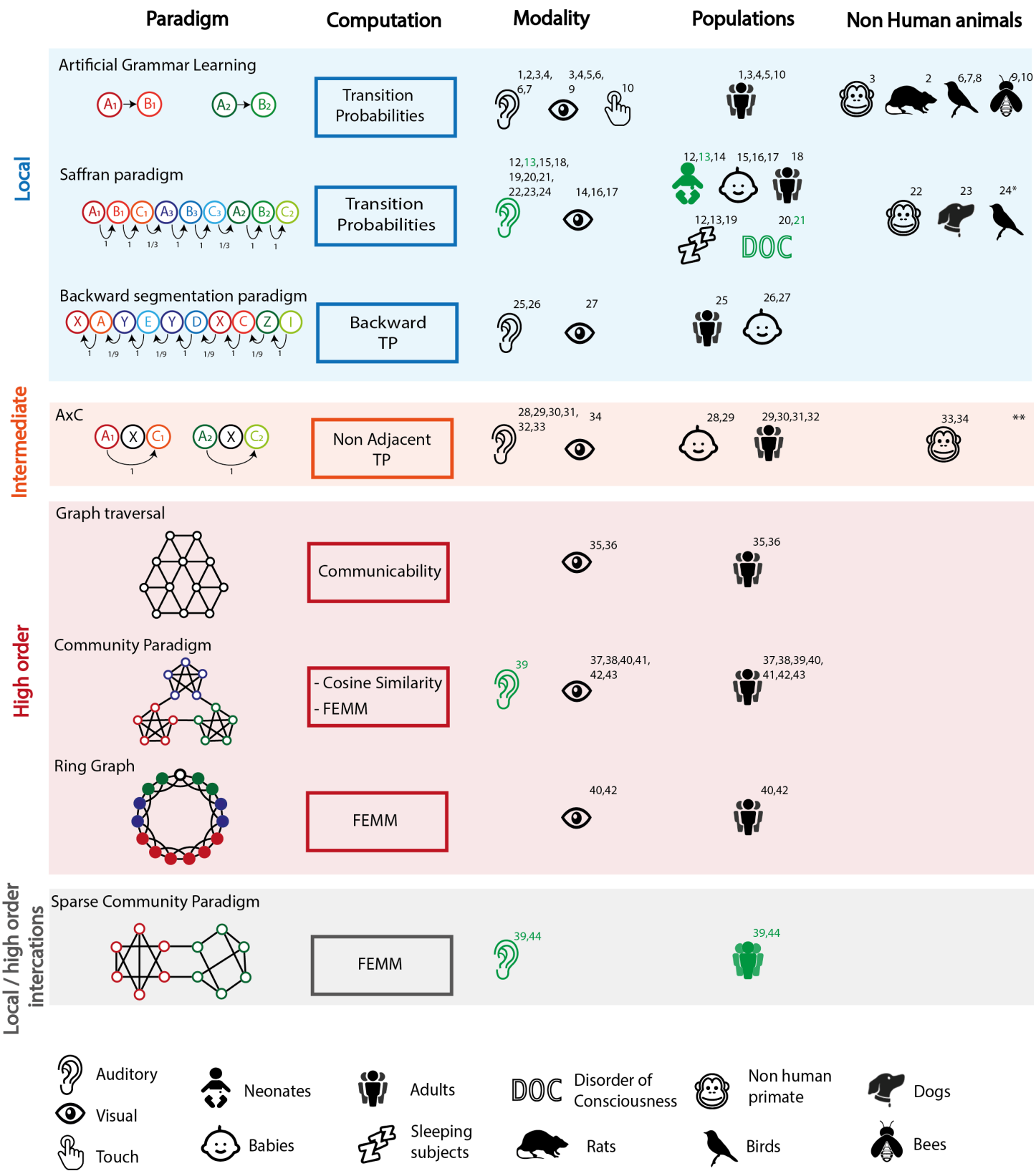

Paradigm
Computation
Modality
Populations
Non Human animals
Local
Intermediate
High order
Local / high order intercations
Artificial Grammar Learning
Transition Probabilities
1,2,3,4, 6,7
3,4,5,6, 9
10
1,3,4,5,10
3
2
6,7,8
9,10
Saffran paradigm
Transition Probabilities
12,13,15,18, 19,20,21, 22,23,24
14,16,17
12,13,14
15,16,17
18
12,13,19
20,21
DOC
22
23
24*
Backward segmentation paradigm
Backward TP
25,26
27
25
26,27
AxC
Non Adjacent TP
28,29,30,31, 32,33
34
28,29
29,30,31,32
33,34
**
Graph traversal
Communicability
35,36
35,36
Community Paradigm
- Cosine Similarity
- FEMM
39
37,38,40,41, 42,43
37,38,39,40, 41,42,43
Ring Graph
FEMM
40,42
40,42
Sparse Community Paradigm
FEMM
39,44
39,44
Auditory
Visual
Touch
Neonates
Babies
Adults
Sleeping subjects
DOC
Disorder of Consciousness
Rats
Non human primate
Birds
Dogs
Bees

Figure 0.3 : Classification of statistical learning paradigms into local, intermediate and high order. *We report here different studies investigation different scales of statistical learning and classify it by modality of the stimuli and populations tested. Studies that are part of this PhD are highlighted in green.*

1. Harris (1955) 2. Toro & Trobalon (2005) 3. Milne et al (2018) 4. Conway & Christiansen (2006) 5. Reber(1967) 6. Santolin et al (2016) 7. James et al (2020) 8. Menhart et al (2015) 9. Avarguès-Weber et al (2020) 10. Comway & Christiansen (2005) 11. Nicholls et al (2014) 12. Fló et al (2019, 2022a) ***13. Benjamin et al (2022) - Chapitre 2*** *14. Bulf et al (2011) 15. Saffran et al (1996a) 16. Fiser & Aslin (2002) 17. Kirkham et al (2002) 18. Saffran et al (1996a, 1999) 19. Batterink & Zhang (2022) 20. Xu et al (2022)* ***21. Benjamin et al (preprint) – Chapitre 1****. 22. Hauser et al (2001) 23. Boros et al (2021) 24. Takahasi (2010) 25. Perruchet and Desaulty (2008a) 26. Pelucchi et al (2009a) 27. Tummelsthammer et al (2017) 28. Buiatti et al (2009) 29. Gomez et al (2017) 30. Peña et al (2002) 31. Endress (2010) 32. Newport & Aslin (2004) 33. Newport et al (2004) 34. Sonnweber et al (2015) 35. Garvert et al (2017) 36. Mark et al (2020) 37. Schapiro et al (2016, 2013) 38. Karuza et al (2019, 2017)* ***39. Benjamin et al (2023a) - Chapitre 3*** *40. Lynn et al (2020a) 41. Ren et al (2022) 42. Stiso et al (2022) 43. Kakaei et al (2021)* ***44. Benjamin et al (preprint) – Chapitre 4***

Statistical learning in the brain

The representation and neural mechanisms underlying such learning is another topic of important research. How can such computation be implemented in the brain ? what is subsequently memorized ?

Several propositions have been explored to account for different type of learning. For the simplest one, the local statistical learning, researchers have proposed several mechanisms to account for it. First, sensory cortices could simply learn to associate consecutive elements by reinforcing their association strength each time they appear together. This very simple idea can be easily implemented using Hebb's learning rule '*Fire together, Wire together*'. Let's consider two neurons, each one firing specifically for a particular input. When two inputs are successively presented the two neurons will tend to fire simultaneously and thus following Hebb's learning rule, wiring each time a bit more. After enough stimulation, the pair will be strongly associated within the sensory cortex with a strong set of connections between neurons specifically tuned for both stimuli of the pair. The full representation of the transition probability matrix of the structure is then encoded in the neuron's connections. Such representation learning of the transition probabilities between elements of the sequence is supported by a signature of expectation violation in the brain around 150ms after the presentation of a stimulus if this stimulus was unpredictable (low transition probability) ***(Maheu et al., 2019; Todorovic and de Lange, 2012)***. Indeed, increased activity in the Superior Temporal Gyrus when listening to statistical compared to random sequences of syllables have been reported using fMRI ***(Cunillera et al., 2009; McNealy et al., 2006)***. Given the predominance of the left hemisphere for speech processing ***(Geschwind, 1970)***, and the described relation between word learning and structural integrity of the left arcuate fasciculus ***(López-Barroso et al., 2013),*** several studies investigated potential hemispheric specificities for statistical speech segmentation task. Left lateralized stroke survivor patients failed to learn statistical regularities in speech

segmentation task, while aged matched control where able to extract the words, suggesting a great importance of the left hemisphere in such computation ***(Fama et al., 2022)***.

However, statistical learning being a cross modal capacity, it is unclear whether the same brain areas are involved in learning and representing the structure independently of the modality, or if multiple modality units have converged though similar mechanisms. Other, modality independents brain regions such as the hippocampus and or inferior frontal gyrus have been implicated in statistical learning ***(Ellis et al., 2021; McNealy et al., 2006; Schapiro et al., 2014, 2012).*** But they have mostly been studied after successful statistical learning and not during learning. It is thus unclear whether they support learning of such structure or subsequent representation of the learnt elements. A recent study ***(Henin et al., 2021)*** explored those questions using intracranial recordings in epileptic patients doing a sequence segmentation task based on local statistical regularities. They measured brain response in different regions for both auditory and visual statistical learning task and showed that several representations of the sequences were maintained in parallel in different brain structures (fig 0.4). Specifically, they found electrodes that responded to both syllables and words onsets in the sensory cortices (auditory and visual) and using RSA they could show that those electrodes mainly encoded transitions probabilities of the sequence (fig 0.4 A,C&E). Some other electrodes more widespread across the brain only encoded for the words onset and maintained the ordinal position of the syllables within words (A vs B vs C) (fig 0.4 B,D&F). Finally, they specifically looked for hippocampal activations and found that in this particular task, the hippocampus represented similarly the syllables belonging to a single word, ignoring transition probabilities and ordinal position. Thus, the representation of a statistical sequence in the brain is not unique and supported by multiple brain regions simultaneously.

Mechanisms to support higher order statistical computations are not well known. Indeed, network learning is often independently studied from statistical learning. It is often considered as a higher-level abstract representation of structures. The hippocampus and frontal regions have then been interesting candidates for such computation ***(Constantinescu et al., 2016; Garvert et al., 2017; Schapiro et al., 2016).*** In support of this abstract map representation hypothesis, late top-down brain signatures have been found in EEG experiments investigating explicit network traversal ***(Ren et al., 2022).*** Some computational work also highlighted the capacity of the hippocampus to represent both local and high order computations within a single structure ***(Schapiro et al., 2017)*** but this result haven't been formally tested yet.

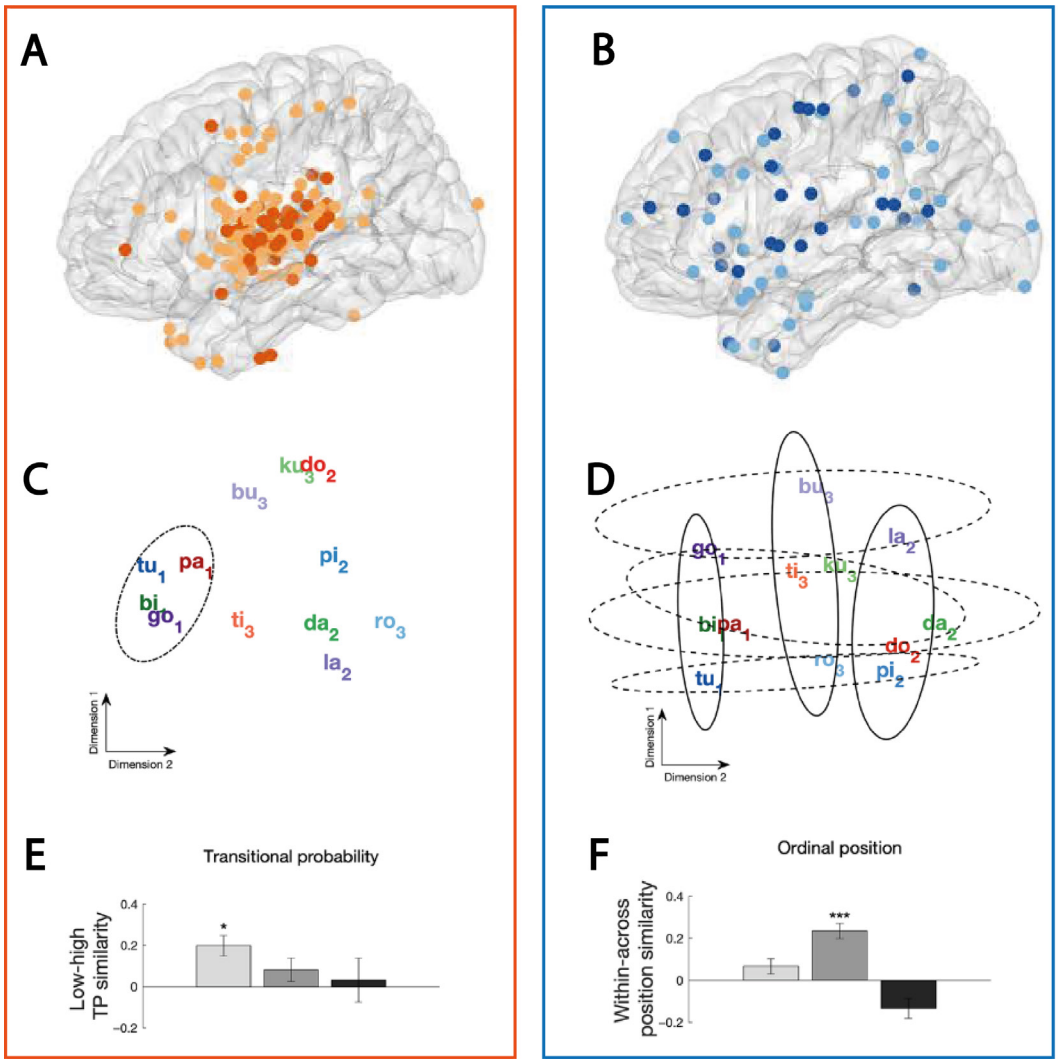


Figure 0.4 : *Adapted from Henin et al 2021. In the orange panel, electrodes that are stimulated both by the syllables and the words (A). The RSA (Representational Similarity analysis) over the representation elicited by each syllable after learning shows that all first syllables of the four words are clustered together (C) revealing an encoding of the transition probabilities (E). In the blue panel, electrodes that only responded to the word rate (B), the RSA showed clusters by ordinal position in the word (first second and last position in the words) (D&F).*

Structure but non-statistical learning

Sensitivity to structures in sequences is not limited to statistical probabilistic properties but also extend to rule learning. We refer to rule learning when the structure is not best described by statistical dependencies between elements but by patterns in the sequence. For example, we can consider item repetition (AABB), alternation (ABAB), repetitions of patterns, or many different arbitrary rules of this kind ***(Al Roumi et al., 2021; Barascud et al., 2016; Marcus et al., 1999; Planton et al., 2021; Southwell and Chait, 2018)***. Other studies from the laboratory explored the learning of such rules and how they combine to form more complex temporal or spatial structure that might be encoded in a language like manner ***(Al Roumi et al., 2021; Dehaene et al., 2022; Sablé-Meyer et al., 2022)***. In this thesis, we will mostly ignore this aspect of sequence learning as it probably relies on another cognitive mechanism ***(Maheu et al., 2020)***.

Purpose of this thesis

In this thesis, we will explore the different scales of statistical learning in humans. For that, we will use EEG and MEG metrics to measure brain correlates of statistical learning. We will also contrast these with behavioral judgment of familiarity and mathematical modelling approach to distinguish between several hypotheses on what is computed.

In a *first part*, we will explore the possibilities and limitations of local statistical learning in non-responsive populations. First, in the introduction, we will describe a study led by A. Flo where we used EEG to test the hypothesis that sleeping newborn can segment statistical sequences to extract and retain words. Based on those results, in **Chapter 1** we re-used the same paradigm to show that statistical learning is partially preserved in some patients with disorders of consciousness. In **Chapter 2**, we explore the dissociation that

exists between learning statistics and successfully segmenting a sequence. To do so, we increased the difficulty by embedding quadri-syllabic words instead of tri-syllabic words in a continuous sequence. We showed that both sleeping neonates and adults were unable to successfully segment such stream while still learning transition probabilities. We also showed that adding a subliminal pause between words helped to recover the segmentation of the sequence showing that learning transition probabilities and segmenting were dissociable processes. We also added **two appendices**. In the first one, we report exploratory behavioral results of segmentation helped by other cues than pauses (visual cue). In the second one, we report a methodological paper that we wrote about the neural entrainment method in EEG that we used to assess segmentation of the sequence in our data.

In a *second part*, we will explore network learning but from the perspective of high order statistical learning. Crucially, to understand if local statistical learning and higher order network learning are similar processes, we proposed a design mixing both. Indeed, we modified the community paradigm and introduced the *sparse community design* where high order and local transition probabilities were simultaneously present (see *fig 0.4*). We behaviorally measured participant learning on this task (**Chapter 3**) and showed that human adults could learn the high order structure and also overgeneralize it, creating fake memories of missing pairs. We systematically compared numerous statistical learning and network models proposed in the literature and showed that only some of them were able to predict the generalization effect that we observed. More specifically, a model of biased transition probability called the *Free Energy Minimization Model (FEMM)* better fitted our data than the other models. We then looked for the neural correlates of the implementation of such computation. We re-used the same mixed local and high order design to measure passive learning while brain activity was recorded with MEG (**Chapter**

4). We showed that the brain signatures of network learning were comparable to the one of local statistical learning and that the brain dynamic was compatible with simple bottom-up associative learning.

Based on the convergence between those two statistical orders, we propose in the *discussion* (**Chapter 5**) that local, intermediate and high order statistical learning rely on a single mechanism. Thanks to the previous studies, we have isolated one candidate model (FEMM) for unification. To support this hypothesis, we systematically collected results from all scales of statistical learning in the literature. We estimated FEMM on many previous experimental designs and showed that it correctly predicted a large variety of results ranging across statistical learning scales arguing for a unified approach of statistical learning at different scales.

Section 1 : LOCAL STATISTICAL LEARNING

Introduction

In this first section, we will explore local statistical learning capacities in non-responsive populations. Indeed, thanks to EEG and neural entrainment method (see appendix 2) we are able to assess the learning of local statistics between elements as well as the ability of non-responsive participants to segment the sequence.

We first explored sleeping neonates' statistical learning capacities on a classical segmentation sequence paradigm. As extracting statistical regularities from the environment is a primary learning mechanism that might support language acquisition, exploring newborns capacities is crucial. For this purpose, A. Fló ***(Fló et al., 2022a)*** led a study on sleeping neonates (1-3 days-old) re-using the classical Saffran paradigm (see fig 0.1 in the introduction). She first presented a sequence of random concatenation of syllables followed by a structured sequence composed of the pseudo-random concatenation of four tri-syllabic words. The electrical brain activity of the sleeping infants was recorded with high density EEG while resting and listening to those two types of sequences. To measure learning during the sequences, we used the neural entrainment approach (see **Appendix 2** for details on the different approaches for this computation). The principle is pretty simple : if the brain receives a rhythmic input, its activity will oscillate at this frequency. Here the syllables are presented rhythmically so that infants will hear four syllables per second. A functional auditory system will then evoke a at 4Hz

activity during both random and structured stream presentation. More importantly, in the structured stream only, three syllables form a word and so word presentation rate is at 1/3 of the syllabic presentation (1.33Hz). Thus, if infants succeed to segment the stream, they should start being sensitive to the word rate and an oscillation at 1.33Hz should appear in their brain data. Measuring the power in the EEG signal during resting state (fig 0.5 A), random sequence (fig 0.5 B) and structured (fig 0.5 C) sequences revealed a successful segmentation of the structured sequence by sleeping neonates. Moreover, analysis of ERP elicited by isolated pseudo-words after learning revealed that infants retained the first syllable of words.

Those results showed that the neural entrainment method was efficient for assessing segmentation sequences in non-responsive populations but also that learning local statistics between consecutive elements was an automatic enough process to be already in place a birth, in sleeping neonates.

We therefore wondered if statistical learning would be preserved in patients with disorders of consciousness as the need of conscious attention is a debated topic in the literature. Thanks to the neural entrainment method, we could test unresponsive patients at different severity of consciousness disorder and assess whether statistical learning is automatic enough to be preserved in a context of degraded consciousness **(Chapter 1)**.
We also explored if successful statistical learning was sufficient to trigger segmentation of the sequence. To do so, we ran a new experiment on adults and neonates with quari-syllabic words (instead of tri-syllabic) to increase the difficulty of word segmentation while keeping the same statistical information in the sequence **(Chapter 2)**.

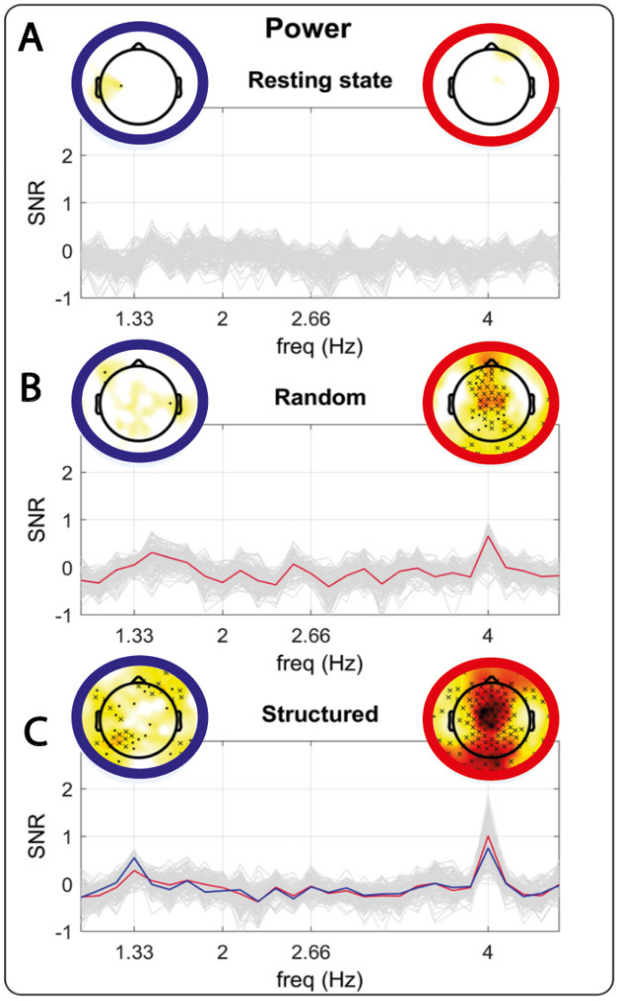


Figure 0.5 : *Adapted from Flo et al 2021 : Neural entrainment to the syllabic rate (4 Hz) and the word rate (1.33 Hz) during the three periods (Resting state, Random stream, and Structured stream). SNR for the power. In light gray, the entrainment for all electrodes. In red, the mean over the electrodes showing significant entrainment ($p < 0.05$, one-sided t-test, FDR corrected) at the syllabic rate. In blue, the mean over the electrodes showing significant entrainment ($p < 0.05$, one-sided t-test, FDR corrected) at the word rate. The topographies represent the entrainment in the electrodes space at the word rate (blue) and at the syllabic rate (red). Asterisks indicate the electrodes showing enhanced neural activity (cross: $p < 0.05$, one-sided t-test, FDR corrected; dot: $p< 0.05$, one-sided t-test, without FDR correction).*

Chapter 1 : Statistical learning is partially preserved in minimally conscious patients

This work is original and will be made available as a preprint in the coming months.

Abstract :

The debate over whether conscious attention is necessary for statistical learning has produced mixed and conflicting results. Testing individuals with impaired consciousness may provide some insight, but very few studies have been conducted due to the difficulties associated with testing such patients. In this study, we examined the ability of patients with varying levels of consciousness disorders (DOC), including coma, unresponsive wakefulness syndrome and minimally conscious patients, to extract statistical regularities from an artificial language composed of four randomly concatenated pseudowords. We used a methodology based on frequency tagging in EEG, which was developed in our previous studies on speech segmentation in sleeping neonates. Our study had two main objectives: first, to assess the automaticity of the segmentation process and, second, to explore a potential new diagnostic indicator to aid in patient management by examining the correlation between successful statistical learning markers and consciousness level. Although we observe some correlation with diagnosis, the high level of interindividual variability raises doubts about the possibility of using this indicator as a stand-alone diagnostic tool. However, we found that segmentation abilities were preserved in some minimally conscious patients, suggesting that statistical learning is a highly automatic low-level process. Furthermore, we found that the temporal accuracy of auditory syllable responses is strongly correlated with coma severity. Therefore, we propose that frequency tagging of an auditory stimulus train, a simple and robust measure, should be further investigated as a possible metric candidate for DOC diagnosis.

Introduction

The analysis of the statistical patterns present in sensory streams, a mechanism called statistical learning, enable us to uncover the structure of the input and to build an efficient mental model of the environment. The possibility of using such a mechanism to discover words in a continuous linguistic stream was first shown between adjacent syllables ***(Saffran et al., 1996a)***. In this type of experiment, four pseudo words are randomly concatenated to form a continuous sequence with high predictability of the next syllable within the pseudo word but a drop of transition probability between pseudo-words. Despite the absence of other cues for word segmentation such as those provided by prosodic indices, participants subsequently distinguished words composed of syllables with high transition probabilities, from part-words comprising a low transition between syllables. This result has been extended to non-linguistic ***(Saffran et al., 1999; Schön et al., 2008)*** but also visual ***(Fiser and Aslin, 2002)*** sequences. It is neither specific to the human species ***(Hauser et al., 2001; James et al., 2020),*** nor limited to adjacent elements, extending to non-adjacent dependencies ***(Kabdebon et al., 2015)***. It has also been shown that humans can learn higher-order statistical structures in sequences ***(Benjamin et al., 2022a; Dehaene et al., 2015; Garvert et al., 2017; Schapiro et al., 2013)***. Thus, statistical learning has been proposed to be a fundamental, and fairly automatic, learning mechanism operating in different modalities and sensitive to different type of statistical regularities.

Whether statistical learning occurs automatically and whether this mechanism requires attention remains a matter of debate, making it unclear under what circumstances such learning can occur. While most studies have used passive exposure to sequences in awake and attentive subjects, the impact of attention on statistical learning has been explicitly questioned by some authors, with mixed results. While some studies have shown a drastic

decrease in performance under divided attention *(Fernandes et al., 2010; Toro et al., 2005)*, others have shown that segmentation of auditory sequence was preserved despite attention diverted from the sequence being learned by asking subjects to focus on an independent visual sequence *(Batterink and Paller, 2019; Benjamin et al., 2021).* Another study even reports improved performances in participants under cognitive fatigue after performing an effortful working memory task prior to the sequence presentation *(Smalle et al., 2022.)*. Thus, studies that directly manipulate attentional focus do not provide any definitive answer regarding the interaction between statistical learning and participants' attentional resources.

Another angle of attack of this question is to study statistical learning in sleeping subjects. Although not directly testing attention, sleep studies address the automaticity of this mechanism. Batterink et al *(2022)* recently tested sleeping adults with pseudo-word segmentation tasks composed of either bi-syllabic or tri-syllabic pseudo-words. They reported that sleeping adults were only sensitive to bi-syllabic pairs but failed to extract tri-syllabic words. It suggests that sleep might reduce the integration period of the statistical computation without preventing the computation between adjacent items. This interpretation appears congruent with Strauss and colleagues's study *(2015)* who observed during sleep, a preserved mismatch response to local auditory violations of a sequence regularity but no reaction to a violation concerning a regularity at a longer time scale, that induces a P300 in awake adults. Interestingly, sleeping one- to three-day-old neonates tested with EEG in the same pseudo-word segmentation paradigm than in Batterink et al *(Benjamin et al., 2022b; Fló et al., 2019, 2022a)* showed similar statistical learning abilities as awake adults, computing statistics between syllables and even segmenting tri-syllabic pseudo-words while sleeping (for quadrisyllabic pseudo-words see *(Benjamin et al., 2022b)*).

An even more radical test of the automaticity of statistical learning would be to observe this learning in comatose patients. Indeed, among Disorders of Consciousness (DOC), patients may present several levels of residual consciousness characterized by diagnostic tools such as CRS-R ***(Giacino et al., 2004)*** or categorical classification into Unresponsive Wakefulness Syndrome (UWS) -no sign of visible awareness- or Minimally Conscious State (MCS) –visible partial awareness- ***(Bruno et al., 2011; Giacino et al., 2002)***. A first attempt to measure statistical learning in DOC patients was recently made by Xu et al ***(2022)*** using bi-syllabic words concatenated in a continuous, monotonic stream. They reported some learning in patients with emerging consciousness. Indeed, the power at the frequency of syllable pairs and its harmonics were significantly above zero, suggesting successful segmentation of sequences into word pairs. Although this study supports the possibility of statistical learning in DOC patients, it had some limitations, which call for further research. First, the bi-syllabic units used in the study, which required only a pairing between two items, do not encompass all aspects of segmentation, as shown in the sleep studies discussed above. Furthermore, the pairs presented in the study could be either frequent and meaningful in natural language or reversed and thus meaningless (e.g., "go home" versus "home go"). However, even in the reversed condition, syllables within a pair are more related to each other in natural language than syllables between pairs. This makes it difficult to differentiate between current learning of statistical regularities in the stream and a memory of previous exposure to natural language. This limitation was partially addressed by obtaining similar results with the learning of artificial tri-syllabic pseudowords, but the sample of *minimally conscious* patients was small (N=8).

In the present study, we used the experimental design and techniques of our previous research in sleeping neonates ***(Fló et al., 2022a)*** (see Methods): Our aim was to investigate the ability of DOC patients to recognize statistical regularities in an artificial syllable

sequence composed of four tri-syllabic pseudo-words that were concatenated in a pseudo-random manner (fig 1.1). The syllables within the words followed each other predictively (transitional probabilities = 1), while between the words, the transitional probabilities dropped to 1/3. In addition, the subjects were also presented with a random sequence composed of the same syllables but without forming pseudo-words, which had a flat transitional probability of 1/11. We also included a group of healthy awake adults as a control to compare with the clinical population.

We utilized high-density EEG recordings (256 channels) to measure sequence segmentation with the frequency tagging method, as it has been demonstrated to be a robust approach for evaluating non-responsive subjects in our previous studies (Fló et al., 2022a). This technique relies on detecting rhythmicity of the brain activity driven by the rhythmicity of the input sequence. It permits the monitoring of sequence segmentation by following the modulation of power and Phase Locking Value (PLV) at the syllabic and word frequencies. Analyzing the power at the syllabic rate allowed us to assess basic auditory processing at the individual level and check which patients have intact auditory perception. Segmentation of the tri-syllabic words is revealed by an increase in the word-rate frequency and eventually harmonics.

Not only does this paradigm address the question of conscious attention in statistical learning, but it also has the potential to improve the clinical characterization of DOC patients. Previous studies have found modification of EEG features with the level of consciousness, such as spectral power ***(Goldfine et al., 2011)***, auditory ERP amplitude and latency ***(Strauss et al., 2015)***, and mismatch responses in EEG oddball paradigms ***(Laureys and Schiff, 2012)***. The depth of language processing during time-locked natural language exposure has also been suggested as a useful metric for predicting patient outcomes ***(Gui***

et al., 2020). With the present study, we also aim to investigate whether neural markers of statistical learning can be utilized for diagnosis and outcome prediction in DOC patients.

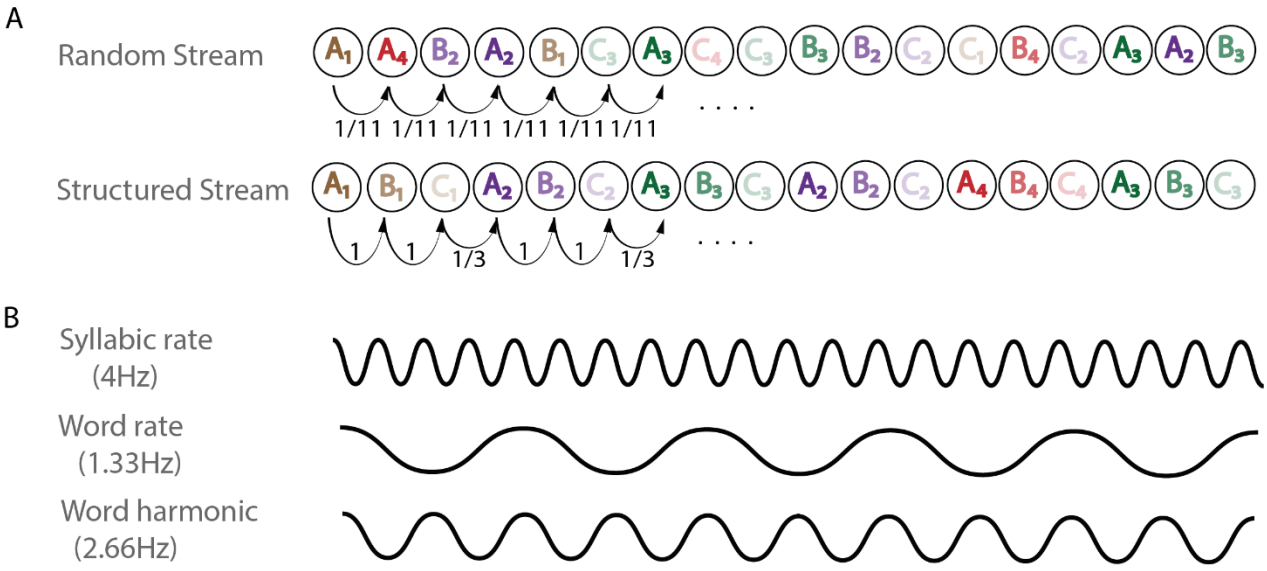


Figure 1.1 : Presentation of the paradigm and dataset*. A) Description of the two streams presented to participants. The random stream consists of syllables that can be followed by any of the other 11 syllables, resulting in a flat transitional probability of 1/11 throughout the sequence. The structured stream is composed of four tri-syllabic pseudo-words with transition probabilities between syllables equal to 1 inside the pseudo-words and 1/3 between the pseudo-words.* ***B)****. Frequency tagging analyses : The syllables were presented at a rate of 4Hz, which is expected to elicit a 4Hz oscillation in the brain of normal hearing subjects. If, and only if, the structured sequence was segmented based on transition probabilities, the phase locking value (PLV) at the word rate (1.33Hz) and its harmonic (2.66Hz) should increase relative to the random stream.*

Methods

Participants

81 patients from Huashan hospital with disorders of consciousness were included in the experiment for a total of 180 recordings. Clinical assessment of the patient state was made

just before each recording. Out of those 180 recordings, 14 were classified as coma, 65 as UWS and 101 as minimally conscious or emerging of consciousness. Due to too high level of noise, 3 recordings were not analyzed. 26 healthy control adults (>18yo) were tested and 24 completed the full experimental procedure (resulting in 72 recordings).

Stimuli

We synthesized the syllables using the open-source text-to-speech synthesizer eSpeaker ***(Jonathan and Reece, 2020)*** with Mandarin Chinese language (zh) and the female voice variant f2. We set the pitch parameter to 70 and adapted the speed to obtain syllables with a duration as close as possible to 250 ms. We further corrected the syllables using Praat ***(Boersma and Weenink, 2020).*** Specifically, we removed silent periods in the beginning and the end to obtain syllables lasting exactly 250ms and removed pitch changes setting a constant pitch of 225 Hz. The syllables audio files were concatenated without pauses to obtain the streams, and the first and last 4.5 s were ramped up and down to avoid the start and end of the stream might serve as perceptual anchors.

The structured streams consisted of a semi-random concatenation of four tri-syllabic pseudowords. The only restriction for concatenation was that the same pseudoword could not appear twice in a row, and the same two pseudowords could not repeatedly alternate more than two times (i.e., the sequence WkWjWkWj, where Wk and Wj are two words, was forbidden). We balanced phonetic features across the three syllables of the pseudowords to avoid they could serve as segmentation cues. In addition, we created three different structured streams by changing the arrangement of the 12 syllables (each syllable occupied a different position within the pseudoword in each stream). The random stream resulted from concatenating the 12 syllables semi-randomly (without syllables repetition), giving an average uniform TP of 1/11. We also created test words to be presented in isolation but we do not present the results here.

Procedure

Scalp electrophysiological activity was recorded using a 256-electrode net (GTEN 200, Magstim EGI) referred to the vertex with a sampling frequency of 1000 Hz. The recording procedure was similar to the one used in neonates ***(Fló et al., 2022a)***. Participants first heard a random sequence of 4 minutes (960 syllables) followed by a structured stream of 4 minutes (960 syllables, 320 words). After, participants listen to 8 repetitions of 30s (120 syllables, 40 words) of structured sequences followed by a block of 16 pseudo-words test trials

All Healthy subjects were tested with the three structured streams (lists) on different days. For the DOC patients, we tried to follow the same pattern by testing each subject on each list as much as the hospital constraints allowed us. On average DOC subjects were tested 2.27 times, with 40 subjects having been tested on the three lists. We obtained 88 recordings with list 1, 50 with list 2 and 46 with list 3.

Data preprocessing

Data were resampled to 250 Hz, band-pass filter 0.3–30 Hz and pre-processed using APICE pipeline for MATLAB ***(Fló et al., 2022b)*** with default rejection strength parameters. This preprocessing pipeline detects artifacts on the continuous signal based on deviance from the distribution of data points. It later corrects isolated artifacts with interpolation methods, recovering part of the signal. Long periods of artifacts, or simultaneous failure of several channels cannot be corrected with neighbors' data and the corresponding epochs are removed from further analysis.

In order to remove eye movements and blinks, we also performed ICA on the healthy control data.

Frequency tagging

The pre-processed data were segmented from the beginning of each sequence into segments comprising 13 words to approach 10s long epochs (13*0.750=9.75s). Segments were not overlapping to avoid an artefact in the frequency domain related to the length of the overlap ***(Benjamin et al., 2021).*** Epochs with artifacts were rejected and sessions with less than 15 good epochs per condition (random and structured) were not included. Data were converted to the frequency domain using the Fast Fourier Transform (FFT) algorithm and the Phase Locking Value (PLV) and Power were estimated for each electrode in both random and structured conditions. The phase locking value ranges from 0 (completely asynchronized data) to 1 (completely timed-locked activity). The value at each frequency bin estimation was then normalized by the mean value of eight neighboring frequency bins on each side.

Statistical analyses

They were performed on the three frequencies of interest: 4Hz corresponding to the syllabic rate, 1.33 and 2.66 corresponding to the frequency of the trisyllabic words and its first harmonic. Results of the analyses performed on the PLV are presented in the main text, and for power in the supplementary material. Results are largely similar for these two metrics.

For each electrode and for each frequency, it was first tested whether the normalized PLV (and power) was above 0 with a one-sample t-test against zero, second whether these values for the word frequencies were larger during the structured stream relative to the random stream (structured > random one-way paired t-test). In all analyses, p-values were corrected for multiple comparisons (256 electrodes) using FDR (fig 1.2).

Modulation of the neural responses by level of consciousness: To examine how the responses were influenced by the level of consciousness in DOC patients, we conducted correlation analyses between the Phase Locking Value (PLV) of each electrode and the Coma Recovery Scale-Revised (CRS-R) score measured prior to the recording in DOC patients only. Importantly, healthy controls were not included in this analysis to ensure that any observed correlation was not solely driven by the differences between healthy and DOC subjects, but rather by the varying degrees of consciousness within the DOC group.

Modulation of words segmentation by the syllabic response : To be able to segment words in such a stream, patient should at least have a minimal auditory function and some patients may not meet this criterion. We considered the responses at the syllabic rate as a proxy of low-level auditory function. Therefore, the previous analyses were done again retaining only patients with a positive average syllabic rate PLV, that is patients for whom the mean PLV across all electrodes for both the random and structured conditions was > 0 (fig 1.3). This rejection metric is quite stringent as it rejected participants with impaired auditory processing but also recording with too low signal/noise ratio to correctly detect this metric. We also re-calculated the correlation between the PLV at the word rates and level of consciousness after regressing out the PLV at the syllabic rate for each electrode: we first performed a regression between word rate PLV and the syllabic rate PLV for each electrode; then the residuals of these regressions were correlated with CRS-R (fig 1.3 panel D).

Results

Evidence of word segmentation in the different experimental groups

To estimate whether patients were able to correctly segment the words concatenated in the structured stream, we measured the normalized PLV at the word rate and its first harmonic and compared the values to those measured in the random stream (fig 1.2 panel A).

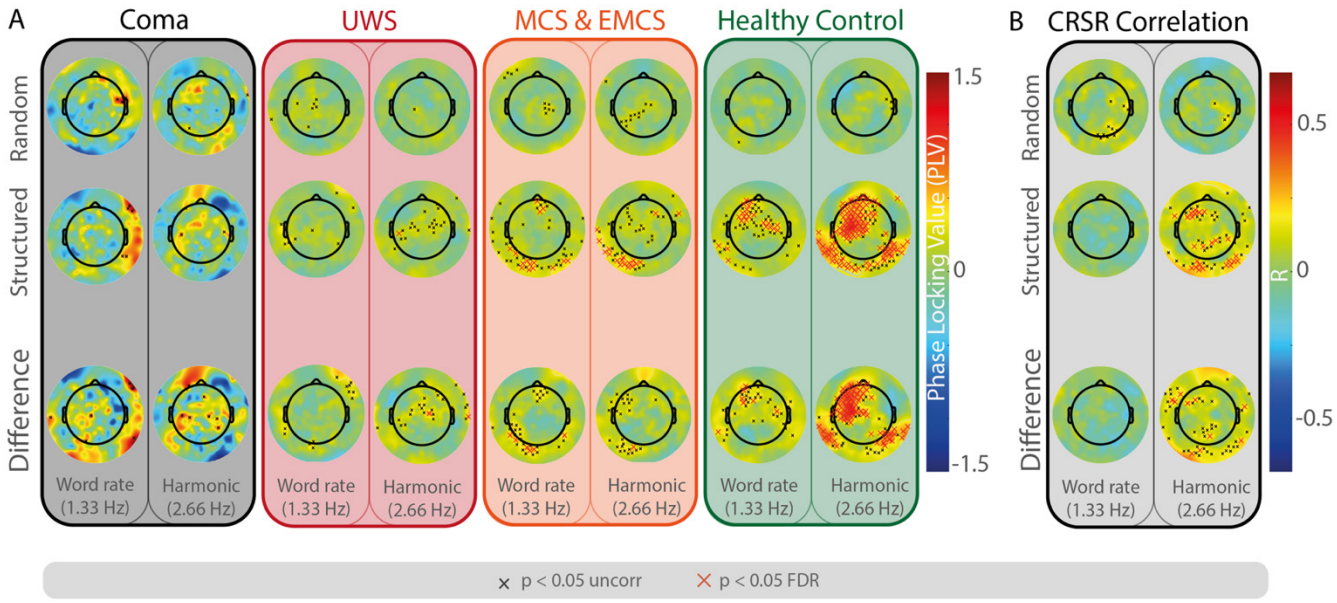


Figure 1.2 : Neural entrainment analysis. A) *Normalized Phase Locking Value (PLV) for each electrode at the word rate (1.33Hz) and its first harmonic (2.66Hz) during the random and structured streams in the recordings of Coma, UWS, MCS or Healthy subjects. The bottom line presents the contrast structured > random. Dots represent electrodes with p< 0.05 before multiple comparison correction and red crosses electrodes significant after FDR correction.* ***B)*** *Correlation of the PLV at word rate and its harmonic with Comatose Recovery Scale Revised (CRSR). The harmonic of the word rate (2.66Hz) significantly correlates with the clinical score only during the structured stream.*

First, in healthy control subjects, we found many electrodes with significant positive PVL at both word rate and its harmonic when they listened to the structured stream (ps < 0.05

FDR corrected) but not to the random stream, with a significant difference between the two conditions in many electrodes (fig 1.2). The same analysis in minimally conscious patients (clinical assessment : MCS or EMCS) showed similar results, although weaker in terms of number of significant electrodes. In the UWS group, a trend in the same directions as the other groups was visible (Structured stream at 2.66Hz : 22 channels with $p<0.05$ uncorrected), but only one channel survived the FDR correction at the word rate harmonic. The comparison with the random stream showed a very modest effect (2 electrodes with $p < 0.05$ FDR corrected). In the Coma patients, no electrode showed a significant segmentation effect.

Correlation of the segmentation performance with CRS-R

We then estimated for each electrode the correlation of the word and harmonic PLV with the clinical Coma recovery scale revised (CRS-R) estimated just before the recording, excluding the healthy participants. We found a highly significant correlation spread on many channels of the neural entrainment (normalized PLV) and the CRS-R score during the structured stream only for the word first harmonic (24 channels with $p < 0.05$ FDR) with a significant difference with the random condition (fig 1.2 B).

Controls taking into account low-level auditory perception as captured by the response at the syllabic rate

The previous correlation between PLV at the word harmonics and CRS-R score might indicate either a modulation in statistical learning and segmentation abilities with DOC severity or a spurious correlation due to a higher number of patients with auditory perception impairment and a severe DOC. Fortunately, PLV at the syllabic rate effectively summarizes basic auditory perception and temporal synchronization of auditory ERP. We presented the distribution of the average PLV at syllabic rate for each recording in figure

1.3A. Therefore, we repeated the previous analyses (fig 1.2) while excluding recordings with negative syllabic rates to ensure that all recordings left are with patients with correct

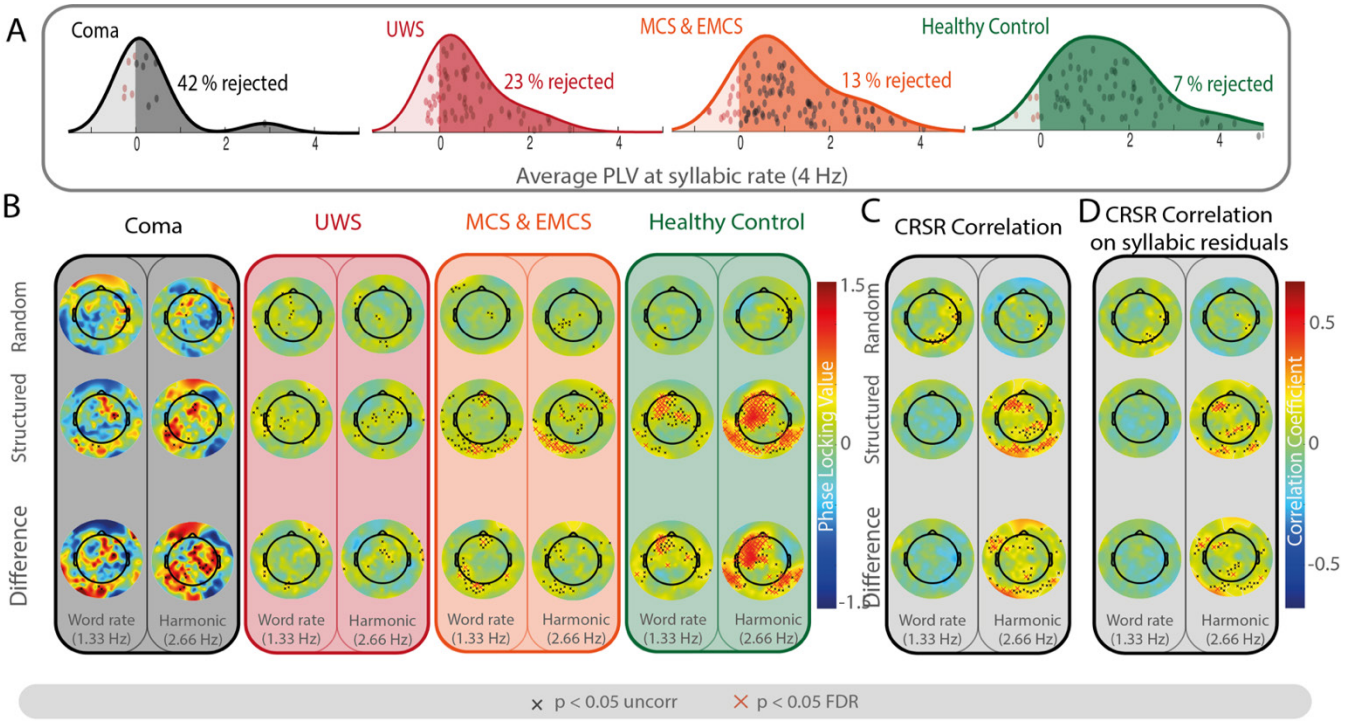


Figure 1.3 : A) *Average PLV at the syllabic rate for each recording. Syllabic rate is a good metric of preserved auditory perception and correct signal/noise ratio in the recording. For the following analysis, we only kept the recordings with a positive average syllabic rate as we cannot be sure that the other participants even heard the stimuli. Red dots represent recording with negative syllabic rates, who were then excluded from the analyses presented in B and C.* ***B)*** *Normalized PLV at each electrode at the word rate (1.33Hz) and its first harmonic (2.66Hz) during the random and structured streams in patients with Coma, UWS, MCS and Healthy subjects, excluding recordings with negative average PLV at syllabic rate (red dots on Panel A). The bottom line is the contrast structured > random. Dots represent electrodes with p< 0.05 before multiple comparison correction and red crosses electrodes significant after FDR correction.* ***C)*** *Correlation of the PLV at word rate and harmonic with CRS-R scores excluding recordings with negative average PLV at syllabic rate. The harmonic of the word rate (2.66Hz) significantly correlates with this score during the structured stream only.* ***D)*** *To account for the variation of neural entrainment at syllabic rate, we computed the correlation after having regressed out the effect of DOC on syllable entrainment, results remained similar*

hearing and a level of signal/noise in the recording that enable correct measure of frequency tagging. The results remained similar, both considering the PLV measures in each group (fig 1.3 B) and the correlation with CRS-R (fig 1.3 C).

Finally, to mitigate the impact of the variation of the PLV at the syllabic rate of each electrode on the segmentation metrics, we regressed out the syllabic rate for each recording for each electrode before computing the correlation of the residuals with CRS-R score. For this analysis, we included all recordings of DOC patients, comprising those with negative average syllabic rates. We obtained results similar to the original analyses: 1) a significant correlation during the structured stream only with the first harmonic, but not at the fundamental; 2) a significant contrast between structured and random sequence streams (fig 1.3 D).

The analyses presented in fig 1.2 and 1.3 have been replicated using power instead of PLV with similar results (fig 1.5 and 1.6 in supplementary material).

Auditory ERP measured with syllabic rate neural entrainment

Independently from statistical learning, we found that the neural entrainment at the syllabic rate was highly correlated with the clinical assessment of the level of consciousness in DOC patients. Indeed, we found many electrodes for which PLV and Power at the syllabic rate (4hz) robustly correlated with diagnostic scales, such as CRS-R and Glasgow Coma Score (fig 1.4). PLV at the syllabic rate significantly correlated with CRS-R on most electrodes. On fig 1.4B, we display the distribution of the average PLV value from the significant electrodes, for each subject by clinical assessment group. We clearly observe higher PLV at the syllabic rate associated with better clinical assessment. We also compared for each group the probability of improved outcome six months later depending on the PLV at the syllabic rate and found no significant differences (recordings associated with an improved six-month outcome are displayed in green on fig 1.4).

Discussion

The first goal of this study was to establish whether statistical learning and auditory sequence segmentation might persist in patients with disorder of consciousness amidst conflicting literature on the role of conscious attention. Our second goal was to explore if statistical learning metrics may help clinical diagnosis and care of these patients.

Statistical learning is partially preserved in DOC patients

The automaticity of statistical learning is still being debated. Indeed, while some studies showed a large decline in performance under divided attention ***(Fernandes et al., 2010; Toro et al., 2005)*** arguing for the need of focused attention on the task, others have reported that sequence segmentation persists even when outside the focus of attention ***(Batterink and Paller, 2019; Benjamin et al., 2021)*** or even improves with cognitive fatigue that impairs focal attention ***(Smalle et al., n.d.)***. Even, sleeping neonates can automatically segment a stream based on its statistical properties ***(Benjamin et al., 2022b; Fló et al., 2022b).*** Similarly, other studies investigating visual statistical learning with hypnosis ***(Nemeth et al., 2013),*** or TMS disruption of DLPFC ***(Ambrus et al., 2020)***, suggest that statistical learning abilities are enhanced when prefrontal activity is reduced. To make progress on this issue, studies in DOC patients provide valuable insight through the test in patients with different level of residual consciousness. In a recent study, ***(Xu et al., 2022)*** showed that some comatose patients were able to extract bi-syllabic real words (or mixed) arguing for a preserved minimal version of statistical learning on pairs of real words. The current study using tri-syllabic artificial pseudo-words further supports that statistical learning and pseudo-word segmentation might occur when MCS patients passively listened to an artificial language. Since minimally conscious patients suffer from severe attention malfunction, this study gives further evidence that statistical learning might largely occur without the full focus of attention.

Why are results more visible on the first harmonic compared to the word rate ?

All our analyses pointed toward a greater neural entrainment at the first harmonic compared to the fundamental frequency at the word rate. In our design, the harmonic of the word rate (2.66Hz) is different from half of the frequency of the syllabic rate (2Hz);

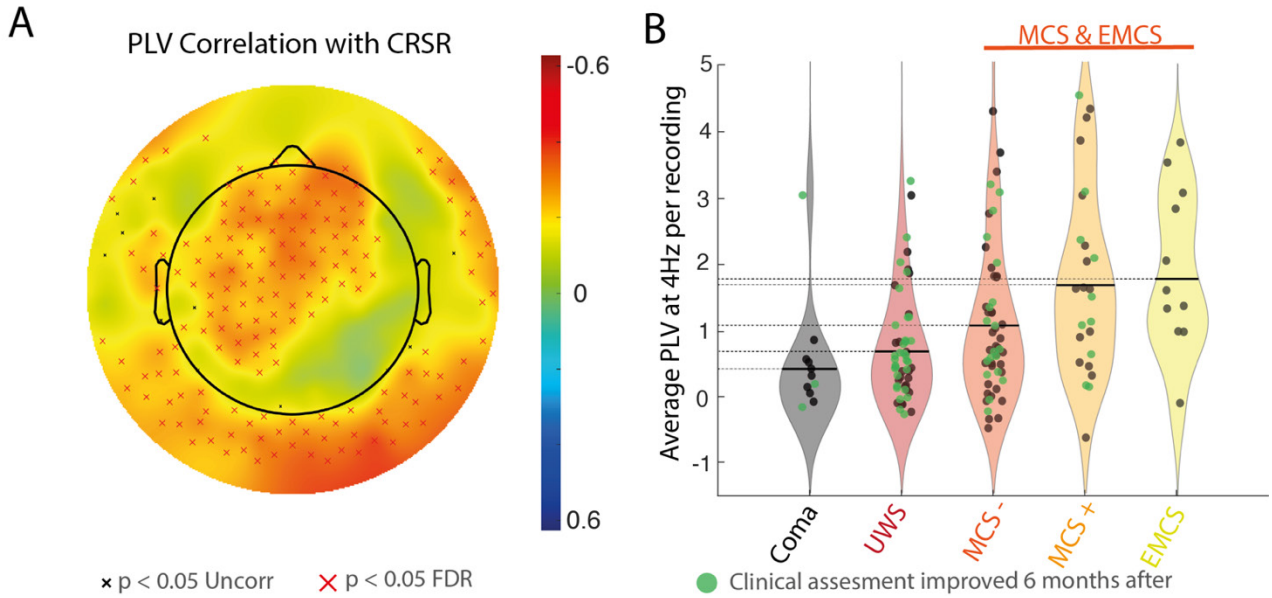


***Figure 1.4 : Syllabic rate correlation with CRS-R**. **A)** For each electrode, we correlated the PLV measure at the syllabic rate (average across random and structured condition) with patients' CRS-R measured just before the recording. **B)** For better visualization of the effect, we report here the distribution of the average PLV at 4Hz across electrodes for each recording and patients' status. The average neural entrainment is modulated by participant's coma depth but the inter-recording variance within each diagnostic group stays higher than the variance explained by the coma depth. Note that this figure is not independent from the analysis done in Panel A and so we did not perform statistical analysis on this. This is just presented for a better visualization and estimation of the inter-recording variance. Recordings associated with an improved clinical assessment 6 months later are displayed in green. We found not systematic relation between PLV at the syllabic rate and probability of improved clinical condition 6 months later.*

thus, the modulation seen here can only be due to the discovery of the word structure and not to the perception of the syllables. The evoked activity by each syllable and word superposes as there is no pause between syllables and words. The shape of this event-

related activity is complex, and the Fourier transform decomposes it into a set of sinusoids with different power depending on the ERP shape. A significant response at the first harmonic argues for a rhythmic response that vanishes faster than the word length, such as a larger response words' first syllable. In contrast, an activity drooling over the following word (e.g. integration of the three syllables) would be more visible at the fundamental frequency ***(Zhou et al., 2016a)***. ERP shapes can explain the different sensitivity of the two measures observed in the above analyses. Further experiments are needed to investigate whether this difference reveals a different encoding of the word in memory. For instance, sleeping neonates segment a tri-syllabic non-word stream but only remember the first syllable of the words, contrary to adults who memorize the entire word ***(Fló et al., 2022a)***.

Correlation between the level of residual consciousness and statistical learning

A significant correlation was observed between CRS-R score and the first harmonic of the word rate in DOC patients, revealing that even though this statistical learning task does not require attention, deep DOC penalizes learning. We thus tried to separate whether impaired learning is related to a drop in statistical computations themselves or rather to a deficit in auditory perception related either to a degraded auditory/phonetic encoding or to a cortical activity badly time-locked to the stimulus. Therefore, we used the 4Hz entrainment as a metric for the auditory low-level processing quality as a proxy of the recording quality which is sometimes impaired by the electrically noisy environment of the hospital wards. We linearly regressed it to CRS-R score. We then used the residuals of this regression in the correlation analysis with the word rate and its harmonic (fig 1.3). Despite this stringent control, a significant correlation remained between CRS-R and the first harmonic only for the structured stream as also a significant difference between the structured and random conditions in this analysis. Thus, statistical learning abilities are affected by the level of residual consciousness, even when the brain responds to the

syllabic tokens. It remains possible that the brain first recovers a syllabic entrainment based on the vocalic nucleus of the CV syllable but without coding the exact phonemes, impairing statistical learning. This should be further explored with musical tones to simplify the identification of the tokens in the stream.

Is there a clinical interest for this type of paradigm ?

Despite the significant result described above, we do not believe that segmentation metrics are usable as a standalone clinical tool, but it could be a relevant measure in a battery of tests. Indeed, the CRS-R effect size observed was smaller than the inter-individual variance and not highly significant. By contrast, the neural entrainment at the syllabic rate was much more informative. Indeed, the correlations with CRS-R were greatly and highly significant and features of the auditory ERPs have been shown to be usable ***(Strauss et al., 2015)***. In our data-set, many showed a significant correlation between the auditory 4Hz neural entrainment and CRS-R measures. Steady-State measure is a more robust and time-economic way to elicit brain responses than isolated ERP, and neural entrainment is a more robust way, to look at the same thing. This is confirmed by the typical auditory topography of the effect size of the correlation (see fig 1.4 A). Further research might be useful to better characterize which frequencies to be entrained are the most sensitive and which electrodes are the most informative in clinical application.

Conclusion

We discovered small but significant markers of preserved statistical learning and word segmentation in a subset of DOC patients using neural entrainment measurements, confirming our hypothesis that attention is not required for statistical learning. Together with the studies of efficient statistical learning abilities in sleeping neonates, it shows that this process is very automatic and constantly scanning the auditory environment even in

condition of disturbed conscious attention. Moreover, we showed that those metrics of statistical learning were significantly correlated with diagnostic metrics such as CRS-R. However, the variability between recordings largely exceeds the part of the variance explained by behavioral ratings, and the stability of the effect is still low despite a very large number of recordings, indicating that segmentation markers in statistical learning paradigms are probably not the best suited methods for clinical assessment of all disorders of consciousness patients. Finally, we showed that basic auditory processing metric such as neural entrainment at the syllabic rate is an interesting predictor of CRS-R score, and propose that neural entrainment robustness could be of interest for better characterization of auditory ERP modification in DOC patients.

Supplementary material:

In the main text, we presented all analyses with PLV as a measure of neural entrainment. We replicated here all analyses with Power as a metric for neural entrainment estimation. The results are mainly similar to those presented in the main text with PLV.

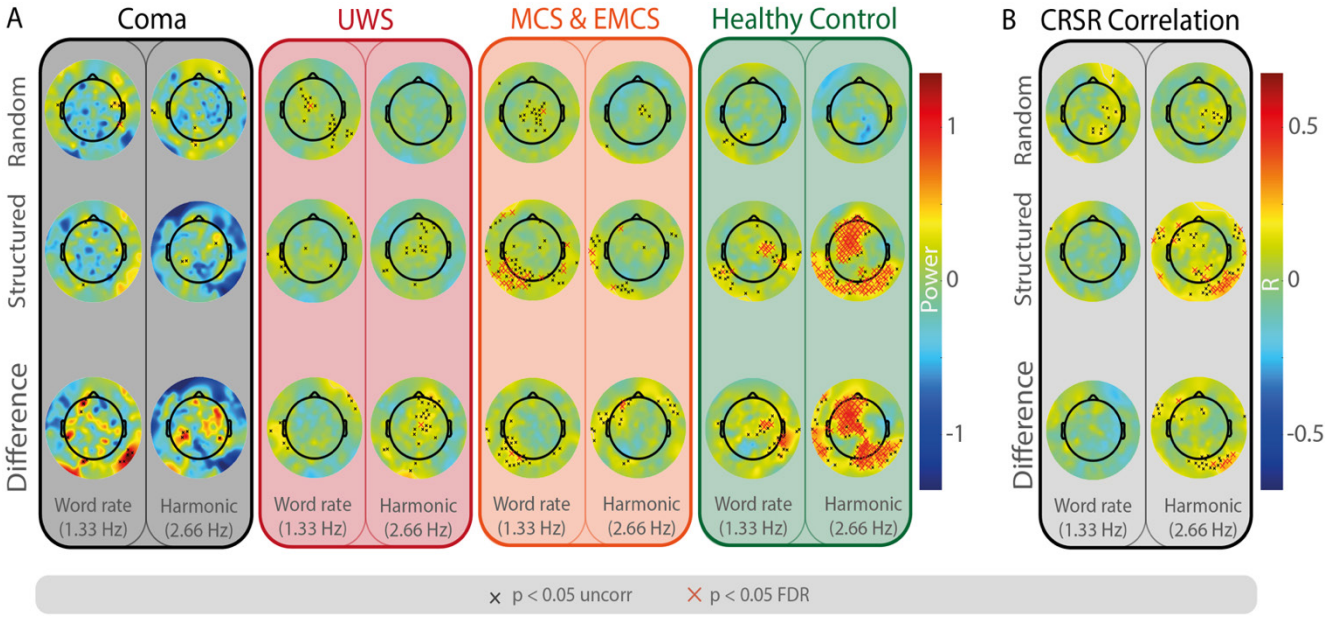


Figure 1.5 *: Replication of results from fig 1.2 with power instead of PVL.* ***A)*** *Power of each electrode at the word rate (1.33Hz) and its first harmonic (2.66Hz) during the random and structured streams. The bottom line present the contrast structured > random. Dot represent electrode with p< 0.05 before multiple comparison correction. The red crosses indicate electrodes significant after FDR correction.* ***(B)*** *Correlation of the Power at word rate and harmonic with Comatose Recovery Scale Revised (CRS-R). The harmonic of the word rate (2.66Hz) significantly correlates with this clinical assessment during the structured stream only.*

Figure 1.5 corresponds to fig 1.2 and presents power analyses. The results were similar with those presented o, the main text with PLV.

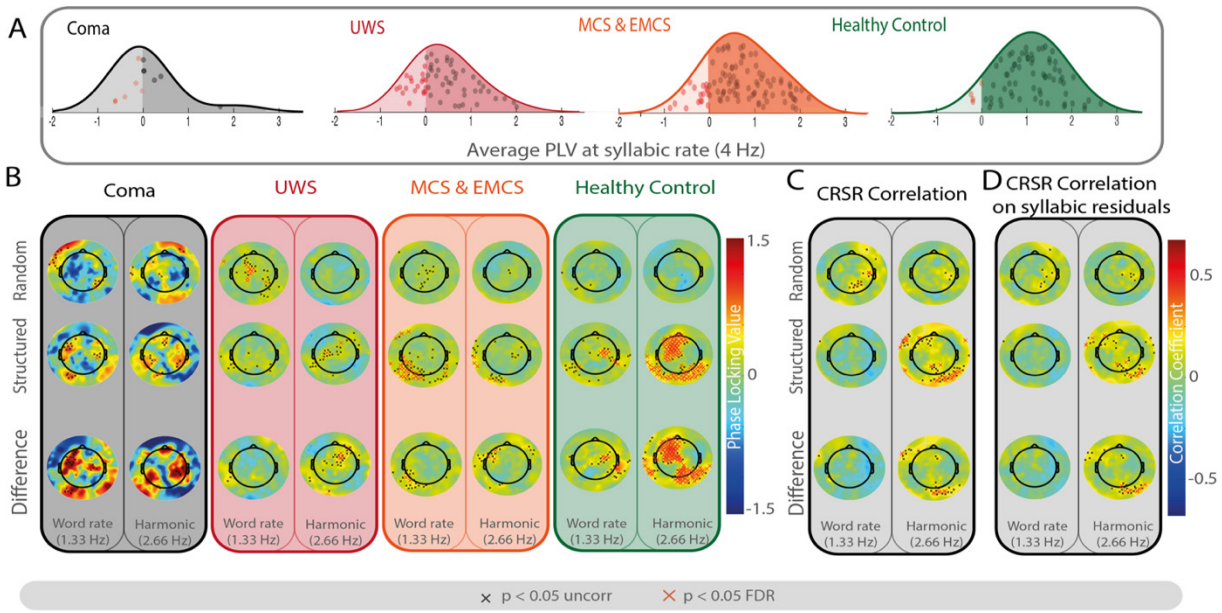


Figure 1.6 : (A) *Average syllabic rate Power measure for each recording. Syllabic rate is a good metric for preserved auditory perception. For the following analysis, we only kept the recordings with positive average syllabic rate as we cannot be sure that the other participants even heard the stimuli. Red dots represents recording with negative syllabic rate and were then excluded from further analysis.* ***(B)*** *Power of each electrode at the word rate (1.33Hz) and its first harmonic (2.66Hz) during the random and structured streams for recording of UWs, MCS or Healthy subjects restricted to recordings with positive average Power at syllabic rate (see Panel A). The bottom line is the contrast structured > random to accesses possible increase of Power during the structured compared to the random stream sequences. Dot represent electrode with p< 0.05 before multiple comparison correction. The red crosses indicate electrodes significant after FDR correction.* ***(C)*** *Correlation of the Power at word rate and harmonic with Comatose Recovery Scale Revised (CRS-R) restricted to recordings with positive average Power at syllabic rate (see Panel A). The harmonic of the word rate (2.66Hz) significantly correlates with this clinical assessment during the structured stream only.* ***(D)*** *To account for the variation of syllabic rate, we first computed a regression between word rate (resp harmonic) and the syllabic rate Power per electrode. The residuals of this correlation are then used to correlate with CRS-R. The results are very comparable to the original correlation without syllabic rate regressed out.*

Figure 1.6 corresponds to fig 1.3 and presents power analyses. The results were similar with those presented in the main text with PLV.

Chapter 2 : Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates

This work has already been published in *Developmental Science* under the reference :

Benjamin, L., Fló, A., Palu, M., Naik, S., Melloni, L., & Dehane-Lambertz, G. (2022). **Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates**. *Developmental Science*, e13300. https://doi.org/10.1111/desc.13300

Abstract :

Since speech is a continuous stream with no systematic boundaries between words, how do pre-verbal infants manage to discover words? A proposed solution is that they might use the transitional probability between adjacent syllables, which drops at word boundaries. Here, we tested the limits of this mechanism by increasing the size of the word-unit to four syllables, and its automaticity by testing asleep neonates. Using markers of statistical learning in neonates' EEG, compared to adult behavioral performances in the same task, we confirmed that statistical learning is automatic enough to be efficient even in sleeping neonates. We also revealed that: (1) Successfully tracking transition probabilities (TP) in a sequence is not sufficient to segment it. (2) Prosodic cues, as subtle as subliminal pauses, enable to recover words segmenting capacities. (3) Adults' and neonates' capacities to segment streams seem remarkably similar despite the difference of maturation and expertise. Finally, we observed that learning increased the overall similarity of neural responses across infants during exposure to the stream, providing a novel neural marker to monitor learning. Thus, from birth, infants are equipped with adult-like tools, allowing them to extract small coherent word-like units from auditory streams, based on the combination of statistical analyses and auditory parsing cues.

Introduction

One of the main challenges encountered by infants to learn their native language and construct their lexicon is that words are rarely produced in isolation. Instead, words are embedded in sentences with no systematic silence or clear acoustic boundaries between them. Subtle acoustical markers such as the lengthening of the last syllable, pitch change, slowing-down of the syllabic rate and less coarticulation between syllables can signal words ending. But adults rely mainly on lexical knowledge and sentential context to retrieve words in their native language ***(Mattys et al., 2005)*** and in an unknown language, they have great difficulty in correctly segmenting sentences into words ***(Wakefield et al., 1974)***. However, when the experimental task is simplified by using an artificial stream of concatenated words, these acoustical cues can be used to discover the possible words as shown by their above chance accuracy in forced-choice tasks ***(Bagou and Frauenfelder, 2018)***. Similarly, neonates can detect these subtle variations in a binary situation in which they have to discriminate pseudo-words constituted of syllables either coming from inside a word (e.g. /mati/ from *mathe**mati**cian*) or from two successive words (e.g. */mati/* from *pyja**ma ti**ssé*) ***(Christophe et al., 1994)***. However, the relative weights of these markers vary across languages ***(Ordin et al., 2017)*** and within a language (i.e. they depend on the position of the word in the sentence and interact with other prosodic features such as lexical stress). Therefore, the robustness of these word-boundary cues is commonly estimated as insufficient for infants to segment natural speech in successive word units.

A second mechanism, based on the analysis of the transitions between syllables, has thus been proposed. Within a word, syllables have a fixed order, whereas any syllable can follow the last syllable of a word. Thus, a word boundary corresponds to

a drop in the transition probabilities (TP) between consecutive syllables. To prove that the concept could apply for word learning in infancy, Saffran, Aslin and Newport ***(1996a)*** used a mini language of 4 tri-syllabic words and tested 8-month-old infants who listened for 3 mins to a continuous stream in which these words were concatenated with a flat intonation. The authors reported that infants were subsequently able to distinguish two different lists of isolated tri-syllables pseudo-words: one corresponding to the words (i.e. ABC : TP were equal to 1 between each syllable) and the other to PartWords formed by the two last syllables of a word and the first syllable of the next word for example (i.e. BCA': TP were equal to 1 and 0.33). This result has been replicated multiple times ***(Black and Bergmann, 2017)*** and extended to non-linguistic stimuli ***(Saffran et al., 1999; Schön et al., 2008)*** and to the visual domain ***(Fiser and Aslin, 2002).*** Sensitivity to statistics in sequences is also observed in animals ***(Hauser et al., 2001; James et al., 2020; Toro and Trobalón, 2005)*** indicating that the capacity of extracting transitional probabilities between subsequent elements is a robust general mechanism. Additionally, it has been reported in asleep neonates ***(Fló et al., 2022a, 2019; Teinonen et al., 2009)*** and to some extent, in inattentive adults ***(Batterink and Choi, 2021; Benjamin et al., 2021; Fernandes et al., 2010; Toro et al., 2005).*** Yet the limits of this mechanism and the influence of development and expertise on the performances are still poorly known.

One of the limitations of statistical learning, already reported in the literature, is its interaction with alternative segmentation cues ***(Black and Bergmann, 2017)***, especially its embedding in prosodic units. A word cannot straddle a prosodic boundary. Therefore, even if two syllables are always presented in succession, they are not attributed to the same word if a prosodic boundary separates them. This property is observed in adults ***(Shukla et al., 2007)*** and in 5 to 8-month-old babies ***(Johnson and Tyler, 2010; Shukla et al., 2011)***. This result should not be surprising given the importance of prosody to structure the speech signal. A hierarchy of prosodic units ***(Nespor and Vogel,***

2006) roughly parallel to the syntactic tree is used to improve speech comprehension in adults and to favor language acquisition in infants. For example, even at two months of age, infants memorize better the phonetic content of a sentence with a well-formed prosodic contour relative to a word list *(Mandel et al., 1994)*. This advantage can be explained because statistical computations are limited to a few elements within the prosodic unit, relieving memory. Prosodic units also provide perceptual anchors, which help infants note the reproducible location of certain words at their edges, such as articles or proper name. Finally, the higher frequency of function words relative to content words has also been proposed as anchors favoring word discovery *(Hochmann et al., 2010).* To succeed in the complex task of constructing a lexicon from natural speech, infants have a toolbox of procedures at their disposal, whose relative contributions are currently underspecified.

Here we investigated another putative limitation of the statistical learning mechanism: the size of the words that can be learned. In fact, most, if not all, studies in infants have used tri-syllabic words. Is it due to particular experimental choices? Or is there a hard limit to segmentation based on statistical computation, especially in immature infants? If the latter, can subtle prosodic cues rescue segmentation and word learning, allowing memory processes to deploy *(Fló et al., 2022a)*? To investigate these questions, we created a first artificial stream consisting of four quadri-syllabic words, pseudo-randomly concatenated without any prosodic cue, and a second one strictly identical to the first one but with a 25ms pause between each word, every four syllables. In previous studies of artificial streams with this short pause, adults reported not noticing it and were at chance when they had to choose which of the two streams had pauses *(Peña et al., 2002)*. Nevertheless, pauses significantly improved their performances *(Buiatti et al., 2009; Peña et al., 2002)*. The pause was probably perceived as a vowel lengthening, a universal ending cue for words and musical segments *(Tyler and Cutler, 2009)*. In adults, final syllable lengthening improved

tri-syllabic word segmentation ***(Saffran et al., 1996b)***. The authors proposed a putative hierarchy in using these cues, i.e., infants first rely on transitional probabilities, then notice that syllable lengthening coincides with a word ending to finally learn this new cue. Yet, this hypothesis remains untested because the relative contribution of transitional probabilities and this subtle prosodic cue was not assessed in this study. We used high-density EEG (128 channels) to evaluate segmentation processes in neonates. EEG allows not only to observe different responses to test-words after learning, but also to track learning while neonates are listening to the artificial stream. As the syllables have exactly the same length, their perception creates a regular evoked response, which is observed as a power increase at the frequency of the syllable presentation. If the syllables are perceived grouped in a quadri-syllabic word, the power should increase at ¼ of the syllable frequency (1/3 if tri-syllabic words are used). Such a power increase at the word frequency has indeed been reported for tri-syllabic words in adults ***(Batterink and Choi, 2021, 2021; Buiatti et al., 2009)*** in 6-8-month-old infants and in neonates ***(Choi et al., 2020; Fló et al., 2022a; Kabdebon et al., 2015)***. We also presented neonates with "random" streams constituted of pseudo-randomly concatenated syllables, with and without a pause every four syllables, to control whether the pause itself was sufficient to induce a 4-syllable-rhythm. In adults, inserting such a pause in a random stream did not produce any increase of power at the pause frequency ***(Buiatti et al., 2009)***. Therefore, in the case of successful segmentation, we expected a significant power increase at the word frequency in the structured streams relative to the random streams. No change, or perhaps a decrease, was expected at the syllabic rate, in line with previous reports in adults in which perceiving the word induced a decrease of the entrainment at the syllabic rate ***(Batterink and Choi, 2021; Benjamin et al., 2021)***.

After the learning phase, three types of test-words were presented in isolation: Words, PartWords, and ShuffleWords (fig 2.1). Successful word segmentation is

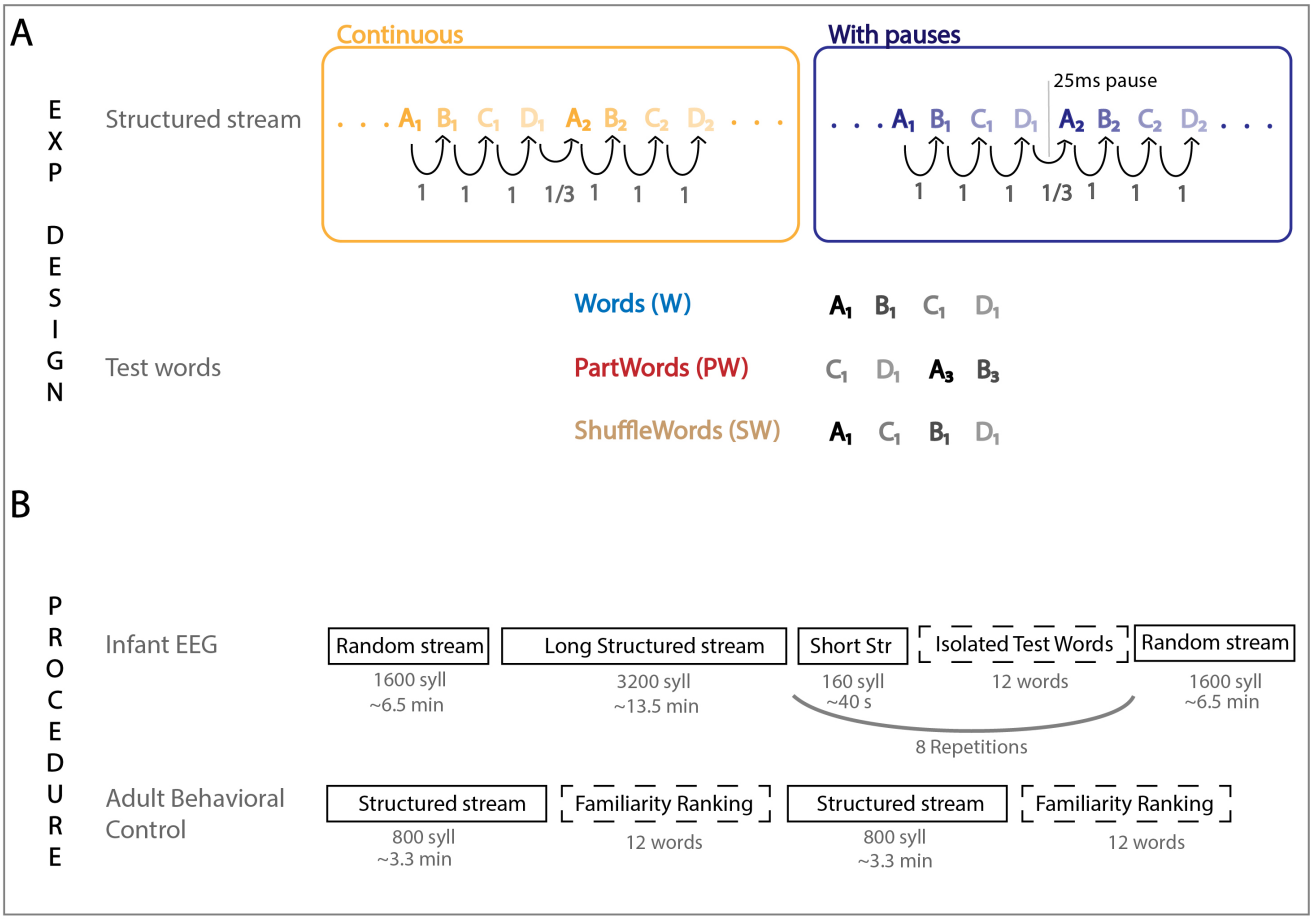
commonly revealed by a significant difference between the measured response to Words and PartWords. In Words, all transitional probabilities between syllables equal 1, while in PartWords (straddling two words), a drop in transitional probabilities indicates an ill-formed word. In ShuffleWords, the two middle syllables of a Word were inverted, violating local position. Thus, while all the transitional probabilities were zero, all syllables were always heard in close proximity during the learning stream. This temporal proximity might induce memory errors and a wrong recognition of ShuffleWords as possible words. Indeed, in longer words of six-syllables, neonates are not able to detect a shuffle of the middle syllables, whereas they detect a shuffle of the edges syllables ***(Ferry et al., 2016)***.

Thus, our experimental design provides several markers of transitional probability computation and word segmentation that might be differently associated, opening the possibility to disentangle several steps or hypotheses of this classical learning task. *(H1: TP computation)* If infants computed TP and memorized the TP matrix, they should reject Words from Shuffle- Words (1+1+1 vs. 0+0+0) but marginally Words from PartWords (1+1+1 vs. 1+0.33+1). *(H2: segmentation)* Stream segmentation should create an increase of neural entrainment at the word frequency. *(H3: complete memorization of the word)* should create a difference between words on one side and PartWords and ShuffleWords on the other side. *(H4: memory errors)* If Words are segmented and swap errors occur, ShuffleWords should not differ from Words due to the temporal proximity of the syllables belonging to the same Word. *(H5: first syllable memorization only)*. This hypothesis could explain why words are preferred over PartWords in many statistical learning studies. As the typical trisyllabic paradigm compares Words (ABC) to PartWords (BCA'), the difference observed could result from the encoding of the first syllable only (A vs B). In a recent study with tri-syllabic words, we indeed observed an ERP difference

between words and PartWords for the first syllable, whereas no difference was recorded when the last syllable was incorrect ***(Fló et al., 2022a)***.

Finally, comparing the two groups of neonates, one listening to the stream without pauses (continuous group) and the other to the stream with pauses (with pauses group), should clarify the relative contribution of auditory parsing cues and transitional probabilities to word segmentation at that age. This comparison should disentangle whether pauses rescue segmentation, subsequently allowing the computation of transitional probabilities on smaller segments, or whether the computation of transition probabilities is done independently of the segmentation process.

Neonates are two-decades far from a mature state in terms of linguistic abilities but also in terms of memory capacities. To our knowledge, no adult equivalent of the paradigm proposed here was available. Thus, we collected adult behavioral data as a mature model of the mechanisms we explored in neonates. We adapted the paradigm to collect behavioral responses on a web-based procedure and shortened the habituation to avoid over-learning already reported in similar experiments in adults ***(Peña et al., 2002)*** (see fig 2.1). Despite different procedures and learning indicators, the results were surprisingly congruent with those obtained in infants, especially showing comparable limitations.



syllabic words (called ABCD) then presented with 3 types of isolated test words. ***B)*** *Experimental procedure: Neonates were tested asleep using high-density EEG (128 channels) while they were presented with random stream, structured stream, and isolated words. Short Str: short structured streams were presented to the neonates to maintain learning. Adults were tested on a web platform. After familiarization with the structured stream, adults were asked to rank the familiarity on a scale (1 to 6). To avoid a bias to quadri-syllabic words, they were also presented with foils corresponding to three other types of bi-syllabic test words (see methods for more details).*

Results

Adults

Two groups of adults were tested online on a web platform (n=43). After having listened to the continuous stream, or to the stream with 25ms subliminal pauses depending on the group, the participants had to judge the familiarity of three types of words (Words, Part-Words and Shuffle-Words). Exposure (3.3mn) and test were performed two times and results of the two tests were aggregated (see method and fig 2.10 for results in each test). We analyzed the responses by items in a linear mixed-effects model in each group with FDR correction. For the stream without subliminal pause at the end of the word, Words and Part-Words were similarly rated ($p=0.26$) and estimated more familiar than the ShuffleWords (W vs. SW $p<0.001$, PW vs. SW $p<0.01$). When subliminal pauses were added at the end of the words in the stream, all types of words were ranked differently (all $p<0.001$) with the following order: Words were judged more familiar than PartWords themselves more familiar than ShuffleWords (fig 2.1 A&B). To better visualize the difference in segmentation performances between the two groups, we calculated the difference in mean familiarity ranking given by each participant to Words and PartWords, and performed an unpaired unidirectional, t-test $t(41)=2.3$, $p=0.013$. The segmentation effect, seen as a positive value on fig 2.2 C, was larger when subliminal pauses were present (fig 2.2 C).

Thus, adults were able to distinguish Words from PartWords, indicating that they had correctly segmented the stream only if helped by a subliminal acoustic cue. Yet even when there was no pause, they rejected ShuffleWords because of null transitional probabilities

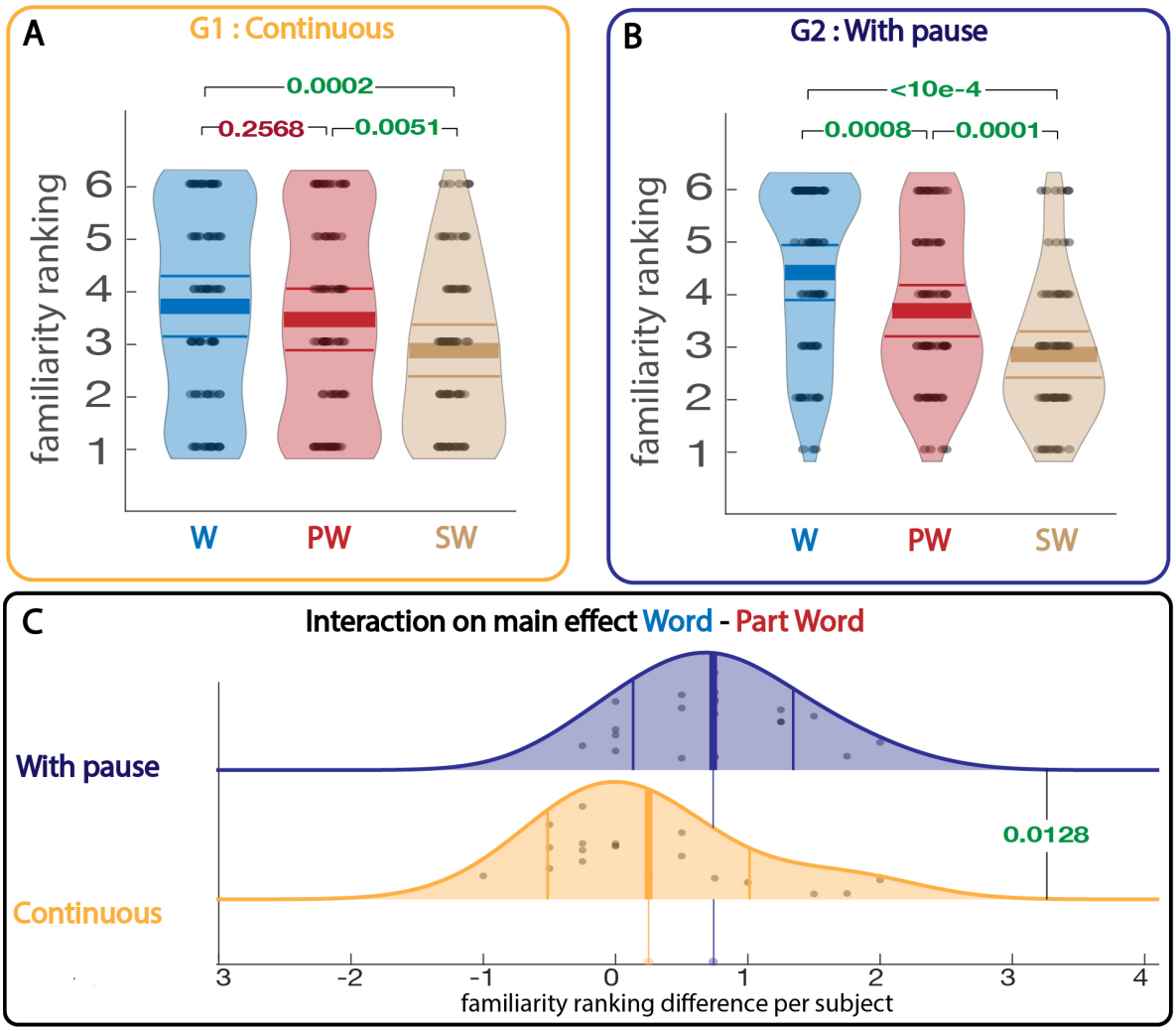


Figure 2.2. A,B) *Results of the familiarity ranking tests in adults' subjects for each item Both test sessions were aggregated. Results for each session are presented in fig S5.* ***C)*** *Interaction at the subject level between both groups on the main effect of se*

Infant EEG Experiment

EEG was recorded in two groups of healthy full-term neonates (n=52 after rejection procedure, see methods) while they were listening to streams without pauses for the first group and with 25 ms pauses every four syllables for the second group. For

each group, neonates were exposed first to ~7 minutes of a ”random” stream in which syllables were randomly concatenated with a flat TP of 0.33, followed by ~13.5 minutes of the word-structured stream. After the exposure learning phase, a test phase followed in which they were exposed to isolated quadri-syllables sequences (Words, PartWords, and ShuffleWords). To avoid interference with learning in the testing phase and to reinforce the learning of the structured materials, 40s-short segments of the structured stream were presented every 12 words during this phase. Finally, another ~7 minutes of the random stream was presented again after this testing phase. The division of the random stream into two periods was done to avoid a time confound in the comparison between random and structured streams. We used a longer exposure time than usual statistical learning experiments in order to perform pattern similarity analyses as done by ***(Henin et al., 2021).***

Neural entrainement

As described in other studies on neural entrainment ***(Fló et al., 2022a; Hochmann et al., 2010; Kabdebon et al., 2015)***, there was a significant increase in power and Phase Locking Value (PLV) at the frequency of the syllable’s presentation compared to neighboring frequencies in both groups (continuous and with pauses) and stream types (random and structured) (all ps<0.05 FDR corrected). No interaction was observed between groups and streams indicating a similar signal-to-noise ratio and comparable experimental conditions in the two groups (all ps>0.05 FDR corrected). The power analysis for each condition and group is presented in the supplementary material.

Stream segmentation should be revealed by a significant increase of power and/or phase locking value at the frequency of the words (i.e., ¼ of the syllabic frequency) relative to the random stream. In the first group (continuous stream), we failed to find this result, contrary to the second group (stream with pause), in whom a significant increase of both power and PLV was observed in several electrodes.

Finally, the interaction between groups and streams was significant for both power and PLV on several electrodes (see fig 2.3 all ps<0.05 cluster corrected, see SI).

It has been described that the power at the syllabic rate decreased when adults segment the stream ***(Buiatti et al., 2009)***. However, we did not find any modulation of the power or PLV at the syllabic frequency in the structured stream compared to the random one. We also performed a time-course analysis of the neural entrainment at the frequencies of interest over sliding time windows of 2 minutes with a 1.5 s step, similarly to Fló et al ***(2022a)***. We observed no change at the word frequency along time for either group (fig 2.8 – supplementary materials). The poor signal/noise ratio at these low frequencies might explain the poor sensitivity of this analysis.

Between-subjects correlation analysis

Because the exact same stream was used in each participant, we were able to analyze whether learning increased the global neural synchronization between neonates beyond neural entrainment. To do so, we tested whether the correlation between participants increased over time more when they listened to the structured stream. We then compared at each time the topography of each subject with the average of the other subjects' topographies at that time. We observed a progressive increase of the mean correlation between subjects in neural activity only in the second group with pauses (fig 2.4 A Left). Indeed, the increase was higher for the second group (with pauses) than the first one (continuous) (cluster corrected p<0.01, time [88-820]s fig 2.4 A). During the random streams, the correlations were flat relative to baseline in both groups (fig 2.4 B Left). To confirm this effect, we computed a linear regression of the variation of subject correlation with the group with time at the subject level during each stream and compared the slopes in the two groups. Only when neonates listened to the structured stream with pauses (second

group), the slope was significantly positive and significantly greater than the same measure in the continuous group (fig 2.4 A Right). No difference was observed during the random streams (fig 2.4 B Right).

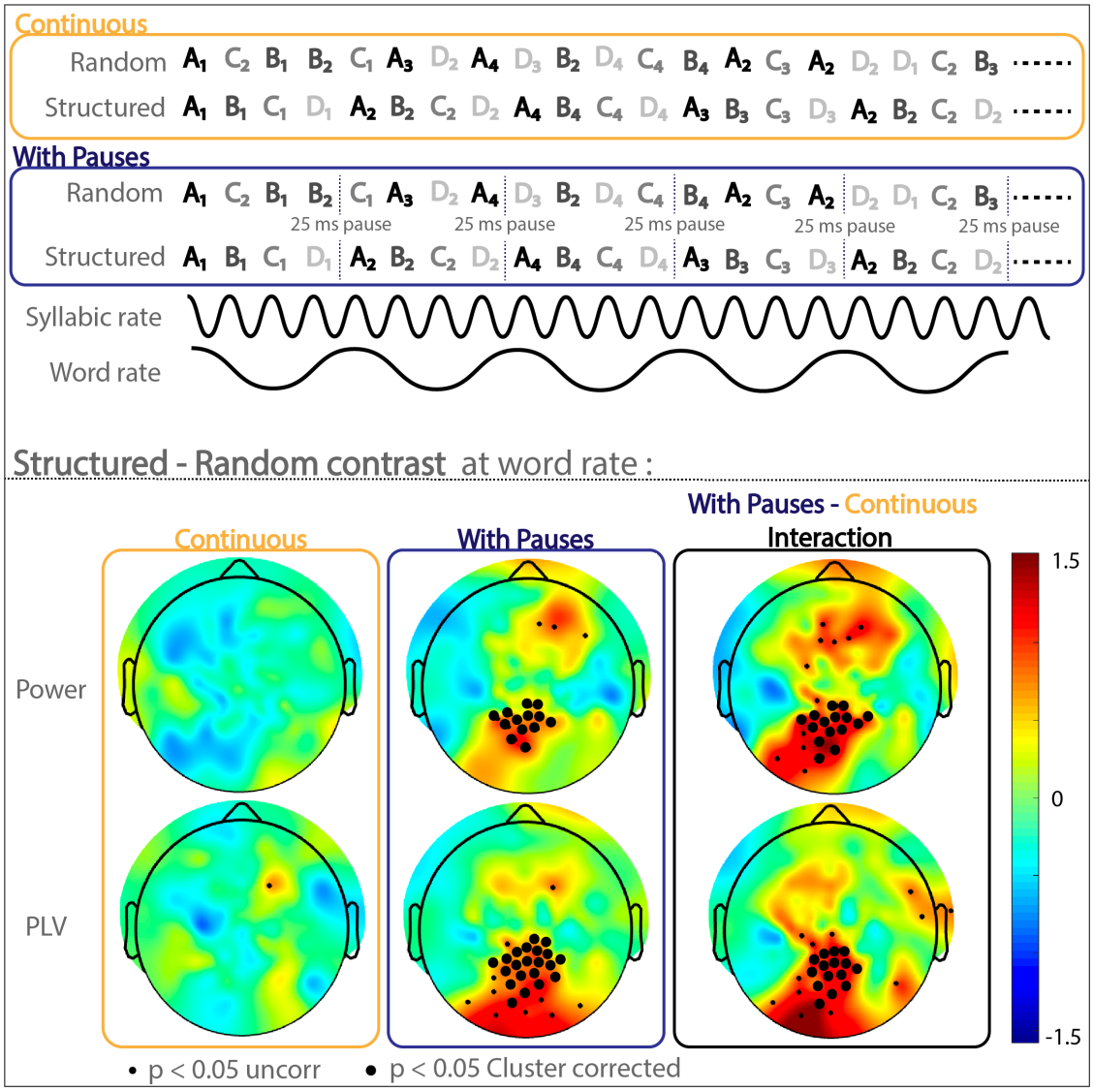


Figure 2.3. *Neural entrainment analyses at the word frequency in the structured stream minus the random stream (Power and phase locking value (PLV)) in the two groups of neonates (continuous and with pauses). Top rows: the presentation of stimuli with a fixed duration evoked a reproducible time-locked neural response that can be recovered as a neural oscillation at the frequency of stimulation. If infants segment the structure streams based on the quadri-syllabic words, an increase at the word frequency should be observed relative to the random stream. It is what is seen in the second group of neonates (with pauses) who listened to the streams with subliminal pauses. The acoustical effect of pauses was controlled by also adding a pause every four syllables in the random stream in this group. The last column shows the interaction between groups and frequencies. Dots locate the electrodes showing a significant result at p<0.05 uncorrected, and larger dots after cluster correction (cluster p < 0.05). Power during structured and random streams are presented separately in each group in fig S2.*

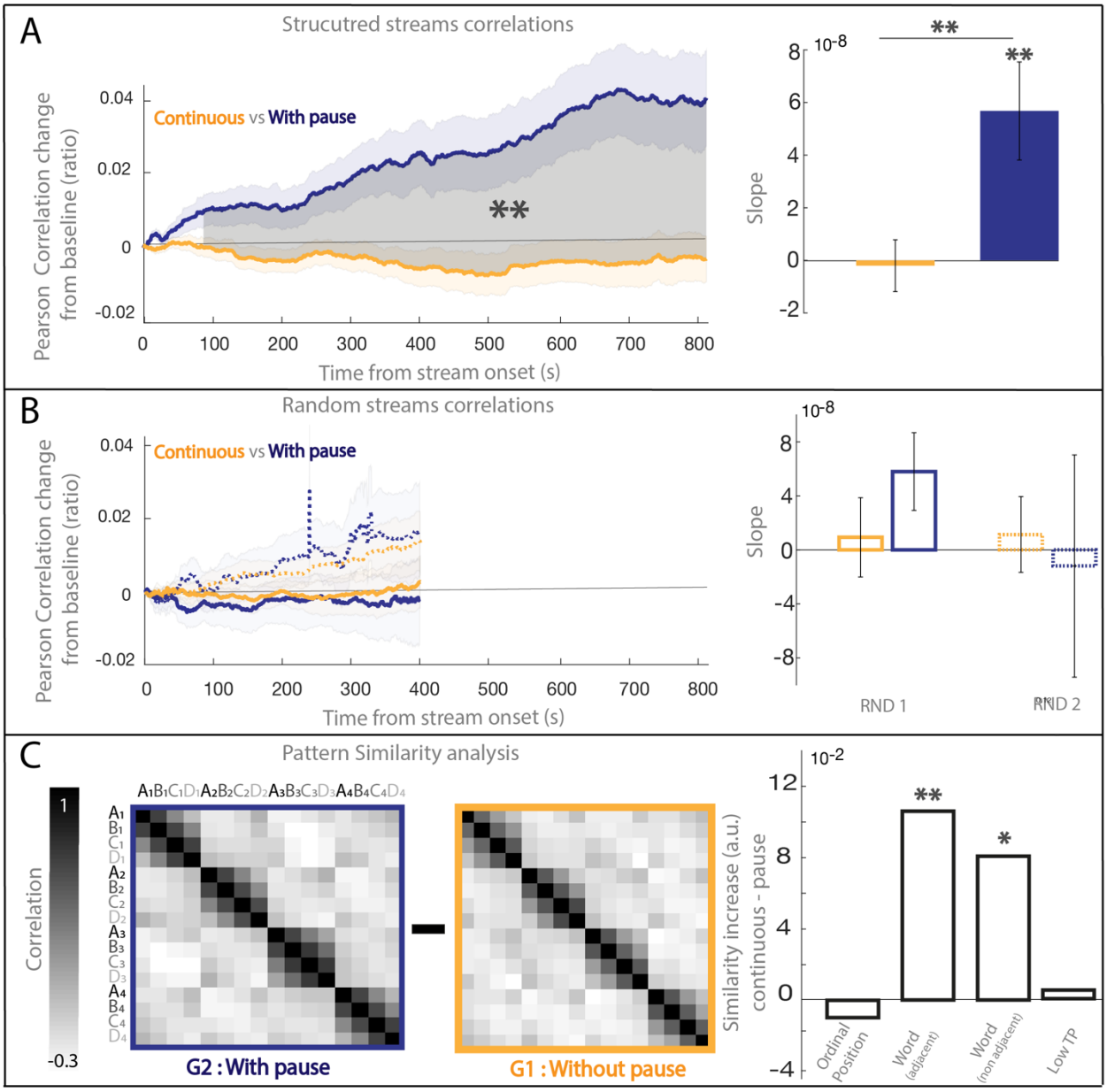


Figure 2.4. Correlation Analysis A) *Comparison of the correlation between neonates in the two groups during the structured stream. Left: Evolution of the correlation across neonates with time. Right: comparison of the slopes of the linear regression with time in each group (orange: continuous, blue: with pauses)* ***B)*** *Similar analysis for the first (plain lines) and second random (dotted lines) streams in both groups. RND=Random.* ***C)*** *Pattern Similarity analysis: We computed the increase of pattern similarity between the EEG response to each syllable in the two groups during the structured stream. The similarity significantly increased for syllables belonging to the same word (for adjacent pairs: AB, BC, CD, and non-adjacent pairs AC, BD and AD).*

Syllable pattern similarity analysis

In a recent paper, Henin et al. ***(2021)*** proposed that pattern similarity between syllables can vary with learning in a similar task. More specifically, using electrocorticography in epileptic adult patients who listened to a structured stream composed of the concatenation of 4 trisyllabic words, they computed different patterns of similarity between the 12 syllables. They took advantage of the high spatial resolution of electrocorticography and observed different clusters of electrodes sensitive either to TP transitions (low vs high TP), the ordinal position (1^{st} vs. 2^{nd} vs. 3^{rd} syllable), or the word identity (word 1 vs. word 2 vs. word 3 vs. word 4) in different brain areas. We computed a similar analysis on the responses to the syllables during the structured stream and showed that the similarity pattern for syllables belonging to the same words was significantly increased in the pause group compared to the continuous group ($p=0.012$). However, we failed to find an increase in pattern similarity for low TP (DA') and ordinal position (AA', BB', CC', and DD') between the two groups. To investigate if the significant increase in similarity for syllables belonging to the same word was only due to an increase in high TP pairs or all pairs belonging to the same word, we separated the word condition in two sub-conditions: Consecutive (AB, BC, CD) and non-Adjacent (AC, BD & AD). Interestingly, Consecutive and non-adjacent pairs showed a significant increase in pattern similarity (both $p<0.05$ FDR corrected) with pauses compared to the continuous group. The differences in pattern similarity between the two groups for each condition are reported in fig 2.4C.

Discussion

In natural speech, many signal-derived cues might assist segmentation ***(Wakefield et al., 1974)***, but none is robustly consistent to be systematically used by infants. Therefore, the computation of the transitions probabilities between syllables has been proposed as a possible solution ***(Saffran et al., 1996a)***. We presented here a stream comprising quadri-syllabic words to investigate the efficiency of this strategy for longer words. Whereas tri-syllabic words are easily extracted from a flat speech stream using TP between syllables in adults ***(Saffran et al., 1996b)***, infants ***(Saffran et al., 1996a)***, and even sleeping neonates ***(Fló et al., 2022a)***, this single cue seemed insufficient here for quadri-syllabic words even in awake vigilant adults, revealing a clear limitation of the statistical learning mechanism in a segmentation task. Whereas the power increase at the syllabic frequency was large during all streams, we did not record a significant neural entrainment at the word frequency in the continuous group contrary to what we obtained in a similar paradigm with tri-syllabic words. Because the "noise" level due to the background neural activity is exponentially growing with low frequencies in EEG data, notably in neonates, the lack of neural entrainment for quadruplets might have been due to a lack of sensitivity of the method at 1Hz, compared to 1.33Hz in the case of triplets. However, the significant word entrainment in the group listening to the stream with pauses and the interaction between both groups confirm that neural entrainment is a sensitive method even at these low frequencies.

The word segmentation failure is also not explained by the higher number of syllables to be memorized (16 syllables here vs. 12 for tri-syllabic words) and the larger word size since the same material, just with the addition of subliminal pauses, rescued the word extraction process. Furthermore, we recorded several other

indicators of learning in the second group who listened to the stream with pauses: First, the increase in power and phase locking-value observed at the word frequency in the structured stream was not related to the mere presence of a pause but to a genuine learning process, as it was not observed for the random stream that also included pauses every four syllables. Second, neural synchrony increased between participants only for the structured stream with pauses, further suggesting that neonates were following a similar learning process constraining their brain state in this condition. Again, this phenomenon was only observed for the structured stream and not for the random stream with pauses. It underscores that it was not a general difference between the two groups of neonates but was related to the learning process engaged when they listened to the structured stream. Third, ERPs to Words and PartWords were significantly different after the stream with pauses, a classical indicator of word segmentation in this type of paradigm. Finally, adults also ranked Words higher than PartWords when they listened to streams with pauses relative to streams without pauses, confirming an undeniable advantage for the former over the latter. All these indicators of successful segmentation were not only lacking when the pause cue was not present, but for all of them, the differences between the two groups were significant in both infants and adults.

Yet, even if participants were not able to segment the words in the continuous structured stream, both adults and neonates rejected ShuffleWords, which contained the exact same syllables as the Words, but in the wrong order. This result reveals that the participants computed Transition Probabilities and were not misled by the temporal proximity of the syllables, but this computation was not sufficient to trigger stream segmentation. Interestingly, attentive adults appeared not better than sleeping neonates in the task: They also failed to segment the stream without the help of acoustical cues. Thus, tracking transition probabilities does not always result in word segmentation.

Word segmentation based on statistical learning is limited by the word size

It was proposed that the computation of the transition probabilities might be used to segment a speech stream, either through boundary markers -a TP drop creates a prediction error, and the surprise allows to memorize the syllable following the drop (i.e. the first syllable of the following word)- or because adjacent events acquired a similar representation. However, neither the local drop of TP nor the temporal proximity within a chunk were sufficient to structure the stream, not even after 13 mins of exposure when the unit size was four syllables (1s long). On the contrary, sleeping neonates perceived a tri-syllabic rhythm in the same circumstances and only after 2 mins of exposure, and memorized the set of possible first syllables ***(Fló et al., 2022a)***. It remains possible that longer exposure to the continuous stream might eventually allow word segmentation. However, compared to the segmentation of tri-syllabic words tested under similar conditions, both neonates and adults had considerable difficulty performing the task with quadri-syllabic words.

The opposite hypothesis is that the neonates might have learned as quickly as in the case of trisyllabic words but that as time passed, this learning faded away because even low probability transitions became familiar. This overlearning effect has been reported in adults ***(Peña et al., 2002)***. The analysis of the neural entrainment along time of exposure was not sensitive to figure out the learning timeline even in the group with pauses, probably because of the very low signal to noise ratio in low frequencies. However, the group difference in the correlation between neonates' recordings increased from around the first minute of exposure showing that the two groups were diverging early on between a learning condition (stream with pauses) and a no-learning condition (continuous stream).

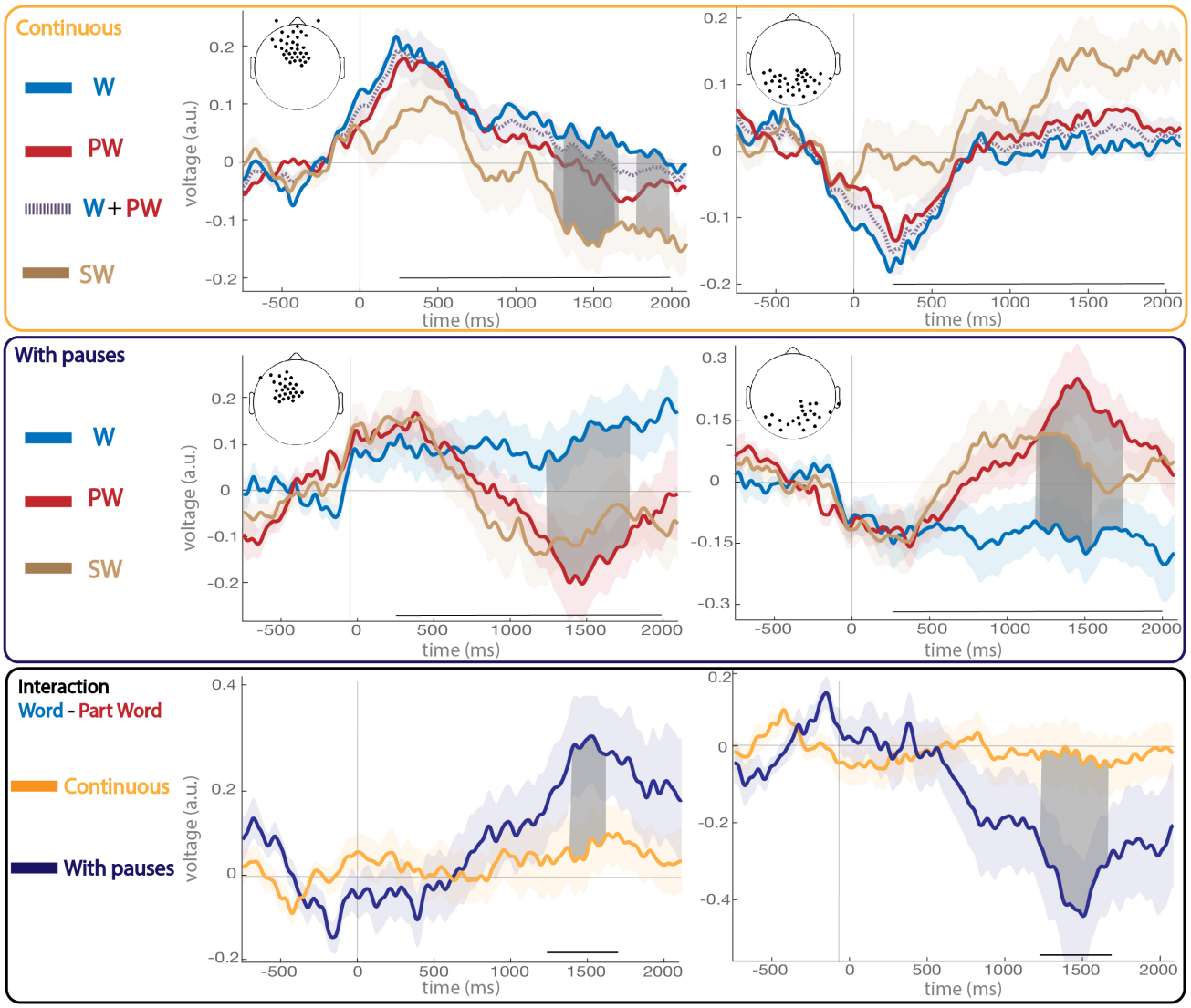


Figure 2.5. *Grand average ERPs to the test-words in both groups and to the word minus part word difference. The ROIs correspond to the two poles of the response to the auditory localizer preceding the test word in each group. Dark grey areas identify significant temporal clusters. Light shaded areas surrounding the thick lines represent the standard error across neonates. W=Words (ABCD), PW=Part Words (CDA'B') and SW=ShuffleWords (ACBD). Gray lines at the bottom of the plots indicate the time windows on which statistics were performed.*

Rescuing segmentation with subliminal pauses

Adding a subliminal pause at the end of the word radically affects the performances at both ages. Although not consciously perceived, pauses act as other word boundary markers (e.g., lengthening of the last syllable, pitch drop) that neonates can perceive *(Christophe et al., 1994)*. Our result demonstrates that such a word boundary marker is not only perceived but is effectively used to segment a stream from birth on, that is, before infants have perceived many isolated words. The use of word-ending cues, at least when it is a pause as here, does not need that infants first learn words as it was postulated *(Saffran et al., 1996b)* but is part of the auditory/linguistic perceptive system. This observation is in agreement with the proposal of a hierarchical framework in weighting the multiple segmentation cues *(Wakefield et al., 1974)* and the subordination of statistical learning to many other cues, such as coarticulation *(Fernandes et al., 2010)*, prosodic contour *(Shukla et al., 2011, 2007)*, and top-down contextual parsing *(Wang et al., 2020)*. However, in adults as exposure lengthens and the absolute frequency of all possible transitions increased, the familiarity advantage for Words relative to Part-Words created by the pauses faded away (Fig 2.10) suggesting that the weight attributed to each parameter might not be strictly hierarchized but dependent on the strength of the evidence provided.

We also observed in infants that similarity between syllables within the word was increased relative to the continuous stream without pauses (fig 2.4 C). Not only similarity between adjacent syllables within a word was stronger in the stream with pauses than without pauses, but similarity also increased between more distant syllables belonging to the same word. We cannot disentangle whether this increase in similarity between syllables in a word induced the segmentation as in a clustering strategy that is opposed to a bracketing strategy in which splitting points are looked for *(Swingley, 2005)*, or the reverse, i.e., because syllables were perceived in the same chunk, their similarity increased.

It is also interesting to note that perceiving the stream at a more complex level of representations increased neural synchrony between the neonates. Whereas the syllabic rate itself, which affects many channels (see fig 2.7 – supplementary information), already creates a strong and similar entrainment across participants, it is not this low-level cue that was predominant in the neural synchrony between neonates but the perception of a higher level of organization of the stream. This synchrony measure probably captures a wider cross-subject convergence beyond neural entrainment at the two frequencies of interest in specific channels. It reveals that neonates' brain states are not purely entrained by the physical features of the stimulation, which remain similar along the stream but also constrained by learning mechanisms that led to more synchronous responses across neonates.

Finally, the performances between the test phase during which isolated quadri-syllabic sequences were presented were also massively affected by the stream condition, suggesting that once segmentation was done, memory encoding was improved. Words and PartWords were indeed only discriminated after the stream with pauses. However, in a similar experimental paradigm but after a stream of concatenated tri-syllabic words, Words were recognized since the first syllable ***(Fló et al., 2022a)***, whereas here, the difference was developing from around 500 ms to become significant only after the end of the word. The lack of first syllable effect was confirmed by the absence of difference between PartWords and ShuffleWords, although the latter started with a correct first syllable. It is also consistent with the lack of similarity increase between both groups within the set of first syllables (fig 2.4 C right), which contrasts with the result reported in adults by Henin et al. ***(2021)***. Thus, contrary to the tri-syllabic stream, the ordinal position of the syllables was not encoded, and the difference between correct and incorrect chunks took longer to develop.

Why a sharp distinction between tri and quadri-syllabic words?

The differences, in terms of neural entrainment during familiarization and ERPs responses during test in infants as the drop of performances in adults, between our two word-segmentation studies ***(Fló et al., 2022a)***, in which we used a similar paradigm except that the word size increased from 3 to 4 syllables, raised interesting questions regarding both word segmentation during the stream and subsequent memory encoding of the word unit.

Although this experiment does not directly test this question, we propose that recovering words in a stream is based on short-term memory (STM). Indeed, if TP between syllables can be locally computed within the auditory cortex, the integration of the successive syllables within a word requires a longer temporal window of integration. The sharp difference between tri- and quadri-syllabic words seems reminiscent of the 4±1 unit limit of the auditory short-term memory ***(Cowan, 2001)*** and suggests that the TP drop leading to word segmentation might only be noticeable when all the elements of a word plus the next syllable are present at once in the STM. Several studies suggest that adults use STM, and more specifically working memory, in such statistical learning tasks. For instance, their performance improves when speech is slowed down, an observation at odds with a decay-time in a purely sensory buffer that should be detrimental as the time between syllables increases, but in favor of maintenance of the successive syllabic items ***(Palmer and Mattys, 2016)***. Performances also drop when participants perform a concurrent two-back task ***(Palmer and Mattys, 2016)*** suggesting competition for general resources. These observations are coherent with the activations reported by Henin et al ***(2021)*** along the dorsal linguistic pathway, and notably in the inferior frontal region. In the neonates, no explicit rehearsal was possible because they were asleep and, in any case, unable at that age to repeat syllables, but even in adults, short-term memory effects may remain implicit ***(Hassin et al., 2009)***. If statistical learning is improved when

adult participants are actively doing the task, the task remains feasible when they are distracted and unaware of the task ***(Fernandes et al., 2007; Palmer and Mattys, 2016)***. The similar drop in performance in neonates and awake linguistically productive adults suggests a structural limitation in the number of items that can be stored in the STM. This limit of 4 in STM has been proposed as explaining several higher order linguistic observations, such as the size of phrasal verbs and idioms predominantly used in spoken languages such as English, the mean length of continuous discourse without pauses ***(Green, 2017)***, and the drop in mutual information scores after four words in many languages ***(Pothos and Juola, 2007)***. It also seems compatible with the observed word length inferior to four syllables in many languages ***(Sigurd and Van De Weijer, 2004; Zipf, 1935)***, suggesting that this chunk size limitation we observe here might be fine-tuned to real language word size. This limitation also reveals that TP computation might not be robust enough to be the proposed general-purpose mechanism for word segmentation in all speech streams without being complemented by other indices.

If neural entrainment during the stream reflects the chunking and word encoding, the ERPs to the isolated chunks in the test phase tested the participants' recognition and familiarity with the different conditions. In the tri-syllabic experiment ***(Fló et al., 2022a)***, neonates during the test-phase were no more sensitive to TP (i.e., no distinctive ERP response for triplets containing a TP=0) and were reacting to an incorrect first syllable. Here, they were rejecting ShuffleWords, thus were still sensitive to TP, but did not react particularly fast to the incorrect first syllable (Word vs. Part-Word), suggesting a more general response to the global familiarity of the word rather than noticing a particular error. In adults, Henin et al. ***(2021)*** confirmed using similarity analyses on ECOG recordings, that the ordinal position of the syllables was encoded. Adults are nevertheless better than neonates, encoding not only which syllables were first but also which were second and which

were last. Here, we tried similar analyses in the neonates' data despite the sparser resolution of EEG. We observed an increase of similarity of the ERPs to the adjacent and non-adjacent syllables belonging to the same word in the stream with pause compared to the continuous stream. However, we found no evidence of an increase of similarity between the words first syllables. Thus, the particular status of the first syllables observed in neonates in the case of tri-syllablic words ***(Fló et al., 2022a)*** had no support in this study when quadri-syllabic words were used. This result might just reflect a lack of power of our analysis, or it might be explained by the difference in perception of the drop of TP in a tri-syllabic word stream. The TP drop, which can induce a surprise following a prediction error, might favor encoding these syllables at a particular position (i.e., the first position of the next word). These results might thus suggest that depending on the segmenting cue, different memory processes are engaged in neonates and that TP computation might favor a more precise encoding of the chunking elements, starting with the first syllable and progressing from one syllable to the next. Such an intriguing hypothesis should be further tested in experiments specifically designed to contrast these two cues and the word-size at this age and also in adults.

Similarity between adults and neonate cognitive abilities

Despite very different measuring methods and attentional state in this set of experiments, the results in neonates and adults pointed to similar successes and failures in terms of TP computation and stream segmentation. This is somehow surprising given the fact that many of the structures that support sequence learning ***(Henin et al., 2021)*** – hippocampus, dorsal linguistic pathway, the superior temporal region– change rapidly in the first year of life; but the classic assumption that immature means poorly functional is increasingly challenged by brain imaging

methods that provide markers of learning in young children. FMRI remains difficult in infants, but some results support the hypothesis of early efficiency despite immaturity. In a recent, paper Ellis et al. ***(2021)*** tested 3 to 24 month-old infants on a statistical learning task in the visual domain with fMRI and reported activation in the hippocampus associated with segmentation. Dehaene-Lambertz et al. ***(2002)*** reported activations in temporal and frontal areas in 3-month-olds listening to speech showing that regions usually reported in adults during statistical learning tasks ***(Henin et al., 2021)*** are, to some extent, already functional in infants.

A major distinction between adults and neonates seems to be the capacity of computing such a task during sleep. Indeed, with both three and quadri-syllabic experiments, we showed that sleeping neonates were able to process and segment the streams, under the correct circumstances. However, recent studies report a learning failure in sleeping adults even for tri-syllabic words ***(Batterink and Zhang, 2022; Farthouat et al., 2018)***, and their learning remained limited to bi-syllabic words, that is, to classical associative learning. Infants might perform better than adults during sleep due to the different organization of sleep-wake cycles. At this age, sleep comprises only two clear stages, quiet (~40% of a sleep cycle at birth) and active sleep (50 of a sleep-cycle at birth) with many micro-arousal periods within and between sleep stages ***(Scher, 2008)*** and near 10% of indeterminate sleep. The short periods of wakefulness are immediately followed by active sleep, which is the equivalent of REM sleep in adults. In adults, learning has been shown to exist during REM ***(Andrillon and Kouider, 2016)*** and also that a task started during wake can continue during REM ***(Andrillon et al., 2016)***, opening the possibilities that neonates might learn and consolidate more efficiently than later, thanks to the closer wake-REM sleep alternations.

Methodological considerations

Together, our results show that behavioral subjective ranking and EEG analyses provide powerful tools to investigate statistical learning and segmenting tasks. There was a neat congruency between the behavioral results in adults and the neural markers observed in neonates. Moreover, EEG data enables the investigation of such questions in preverbal and non-verbal subjects with different levels of attention (e.g., neonates, sleeping subjects, comatose patients). Power and PLV during the stream as well as ERP during isolated test words, were already proposed as reliable neural markers in this task ***(Fló et al., 2022a; Kabdebon et al., 2015)***. However, to our knowledge, between-subject correlation as a function of time had not been shown to capture learning in infants successfully. Our results confirm that despite the noise in infant EEG data, a significant part of the variance cannot be only explained by low-level bottom-up activation to external stimuli but instead by a more sustained learning effect. Although this first attempt might have been still noisy, we might hope that this method could more accurately quantify the average amount of learning of a group over time or even characterize learning at the subject level. One drawback of this method is that, to compare across subjects, all subjects have to be exposed to the exact same stimuli, which presents a risk of confound in the experimental design. Here we minimized this risk by taking two precautions. We first carefully designed and balanced the auditory material on acoustic aspects (see SI). Secondly, we ran two groups with a minimal change (a subliminal pause every four syllables) so that, if any bias persists, it would be the same in both groups and thus cannot explain differences between groups.

We also implemented what we believe to be an improvement for ERP analysis. Before the presentation of isolated words, we presented a short audio click as an auditory localizer. In this way, we were able to extract ROIs for analysis with a data-driven approach instead of literature driven. We performed a cluster based-

permutation analysis on all data against zero during the click presentation to extract the auditory ERP regions of interest (ROI). Moreover, this localizer cluster was representative of the auditory response in this particular group of subjects taking into account non-relevant variations due to 1) Experimental conditions such as the placement of the net on the infant's head which is more variable at this age due to birth-related head deformation, and can introduce between groups differences 2) eventually long-tail effects of the previous trials on the topography that can affect the baseline.

Conclusion

Human neonates display sequence learning abilities even during sleep, based on TP computations and segmenting helped by acoustic/prosodic cues. The similarities with adults' successes and failures were remarkable, revealing early powerful capacities to process speech. A speech stream is not a uniform landscape for infants, but different cues might help them to chunk it into smaller units, opening the possibility to discover the linguistic regularities and the productive properties of speech.

Materials and Methods

Behavioral experiment

Participants

A total of 43 adults were recruited via social media and mailing (21 males, age distribution = [18-25] : 9, [25-40] :16, [40-60] : 17, 60+ : 1]) with no reported auditory issue or language related troubles. They were randomly assigned to one of the streams with the instruction to carefully listen for ~3 minutes to a nonsense language composed of nonsense words that they have to learn because they will have to

answer questions on the words afterward. The learning/test procedure was repeated twice.

The study was coded in javascript using jspsych toolbox ***(de Leeuw, 2015)*** and played audio mp3 pre-loaded and pre-created in MATLAB (see below) to avoid latencies during the presentation. Subjects voluntarily participated on their computer. They were asked to wear headphones, sit in a quiet environment, and stay focused during the whole task.

The Ethical research committee of Paris Saclay University approved the protocol under the reference CER-Paris-Saclay-2019-063.

Stimuli

All speech stimuli were generated with the MBROLA text-to-speech software ***(Dutoit et al., 1996)*** using French diphones. The duration of all syllables was equalized to 250 ms with flat intonation and no coarticulation between syllables. Each experiment was composed of 800 syllables (3.3 mn) of an artificial monotonous stream of concatenated syllables that correspond to the four possible words randomly concatenated with the only restriction that the same word could not be presented twice in a row. The same vocabulary (sixteen syllables) was used in the two streams, with and without pause. In the stream with pause, a 25-ms pause was inserted every 4-syllables (total duration 3.4 mn). All streams were ramped up and down during the first and last 5 s to avoid the start and end of the streams serving as perceptual anchors. We used the same syllables and words for the infant experiment. To avoid phonological similarity effects that could bias toward one or the other condition, Words and PartWords were reversed for half of the subjects.

In a previous experiment with similar streams with and without 25 ms pauses, ***(Peña et al., 2002)*** showed that participants were at chance when they had to choose which of the two streams had pauses. To confirm that the pauses were not consciously

perceived, 8 adults listened to 20 streams (40 syllables – 10 seconds) presented randomly (10 without pauses and 10 with a 25 ms-pause every four syllables) and were unable to indicate which stream had pauses or not (mean = 49% (range [40, 59]%); p = 0.89).

Procedure

After listening to the structured stream, participants were asked to rank the familiarity of the individual words (from “Completely unfamiliar” to “Completely familiar” on a 6-step scale). Learning and test phases were repeated twice. Data of the two tests sessions were aggregated in the main analysis (see separated analysis of each session in fig 2.10). Six conditions (three bi-syllabic as foils and three quadri-syllabic conditions) were used to avoid any bias based on the length of the test words, with 4 trials in each of the 6 conditions. To avoid phonological similarity effects that could bias toward one or the other condition, participants were assigned to one of two groups where conditions were reversed. Four different pairs of structured streams per group were also generated, and participants were randomly assigned to one pair to avoid any given particularity of a stream driving the results. Three conditions were studied: Words, PartWords, and ShuffleWords. Words corresponded to the words that were embedded in the structured streams (ABCD), while PartWords corresponded to the two last syllables of a word and the two first of another word (CDA’B’). Thus, although PartWords were heard during the structured stream, they violated chunking based on TP. ShuffleWords corresponded to words in which the second and third syllables were inverted, creating a null TP between all syllables (none of the transitions were heard during the structured stream). However, the first and last syllables were correct.

Data Processing

Each answer was converted to a numerical value from 1 (completely unfamiliar) to 6 (completely familiar). The responses to the bisyllabic trials were not considered. All data, from both test sessions, were aggregated together in each group to compute a linear mixed-effects model on items (y ~ condition + (1|subject)) to take the subject effect into account. The p-values were then FDR corrected. To compare subjects' segmenting performances for both streams, we computed the mean familiarity ranking for each condition in each subject and subtracted the PartWord ranking from the word ranking within each subject. We then performed a one-way unpaired t-test between the two groups.

Infant EEG Experiment

Participants

Two groups of healthy full-term neonates were tested between days 1 and 3. There was no problem during pregnancy and delivery, birthweight > 2500 g, term > 38 wGA, APGAR ≥ 6 and 8 at 1' and 5', normal audition tested with otoacoustic emission. Parents provided informed consent, and the Ethical Committee (CPP Tours Region Centre Ouest 1) approved the study (EudraCT/ID RCB: 2017-A00513-50). In the first group (continuous), 34 neonates were tested. Among them, seven were excluded because they did not complete the experimental protocol or technical issues leaving 27 infants (14 males). In the second group, 34 infants were tested (with pauses). Nine were excluded because they did not complete the experimental protocol or technical issues, leaving 25 infants (13 males).

Stimuli

We used the same 16 isolated syllables generated with MBROLA as in the adult experiment to construct 4 different streams (structured and random, with and without pause). The random stream consisted of 1600 pseudo-randomly concatenated syllables (6.7 mn). Each syllable could be followed by three others from the pool leading to a flat TP during the stream. This pseudo-random stream offers a more controlled stimulus than the random streams used previously because the TPs were fixed to 1/3 (instead of 1/15), a similar value than the TP between words in the structured stream. The structured stream was comprised of 3200 syllables (13.3 mn). All streams were ramped up and down during the first and last 5 s to prevent the beginning and the end of the streams from being used as anchors. We created only one syllabic order for each stream to obtain learning markers better comparable between infants. For the second group, a pause was added every 4 syllables in both the structured (duration: 13.7 mn) and the random streams (duration: 6.8 mn). Thus, the sequences were identical for all infants in both groups except for the 25-ms subliminal pauses every 4 syllables in the second group. Because all subjects had the same auditory materials, we carefully controlled for low-level acoustic-phonetic properties. We equilibrated the characteristics of consonants and vowels in the different words and at the different syllabic positions within words to avoid learning based on low-level acoustic cues (See fig 2.6 – supplementary information for more details). As in adults, three types of test words were created: Word (ABCD), PartWord (CBA'B'), and ShuffleWords (ACBD). (Table 1 and fig 2.6 supplementary information).

Following a reviewer's requirement, we tested adults on the same material with the same exposure duration (~13mn). Their behavioral results are presented in fig 2.10 – supplementary information.

Procedure

EEG was recorded with 128 electrodes (EGI geodesic sensor net), carefully placed on the neonates' heads by trained researchers to increase the consistency of the net placement. Three nets with different radii were used to fit infants' heads. For the continuous group, infants were tested while asleep in the experimenter or parent's arms. Due to COVID restrictions, the second group of babies was tested asleep in the crib. This slightly increased the noise level in the second group and might have marginally decreased the sensitivity of our analysis for this group.

Both groups followed the same procedure (fig 2.1). A first control stream of a pseudorandom concatenation of 1600 syllables was followed by a structured stream composed of 3200 syllables grouped in words of 4 syllables. Infants then heard eight repetitions of short structured streams (160 syllables) followed by 12 test words presented in isolation (4 in each condition: Word, PartWords and ShuffleWords, ISI 2-2.5s) for a total of 96 test-words (32 in each condition). The short streams were added to maintain learning because 2/3 of the test-words violated the learned structure. Each test word was preceded by a short click 200 ms before its onset. The click was added as a task unrelated auditory localizer and to reset the baseline with a neutral event to avoid long-range drifts following the words. Finally, a second control stream was presented. Thus, the two random-streams were sandwiching the structured stream to control for habituation to the auditory stimulation, change in sleep stage, and any confounding time effect.

Data processing

EEG recordings were band-pass filtered between 0.2 and 15Hz for all analyses. Artifact rejection was performed on the non-epoched recording session using APICE pipeline ***(Fló et al., 2022b)*** based on the EEGLAB toolbox ***(Delorme and Makeig, 2004)***.

Artifacts were identified on continuous data, based on voltage amplitude, variance, first derivative, and running average. The variance algorithm was applied in sliding time windows of 500 ms with 100 ms steps. Adaptive thresholds were established for each subject and electrode as two interquartile ranges away from the 3rd quartile. This gave a logical matrix of the size of the recording, indicating bad data. Electrodes were definitely rejected if they were marked as bad more than 50% of the recording time, and time-samples were marked as bad if more than 35% of the electrodes were marked bad at this time-sample. For the ERP analysis, we then performed spatial interpolation of missing channels, and the data were mathematically referenced to the average of the 128 channels.

Neural entrainment

The recordings from the structured and random streams were segmented into consecutive non-overlapping epochs of 15 words (corresponding to 15s in the continuous group and 15.375s in the pause group). All subjects having 10 good epochs or more in each condition were included in this analysis (25 neonates in the continuous group, 21 in the pause group). We averaged the activity over artifact-free epochs for each neonate and electrode and computed the Fourier Transform using the fast Fourier transform algorithm (FFT) as implemented in MATLAB. We then computed the power of the FFT. The Phase Locking Value (PLV) between trials was computed on the FFT of single trials. Those values were normalized with neighboring frequency bins ([-8:1,1:8]). The frequencies of interest were selected as the inverse of the duration of a word (f = 1Hz for the continuous group f=0.975 for the second group with pauses) and one-quarter of a word (i.e., roughly a syllabic rate, f = 4 Hz for the first group, f = 3.9 Hz for the second). To assess the significance of the power/PLV at the two frequencies of interest, we computed a contrast between the

power/PLV during the structured stream compared to the random streams for each electrode. As we expect learning during the structured stream to elicit a word rate oscillation, we computed a one-way (structured>random) paired t-test on each electrode. We corrected for multiple comparisons using cluster corrected approach (alpha = 0.05). To look for a potential difference between groups, we computed an interaction between the previously described contrasts of both groups (difference of the structure minus random contrast in each group). Specifically, we ran a one-way unpaired t-test on each electrode and the clustering approach for the interaction.

Additionally, we also replicated the neural entrainment effects with a slightly different method as proposed in Fló et al. ***(Fló et al., 2022a)***. With this approach, the signal is decomposed on 1s long epochs and reconstructed in longer meta-epochs composed of several non-necessarily consecutive segment. It allows to save more data for shorter experiments at the expense of more data manipulations. Both approaches were quite similar, confirming the validity of both that can be better adapted depending on the amount of available data.

Correlation Analysis

In both experiments, all subjects heard the exact same auditory material avoiding differences in stimulation between participants. We could thus compute the instantaneous correlation between each participant and the others. For each subject at each time during the streams, we computed the correlation at the topographical level between the topography of subject i at time t (a vector of 128 voltage values at time t corresponding to the 128 electrodes) and the topography of the grand average excluding subject i at time t (a vector of 128 values corresponding to the average across the other subjects at time t for each of the 128 electrodes). It gave, for each

subject, a vector of correlation between its own topography and the mean topography of all other subjects throughout time. Bad data were replaced by zeros and not taken into account for the average topographies across subjects. Time points with only bad data gave NaN correlation results. We hypothesized that learning should lead to an increase with time in the correlation between neonates as they learn the same material. To test it, we used two different methods. In the first one, we smoothed the correlation signal using a 400s-sliding-average-window in each neonate and stream, then computed a cluster-based analysis to reveal a significant cluster of time during which one stream showed a greater correlation than the other one. In the second one, we computed the slope of the linear regression with time in each subject and then considered the slope as a variable for the structured and random conditions in t-test comparing both groups.

Pattern Similarity Analysis

To compute pattern similarity between syllables, we epoched each syllable from the structured stream from -100 ms to 350 ms. We removed the 100 first syllables to give enough time for participants to learn the task. The remaining epochs were averaged by syllables for each subject and a correlation matrix between each pair of syllables was computed with all the electrodes between 0 and 350 ms. We then separated the pairs of syllables into 5 conditions: First syllable (AA'), Ordinal position (BB' or CC' or DD'), Word and TP (AB or BC or CD), Word only (AC or AD or BD) and Low TP (DA'). We then averaged the similarity per condition and subtracted the correlation between all the other pairs. We then compared if pattern similarity between groups of syllables was increased differently across groups (One-way t-test with pauses>continuous).

ERP Analysis

Data were segmented in 2850 ms long epochs ([-750 +2100]ms relative to word onset), averaged in the three conditions (Words, PartWords, and ShuffleWords), and baseline-corrected with the mean voltage value in the interval [-750 – 0] in each neonate. Neonates with less than 20 remaining trials in total were excluded from analysis (none in the continuous group, 1 in the pause group).

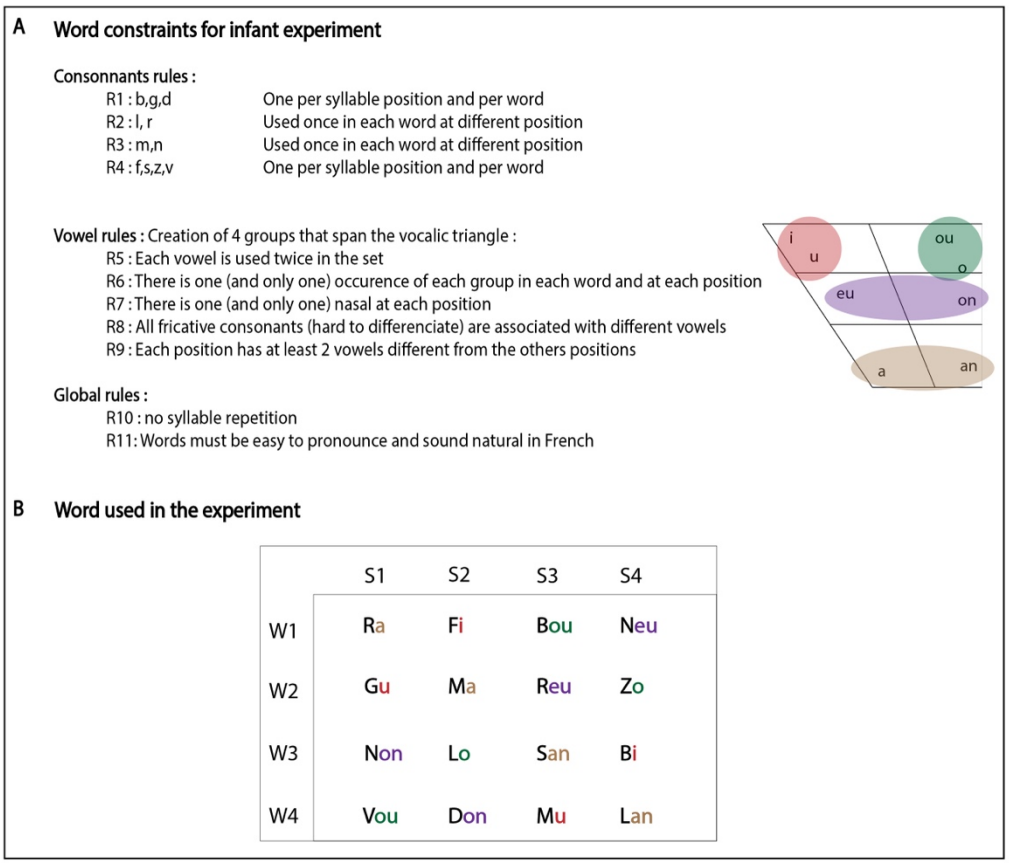
To extract ROI corresponding to the functional auditory localizer of each group, we measured the auditory event-related potential associated with the click presentation at the beginning of each trial by running a cluster-based analysis against zero to extract auditory ERP (5000 randomizations, two tailed t-test, alpha < 0.01, cluster-alpha < 0.01, between -200 and 0 ms). This procedure identified a positive frontal and a negative occipital cluster in each group, on which we restricted the ERP analyses. Therefore, the voltage was averaged across electrodes in each of the two clusters in each neonate and condition.

A cluster-based analysis was performed on the obtained time-series (10000 randomizations two tailed t-test alpha < 0.05, cluster alpha < 0.05) between 250ms (end of the first syllable) and 2000 ms to compare all pairs of conditions. Because of the adults' behavioral results, we added the contrast 'heard' (average of Word and PartWord) vs. 'non heard' (ShuffleWord) in the continuous group. Finally, we computed the interaction between groups and conditions (Word-PartWord) during the time window in which the previous analysis revealed a significant effect.

Supplementary information

Constraints followed to create the set of words

In order to use the correlation between subjects' EEG as a measure of learning during the stream, we exposed all participants to exactly the same auditory material in the same order. This choice presented a risk of bias if, by some bad luck, there was a noticeable structure in the streams or acoustical bias in the words we used. Furthermore, we wanted to obtain 16 clearly different syllables. Because neonates might be particularly sensitive to the vowel content, we used 8 different vowels spanning the vocalic triangle to avoid any low-level confusion. Finally, ERPs are sensitive to low-level temporal properties in the stimuli (e.g. the difference between



A **Word constraints for infant experiment**

Consonnants rules :

R1 : b,g,d	One per syllable position and per word
R2 : l, r	Used once in each word at different position
R3 : m,n	Used once in each word at different position
R4 : f,s,z,v	One per syllable position and per word

Vowel rules : Creation of 4 groups that span the vocalic triangle :
R5 : Each vowel is used twice in the set
R6 : There is one (and only one) occurence of each group in each word and at each position
R7 : There is one (and only one) nasal at each position
R8 : All fricative consonants (hard to differenciate) are associated with different vowels
R9 : Each position has at least 2 vowels different from the others positions

Global rules :
R10 : no syllable repetition
R11: Words must be easy to pronounce and sound natural in French

B **Word used in the experiment**

	S1	S2	S3	S4
W1	Ra	Fi	Bou	Neu
W2	Gu	Ma	Reu	Zo
W3	Non	Lo	San	Bi
W4	Vou	Don	Mu	Lan

***Figure 2.6** : **A)** Constraints followed to create an unbiased set of words; **B)** Words used in the experiments*

fricatives and plosives); thus, we wanted to control for the position of the type of consonants to avoid biases in perception and in the EEG responses.

To minimize this risk, we used several methods. First, both streams, with and without pauses, were exactly the same, except for the subliminal pauses added at the end of the words. Thus, any bias should have affected both groups equally. Second, we carefully designed the auditory material to prevent a particular low-level acoustic feature from causing the effect. Finally, we ran the same experiment with adults using a more randomized approach between participants and replicated the results. Overall, this great attention to stimuli design allowed us to perform a new between-subjects analysis and find a novel neural correlate of learning in the infant EEG data. The balance of low-level acoustic features is presented in figure 2.6 : we applied many rules to ensure that consonants and vowels with different acoustic properties were balanced across words.

Neural entrainment power

As described in the main text, we computed the power of the neural entrainment at the syllable and word frequency for both structured and random streams. In the main figure (fig 2.3), we presented the contrast at the word rate between the structured and random streams. For the sake of completeness, we present here the results of the power at the word and syllabic rates for structured and random streams in both groups. For each frequency of interest, we estimated the significance of the power for each electrode by comparing to the power of the other frequencies.

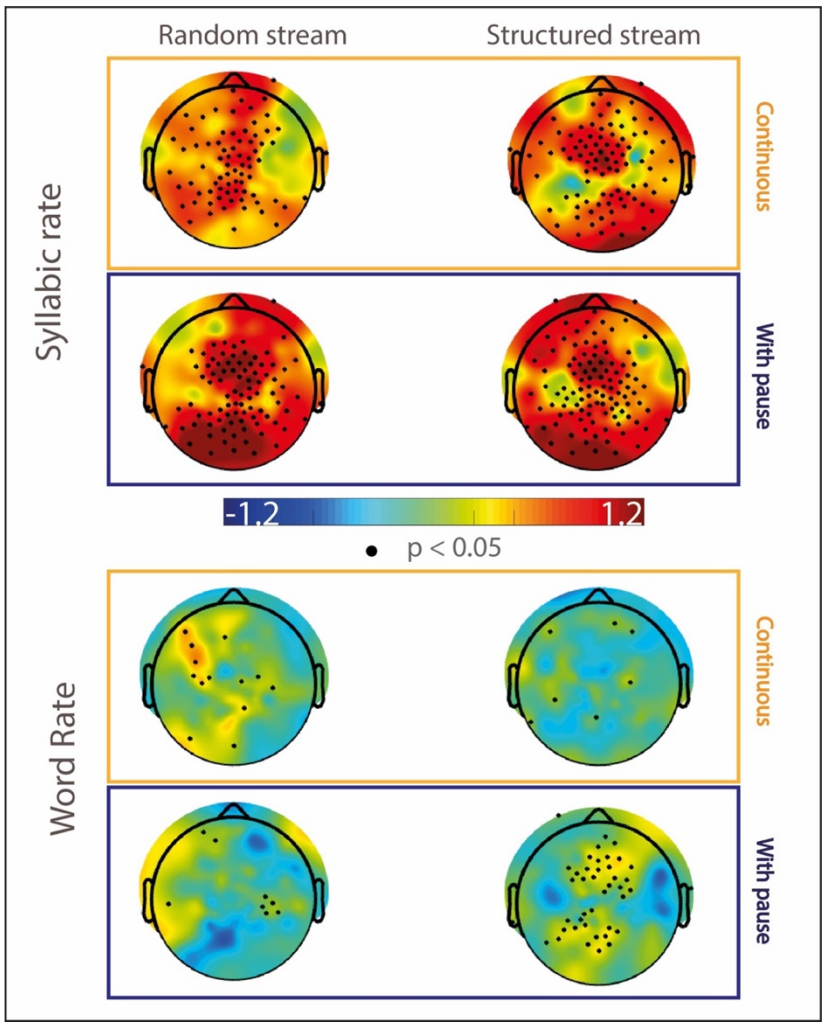


Figure 2.7 : Power of the neural entrainment *compared to other frequencies at the syllable (top) and word (bottom) frequency. The left column corresponds to the random stream and the right column to the structured stream. Each row corresponds to one neonate group listening to the continuous stream (orange) and to the stream with pauses (blue). Dots locate the significant electrodes among the 128 of the geodesic net (p<0.05).*

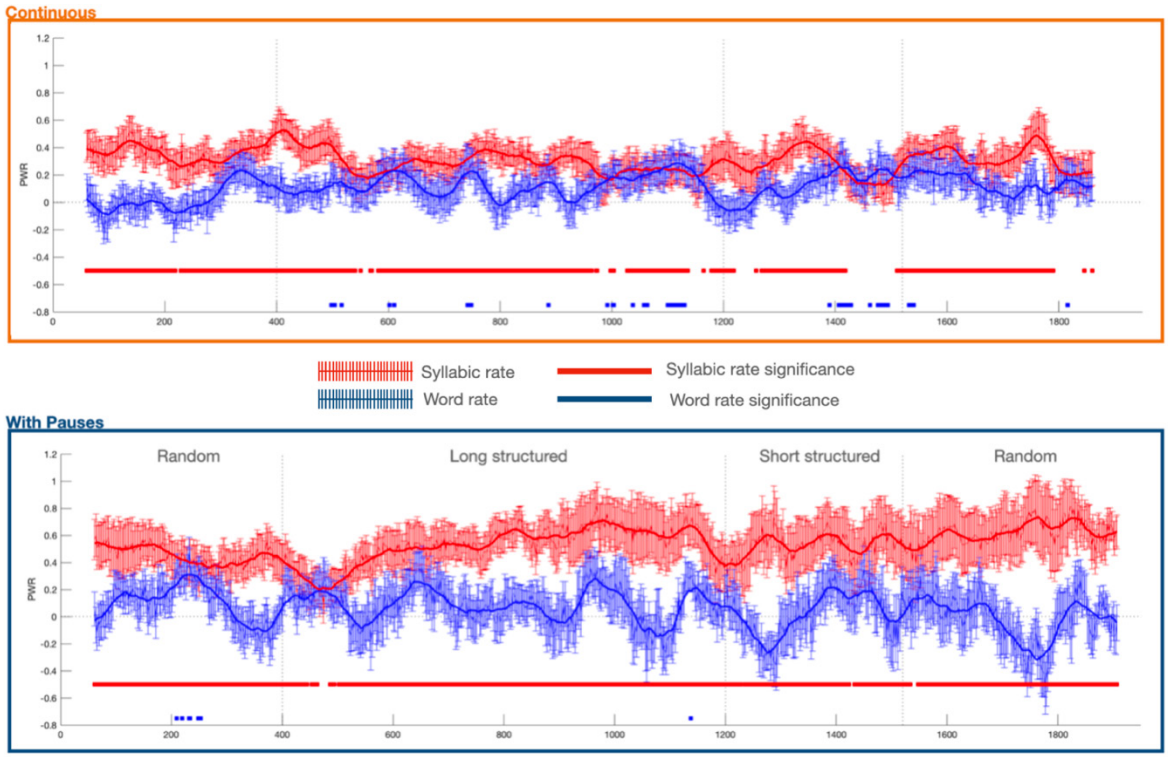


Figure 2.8 : Time course analysis of the power entrainment at the word and syllabic rates*. Contrary to Flo et al., we did not observe a clear time course of the power increase at the word frequency probably due to a lower signal to noise ratio at 1hz. Solid lines represent the time where entrainment is significantly greater than zero.*

ERP analyses using cluster-permutation-based approach

In the main text, we presented ERP differences based on responses restricted to ROIs defined by an auditory localizer. We believe that this approach may be an improvement over cluster-based methods because it increases sensitivity within this localizer ROI and decreases the risk of false positivity by restricting the analysis to these localized areas. The sensitivity of cluster-based approaches have been raised (1). In addition, contrary to other neuroimaging methods such as fMRI or MEG, EEG is not realigned on a template while there is variability in the placement of each electrode due to net placement, head shape molded by the birth canal, some head

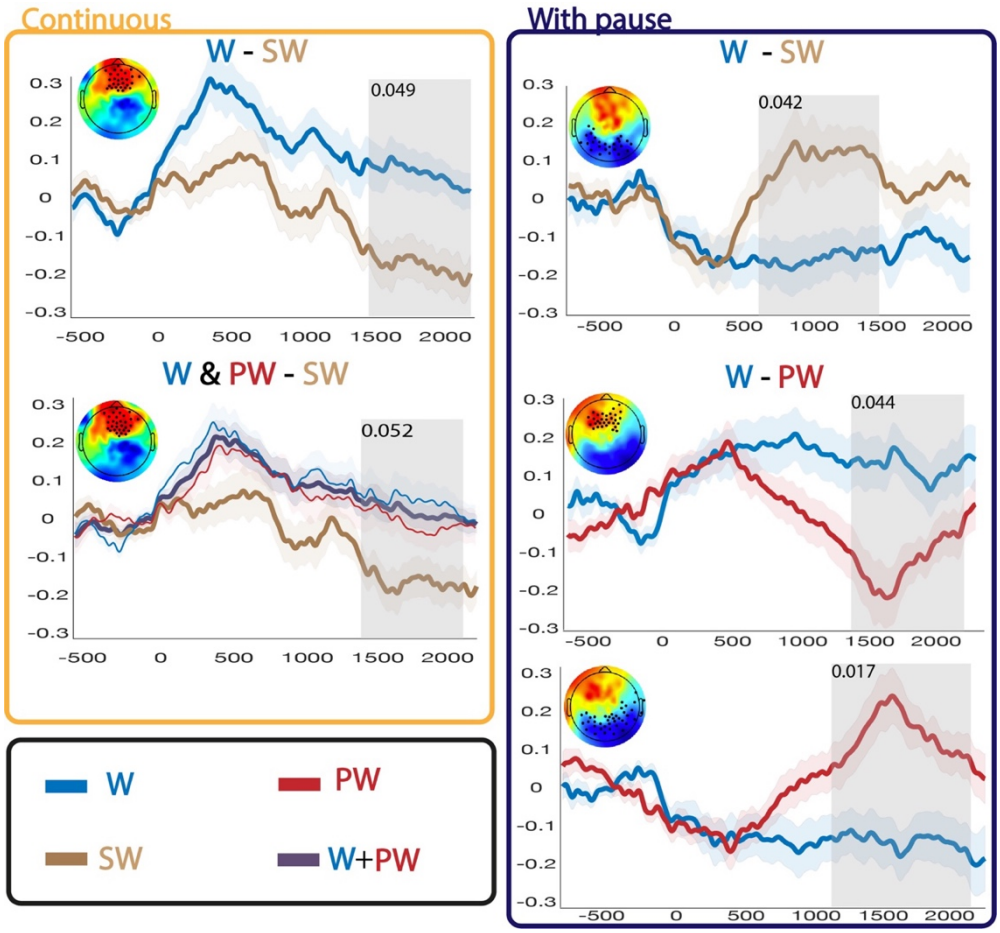


Figure 2.9 : *Re-analysis of the ERP results using fieldtrip cluster -based method : overall, the results are qualitatively similar to what we reported in the main text. Estimated p-values are reported with 10000 bootstrap iterations.*

oedema due to this “difficult” moment, but also variability in the response itself due to brain immaturity (2). The addition of a localizer allows to partially deal with this variability and to estimate the auditory response on a particular group. Therefore, we believe that our approach using data driven ROIs from a localizer is more suitable for infant EEG data than clustering methods. However, we replicated those results using a classical cluster-based approach (fieldtrip). Results are presented in fig 2.9.

We used fieldtrip method with alpha = 0.05 in [250-2000]s time window. Overall, the results remain qualitatively similar to those presented in the main text, although slightly less sensitive.

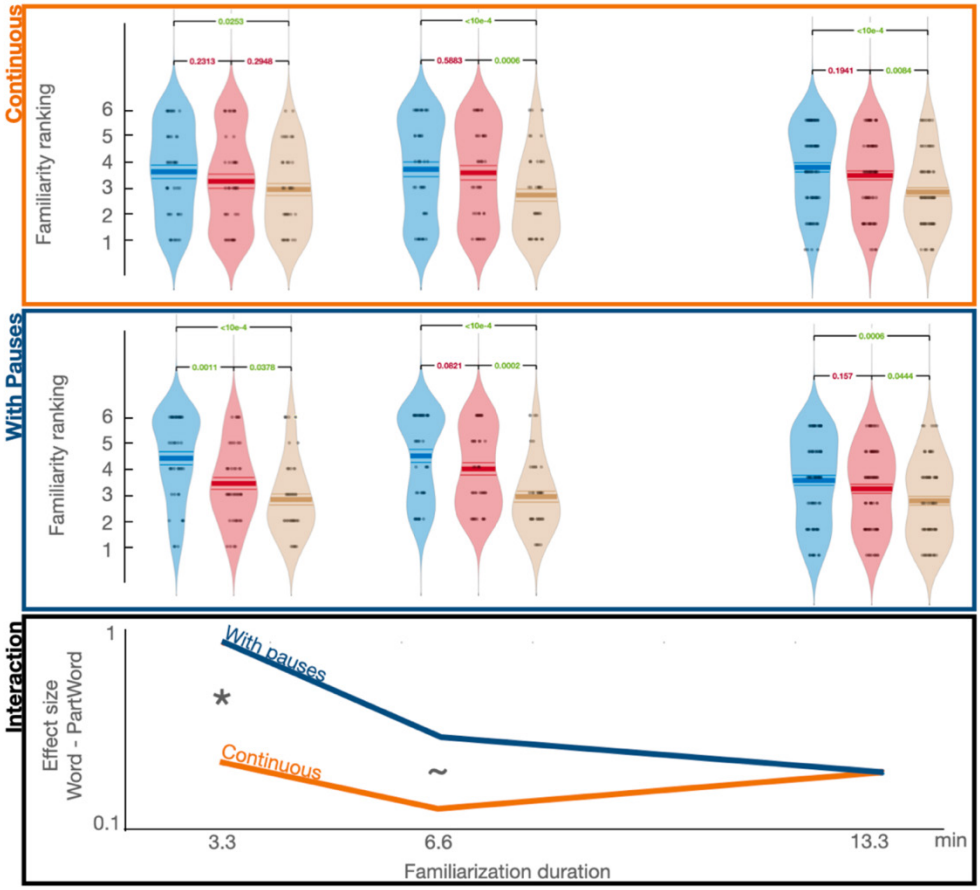


Figure 2.10 : Results of the adults' behavior as a function of the duration of exposure: *The main results reported in the text comprised the two test phases combined i.e. after 3.3 and 6.6 minutes of familiarization, respectively. Here we separated the results of the two tests. We also ran the task with a 13.3 minutes exposure using the neonates' habituation stream (right plots). For this log duration, there was no more difference between words and part-words suggesting that all transitions were learned despite the pause. Pvalues are estimated running a linear mixed effect as described in the main text*

Adults' behavioral results are modulated by the length of the familiarization

Concerning the adult experiments, we used a shorter 3.3-mn exposure (two times) compared to the neonate experiments because previous studies have shown over-learning when the exposure was too long (3). Furthermore, we thought that it was difficult to control participants' attention on a web platform, and the role of attention in statistical learning in adults compared to infants is debated (4,5,6).

The data reported in the main text were collected in two 3.3-minute exposures followed by a test. The main figure in the manuscript was the aggregation of the two tests sessions, thus after 3.3 and 6.6 minutes respectively. If we split them (fig 2.10) participants exposed with the continuous stream reported a different familiarity with words and part-words, contrary to shuffle words, neither after 3.3 nor after 6.6 minutes. In the group listening to the stream with pauses, the word-partword difference decreases after 6.6 minutes of familiarization probably because of over-learning and lack of attention as discussed above.

As proposed by a reviewer, we ran a control experiment without pauses with 13 minutes exposure (n=20) to test if longer exposure might be necessary for successful segmentation in the case of quadruplets and found no supplementary evidence of learning. As can be seen in figure 2.10, increasing the duration of familiarization did not improve the difference in adults' familiarity for words and part-words while shuffle-words remained classified as unfamiliar. To be systematic, we also run the same 13mn familiarization with the stream with pauses in another group of adult participants (n=20). In this experiment, the difference between words and part-words was erased. It is interesting to note that the pause was no more sufficient to chunk the stream once every existing transition had been sufficiently repeated along the 13 mn to become familiar (i.e. the absolute frequency of each transition was high, independently of its relative probability).

Appendix 1 : Are pauses the only way to recover segmentation ?

This work is original and will probably not be published as it is more an exploration and pilots for possible future research questions than a proper study.

Are pauses prosodic or contextual cues ?

Given the previous results on both adults and infants failing to segment quadrisyllabic pseudowords in continuous sequences, we planned to investigate under which conditions we might recover sequence segmentation in order to better understand the limits and constraints on statistical learning. We showed earlier that a subliminal pause between words was sufficient for both adults and infants to recover segmentation of the sequence into word like units. We interpreted this pause as a possible prosodic cue.

Another, and more general possibility might be that pauses act as contextual boundaries and that any contextual change would trigger segmentation. To test this hypothesis, we decided to measure the behavioral effect of another possible contextual boundary but, this time in the visual domain, and a priori irrelevant for language learning. Specifically, we took advantage of the fixation cross usually present in cognitive experiments to help participant stay still and look at the center of the screen. We re-used the exact same procedure as the adult online behavioral test with and without pauses and collected 30

subjects on a sequence where the fixation cross was turning of 45° every four elements in a congruent manner with the word segmentation. Importantly, no information was given to the participant about the fixation cross and its relevance for the experiment. Subjects that previously participated to other segmentation experiments from the lab could not

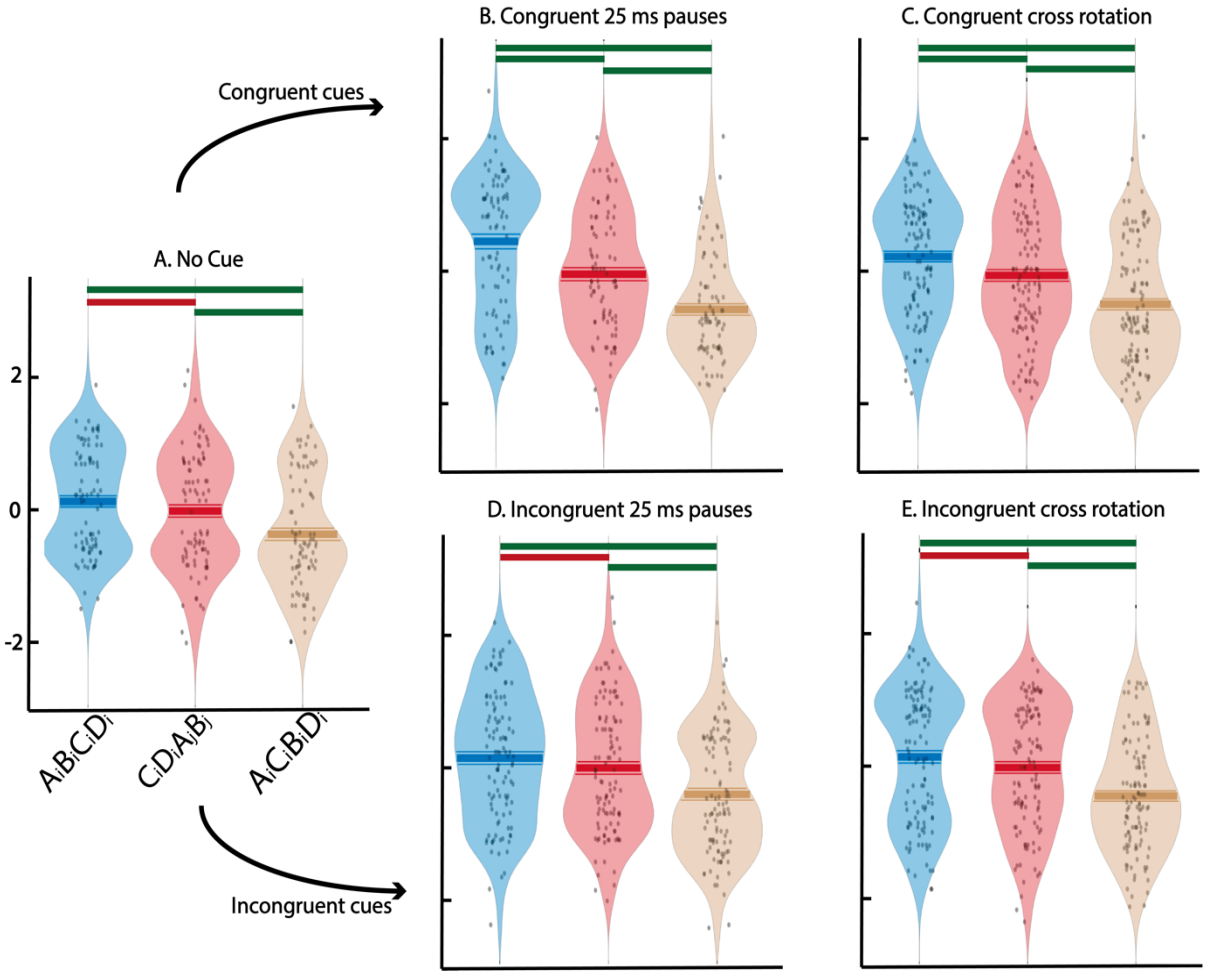


Figure A1 *: Normalized adult behavioral familiarity rating in our 5 experiments. Green bars indicate a significant difference between the two conditions (p<0.05) while red bar indicate non-significant difference (p>0.05). Without any additional cue, adult failed to segment quadri-syllabic sequence into word like unit as revealed by the absence of difference between Words and Part-Words. However, they have successfully learnt transition probabilities as they rejected the Shuffle Words. When adding a contextual cue such as a subliminal pause or a rotation of the fixation cross congruent with the word-boundaries, participants recovered the segmentation of the structure and reported significantly higher familiarity to Words than PartWords. To control for attentional effect of the cues, we also ran controls with the cue (pause or cross rotation) incongruent with the word boundaries. Like in the no cue condition, participants could learn the transition probabilities but failed to segment the sequence*

participate in this study to avoid any possible prior on the task to achieve. Amazingly, participants showed a similar pattern to the group with pauses and could significantly differentiate Words from Part-Words (see fig A1 for normalized familiarity responses). To ensure that the pauses or rotating fixation cross were not just enhancing participant attention to the sequence, we also performed two control groups (N=27 & N=29) with a cue every 3 elements and thus incongruent with pseudo word segmentation (resp 25 ms pauses or fixation cross 45° rotation). In both control groups, participants could still significantly compute transition probabilities (they systematically rejected the Shuffle-Words compared to Word and Part-Words) but failed to differentiate Words from Part Words in both cases (see fig A1 for all data).

However, these results should be taken cautiously as the interaction between the visual cue group and the continuous group did not reach significance (unlike pause vs continuous). Deeper exploration of this question would require more participants and maybe other contextual cues to better explore the impact of each (male/female voice, change in sound velocity, change in background colors…)

Appendix 2 : Remarks on the analysis of steady-state responses: spurious artifacts introduced by overlapping epochs

This work was published in Cortex under the reference :

Lucas Benjamin, Ghislaine Dehaene-Lambertz, Ana Fló, **Remarks on the analysis of steady-state responses: Spurious artifacts introduced by overlapping epochs**, *Cortex*, Volume 142,2021,Pages 370-378,ISSN 0010-9452

Abstract :

Periodic and stable sensory input can result in rhythmic and stable neural responses, a phenomenon commonly referred to as neural entrainment. Although the use of neural entrainment to investigate the regularities the brain tracks has increased in recent years, the methods used for its quantification are not well-defined in the literature. Here we argue that some strategies used in previous papers, are inadequate for the study of steady-state response, and lead to methodological artefacts. The aim of this commentary is to discuss these articles and to propose alternative measures of neural entrainment. Specifically, we applied four possible alternatives and two epoching approaches reported in the literature to quantify neural entrainment on simulated datasets. Our results demonstrate that overlapping epochs, as used in the original Batterink and colleagues' articles, inevitably lead to a methodological artefact at the frequency corresponding to the overlap. We therefore strongly discourage this approach and encourage the re-analysis of data based on overlapping epochs. Additionally, we argue that the use of time–frequency decomposition to compute phase coherence at low frequencies to reveal neural entrainment is not optimal.

Introduction

Neural Entrainment

It is well established that a stable and rhythmic stimulation elicits a stable and rhythmic neural response *(Picton et al., 2003; Regan, 1977)*, a phenomenon usually referred to as neural entrainment or steady-state responses. An enhanced cortical response, measurable with electro or magneto-encephalography (EEG or MEG), appears at very specific frequencies in function of the periodicity of the stimulation. Although it is still discussed whether it results from multiple evoked responses to the stimulus train *(Capilla et al., 2011)* or from the alignment of intrinsic brain oscillation to the input *(Doelling et al., 2019)*, this enhanced response indicates which event the brain is tracking in a train of stimuli. It can be low-level properties of the sensory input (the "on- effect" of an image for example), or more abstract regularities, such as a face presented every four images *(de Heering and Rossion, 2015).* Therefore, this phenomenon receives more and more attention due to its theoretical implications *(Buzsáki, 2006; Giraud and Poeppel, 2012)*, and also to its promising uses in special populations such as infants. For example, it has been used to study categorization in infants *(de Heering and Rossion, 2015; Peykarjou et al., 2017)*, the tracking of linguistic structures in adults *(Ding et al., 2017, 2016)*, and speech segmentation in adults *(Batterink and Paller, 2019, 2017; Buiatti et al., 2009)* and infants *(Choi et al., 2020; Kabdebon et al., 2015)*.

To quantify neural entrainment, frequency-domain analyses are preferred to conventional time-domain analysis for two reasons. First, it is generally easy to predict the frequencies at which the increased response should occur *(Zhou et al., 2016b)* reducing the number of statistical comparisons. Second, frequency-domain analysis provides a better signal-to-noise ratio *(Norcia et al., 2015)*. However, different methods can be used and have been used in the literature, some presenting a risk of bias when overlapping data are used *(Batterink*

and Paller, 2019, 2017; Choi et al., 2020). Thus, our goal is to discuss the pros and cons of the different methods and then provide evidence with simulated data.

Although our point can be extended to all studies using neural entrainment, we decided to simulate a speech segmentation study. Neural entrainment is an adequate and extensively used tool to recover speech perception units in adults within natural sentences *(Ding et al., 2017, 2014)* or artificial streams *(Buiatti et al., 2009; Kabdebon et al., 2015)*. Because speech is a continuous stream of phonemes, syllables, and words, an important question notably concerns how infants succeed to chunk this stream in candidate words. It has been proposed that infants can take advantage of the probabilities of transitions between successive syllables that are constant within a word and drop at its boundaries *(Saffran et al., 1996a)*. Thus, if a stream is composed of four tri- syllabic words randomly concatenated without repetition, the transition probability is 1 within a word and 0.33 between words. Neural entrainment appears as a suitable approach to investigate this experimental paradigm. As the syllables have equal duration, the neural response is entrained at the syllable rate, as many segmentation experiments report *(Batterink and Paller, 2019, 2017; Buiatti et al., 2009; Choi et al., 2020; Kabdebon et al., 2015)*. If the participant discovers the word regularity, the neural response should also be entrained at the word rate as also reported in the same studies.

Methodological considerations

When using rapid periodic stimulation, a sustained oscillatory activity is expected to emerge at the stimulation frequencies. The most direct analytical approach would be to transform the entire dataset in the frequency domain by applying, for example, the Fast Fourier Transformation (FFT) algorithm. However, this procedure is not the most commonly used because, given the signal's stationary nature, averaging across multiple measures (implying multiple shorter segments) should provide a better signal to noise

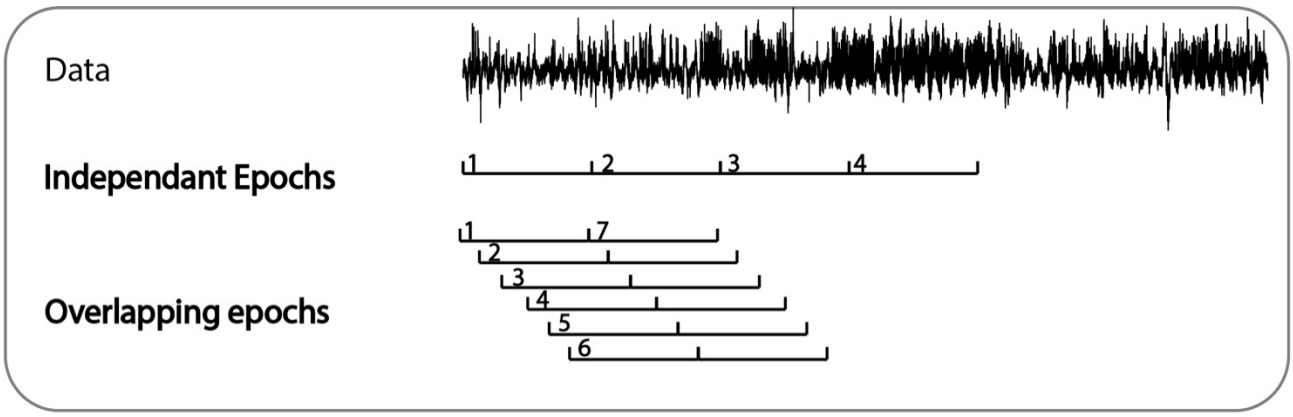
(SNR) than a single measure on the whole recording. Two constraints are therefore opposed for a correct analysis of neural entrainment: on the one hand, to have long segments and on the other hand, to have many of them.

First, to obtain an optimal description in the frequency domain, each epoch should be long enough for the signal and noise to be well represented. The epoch's length determines the frequency resolution (fres = 1/epoch length), and the frequency resolution has to be at least half the frequency of the slower steady-state response to be able to see it in the frequency domain. In other words, the slower response must repeat at least twice to be detected in the frequency domain. Moreover, longer epochs also provide a better SNR because, with a higher frequency resolution, the noise is spread over more frequency bins, while the signal remains restricted to specific frequencies ***(Norcia et al., 2015)***.

Second, in steady-state responses, a common evoked activity is expected after each repetition of the stimulus. Thus, dividing the recording into sub-segments aligned with each stimulus onset and averaging the resulting epochs, should enhance phase-locked activity and reduce non-phase locked activity, as it is done to recover the classical event-related potentials (ERP). Subsequently, entrainment is detected as an increase in power at specific frequencies in the spectrum of the average, even though the steady-state responses are much weaker than the background noise. Furthermore, phase coherence across epochs can also be computed.

While most studies use multiple epochs to estimate the power spectrum or the phase coherence across them, the computations are differently implemented in practice. Here we have isolated two epoching approaches, and four different calculations to estimate neural entrainment

(see Fig. A3). We applied these methods to simulated data to evaluate their performance and potential biases.



Method 1 FFT on each epoch

Mean power of the FFT of each epoch

$$P(f) = \frac{1}{N}\sum_{i=1}^{N}\left|\mathcal{F}(f,i)\right|$$

Method 2 FFT on the average of all epochs

Power of the FFT on the avergage of epochs

$$P(f) = \left|\mathcal{F}(f,\frac{1}{N}\sum_{i=1}^{N}i)\right|$$

Method 3 Inter Trial Coherence

Inter Trial Coherence between epochs

$$ITC\,(f) = \frac{1}{N}\left|\sum_{i=1}^{N}e^{i\phi(f,i)}\right|$$

Method 4 Inter Trial Coherence using newtimef

Inter Trial Coherence between epochs using time frequency wavelet decomposition

$$ITC_{\,\mathrm{newtimef}}(\mathrm{f}) = \frac{1}{\mathrm{T}}\left(\sum_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{PLV}(\mathrm{f},\mathrm{t})\right)$$

N : Number of epochs
ϕ(f,i) : Phase at freq f & trial i
$\mathcal{F}$(f,i) : Fourier transorm at freq f & trial i
T : Number of time bins for time freq analysis

***Figure A3**: **Description of the two epoching approaches** and the four methods used to estimate neural entrainment. Note that methods 3 and 4 use the same approach (ITC) but on the whole epoch (method3) or on the time average of time frequency analysis (method 4).*

Methods used in the literature

As mentioned before, the definition of epochs plays a crucial role in these analyses. Epochs have to be long enough to represent properly signal and noise, and at the same time, the more epochs, the better to detect phase-locked activity. In most of the papers, the segment's length is calculated to include 10-25 repetitions of the slower response. Although the epoch's length may affect sensitivity, it does not lead to artifactual measures of entrainment. Most papers have used independent epochs ***(de Heering and Rossion, 2015; Ding et al., 2016; Kabdebon et al., 2015)***, but some have used overlapping epochs to increase the total number of epochs ***(Batterink and Paller, 2017; Buiatti et al., 2009; Choi et al., 2020)*** (fig A3). We believe that overlapping epochs entails a methodological artefact; so, we decided to test both approaches.

The four methods described in the literature further differ in the operations' order or in the measure used to estimate neural entrainment (power spectrum or phase coherence). In the first method (method 1) the power spectrum is computed for single epochs and then averaged across epochs to obtain an averaged power spectrum. Buiatti and colleagues used this approach in a speech segmentation task in adults ***(Buiatti et al., 2009)***. In this case, the entrainment measure depends on the presence of induced and evoked activity that is strong enough to emerge from the noise level. In the second method (method 2), the average across epochs is obtained first, and its power spectrum is computed afterwards. Here, non-phase locked activity is first reduced by the averaging step, which increases sensitivity to phase-locked activity. This is the most common approach used in numerous papers ***(de Heering and Rossion, 2015; Ding et al., 2016)***. The third method (method 3) calculates the inter-trial coherence (ITC), or phase-locking value (PLV) across all epochs. Because it measures phase-locked activity, it is in principle similar to method 2 in terms of sensitivity and neural bases. This approach is widely used in the literature as well ***(Kabdebon et al., 2015)***. The fourth method uses a wavelet decomposition

to estimate ITC, and has been implemented by Batterink and colleagues in speech segmentation experiments ***(Batterink and Paller, 2019, 2017; Choi et al., 2020)***. While this method is in theory applicable, we argue that it is not recommended in this context.

Methods

Creation of the stimulated data

To compare the different methods and their risks of bias, we calculated the results of the four methods described above with independent and overlapping epochs on simulated data. To have datasets with a realistic noise structure, we used real EEG data, to which we added (or not) simulated steady-state responses.

We used EEG data from 27 asleep 5-month-old infants without stimulation. The data were recorded using a 128-electrode net (Electrical Geodesics, Inc.) referred to the vertex with a sampling frequency of 500 Hz, then bandpass-filtered [0.2, 40] Hz. The segments containing motion artefacts were removed. Only data from one frontal electrode (Fz) was kept for the simulations. For each infant, we had between 25 and 48 minutes of artefacts-free data. From these artefacts-free data, we recreated 40 recordings of 720 s each (12 mn).

First, we simulated a study in which infants listen to a random concatenation of syllables with a fixed duration of 300 ms; thus, the neural signal should be entrained at 3.33 Hz. We did so

by adding a half-sinusoidal wave (frequency 2.5 Hz) every 300 ms (Figure A4 A) to each of the 40 recordings. Second, we then simulated a study in which infants discover the three-syllabic words (900 ms) embedded in the stream; thus, the neural response is entrained both by the syllable regularity and the word regularity (i.e. expected neural entrainment at 3.33 and 1.11 Hz). We created the simulated data by adding a half-sinusoidal wave (frequency 0.8 Hz) every 900 ms (Figure A4 B) to each of the 40 recordings. The signal to

noise ratio was 1/40 for the "syllabic" steady-response (300 ms), and 1/20 for "word" steady-response (900 ms) (the signal to noise ratio was estimated as the signal amplitude divided by the standard deviation of the original EEG data).

Thus we analyzed three sets of data, each one with 40 recordings: (i) original data, (ii) Simulated steady-state response at 3.33 Hz and (iii) Simulated steady-state response at 3.33 Hz and 1.11 Hz. Note that the third set corresponds to Batterink and colleagues' studies ***(Batterink and Paller, 2019, 2017; Choi et al., 2020)***. In their studies, the syllabic rate was 3.33 Hz and the word rate 1.11 Hz.

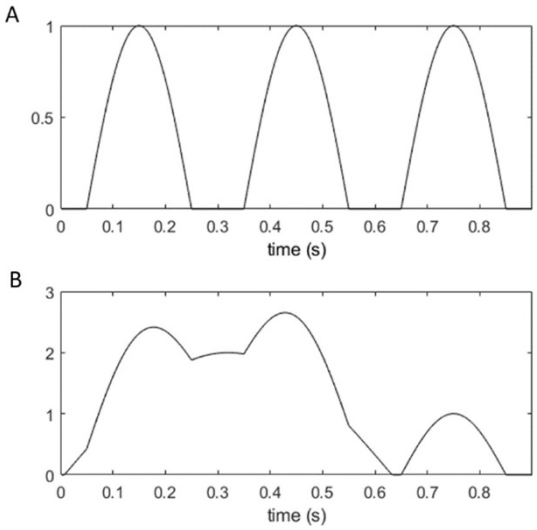


Figure A4. Simulated steady-state response *(A) to the syllables (one every 300 ms), and (B) to the syllables and the words (one every 900 ms).*

Data preprocessing

The three sets of data were segmented in epochs of 10.8 s that either did not overlap (66 epochs), or have an overlap of 11/12 (788 epochs), as in Batterink and colleagues' studies ***(Batterink and Paller, 2019, 2017; Choi et al., 2020)***. Notice that 10.8 s corresponds to 12 times the slower expected steady state-response response (i.e. 900 ms).

The analysis for each method was as follows. In Method 1, the FFT was applied to the data, the power spectrum was calculated for single epochs, and finally, the average power spectrum was obtained. Afterwards, the power spectrum at each frequency bin was normalized by the six adjacent frequency bins. In Method 2, the FFT was applied to the data, the average across epochs was computed, and the average response's power spectrum was computed. As in method 1, the power spectrum at each frequency bin was normalized by the six adjacent frequency bins. In Method 3, the FFT was applied to the data, and the PLV was computed. The PLV was normalized by computing at each frequency bin the difference with the PLV at the adjacent six. In Method 4, a Morlet wavelet transformation was used to estimate the ITC across epochs, by using the *newtimef* function of EEGLAB ***(Delorme and Makeig, 2004)***. We used the same parameters that Batterink and colleagues ***(Batterink and Paller, 2019, 2017; Choi et al., 2020)*** (0.1 Hz step, with 1 cycle at 0.2 Hz linearly increasing until 45 cycles at 20.2 Hz). To obtain a single value of ITC at each frequency bin, the ITC was averaged across time.

Results

Original dataset

In the original dataset, no indices of neural entrainment are expected. It is indeed the case when non-overlapping epochs are used whatever the method: the response is flat across frequency bins (blue lines in Figure A5 left column for methods 1 to 4). By contrast, using overlapping epochs (blue lines in Figure A5 right column), creates a pronounced peak at the frequency corresponding to the overlap, for methods 2, 3 and 4, demonstrating a methodological artefact of the approach. The origin of the methodological artefacts relies on the multiple uses of the same data. Averaging the epochs (method 2) creates an artifactual periodic signal at the overlap frequency because the same data appears periodically. Thus, the mean signal power spectrum shows a peak at the frequency

1/overlap and its harmonics. A similar effect occurs with the phase coherence is estimated either by using the FFT or wavelet transform. Comparing the phase between epochs that share a large amount of data only shifted by a fixed value will result in an artificially high phase coherence at the frequency (and harmonics) of this overlap shift value. Note the particular case of method 1 where there is no visible artefact. Because the power is estimated over single epochs, the phase information is lost before the data of different epochs are averaged.

Simulated entrainment

When simulated steady-states responses are added, neural entrainment is detected at the syllable and word frequencies by methods 2, 3, and 4 (green and red lines in Figure A5). The peaks clearly appear relatively to the no-stimulation case (blue line) in the case of independent epochs. When using overlapping epochs, the entrainment adds to the intrinsic bias due to overlapping epochs.

Method 1 was not sensitive enough to detect neural entrainment, which can be understood if we consider that the power is computed on one single epoch. Thus, the steady-state response has

to be strong enough relative to the noise to emerge at the single epoch level. In other words, the method does not take advantage of the phase coherence across trials of the steady-state response. The sensitivity is slightly increased when the number of epochs is increased by using overlapping epochs but this method remains sub-optimal.

The ITC computed using the wavelet decomposition shows a particular behavior. Entrainment appears at the frequency of the overlap and its harmonics, but the signal is smoother. The smooth of the signal affects both the neural entrainment resulting from the real steady-state response and the artefact due to overlapping epochs. This phenomenon is a consequence of the uncertainty principle in signal processing. The

wavelet transform enables a time-frequency decomposition; therefore, making it possible to investigate modulations either in the intensity or in the synchronization of the activity across trials over different frequency bands and at different time points. However, the uncertainty principle exposes a trade-off between frequency and time resolution (dt*dw>=1/2), that results in a smooth of the measure of entrainment. With the parameters used here, at 3.33 Hz the width of the wavelet is 2.36 s, and at 1.11 Hz its width is 2.71 s, in both cases much smaller than the duration of a whole epoch (10.8s). The gain in time resolution, which has no interest here, comes with a loss in frequency resolution explaining the signal's smoothing.

The time-varying information provided by wavelet decomposition is valuable in many contexts, but not meaningful in most steady-state stimulation experiments. In a steady-state experiment, the stimulus appears repeatedly, and the data are divided into sub-segments containing multiple instances of each stimulus. Therefore, looking for a time dimension within epochs is meaningless when all epochs are analyzed together. The strength of steady-state stimulation is that the neural response is periodic and stable, justifying a frequency domain analysis. To obtain a measure of phase coherence at each frequency using wavelet decomposition, the measure must be

averaged across the time dimension ***(Batterink and Paller, 2019, 2017; Choi et al., 2020)***. This step directly shows the lack of meaning of the time dimension in this context. Additionally, the computational cost of the wavelet transform on our simulated data set relatively to the other approach was of the order of 10^4 times higher.

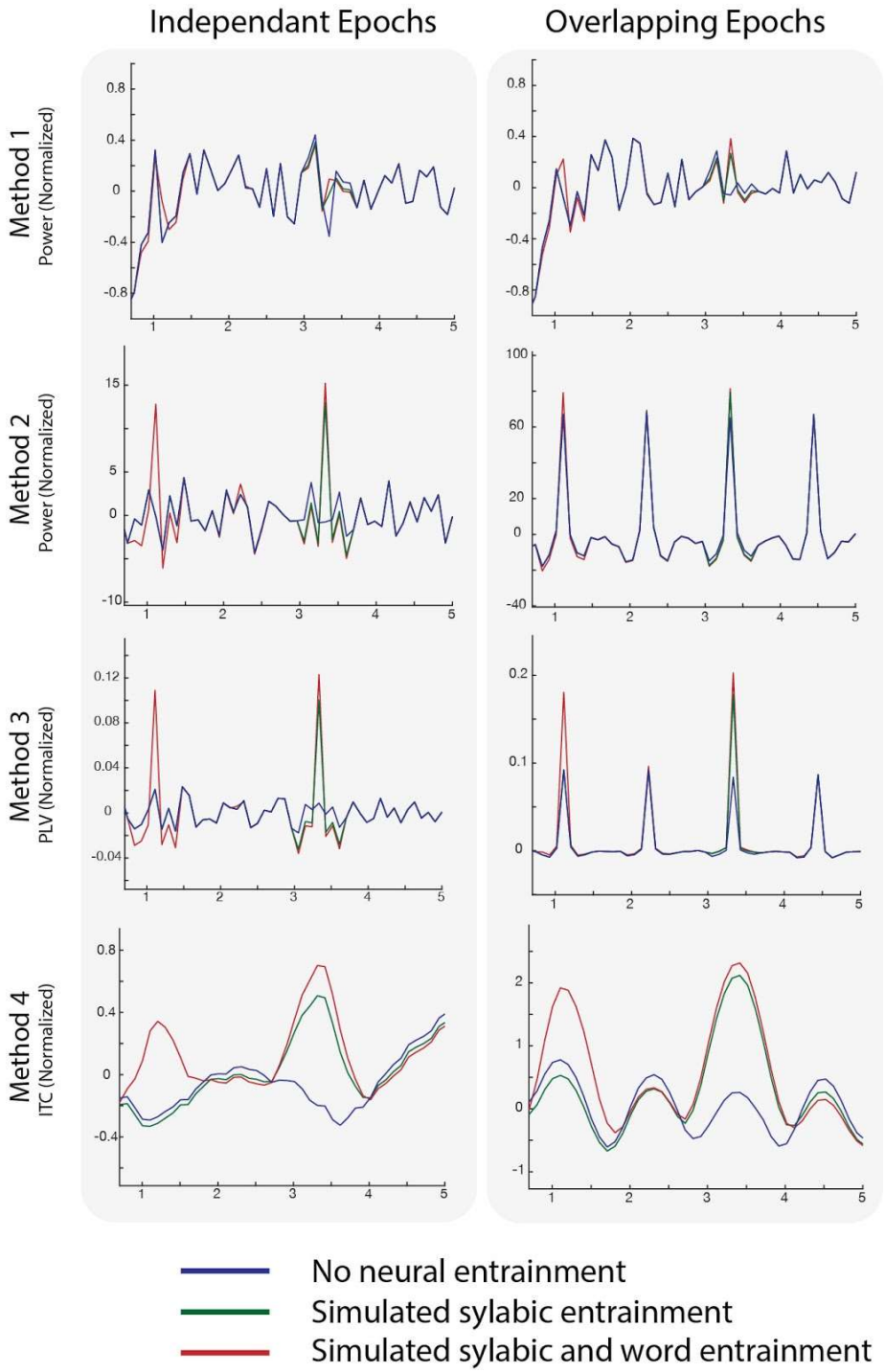


Figure A5. *Results of the analysis with the different methods and simulations. The expected peaks at the syllable (3.33 Hz) and word (1.11) frequencies are clearly distinguishable from the control situation (blue line) when epochs are independent (left column, line 2 to 4). When overlapping epochs are used, artifactual peaks are present (right column). The wavelet method used in method 4 (last line) inadequately smooths the frequency profiles.*

Conclusions

Our simulations show that overlapping epochs unavoidably lead to artifactual entrainment at the overlapping frequency. Nevertheless, this does not mean that overlapping epochs prevent the measurement of brain oscillations. The real neural entrainment adds to the methodological artefact. Thus, the main results of previous works using this approach might be valid especially when there was a difference with a control condition exposed to the same bias. In principle, if an adequate control condition is used, the entrainment derived from brain activity could be quantified. This is the case of one adult's segmentation experiment ***(Batterink and Paller, 2017)***, where the entrainment at the word and syllabic rates during the presentation of a structured stream of syllables was compared with the entrainment during a random stream of syllables. In a second paper, the authors investigated top-down attention's effect by comparing speech segmentation while manipulating the attention to a competitive task in two groups of subjects ***(Batterink and Paller, 2019)***. While there was no within-subjects control condition, we could assume that the methodological artefact is comparable between groups; thus, the comparison across groups would remain valid. In an infant segmentation task experiment using the same methodological approach ***(Choi et al., 2020)***, the authors report an increase in the word rate's entrainment relatively to the syllabic rate entrainment during the familiarization. While the experiment did not have a control condition, in principle, the increase cannot derive from the methodological artefact.

Based on the above considerations, we do not question the data set's validity of Batterink and colleagues and their main results ***(Batterink and Paller, 2019, 2017; Choi et al., 2020)***. However, we believe that overlapping epochs should be avoided because they introduce an artefact that compromises the interpretation of the results, while it does not provide a clear advantage. For data transparency, we encourage the re-analysis of the dataset.

Additionally, the wavelet decomposition has an enormous computational cost compared to the FFT, and the useless time dimension comes with a loss of frequency resolution. Therefore, we see no theoretical advantage in using the wavelet decomposition for a classical analysis of steady-state experiments. For the analysis of classical steady-state experiments, we recommend using FFT on non-overlapping epochs.

NB : The authors responded to this comment and re-analyzed the data following the guidelines we proposed, confirming most of the results they had obtained in their previous studies. ***(Batterink and Choi, 2021).***

Section 2: Higher Order structures and Network Learning

Introduction

IN the previous section, we explored *local* statistical learning regularities in non-responsive populations. We showed how automatic this process was and that both sleeping newborns and comatose patients showed evidence for learning local statistical regularities in auditory sequences. We also explored the limits of statistical learning as a segmentation cue in language-type sequences, showing that for long words (four syllables) , an additional contextual cue is required for successful segmentation of embedded words. As described in the introduction, local transition probabilities between adjacent elements are not the only statistical regularities that human can extract from sequences. In particular, a recent body of work showed human sensitivity to network structures and high order organization of these networks. Among the paradigms proposed to explore this question, the community paradigm is particularly interesting from our point of view, as participants manage to grasp the structure despite all the local transition probabilities being equalized. This paradigm is briefly described in the introduction, and forms the basis of the following studies.

Network science provides powerful tools for describing the structures underlying Markov sequences. In this framework, every stimulus is represented as a *node* and every

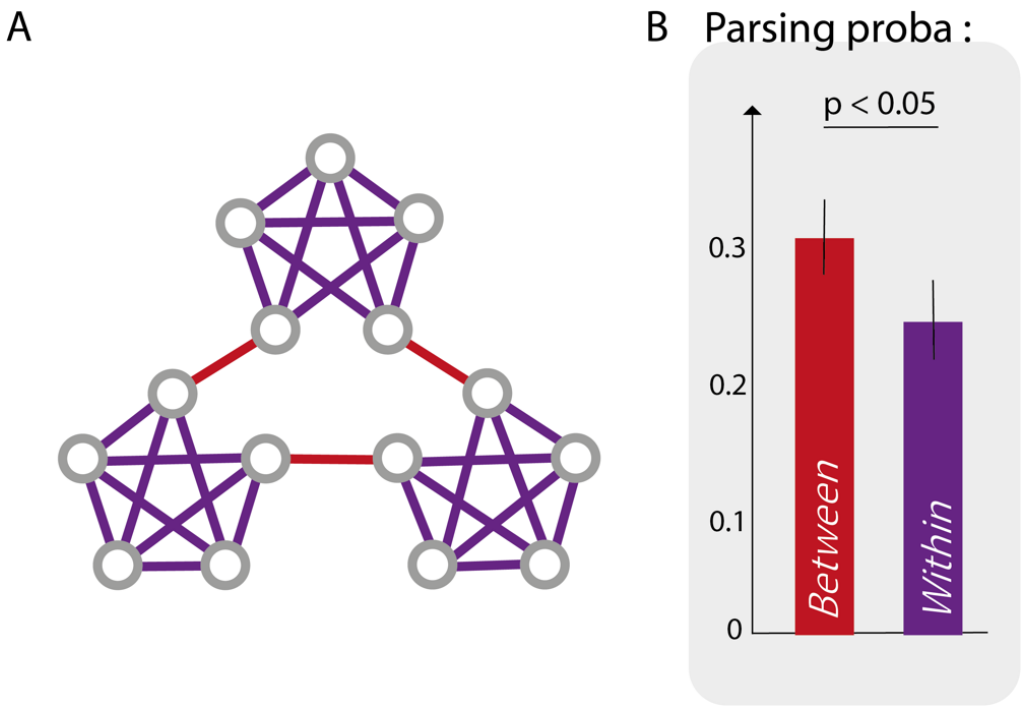


***Figure 0.2: Description of the community paradigm proposed by Schapiro and colleagues (2016). A** : representation of the community structure, each node of the network represent a stimulus and each edge a possible transition between two stimuli. The fifteen stimuli are organized into three communities where all nodes are connected to all the other nodes of the community (except the two nodes at the edge of the community). Only one transition is possible from one community to another. Each node exactly has four neighbors so that local transition probabilities between elements are always ¼. **B** When exposed to a sequence derived of this network and asked to press a button when they feel a natural break in the sequence, subjects pressed more when switching communities compared to staying within a community. As both Within and Between community transitions have the same local transition probability of 1/4 , this shows as sensitivity to the high order structure of the sequence, beyond local properties.*

possible transition from one stimulus to another is presented by an *edge*. In the original community paradigm proposed by Schapiro and colleagues ***(2013)***, a corpus of fifteen visual stimuli was randomly separated into three groups of five stimuli (called a community). All the elements belonging to the same community are linked so that every

possible transition from two stimuli belonging to the same community are possible. However, only two of the elements (at the border of the community) are linked with an element at the border on another community. To maintain the same number of possible transitions from each stimulus, the transition between the two border elements of a community was removed (See Fig 0.2 A for the graphical representation of the community network). By building such a network, all stimuli were associated with four possible neighbors with an equiprobable transition probability of ¼. Crucially, the transition *within* a community (between two elements belonging at the same group) and *between* communities (between two border elements of two groups) are associated with the same local transition probability. If computing transition probability between adjacent elements is the only statistical regularity human can grasp, the community structure should remain undetected by subjects. In particular, transition *within* and *between* communities should be equally perceived by participants. To test this hypothesis, they exposed subjects to a sequence of input following a random walk (and controlled with Hamiltonian walk) in this network and simply asked participants to press a button when they feel a natural break in the sequence. Interestingly, participants pressed the button significantly more after a change of community (*between* community transition) compared to transitions staying inside the same community (*within* community transition) – see fig 0.2 B. As this result cannot be accounted by local statistical learning (the two types of transition shared the same local property), it revealed that humans were sensitive to high order statistical properties such as clustering properties in network structures.

Thus, humans are sensitive to local and higher-order statistical properties in sequences. However, it remains unclear whether these two abilities are supported by similar cognitive and brain mechanisms, or whether these two lines of research builds upon different cognitive properties. We will explore this question in the second and third sections of this thesis. In the second section, we propose a variation of the community

paradigm in which we mix local and high order statistical properties, and in the last section, we argue for the possibility of a unified mechanism by reconsidering with a common model previous published results studying statistics at different scales and explained with distinct models.

Chapter 3 : Humans parsimoniously represent auditory sequences by pruning and completing the underlying network structure

This work has already been published in eLife under the reference :

Benjamin L, Fló A, Al Roumi F, Dehaene-Lambertz G (2023) **Humans parsimoniously represent auditory sequences by pruning and completing the underlying network structure** eLife 12:e86430.

Abstract : Successive auditory inputs are rarely independent, their relationships ranging from local transitions between elements to hierarchical and nested representations. In many situations, humans retrieve these dependencies even from limited datasets. However, this learning at multiple scale levels is poorly understood. Here we used the formalism proposed by network science to study the representation of local and higher-order structures and their interaction in auditory sequences. We show that human adults exhibited biases in their perception of local transitions between elements, which made them sensitive to high-order network structures such as communities. This behavior is consistent with the creation of a parsimonious simplified model from the evidence they receive, achieved by pruning and completing relationships between network elements. This observation suggests that the brain does not rely on exact memories but on a parsimonious representation of the world. Moreover, this bias can be analytically modeled by a memory/efficiency trade-off. This model correctly accounts for previous findings, including local transition probabilities as well as high-order network structures, unifying sequence learning across scales. We finally propose putative brain implementations of such bias.

"The fact then that many complex systems have a nearly decomposable, hierarchic structure is a major facilitating factor enabling us to understand, describe, and even "see" such system and their parts"

H.Simon, The architecture of complexity (1962)

Introduction

To interact efficiently with their environment, humans have to learn how to structure its complexity. In fact, far from being random, the sensory inputs we face are highly interdependent and often follow an underlying hidden structure that the brain tries to capture from the incomplete or noisy input it receives. For instance, ***Tenenbaum et al (2011)***, proposed that learning implies building the simpler underlying relational model which can explain the data. Indeed, evidence suggests that humans can infer structures from data at different scales, ranging from local statistics on consecutive items ***(Saffran et al., 1996a)*** to local and global statistical dependencies across sequences of notes ***(Basirat et al., 2014; Bekinschtein et al., 2009)*** or more high order and abstract relationships such as pattern repetitions ***(Barascud et al., 2016)***, hierarchical patterns and nested structures ***(Dehaene et al., 2015)***, networks ***(Garvert et al., 2017; Schapiro et al., 2013)*** and rules ***(Maheu et al., 2020)***.

At first, the extraction of local regularities in auditory streams was proposed as a major mechanism to structure the input, available from an early age since ***Saffran et al (1996a)*** showed that 8-month-old infants can use transition probabilities - $P(E_t|E_{t-1})$ - between syllables to extract words from a monotonous stream with no other available cues. Since then, the sensitivity of humans to local dependencies has been robustly demonstrated in the auditory and visual domain ***(Fiser and Aslin, 2002)*** without the focus of attention ***(Batterink***

and Choi, 2021; Batterink and Paller, 2019; Benjamin et al., 2021) and even in asleep neonates ***(Benjamin et al., 2022b; Fló et al., 2022a)***. Moreover, it is not limited to adjacent elements but can be extended to non-adjacent syllables - $P(E_t|XE_{t-2})$ - that could account for non-adjacent dependencies in language ***(Peña et al., 2002)***.

However, the computation of transition probabilities (TP) between adjacent $P(E_t|E_{t-1})$- and non-adjacent elements - $P(E_t|XE_{t-2})$ - seems too limited to allow the extraction of higher-order properties without an infinite memory that the human brain does not have. Network science - an emerging interdisciplinary field - thus proposed a different description to characterize more complex streams ***(Lynn et al., 2020a)***. In this framework, a stream of stimuli corresponds to a random walk in the associated probabilistic network. Several studies used this network approach to investigate how humans encode visual sequential information ***(Garvert et al., 2017; Mark et al., 2020)***. Shapiro and colleagues ***(Schapiro et al., 2013)*** tested human adults with a network consisting of three communities (i.e. sets of nodes densely connected with each other and poorly connected with the rest of the graph - ***(Newman, 2003)***) where transitions between all elements were equiprobable (each node had the same degree). This community structure is an extreme version of the communities and clustering properties that are often found in real-life networks, whether social, biological or phonological ***(Girvan and Newman, 2002; Karuza et al., 2016; Siew, 2013)***. The authors reported that subjects discriminated transitions between communities from those within communities. Since local properties (TP) were not informative, this result revealed participants' sensitivity to higher-order properties not covered by local probabilistic models. This sensitivity seems already to be in place at 6y-o ***(Pudhiyidath et al., 2020)*** . Recently, Lynn and colleagues ***(2020a)*** replicated a similar effect with a probabilistic sequential response task. They presented subjects with sequences of visual stimuli that followed a random walk into a network composed of three communities. After each stimulus, subjects were asked to press one or two computer keys, and their reaction time

was measured as a proxy of the predictability of the stimulus. To explain the response pattern, the authors proposed an analytical model that optimizes the trade-off between accuracy and computational complexity by minimizing the free energy function. This model allows taking into account the probability of memory errors in the computation of the transition probabilities between the elements of the stream. From now on, we will refer to this model as the Free-Energy Minimization Model (FEMM: Model D, explained below).

In this paper, we aim to merge these two lines of research and validate a model that can explain how humans learn local and high-order relations simultaneously present in sequences generated from noisy or incomplete structures. Moreover, we propose that adults do not encode the exact input but a parsimonious version based on the generalization of the underlying structure. To this end, we leveraged the community network framework and adapted it to expose adult participants to rapid sequences of sounds that followed a random walk through a network, building on the studies described above ***(Lynn et al., 2020a; Schapiro et al., 2013)***, but using sparse communities with missing transitions between elements of the same community (see fig 3.1). This design allows investigating whether participants are able to complete the network according to the high-order structure or if, on the contrary, they rely on local transitions and reject impossible transitions ignoring the high-order structure. In other words, after training with an incomplete network, if new ("unheard") transitions are presented, are participants more willing to accept them if they belong to the community (i.e., within-community transitions) than if they occur between communities? Moreover, while several papers have studied network learning in the visual domain ***(Karuza et al., 2019; Lynn et al., 2020a; Schapiro et al., 2013)***, to our knowledge, it has never been tested in the auditory domain despite the better statistical learning capacities in the auditory modality ***(Conway and Christiansen, 2005)***, the sophisticated auditory sequence processing abilities observed in

humans compared to other primates ***(Dehaene et al., 2015)*** and their potential importance in language acquisition. In addition, the original design was at a very slow rate, allowing for possible conscious decision to take place on the adequation of each element of the sequence to the structure. Here, we used a 4Hz presentation, typically used in auditory sequence learning tasks, in order to force rapid processing of each element of the sequence and to be more comparable to the sequence learning literature. Finally, we compared how the different models proposed in the literature might fit our data and proposed a unified hypothesis of how any structure (local or global) might be extracted from a sequence.

For this purpose, we tested three different experimental paradigms in an online task, using sequences of pure tones or of syllables (~240 adult participants tested in each paradigm). The first paradigm - Full Community – tested a network composed of two communities of 6 elements each, with all nodes within a community connected with each other (except two nodes at the border of the community to keep an equal degree for each node). In the second and third paradigms, the communities were incomplete, some connections being never presented during the exposure to the continuous sequence: In the Sparse and High Sparse Community paradigms, respectively one and two possible edges for each node were removed. The performances in these two "sparse" designs, relative to the full-community design, are crucial to investigate the participants' underlying representations of the sequences. In each paradigm, participants were first asked to carefully listen to a continuous sequence for about 4 mn and then to press a key when they felt there was a natural break in the sequence (~2mn). This task allowed measuring participants' ability to parse the sequence and to compare their performances in the auditory domain with those published in the visual domain. In a following test phase, they were asked to choose between two isolated quadruplets, the most congruent with what they had heard before, during the familiarization sequence. With this test phase, we could present previously

unheard transitions ("new transitions") and study whether participants were able to generalize the network structure (fig 3.1), notably in the two incomplete networks (sparse and high sparse paradigms). These two tasks were done twice.

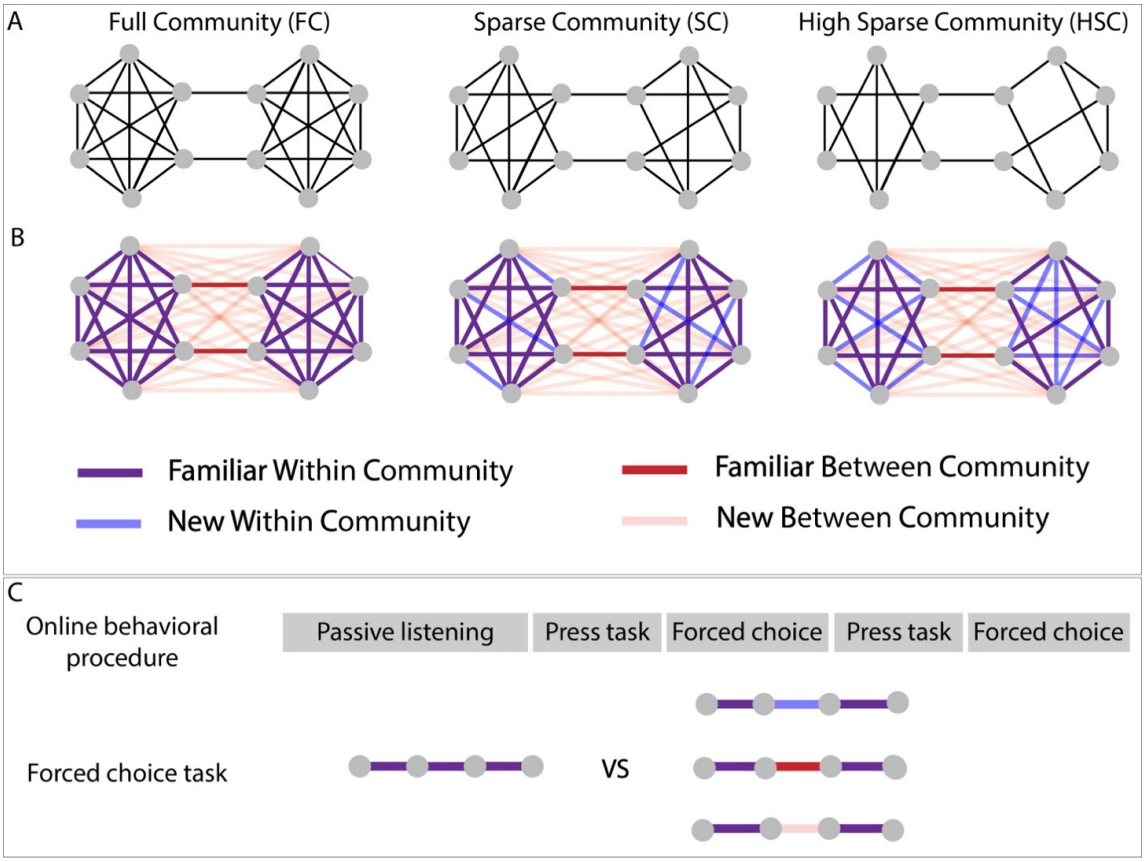


Figure 3.1 *: A: Graph structure to which adult subjects were exposed in three different paradigms B: Graph design with color coded conditions. Blue and pink lines represent transitions that have never been presented during the stream presentation but only during the forced-choice task. C: Test procedure used for behavioral testing. In the press task phase, participants had to press a key when they felt there was a natural break in the sequence. In the forced-choice task, they had to choose between two quadruplets, the most congruent with the sequence they had heard. In the proposed pair, one was always a Familiar Within Condition transition (purple transitions), and the other, one of the three other conditions.*

In the forced-choice task between the isolated quadruplets, we tested each other conditions against the *Familiar Within Community transitions* (condition considered as the reference (fig 3.1 C)). If participants did not learn the graph structure of the sequence, they had to be random in their familiarity choice between *Familiar within* and *between Community* transitions because all quadruplets have been presented and had the same local transition probabilities between their elements. By contrast, if they had indeed learned the graph, their familiarity score should be below 50% denoting their preference for the *Familiar Within Community transitions* (i.e. reference). The performances for the unheard transitions, which can be either within or between community transitions (i.e. *New Within Community* condition and *New Between Community* condition) relative to the reference should allow to separate the different models proposed in the literature to explain how structures are perceived. Therefore, we compared the participants' behavior (i.e. their familiarity rating for the presented transition relative to the reference) to the predictions of different theoretical models proposed in the stream processing and graph learning literature (fig 3.2).

- ***Model A: Transition Probabilities (TP) and Ngrams:*** Local transitions between consecutive elements - $P(E_t|E_{t-1})$ - have been proposed as an efficient learning mechanism to structure streams of input. We tested the limits of this simple local learning computation in the presence of a high-order structure. Ngrams are similar to TP but take into account n previous items in the computation of the transition. For example, for trigrams, $P(E_t|E_{t-1}E_{t-2}E_{t-3})$. Note that because our designs are random walks into Markovian networks, the transition probabilities and Ngram models are identical, $P(E_t|E_{t-1}E_{t-2}E_{t-3}) = P(E_t|E_{t-1})$. Chunking based models, such as PARSER ***(Perruchet and Vinter, 1998)*** rely on the repetition of chunks of

consecutive elements and, as TP and ngrams, would reject any chunk with new transitions as they never occurred during familiarization.

- ***Model B: Non adjacent TP:*** This metric is similar to the transition probabilities but on non-consecutive items $P(E_t|XE_{t-2})$**.** We included it in our analysis because several studies have shown human sensitivity to such properties in streams ***(Peña et al., 2002)*****.**
- ***Model C: Graph Communicability:*** This model comes from the network science literature and computes the relative proximity between nodes in the network, making it sensitive to cluster-like structures like communities. Interestingly, a recent study shows that this measure correlates with fMRI data ***(Garvert et al., 2017)*** suggesting a potential relevance in human cognition.
- **Model D: Free Energy Minimization Model (FEMM):** This model, recently proposed by Lynn et al ***(2020a)*** to account for community sensitivity by humans, is a trade-off between accuracy and computational complexity. It can be explained by memory errors while computing Transition Probability (TP) between elements in a stream. Participants exposed to a stream of elements reinforce the association between element i and i-1. However, errors in this process may lead participants to sometimes bind element i with element i-2, i-3, i-4…. with a decreasing probability (for a full description of the model see ***(Lynn et al., 2020a)***. Mathematically, the distribution of the error size that minimizes the free energy function is a decreasing exponential (Boltzmann distribution). Therefore, the estimated mental model of transition probability is biased compared to the streams' objective transition probabilities enabling participants to encode high-order structure. In more detail, the mental model is a linear combination of the transition probability matrix (A) and non-adjacent transition probabilities of every order $(A^{\Delta t})$

with a weight of $P(\Delta t)$ where Δt is the order of non-adjacency (or size of the memory error, ie $\Delta t = n$ corresponds to $P(E_t|X.....XE_{t-n})$). The estimated model can then be written as:

$$\hat{A} = \sum_{\Delta t=0}^{+\infty} P(\Delta t) A^{\Delta t+1}$$

With

$$P(\Delta t) = \frac{1}{Z} e^{-\beta \Delta t}$$

where A is the transition probability matrix of the graph. β was previously estimated to 0.06 in a comparable task with human adults ***(Lynn et al., 2020a)***. We therefore first used this value to test this model on our behavioral data and later confirmed this estimation with our data (see SI).

In the reinforcement learning literature, the hippocampal place cells have been proposed to represent maps of probabilistic future states and reward by encoding *successor representation* instead of positional cognitive maps ***(Dayan, 1993; Stachenfeld et al., 2017)*.** Successor representation has been formally defined as the sum of probabilistic future state and can be written SR = $\sum_{\Delta t} \gamma^{\Delta t} A^{\Delta t}$. This approach is very similar to FEMM with an infinite sum of all power of the transition matrix, pondered by an exponentially decreasing factor. Here the factor is $\gamma^{\Delta t}$ with $0 < \gamma < 1$ and generally $\gamma = \frac{0.85}{\lambda max}$ with λmax the largest eigenvalue of the transition matrix ***(Garvert et al., 2017)***. This approach has been proposed to account for community perception ***(Pudhiyidath et al., 2022)*** but here we only included FEMM in our

study, as the two models are identical with $\gamma = e^{-\beta}$ (with a different constant).

Another metric computing the same property but from a sequence point of view is the Hitting time.

- ***Model E: Hitting Time:*** This metric, also coming from network science, estimates the distance between two nodes in a graph as the average number of edges needed (path length) to move from one node to another during a random walk. Similar to Communicability (model C) and FEMM (model D), it measures a 'proximity' between nodes in a network. To make it more comparable with the other models we computed its inverse value.

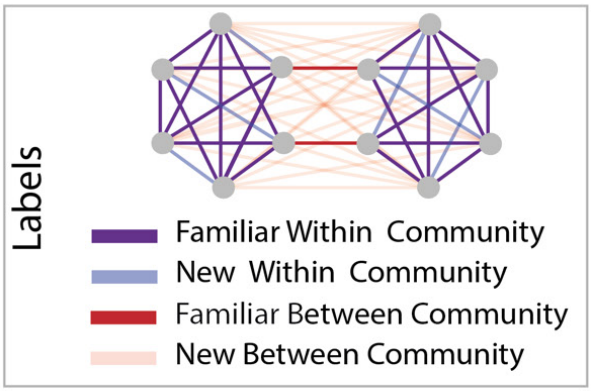
Although the different models are partially correlated with each other, they give different predictions about participants' familiarity responses. First, they were two kinds of local transitions: Familiar transitions and New transitions (TP = 0). Since the TP calculation does not consider the community structure (model A), participants should equally reject new transitions regardless of their relation with respect to communities (New Within Communities = New Between Communities). Second, concerning the new transitions, FEMM and Hitting Time models predict that participants should better detect *New Between Community* than *New Within Community transitions (completion effect).* It is also partly the case for the Communicability model, but not for the TP and non-adjacent TP models (models A and B). The similarity of the predictions of FEMM, Hitting Time, and Communicability models is not surprising as they all describe the same property of the network: proximity between nodes. Intuitively, items from the same community will appear closer together than items from different communities, even if the two nodes are not connected. In fact, FEMM and Communicability are mathematically very close but with a different decay (exponential vs. factorial). However, they can still be differentiated

thanks to the high sparse paradigm were the relative predicted familiarity of *New Within* and *Familiar Between* transitions are different between the two models.

In addition to those theoretical models, we considered two putative brain implementations using biologically realistic neural networks:

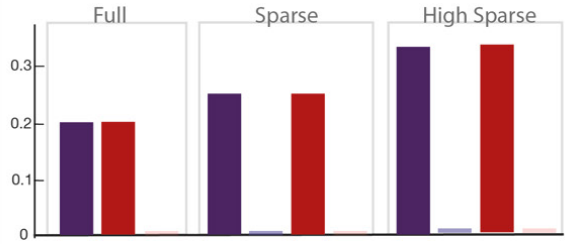
- **Model F: Hippocampus CA1 similarity:** This neural network aims to reproduce the hippocampus structure ***(Norman and O'Reilly, 2003)***, which is often described as a key structure in statistical and structure learning ***(Henin et al., 2021; Schapiro et al., 2017, 2016)***. We compute here the similarity in CA1 layer as it has been proposed to capture community-like structures in previous studies ***(Schapiro et al., 2017)***. Indeed, thanks to its ability to have overlapping representations of the input and direct connection with the entorhinal cortex through the monosynaptic pathway, CA1 structure is also sensitive to long-distance dependencies allowing high order structure learning.
- **Model G: Hebbian learning with decay:** Hebbian learning is a biologically plausible implementation of associative learning. Some neurons fire specifically to some objects in the environment. When two of those neurons co-fire, the pair is reinforced. It has been suggested that learning transition probabilities is based on such a mechanism in the cortex. Here we adapted this idea to implement the FEMM computation instead of TP, specifically by adding a temporal exponential decay in the probability of a neuron firing after a stimulus's presentation. When the exponential decay has the same β parameter as the FEMM, the results of the FEMM and the Hebbian learning with decay are mostly similar.

Theoretical Models



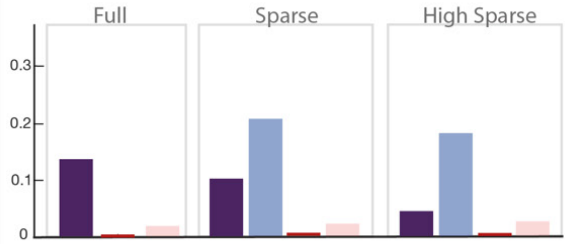
A : Transition probability (& n-grams) Saffran et al 1995

$$\hat{A} = A$$



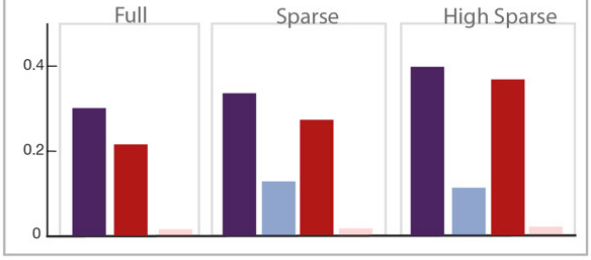
B : Non Adjacent Transition proability Peña et al 2002

$$\hat{A} = A^2$$



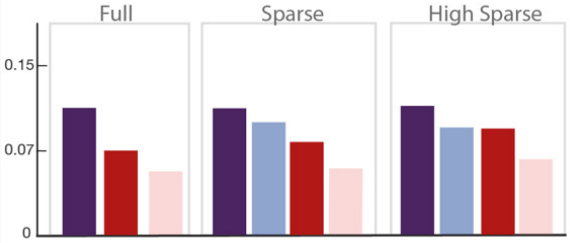
C : Communicability Gavert et al 2017

$$\hat{A} = \sum_{\Delta t} P(\Delta t) A^{\Delta t + 1} \; ; \quad P(\Delta t) = \frac{1}{\Delta t!}$$



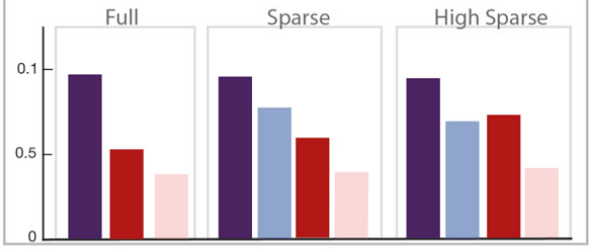
D : FEMM Lynn et al 2020

$$\hat{A} = \sum_{\Delta t} P(\Delta t) A^{\Delta t + 1} \; ; \quad P(\Delta t) = \frac{1}{Z} e^{-\beta \Delta t}$$



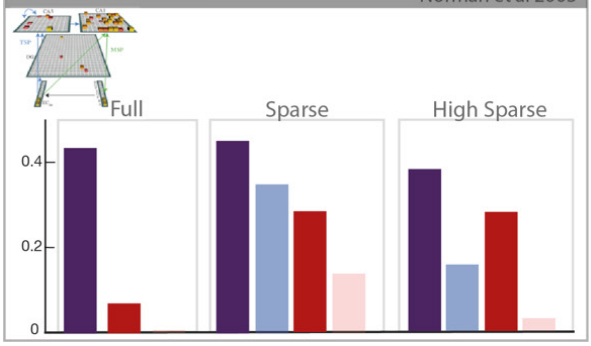
E : Hitting Time

$$\hat{A} = \frac{1}{\mathrm{HT}(A)}$$



Neural Models

F : CA1 Similarity Schapiro et al 2017 Norman et al 2003



G : Hebbian Learning

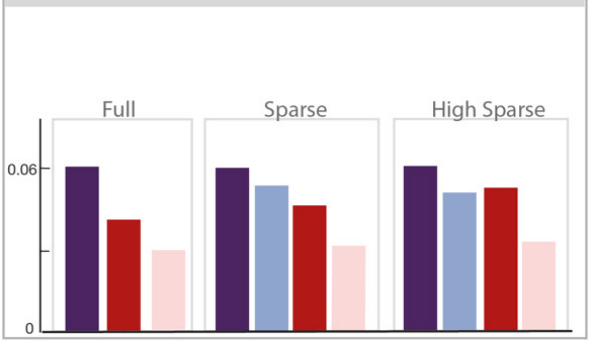


Figure 3.2 *: Model description and predictions for the three paradigms tested. For each model, we computed the estimated familiarity (a.u) predicted for each condition in the Full, Sparse, and High Sparse paradigms. Although the models are partially correlated, they differ in their prediction about the familiarity of New Within Community transitions (light blue) which allow to separate the different models. Model D and E (FEMM and Hitting Time) are two variations of the same sequence property from a statistical modeling or sequential point of view. Their predictions are then almost identical. Models A, B, C, D ,and E are theoretical metrics over the graph structure that predict more or less familiarity with the different types of transitions. Model F and G are biologically plausible neural encoding of those metrics. The box colors correspond to the conditions labeled in the top-left panel.*

Results

Human Behavior

Key presses distribution during active listening

All participants were exposed to a stream of either tones or syllables adhering to one of three possible graphs (fig. 3.1 A&B). After a 4-mn-familiarization period, they were instructed to press the spacebar when they felt the impression of a natural break in the sequence (2 mn). This task was a sanity check to corroborate that participant were listening to the stream and that their performance was comparable to previous studies testing graph learning using the visual modality at a much slower pace than we used here. fig 3.3 top row shows the normalized distribution probability of key presses after a transition, using a kernel approach (see methods for detailed computation). In all three paradigms (each corresponding to a graph in fig 3.1), the significant increase in key presses after Between Community vs. Within Community transitions ($p<0.05$ are indicated in bold

lines), reveals that participants were sensitive to the switch between sound communities. Full Community and Sparse community designs showed a similar effect size, while the High Sparse Community design elicited a small but significant effect. Unpaired t-tests every ms in [-0.1, 2.750] s window, contrasting the Full Community vs. High Sparse Community, show a significant difference between 1 s and 2.6s post-transition ($p<0.05$ Bonferroni corrected). Similarly, Sparse Community vs. High Sparse Community differed between 0.8 s and 2.5 s ($p<0.05$ Bonferroni corrected).

Two-forced-choice task

Participants were given a two-forced-choice task, in which they had to choose between two sequences the one that best matched the structure of the stream they had listened to (fig. 3.1 C). This task is the crucial test for comparing models because it allows to present new transitions that matched, or not, the familiar structure and thus to assess the representation of the memorized graph. We report the results at the end of the learning (second block). Results separated by groups and testing block are presented in SI. It can be seen that in contrast with the three other data points, participants' choice were close to random after the first block in the syllable experiment and their performance could not be explained by any of the models. As pointed in other experiments on statistical learning using syllables ***(Elazar et al., 2022a; Onnis and Thiessen, 2013b; Siegelman et al., 2018)***, the familiarity with speech and the phonetic rules of the native language, create priors on the probability of sequences of syllables, that might compete with the real syllable distribution in the task. At the end of learning, no difference was found between the groups using tones and syllables (unpaired t-test for each condition, all $ps>0.2$), we thus merged the data of the tone and syllable groups.

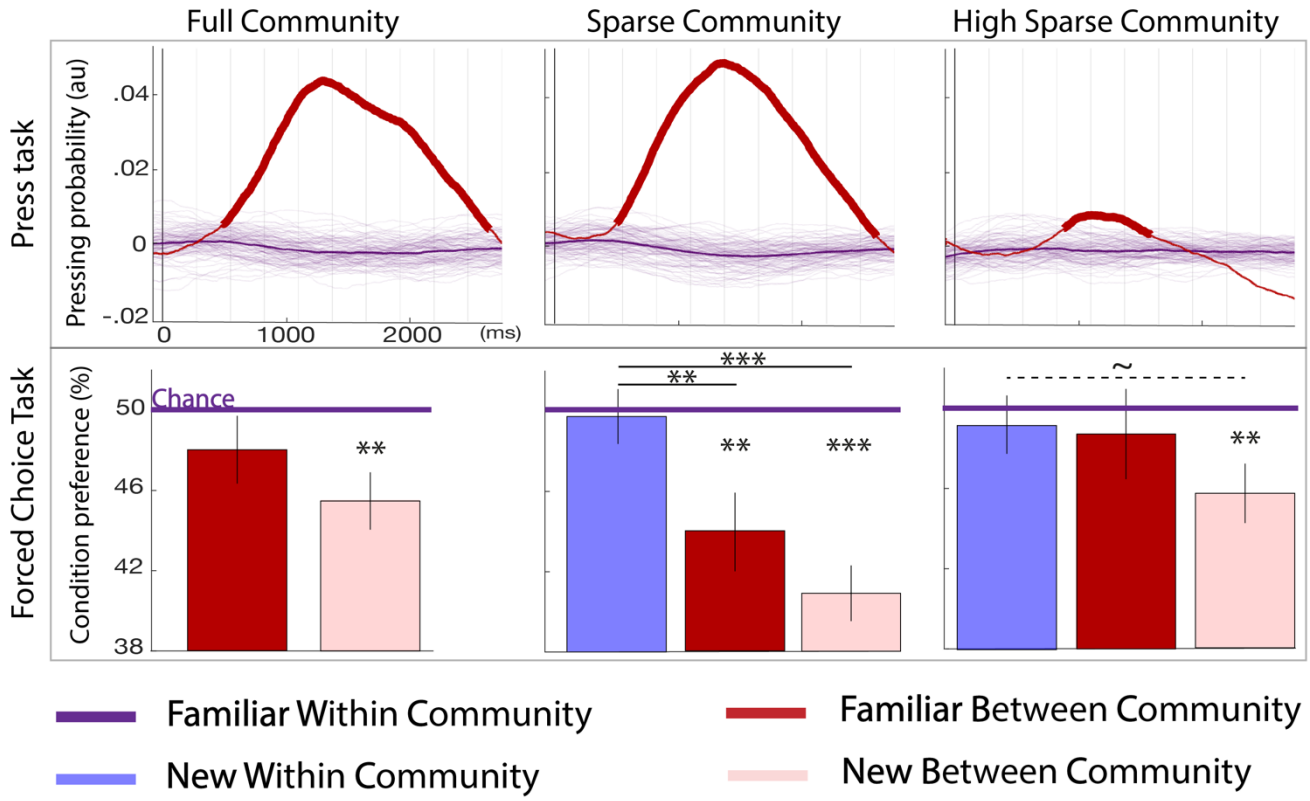


Figure 3.3. *Top panel: parsing probability during the active listening phase (distribution of key presses after the offset of a given transition) purple lines: Familiar Within Community transitions, red line: Familiar Between Community transitions. Thin purple lines each represent a bootstrap occurrence of the parsing probability for the Familiar Within Community transition. The bold red line indicates the time-points where there was a significant increase of parsing probability after a Familiar Between Community transition compared to a Familiar Within Community transition. Bottom panel : Familiarity measure in each paradigm: percentage of responses for each condition during the forced-choice task. By design, the chance level (50%) represents the Familiar Within Community estimated familiarity (reference). The stars indicate significance against the reference and between conditions (pval<0.05 FDRcorr) the dotted line marginal significance (pval = 0.046 uncorr)*

In this task, scores below 50% indicate that the reference (*Familiar Within Community* transitions) was judged more familiar than the tested condition. We

postulated that if participants were only sensitive to familiar transitions, any novel transitions should be judged less familiar than the *Familiar Between Community* transition. On the other hand, if participants encoded the underlying structure of the communities, they should not notice the novelty of the *New Within Community* transitions and reject the two between-community conditions (familiar and new).

As can be seen in fig 3.3, participants significantly rejected the *New Between Community* transitions in each paradigm (ps<0.01 FDR), this transition is both novel and jumping across communities. The *Familiar Between Community* transition condition was only significantly rejected in the Sparse Community paradigm (p<0.01 FDR). Second, the *New Within Community* transitions were chosen/rejected at chance in the Sparse and High Sparse Community paradigms indicating a similar perception of familiarity for these never heard transitions and the reference. Third, in the Sparse Community paradigm, the familiarity score was larger for the *New Within Community* transitions than for both between community transitions (new: p<0.01 FDR; and familiar: p<0.05 FDR). These comparisons were only marginally significant in the High Sparse paradigm (uncorrected p = 0.046). In other words, the participants encoded the graph structure as revealed by the difference in familiarity between within- and between-community transitions and naturally completed the graph as indicated by the scores at chance for never heard transitions compatible with the graph structure.

Which model best fits the participants' behavior

Correlation between human data and theoretical model predictions

To estimate the adequacy of the theoretical models to explain the behavioral data, we pooled together the three paradigms and estimated the correlation with each model. We normalized each model prediction by the model's value for *Familiar Within Community* transitions to be comparable with the behavioral results of the two-force-choice task. It is

worth noticing that models A, B, C, and F predict differences in Familiar Within familiarity between the three paradigms; however, our experimental design does not allow us to estimate differences in these transitions between paradigms but only relative differences to the *Familiar Within Community* condition within paradigms. To estimate the significance of the correlation differences, we used a bootstrapping approach with subjects (with replacement) and estimated the number of bootstrap occurrences in favor of one model against another. Fig 3.4 A shows the correlations' distribution between the data and each model (presented on the diagonal) and between pairs of models. We estimated the significance of the correlation strength between the data and model i or j by counting the percentage of occurrences in which model i had a stronger correlation with the data than model j. All models were significantly correlated with the data (all p< 0.01 FDRcorr), with a correlation strength following the order FEMM ≈ Hitting Time > Communicability > Non-adjacent TPs ≈ TPs (fig 3.4 C). Note that the FEMM and Hitting Time are similar models, and thus predictions are almost identical. They had the best correlation with the data (81%) and were significantly better than all the other theoretical models (p<0.05 FDR).

Correlation between human data and neural model predictions

As the FEMM computation and the Hitting Time were the best theoretical models, we translated them into a realistic biological architecture using Hebbian rules. We estimated this implementation on a 50 000 item-long stream for each paradigm. The correlation between the analytical computation and the Hebbian learning implementation exceeds 99%. Using the same bootstrap approach, we compared this Hebbian approach with a neural network reproducing hippocampus architecture proposed by ***Norman and O'Reilly (2003)***. Both models were significantly highly correlated with the data and with each other.

However, the Hebbian implementation of FEMM was slightly but significantly more correlated to our data than the hippocampus model (fig 3.4) typically because of the lack of agreement between the hippocampus model and the data in the High-Sparse paradigm. However, because the hippocampus model highly fits our data, we cannot rule out the hippocampus as a potential crucial structure for such tasks.

Estimation of the ceiling correlation with our data

We also used the same bootstrapping approach to estimate the noise ceiling for the model fit. For each bootstrap, we randomly selected n subjects with replacement twice and correlated the data of those two random samples. We find an average of 84% correlations as a noise ceiling for those data. Our best fit with any model is the 77% average bootstrap correlation between our data and the FEMM, which is relatively close to the ceiling fit given this dataset, showing a very high relevance of the FEMM to account for the data.

Discussion

Transition probabilities between elements of the sequence are biased by the structure of the underlying generative network.

Our results show that human adults do not encode transition probabilities objectively when familiarized with a stream of sounds. Instead, they seem to have a systematic bias to complete the transitions within a community suggesting a subjective internal representation that differs from the objective distribution of the transitions they heard. This behavior is compatible with two proposed theoretical models: the Free Energy Minimization Model (FEMM) and the Hitting Time.

The high agreement between the Free Energy Minimization Model and the data we observed, suggests that the bias can be analytically estimated using the FEMM $\hat{A} =$

$\sum_{\Delta t=0}^{+\infty} P(\Delta t)A^{\Delta t+1}$ with $P(\Delta t) = \frac{1}{Z}e^{-\beta \Delta t}$. ***Lynn et al (2020)*** proposed that this bias corresponds to memory errors when recalling the previous item of the stream during the TP computation. The bias in the encoding of Transition Probability between successive elements enabled the extraction and encoding of high-order structures in graphs, i.e., a community structure. We can distinguish two distinct bias effects: First, the ***pruning*** of familiar transitions that do not conform to the community structure (i.e., *Familiar Between Community* transitions are rejected). Second, the ***completion*** of the structure by over-generalizing new transitions when they are compatible with the high-order structure (i.e. *New Within Community* transitions are accepted). These perceptual biases lead to a more parsimonious internal representation of graphs.

Putative brain implementation of such computation

We showed that the computation of transition probabilities is biased in humans, and analytically, this bias is characterized as an optimal trade-off between accuracy and computational complexity. Indeed, perfect accuracy in the encoding would result in no sensitivity to the high-order structure, while too low accuracy would result in no learning at all. We also presented putative brain implementations and tested to what extent two previously described mechanisms might explain our results: Hebbian learning and Hippocampus episodic memory.

Hebbian learning is a very simple mechanism that consists of reinforcing co-occurrences in a signal. It has been proposed as a learning mechanism in statistical learning tasks ***(Endress and Johnson, 2021)***. Here, we minimally modified it as described above to introduce the bias in TP computation. Such learning could be implemented in many brain regions through learning-induced synaptic plasticity and does not require any specific structural organization of neurons. In contrast, the CA1 similarity model relies on the specific architecture of the hippocampus. Testing a hippocampus specific model is essential

because several authors have proposed that statistical learning and graph learning might be represented as the construction of an abstract map of relational knowledge, analogous to topographic maps ***(Constantinescu et al., 2016; Garvert et al., 2017)***, which are known to involve the hippocampus. Moreover, the hippocampus has also been proposed as a good candidate for the implementation of the successor representation, giving this structure the role of a predictive map unifying temporal and spatial relational knowledge under a common framework ***(Stachenfeld et al., 2017)***.

A recent experimental study ***(Henin et al., 2021)***, showed that when exposed to statistically organized auditory or visual streams, the hippocampus activity measured with ECoG exhibited a cluster-like behavior, with all elements belonging to the same group being similarly encoded. Using the community paradigm with fMRI, ***Schapiro et al (2016)*** also reported an increased pattern similarity in the hippocampus for elements belonging to the same community (see also ***(Pudhiyidath et al., 2022)***). Another piece of evidence comes from modeling the hippocampus activity in different statistical learning tasks ***(Schapiro et al., 2017)***. In this study, the authors used a neural model mimicking the hippocampus architecture and trained it on different statistical learning tasks including community structure learning. They showed that the pattern of activity in CA1 might account for both pair learning (episodic memory) and community structure learning, and thus is partially consistent with two mechanisms observed in the hippocampus: pattern completion (i.e. the similarity of the neural representations of close stimuli increases, which allows generalization) and pattern separation (i.e. the similarity of neural representation of close stimuli decreases, to disambiguate them) ***(Bakker et al., 2008; Liu et al., 2016; Yassa and Stark, 2011)***.

Here, we showed that both a general Hebbian model and a more specific hippocampal model fit very well the pattern of familiarity scores given by the participants with a slightly better result, yet significant, for the Hebbian learning approach. Since we only have

behavioral results, it is difficult to conclude on the exact brain regions involved, especially since recent work proposed the joint use of several computation involving cortical and hippocampal learning in similar tasks ***(Varga et al., 2022; Whittington et al., 2020)***. In any case, the agreement between the behavioral data and two brain models shows that the FEMM model (an analytical model), does not only explain behavioral data but also has biologically valid candidates.

A general model of statistical learning for sequence acquisition

Statistical learning has been proposed as a powerful general learning mechanism that might be particularly useful in language acquisition in order to extract words from the speech stream ***(Saffran et al., 1996a)***. However, the exact model explaining statistical learning remains under-specified: What is computed remains unclear ***(Fló et al., 2022a; Henin et al., 2021)*** and authors often tailored the computation to suit the paradigm (transition probabilities in some studies, non-adjacent or backward transition probabilities in others, biased transitions probabilities in network studies...). We argue that the Free Energy Minimization Model (FEMM) is a more general model that, beyond explaining community separation, as shown above, can also account for results traditionally explained by the computation of local transitional probabilities and those that require the computation of long-distance dependencies. Indeed, the first-order approximation of the Free Energy Minimization Model corresponds to the objective Transition Probabilities model ($\widehat{A_0}$ see SI). Thus, the predictions of the FEMM model are the same as those of the transition probability model in many tasks, notably in classical speech segmentation experiments, where a drop in TP signals word edges ***(Saffran et al., 1996a)***. Another approach in the literature about sequence learning, considers the recognition of chunks more than statistical learning as a primary mechanism for segmenting sequences. Based on this approach PARSER and TRAXCS, detect often occurring chunks in sequences but do not

associate a familiarity rating with each transition. In a previous experiment ***(Benjamin et al., 2022b)***, we showed that familiarity based on statistical learning does not always lead to sequence chunking and here we focused on this sense of familiarity which does not require the construction of a repertoire of possible chunks postulated by chunking models. Therefore, we did not consider these models here.

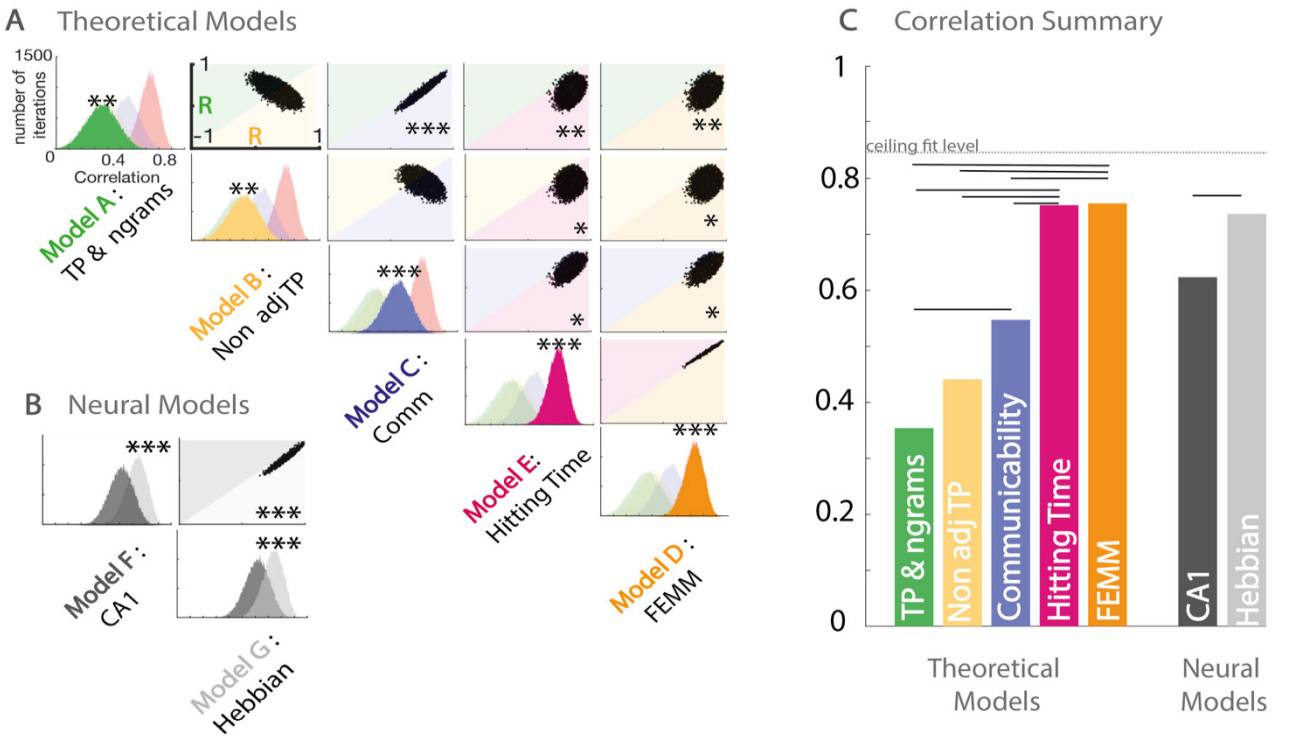


Figure 3.4 : *A: Estimation of the correlation of the participants' familiarity score pattern with each theoretical model (A to E) using bootstrap re-sampling. The diagonal of the matrix displays the distribution of correlations between the participants' familiarity pattern across conditions and the predictions generated by each model (A) theoretical models A-E and (B) neural models F&G). Each panel of the diagonal presents the same result, the color of the relevant model being highlighted to facilitate the comparison between models. For each pair, the significance between models (indicated by stars) is estimated by counting the number of bootstrap occurrences for which one model was more correlated with the data than the other. We plotted this bootstrap as a cloud of dots in the Correlation with Model1 x Correlation with Model2 subspace. Significance is then represented by the percentage of dots above the diagonal. Models with similar predictions display a line style cloud of dots aligned along the diagonal. B: We did the same comparison with the two neural models (F&G). C: Summary of the correlations between each model and the behavioral data. Plain lines above the boxes indicate the significant differences between models. FEMM and Hitting Time (D&E) are equivalent and equally good and significantly better than all other theoretical models. For neural models, the Hebbian model (G) shows a slight, but highly significant, better fit with the participants' scores. The dotted line indicates the ceiling fit level estimated for this dataset.*

Another part of the statistical learning literature focuses on AxC structures, in which the first syllable of a triplet predicts the last syllable ***(Buiatti et al., 2009; Endress and Johnson, 2021; Kabdebon et al., 2015; Marchetto and Bonatti, 2015; Peña et al., 2002)***. The computation of first-order TPs is insufficient to solve this task, which requires the encoding of non-adjacent TPs. However, a bias estimation of TPs following the FEMM is sensitive to non-adjacent dependencies and can explain the emergence of AxC structures. Additionally, as previous papers and our results show, the FEMM can also explain subjects' behavior in different kinds of network learning ***(Karuza et al., 2016; Lynn et al., 2020a; Schapiro et al., 2013)***. ***Lynn and colleagues (2020a)*** interpret the FEMM as errors in the associations between elements, whose probability decays with the distance between associated elements. We proposed that implementing the TPs computation through Hebbian learning with a firing decay results in a comparable computation to the Free Energy model.

Finally, a similar Hebbian learning approach enables to explain the sensitivity to backward TP reported in the literature ***(Endress and Johnson, 2021; Pelucchi et al., 2009a)***. A similar idea has recently been proposed by Endress and Johnson ***(2021)***. However, the authors did not refer to free energy optimum or provide an analytical approach. Instead, they proposed a Hebbian learning rule with the same idea of mixing TP with non-adjacent TP (which corresponds to a second order approximation of the Free Energy Minimization Model that we propose here, see $\widehat{A_1}$ in SI). Like we do here, they argued that this mechanism could account for results currently explained by different models in the literature. Thus, the FEMM model and its putative neural implementation through Hebbian rules unifies different proposals concerning statistical learning on the one hand and network learning results on the other hand, under a common principle. It is important to note that we investigated how the FEMM model –and the other models– account for the extraction of regularity from a sequence, which is the first needed step of many other processes. We

did not test for further abstract representations of the sequence that could be subsequently computed.

Information compression and stream complexity

Our results showed that adult humans have a biased subjective representation of first-order transition probabilities compared to the actual transition probabilities, which makes them to be sensitive to high-order structure in the underlying graph and to overgeneralize transitions that they never experienced. What is the advantage of such a computational bias for human cognition? We postulate three main advantages.

Higher-order structures and generalization can be relevant information to learn. Unlike random networks, many real-world networks have transitivity properties ***(Girvan and Newman, 2002; Newman, 2006, 2003)*** - if A is connected to B and B to C, there is a high chance for A and C to be connected (a friend of my friend is likely to be my friend).

Overgeneralizing enables faster learning. Overgeneralizing means accepting transitions congruent with the structure even before they appear in the stream. Thus, for short exposures, the estimation of the Free Energy Minimization Model is closer to the real transition probability matrix than the estimation of the Transition Probability model based on the input because it infers transitions that have not been presented yet. This fast learning might be of importance, for example, for language acquisition, given that human infants are exposed to a limited amount of speech.

Adding to why humans have biased statistical learning, we propose that this learning bias in extracting statistical information might subsequently be used to form abstract condensed network representations. In fact, the extraction of high-order structures might enable information compression in long term memory. Because of the computational cost and the pressure on memory to encode long sequences, compressing information is a major advantage. In a community paradigm, the learned representation could be later

simplified to reduce the stream complexity to a binary sequence with a certain probability of changing between communities A and B (fig 3.5). Instead of remembering all the transitions of the stream, remembering community labels and the probability of transition between communities is sufficient. Recent data ***(Al Roumi et al., 2021; Dehaene et al., 2014; Planton et al., 2021; Sablé-Meyer et al., 2022, 2021)*** showed that in some circumstances, humans' performances were highly sensitive to input compressibility, arguing for a condensed encoding of inputs. Note that the familiarity measure we report here does not show compression of the structure. Still, the familiarity bias could be at the basis of a later abstract condensed network representation (this hypothesis is presented in fig 3.5). In the same line, a recent study using a graph perspective ***(Whittington et al., 2020)*** proposes that the representation of the abstract relational structure of a sequence and the mapping between node and stimuli identity could be factorized. In the case of community paradigm, ***Pudhiyidath et al (2022)*** even proposed that the formation of such an abstract structure could allow humans to transfer learnt properties between elements belonging to the same community***.*** ***Mark et al (2020)*** showed that the learning of the structure of a network could be re-used on the next day to allow fast and generalizable learning arguing for a factorized brain representation between the stimuli mapping and the abstract network encoding. This compressibility hypothesis, represented in fig 3.5, needs formal testing to be confirmed or infirmed.

Finally, the human sensitivity to community is in line with Simon's postulate that the complexity of a system can only be handled thanks to its hierarchical nearly decomposable property ***(Simon, 1962)***. In other words, a complex structure is no more than the sparse assembly of less complex dense substructures. Here we propose empirical arguments by demonstrating that human adults are sensitive to the decomposition of a complex network into two simpler sub-networks.

Methodological remarks

In this study, we used two different metrics. The press bar task during attentive listening showed high sensitivity, but it only allowed testing within vs between community transitions during learning and thus assessing clustering (different perception of *Familiar*

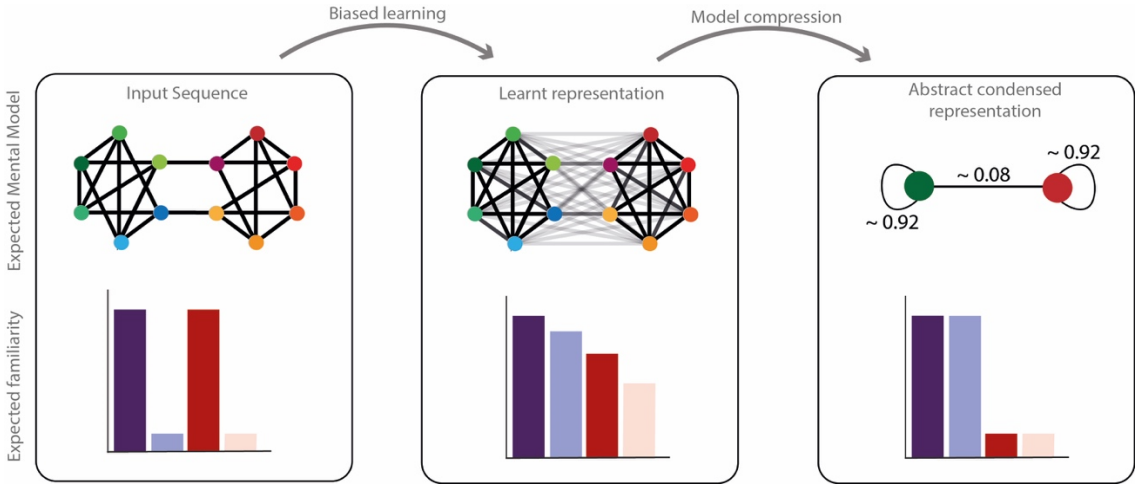


Figure 3.5 : Compressibility hypothesis*. In the left panel, the real underlying structure of the input presented. In the middle the learned representation by humans. As described above, this representation does not completely reflect the real input structure but a biased parsimonious version of it, including pruning and generalization of transitions. In the right panel, we hypothesized a condensed representation that might be formed subsequently to simplify and compress the information. In this representation, the identity of the elements would be ignored in favor of their community label. The familiarity of each transition is represented with transparency of the edge in the network representation and each condition familiarity pattern is represented with barplots below.*

Within and *Familiar between* transitions). The forced-choice task on the isolated quadruplets allowed testing for more conditions after learning and thus to distinguish between models. However, this second metric had a low sensitivity because only a few trials could be collected resulting in high error variance that was compensated by a very large sample of participants (N=727).

This design also did not allow us to efficiently study the dynamics of learning. We had only two points for the estimation of the learning of the graph by explicitly detecting

quadruplets familiarity. This is particularly insufficient when, as here, the speech or non-speech nature of the stimuli modulate performance because of different priors on the possible composition of the sequences. Even for tones, we could not determine when learning took place as it seems stable from the first measure point.

Conclusion

The results shown in this study reveal 1) Community representation in the auditory domain; 2) The persistence of a biased, subjective transition probabilities representation after learning; and most importantly, 3) pruning and completion effects allowing to build a parsimonious representation of the underlying network structure. Transition probabilities are thus not exactly encoded by the participants but biased in a way that can be predicted by the Free Energy Minimization computation. Importantly, the same model might explain human sensitivity to local and high-level regularities without the need for specific models for each task.

More research is needed to characterize how and where such computations take place in the human brain and how this bias varies across individuals and with development. However, Hebbian rules in the cortex and/or hippocampus might be plausible candidates for a biological implementation of this analytical model. Finally, finding appropriate metrics to cluster graphs is a current research topic in applied mathematics ***(Newman, 2006)***. Thus, we believe that understanding the cognitive processes at stake when humans are exposed to such structured networks might provide insight to cognitively and biologically plausible computations.

Materials and Methods

Behavioral Task

Participants

A total of 727 French adults were recruited via social media (424 of which were retributed 2.5$ on Prolific platform). They had to have no hearing or language problems and French had to be their first language. They were assigned to one version of the experiments and instructed to carefully listen for 4.4 minutes to a nonsense language composed of nonsense words that they had to learn because they would have to answer questions on the words afterward. Participants were either exposed to the Full Community (N= 250), the Sparse Community (N= 249), or the High Sparse Community (N= 228) paradigms with either pure tones or syllables as stimuli.

Ethic approval

All participants gave their informed consents for participation and publication and this research was approved by the Ethical research committee of Paris-Saclay University under the reference CER-Paris-Saclay-2019-063

Stimuli

We generated twelve tones of 275ms duration, linearly distributed from 300 to 1800 Hz. We also generated syllables with the same duration and flat intonation using the MBROLA text-to-speech software ***(Dutoit et al., 1996)*** with French diphones. There was no coarticulation between syllables.

Each experiment was composed of 4.4 minutes of an artificial monotonous stream of concatenated tones (or syllables) without any pause, resulting from a random walk into the tested graph. The graph was either complete (full community), with one missing transition (sparse community) or two missing transitions at each node (high sparse

community) creating three experimental paradigms. To avoid any putative acoustical bias, we collected 8 groups of subjects for each paradigm. For each of the 8 groups, we randomly generated a new graph (except for the full community graph, for which only one graph was possible), a new correspondence between the alphabet of tones and the nodes of the graph and finally new random walks into the graph.

In the original study ***(Schapiro et al., 2013)***, the authors explored different graph traversal: Random Walk and Hamiltonian Path. In the Hamiltonian path, each node is presented only once, avoiding short distance repetitions and thus controlling for a putative novelty effect when there is a change of community which could potentially serve as a parsing cue in a random walk. However, participants did not parse the sequences better in the case of Random walks relative to Hamiltonian walks (figure 2 in ***Shapiro et al, 2016***) minimizing the concern of a possible habituation effect if random walks are used. Here, we chose a Random walk because the Hamiltonian path introduces more predictability to the sequence. As previously presented stimuli of the community can no longer be presented, the predictability of the next element increases with the length of the path within a community until a perfect predictability for the 5th and 6th elements (node at the border of communities) and the next element in the other community whereas a random walk keeps the prediction flat. Thus, learning a graph through a Hamiltonian walk can be fully explained with n-gram approaches and cannot disentangle the different learning models proposed. Moreover, the number of Hamiltonian paths available drastically decreases with sparsity up to the point where, in the high sparse paradigm, a single sequence is possible of a given first element leading to a trivial pattern of repetition of twelve elements.

With a random walk, the tones belonging to the same community are presented on average closer in time than those belonging to different communities. However, the length of the walk within one community can be short without repetition or without going

through all the tones of the community, or longer with repetition of some tones at a random distance. Therefore, there is no consistency over time that could allow to capture a repetition pattern. Furthermore, the absolute frequency of each tone is equal within the stream, which avoids long-term habituation effects, and the local transition probability is flat, which avoids the possibility of predicting the next tone. Finally, the tones frequency was distributed between the two communities, to prevent a separation based on an auditory spectral partition. However, due to the design reasons explained before, Halmitonian walks are not usable and thus we could not formally control for potential habituation effect in our design. The key-press results of this study (but not the 2-forced choice results) are therefore potentially subject to confounding by habituation .

For the isolated quadruplets, we concatenated four sounds so that the first and last transition were always non deviant (Familiar Within Transition) but that the transition in the middle would be of each type of transition. We used quadruplets in this study for consistency with previous work of the team and especially for comparing latencies of developmental ERPs in possible future electrophysiological work.

Procedure

Participants started with a 4.4-mn familiarization phase of exposure to the stream (960 items). Then learning was tested with two tasks. First, participants were told that the order of the tones/syllables was not random and that they had to press the spacebar when there was a noticeable change in the tones (or syllables) group used in the stream. Second, they were presented with a two-forced-choice task in which they had to choose between two quadri-elements sequences, the most likely sequence, part of the language they learned. The two-forced-choice trials always comprised a *Familiar Within Community* transition and one representing the other conditions. These conditions were *New Within Community t*ransitions*, New Between Community* transitions, and *Familiar Between Community*

transitions (fig 3.1). Participants were exposed to 8 trials per type (with different sounds each time) except for the *New Within Community* type, where they were only exposed to 4 trials because, by design, there are only 4 of those transitions in the graphs. Each transition used in the set was presented in both directions (AB and BA). Four catch trials were also included to control participants' engagement in the task. These catch trials were two consecutive identical quadruplets that subjects had to detect. Then, they were again exposed to a random walk stream for 2.2 min (active listening – 479 transitions) followed by the same forced-choice task as before.

Data Processing: Active listening task

Participants who pressed less than 10, or more than 200, times during the experiments were excluded from further analysis (FC: 52/250; SC: 24/249; HSC: 23/228). A null array of the stream size was built and filled with ones at times when participants pressed the space-bar (Dirac impulses). To convert it into a continuous signal, we convoluted it with an exponential window. Then, we epoched this continuous signal from -2.75 to 2.75 seconds after each transition's offset. Finally, we averaged all the epochs corresponding to the four Familiar Between Community transitions and four out of all Familiar Within Community transitions, and compared them. We repeated this with 1000 random groups of four Familiar Within Community transitions in each subject. By normalizing and averaging across subjects, we were able to estimate the increase of the pressing probability after a Familiar Between Community transition compared to a Familiar Within Community transition at each time point. This method is similar to the kernel approach for estimating probability density from discrete observations.

Data Processing: Forced-choice task

Participants that failed on more than two catch trials (two identical quadruplets) out of 4 were excluded from further analysis (FC: 35/250; SC: 45/249; HSC: 34/228). For each subject, we computed a percentage of preference for the tested transition relative to the reference (*Familiar Within Community* transition) in each condition (i.e., the ratio between the number of trials where the subject chose the tested sequence and the total number of trials of this condition). The measure ranges from 0 (the *Familiar Within Community* transition is always selected) to 100 (the other transition is always selected) with a chance level of 50%. We estimated the familiarity score of each condition vs the chance level (50%) using paired t-tests. We report the data from the second forced-choice-task session, corresponding to the maximum exposure to the streams. For the tone stream, results were similar in the first and second sessions. For the syllable stream, results from the first session were poorly consistent across participants, probably because the task was more difficult in the case of syllables. Indeed, flat transitions between syllables violate language structure and participants' priors on syllable sequences. The conflict between priors and the real structure of the sequence might need a variable time to be resolved by each participant ***(Elazar et al., 2022b; Lew-Williams and Saffran, 2012; Onnis and Thiessen, 2013a; Siegelman et al., 2018)***. For completeness, we performed the correlation analysis with each sub-group of data (first vs. second session and tones vs. syllables). These analyses are presented in figure 3.7 (SI). None of the models could adequately explain the first session of the syllable group. To further investigate the learning dynamics and in particular the influence of priors, another paradigm should be proposed, which is beyond the scope of the present study.

Modelling

Theoretical models

For the four models that could be analytically computed from the transition probability matrix (A, B, C, and E), we computed the predictions made by the models for each of our graphs (8 with syllables, 8 with tones). Given A the transition matrix of the graph, models were computed using the analytical description:

- ***Model A: Transition Probabilities (TP) and Ngrams:*** By construction of the Transition matrix, the transition probabilities between nodes are the elements of A.

$$\hat{A} = A$$

- ***Model B: Non adjacent TP:*** Non-adjacent TP are computed by taking the square of the transition matrix.

$$\hat{A} = A^2$$

- **Model C**: Communicability:

$$\hat{A} = \sum_{\Delta t=0}^{+\infty} P(\Delta t) A^{\Delta t} \qquad with \qquad P(\Delta t) = \frac{1}{\Delta t!}$$

Thus, $\hat{A}$ corresponds to the exponential serie: $\hat{A} = e^A$. We use Matlab function 'expm' to compute this value.

The communicability model as described in ***Garvert et al (2017)***. uses the adjacency matrix. Here we used the transition probability matrix. We believe it is more appropriate to consider the relative weights of each transition and not only its existence or not, because a random walk into a weighted graph follows the transition matrix and not the adjacency one. It makes it also more comparable with the other models.

- **Model D**: Free Energy Minimization Model (FEMM):

  $\hat{A} = \sum_{\Delta t=0}^{+\infty} P(\Delta t) A^{\Delta t+1} \quad with \quad P(\Delta t) = \frac{e^{-\beta \Delta t}}{\sum_{\Delta t}^{+\infty} e^{-\beta \Delta t}}$

  which can be re-written:

  $$\hat{A} = \left(1 - e^{-\beta}\right) A \left(I - e^{-\beta} A\right)^{-1}$$

  We then computed the average estimate for each of the conditions for each design. Only the Free Energy Minimization Model (Model D) had one free parameter in its equation. To remove this free parameter and make the model more comparable to the others, we used a previously estimated value of ß=0.06 reported in the literature ***(Lynn et al., 2020a)***. To confirm that this estimation corresponded to our data, we computed the correlation between the subjects' data and the predictions for β ranging from 10^{-15} to 10^{15}. We smoothed this correlation vector to avoid local variations and found a plateau of high correlation for $\beta = [10^{-4}; 10^{-1}]$ with a maximum for $\beta = 0.049$ (Correlation 81%). ,Similarly, we computed the correlation between the FEMM and the hitting time estimation as a function of β. Here again, following the same procedure, we found a plateau of high correlation from $\beta = [10^{-4}; 10^{-1}]$ with a maximum for $\beta = 0.053$ (Correlation = 99.3%). The two models can then be considered quasi equivalent with the β parameter considered in this paper (0.06).

- ***Model E: Hitting Time:*** For this Model, we approximated its value by creating 50000 items long streams corresponding to each graph and computing the average number of elements between each pair of stimuli. We took the inverse of this value to make it more directly comparable with the other models.

Neural models

- **Model F: CA1 similarity:** We used the neural network and the procedure explained in ***(Schapiro et al., 2017)*** originally published by ***(Norman and O'Reilly, 2003)***. We did not change any parameter from this original study because our goal was to see how predictable this model was for our paradigms. We trained it 25 times on each of our graph structures (for each paradigm, 25 batches for 8 groups with Syllables and 8 groups with tones: 25*8*2 = 400 replications). We then presented after each training each node as input in isolation and recorded the pattern of activity in the CA1 layer. To estimate the similarity in nodes' encoding, we computed the correlation between the pattern of activity in CA1 for pairs of elements. Finally, we then made predictions on our task by comparing the similarity between two nodes linked by our four types of transitions.

- **Model G: Hebbian Learning with decay:** This model aim to implement the FEMM computation with an adaptation of the Hebbian approach proposed for associative learning. To achieve that, we declared a layer of neurons with at least one neuron per node of the graph (it can contain more for generalization to bigger networks). The neurons started firing with an exponential decay corresponding to the FEMM decay for each sound in the sequence. Thus, if another sound was presented before the previous neuron stopped firing, several neurons encoding for different nodes co-fired simultaneously. It biased the estimation of TP between two elements. This co-firing behavior can be computed using Hebbian learning rule to update the weights between the neurons. This weight Matrix is then an estimation of the Free Energy Minimization Model that will converge as the length of

the input stream increases. To estimate this model, we followed the same procedure as for the Hitting Time. We created 50 000 item-long streams corresponding to each graph and used those streams as inputs of the neural network. We updated the weight matrix at each step using Hebbian rule as described before. The weight matrix after the 50 000 items was used as an estimation of the model.

Model Comparison

To compare models and data, we considered all experimental paradigms together. To make it comparable with the two-forced-choice data, we normalized each design prediction by the model's value for *Familiar Within Community* transitions. We then pooled all data from all paradigms and estimated the correlation between the data and the models' predictions using 5000 bootstrap re-sampling occurrences. The p-values were estimated by counting the percentage of bootstrap occurrences correlating more with one model compared to another. All the bootstrap occurrences and their correlation with each pair of models are presented in Fig 3.4 B. Each dot represents one bootstrap occurrence. The distribution of these dots below and above the diagonal indicates the comparison between two models. The scatterplot's shape shows the correlation, independence, or anti-correlation between two models. This main analysis of data and model comparison have also been performed for each subgroup of data (first/second session; tones/syllables) and are presented in figure 3.7 (SI). To try better differentiate Communicability with the other models, we recomputed the same correlation analysis but restricted to conditions were communicability makes qualitatively different predictions (*New Within* vs *Familiar Between* transitions in the sparse and high sparse designs). By doing so, we reduced most of the correlation between models and only tested for specific

contradictory predictions. We again find that Hitting Time, FEMM and Hebbian models are equivalent and better than the other models (see fig 3.8 – SI).

Data and analysis availability

All Data and analysis are publicly available at ***https://osf.io/e8u7f/***

Supplementary Information

Free energy approximation of order n

As described in the main text, FEMM is an infinite weighted sum of Transition Probabilities ($P(E_{t-1})$), Non-Adjacent transition probabilities ($P(XE_{t-2})$), Non Adjacent transition probabilities of second ordre ($P(XXE_{t-3})$) etc... which can be formally written $\hat{A} = \sum_{\Delta t=0}^{\infty} P(\Delta t) A^{\Delta t+1}$ with $P(\Delta t) = \frac{1}{Z} e^{-\beta \Delta t}$ and A the transition probability matrix between all nodes of the graph. An alternative explanation of our data could be that subject compute only TP (A)and non-adjacent TP (A^2) and combine both evidences later on for decision making. We investigated this, and more generally the nth order approximation of our model, by decomposing $\hat{A}$ in two parts: the first n elements of the sum and the others. We can then write: $\hat{A} = \sum_{\Delta t=0}^{n} P(\Delta t) A^{\Delta t+1} + \sum_{\Delta t=n+1}^{\infty} P(\Delta t) A^{\Delta t+1}$ Let's call $\widehat{A_n}$ the first n elements of the sum: $\widehat{A_n} = \sum_{\Delta t=0}^{n} P(\Delta t) A^{\Delta t+1}$. $\widehat{A_n}$is then converging toward $\hat{A}$ as n increases. Note that A_0 is equivalent the computation of transition probability. We estimated the correlation between the behavioral data and $\widehat{A_n}$ as a function of n ranging from 0 to 15 to estimate which approximation of this infinite sum is enough to correctly represent the data. The correlation between our data and the Free Energy Minimization Model $FEMM = \widehat{A_\infty}$ is 81%. We found that the correlation logarithmically converged toward $\widehat{A_\infty}$ with 95% of the final value reached for n=4 and 99% for n=8 (fig 3.6). The

second order approximation including only TP (A) and non-adjacent TP (A^2) seems then not sufficient to fully explain our data.

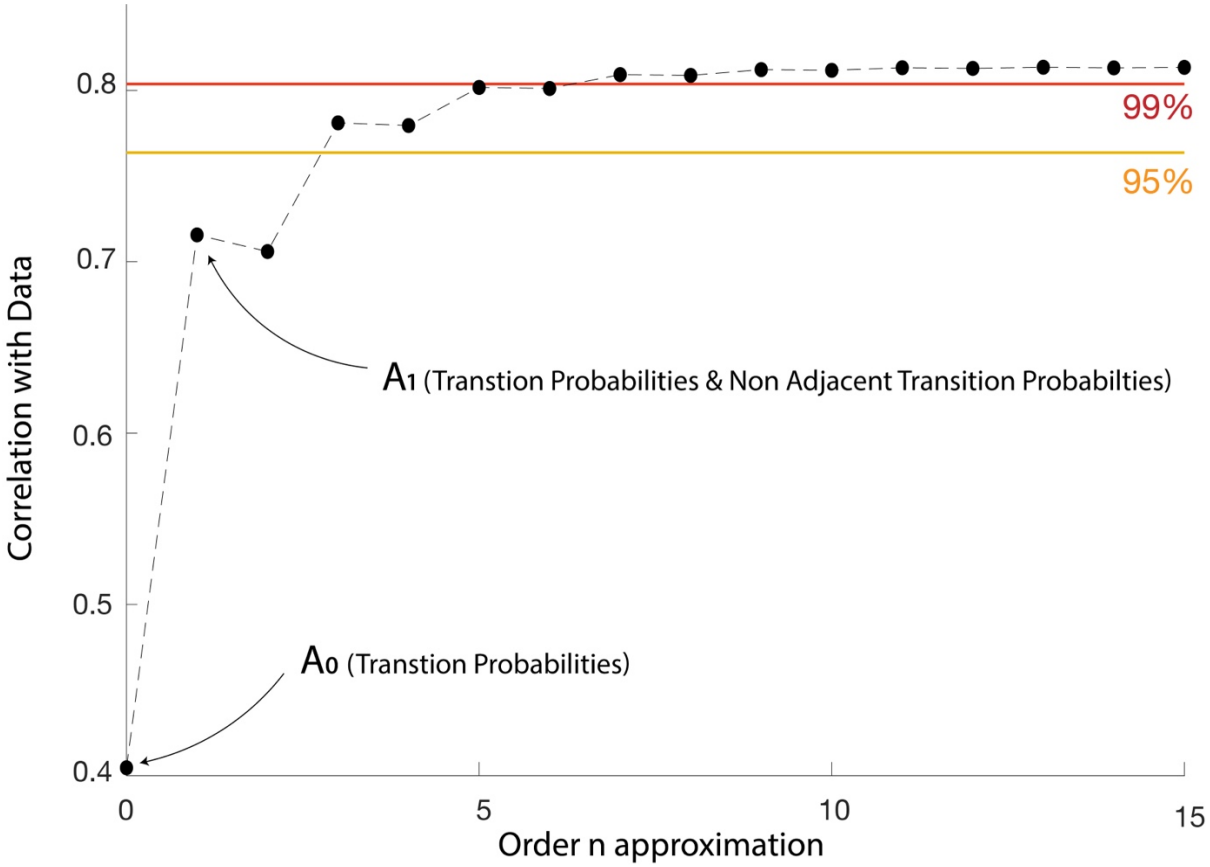


Figure 3.6 *: Correlation of with behavioral data as a function of n. 95% of the maximum correlation is obtained for n=4, 99% for n=8.*

Data and Correlation between human data and model by subgroup

In fig 3.3 and fig 3.4, we presented the results of the two forced choice task and the correlation between human familiarity scores and the different models when mixing tone and syllables streams together. In the main text, we only used data from the second two-forced choice task session, corresponding to the maximal training. Below we show the same analysis with tone and syllables groups separated for the first and second two-forced

choice sessions (Fig 3.7). The first session in the syllable stream did not show any sensitive pattern congruent with any of the learning model. Moreover, the data from the three

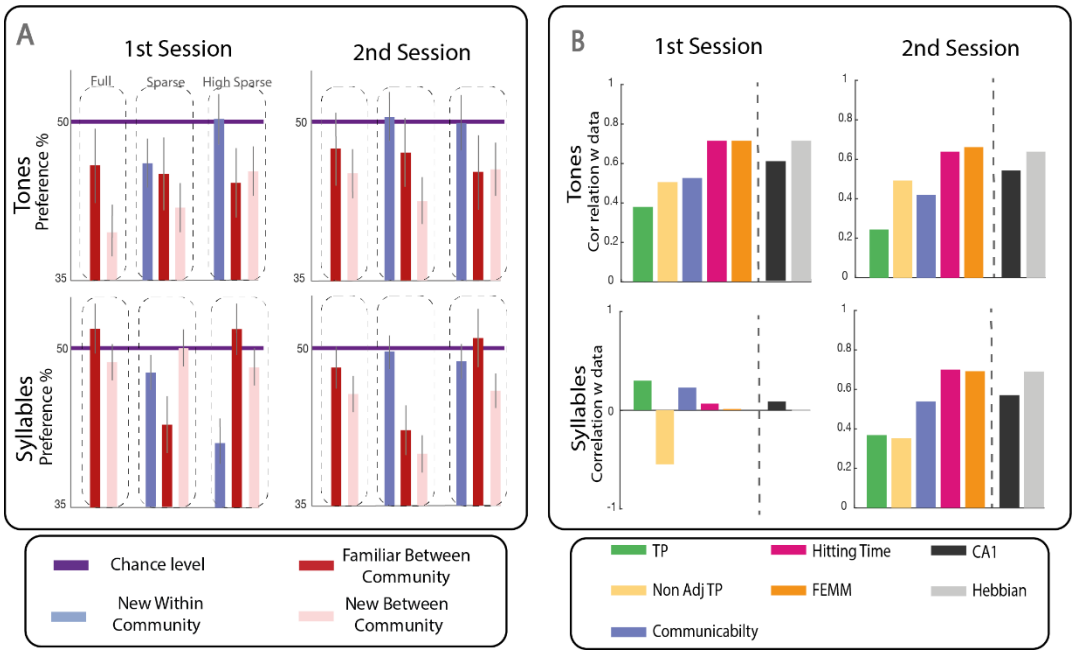


Figure 3.7 : *A : 2-forced choice results per group (Tones & syllables) and session (1st & 2nd) for the full, sparse, and high-sparse communities. B: Correlation between models and data divided by test sessions and subgroups.*

different paradigms were not always in accordance. This observation might be related to participants' priors on syllable transition in sequences ***(Elazar et al., 2022b; Onnis and Thiessen, 2013b; Siegelman et al., 2018)***. Thus, a random walk among syllables strongly contradicts the usual organization of speech. Such a prior would conflict with correct learning of the structure. For tones, the range we used did not belong to any musical scale, limiting priors and making them more appropriate for such task. Nonetheless, after sufficient exposure (2^{nd} block), the data from the syllable and tone groups no longer differed (none of the conditions were significantly different between the two groups), revealing that

participants were able to overcome their priors about the organization of syllable sequences and learned the graph structure in the same way as in the tone experiment.

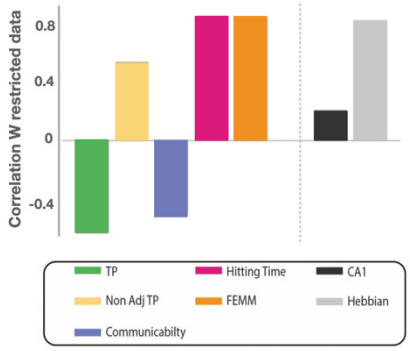


Figure 3.8 *Correlation between data and models restricted to the conditions where predictions between the models are largely different (New Within and Familiar between in Sparse and High Sparse paradigms)*

We also compared the different models restricted to conditions for which the four correlated models and communicability make qualitatively different predictions (*New Within* vs *Familiar Between* transitions in the sparse and high sparse designs). By doing so, we reduced most of the correlation between models and only tested for specific contradictory predictions. We again find that Hitting Time, FEMM and Hebbian models are equivalent and better than the other models (see figure 3.8).

Chapter 4 : Associative learning explains human sensitivity to statistical and network structures in auditory sequences

This work is original to this thesis and is under review for publication in a slightly different version.

Abstract : Networks are a useful mathematical tool for capturing the complexity of the world. In a previous behavioral study, we showed that human adults were sensitive to the high-level network structure underlying auditory sequences, even when presented with incomplete information. Their performance was best explained by a mathematical model compatible with associative learning principles, the Free-Energy Minimization Model (FEMM) which is based on the integration of the transition probabilities between adjacent and non-adjacent elements with memory decay. In the present study, we explored the neural correlates of this hypothesis via magnetoencephalography (MEG). Participants passively listened to sequences of tones organized in a sparse community network structure, comprising two communities. First, the comparison of the brain responses to tone transitions within and between communities revealed an early difference (~150 ms). It is consistent with a mismatch response, implying a rapid, automatic encoding of sequence structure. Second, we used time-resolved decoding to determine the duration of the representation of each tone. The decoding performance exhibited exponential decay, resulting in a significant overlap between the representations of successive tones. We estimated the Hebbian novelty for each transition based on this extended decay profile, and found a noteworthy correlation of this measure with the MEG signal. Overall, our study sheds light on the neural mechanisms underlying human sensitivity to network structures and highlights the potential role of Hebbian-like strategies in supporting learning at various complexity levels.

Significant statement

We conducted a MEG study where human adults were exposed to passive sequences of tones organized in a sparse community network structure. Despite the uniform transition probabilities between tones, participants' brain activity exhibited sensitivity to the network structure. Notably, a consistent response was observed at ~150 ms when switching between communities. This rapid bottom-up surprise resembles the mismatch response based on transitional probabilities. The presence of a long-tail exponential decay in tone representation allowed for overlapping representations of successive sequence elements, facilitating long-range associative mechanisms. This binding mechanism adequately accounted for various scales of sequence learning, bridging the gap between statistical and network learning approaches.

Introduction

Understanding the structure of the input sequences we encounter is fundamental for developing a comprehensive mental model of our environment ***(Dehaene et al., 2022, 2015)***. The capacity to detect first-order relationships between successive events (i.e. transition probabilities) and its limits have been extensively studied in humans at the behavioral and neural levels ***(Benjamin et al., 2023b, 2021; Fló et al., 2022a; Henin et al., 2021; Maheu et al., 2019; Saffran et al., 1996a)*** as well as in non-humans animals ***(Boros et al., 2021; James et al., 2020; Toro and Trobalón, 2005)***. Higher-order statistical relations between elements of a sequence are also detected by human adults and children ***(Karuza et al., 2019, 2019; Lynn et al., 2020a; Mark et al., 2020; Schapiro et al., 2013)***, but only a limited number of neuroimaging studies have explored the neural correlates of this learning ***(Ren et al., 2022; Schapiro et al., 2016; Stiso et al., 2022)***.Therefore, we still do not know if a common mechanism can adequately explain both

first order (local transitions) and network structure learning or if these computations require distinct cognitive and brain processes.

To bridge the gap between local statistical and network-level learning studies, we previously proposed the *sparse community paradigm* to simultaneously characterized these aspects on auditory sequences ***(Benjamin et al., 2023a)***. Building upon the community network paradigm utilized by ***Schapiro et al (2013)***, we created a network consisting of two densely but incompletely connected clusters (called communities) of six elements each, exemplified in fig 4.1 A. This enables a 2-by-2 design wherein transitions can be locally congruent or not (i.e. the tones are connected or not) or adhere to higher-order community membership (the tones belong to the same community or not) ***(Benjamin et al., 2023a)***. After being exposed to this sparse structure, participants were asked to judge their familiarity with various tone transitions. Interestingly, participants judged as highly familiar new transitions they had never heard before if they were between elements belonging to the same community. This *completion* effect demonstrated that they generalized the structure to missing data. Conversely, they judged transitions between communities less familiar than within communities, despite the absence of any difference in local transition probability. This *pruning* effect decreased the familiarity of actual transitions, but inconsistent with each community. Among the various models proposed in the statistical and network learning literatures, the free energy minimization model (FEMM) initially proposed by Lynn et al. ***(2020a)*** and conceptually related to the successor representation, provided the best fit to participants' behavior. According to this model, participants did not solely compute adjacent transition probabilities but a linear sum of transition probabilities at all orders (adjacent, first-order non-adjacent, second-order non-adjacent and so on), weighted by a decreasing exponential factor. This explains how local transitions but also network structures can be perceived. Indeed, this model successfully accounts for behavioral results across different network types, including community and

sparse community networks ***(Benjamin et al., 2023a; Lynn et al., 2020a)***, ring and lattice networks ***(Lynn et al., 2020a)***, but also results from local statistical learning literature. The FEMM might therefore provide a unifying framework for understanding sequence learning.

However, a common model is insufficient to postulate a common implementation ***(Marr, 1982)*** and there is still no consensus on how the brain might implement these computations. On one hand, the sensitivity to networks structure is often described as an abstraction of the structure involving top-down processes with late brain signatures ***(Ren et al., 2022)*** typically in the prefrontal cortex ***(Stiso et al., 2022)***. On the other hand, we previously postulated that low-level associative learning ***(Benjamin et al., 2023a; Endress, 2010; Endress and Johnson, 2021; Schapiro et al., 2017)*** was sufficient for both local and higher-order learning. To disentangle those two hypotheses, we tested whether passive exposure, without explicit indications, to a rapid auditory sequence can lead to successful learning.

Materials and methods

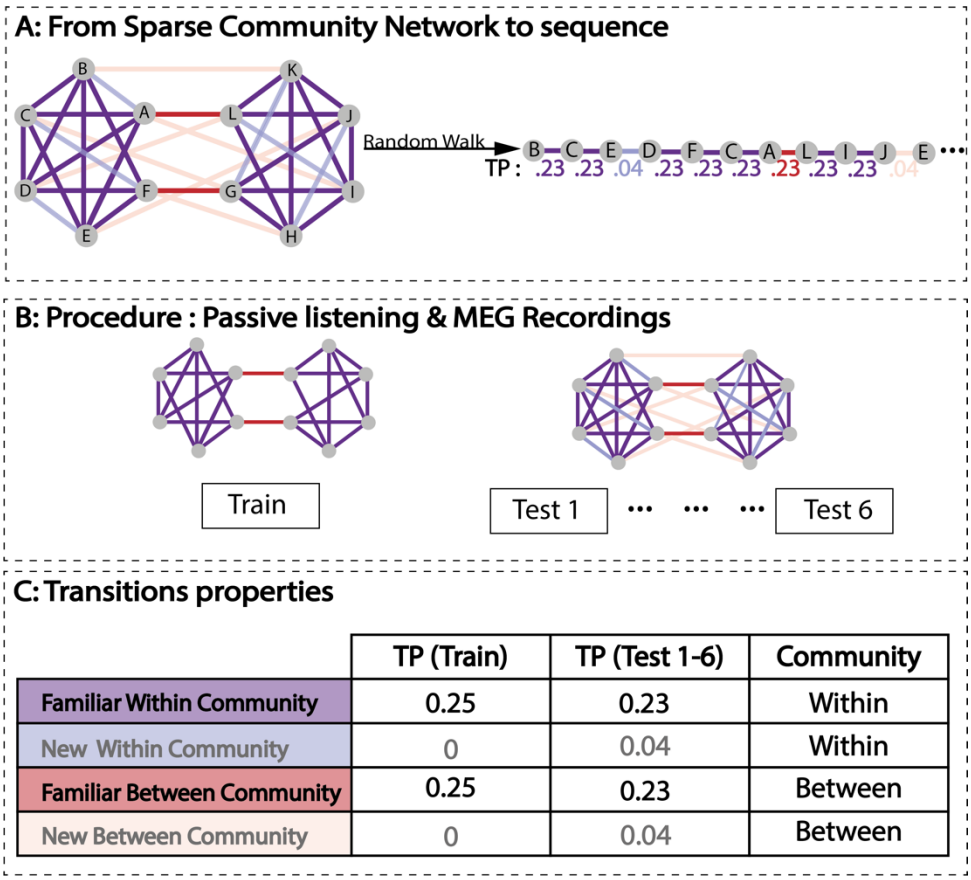
Stimuli and procedure

We generated twelve tones of 50ms duration, logarithmically distributed from 300 to 1800 Hz. For each participant, the twelve tones were randomly assigned to the twelve nodes of the sparse community network (see fig 4.1 for a complete description of the network structure). The sparse community network comprised two communities (i.e. clusters) made of six nodes, densely connected to each other but poorly connected to the nodes of the other community. However, in this sparse design, the communities were still incomplete: some transitions between two tones of the same community were removed creating a 2x2 design. Critically, transitions could be locally connected or not (Adjacent Transition Probability, which we refer to as *Familiar* vs *New*) but also globally congruent

or not in respect to the communities (Community membership, which we refer to as *Within* vs *Between*) transitions properties.

To ensure that participants learn the network structure, we began the experiment with a training sequence obtained from the sparse community network. Then in test blocks, we introduced with a low transition probability (4%), *New Within* and *New Between* Transitions. The 12 *New Between* Transitions (one per node), violated both local transition probabilities and the structure. The New Within and New Between transitions were randomly drawn for each subject to add variability in the network structure (see fig 4.1 A for an example of one structure and the associated sequence used for one participant). We then performed random walks in the sparse community graph and derived six 960 items-long sequences with 200ms SOA between each tone (see fig 4.1 A). To summarize, the first run only comprised *Familiar Within* and *Familiar Between* transitions. For the next six runs (Test 1-6), we introduced infrequent (transition probability of 4%) New Within and New Between transitions to be able to measure brain activity for each type of transition.

Crucially, the experiment was completely passive and participants were unaware of the structure of the auditory sequence. They were only instructed to pay attention to the sound sequences and to stay still while looking at a fixation cross displayed at the center of the screen. The experiment lasted around 45 minutes and a small break inside the MEG was possible between each run.



	TP (Train)	TP (Test 1-6)	Community
Familiar Within Community	0.25	0.23	Within
New Within Community	0	0.04	Within
Familiar Between Community	0.25	0.23	Between
New Between Community	0	0.04	Between

Figure 4.1 : Design and procedure. *A) Example of a sparse community network for one participant. All community networks are similar in terms of properties but New Within and New Between transitions are randomly drawn for each participant. Purple lines correspond to Familiar Within-community transitions and red lines to Familiar Between community transitions, and blue and pink lines correspond respectively to New Within and New Between transitions. We can derive a sequence by performing a random walk into this network. Here we display an example of a test sequence derived from this structure. B) Experimental procedure. First, participants passively listened to a sequence from a sparse community network (Train). Then they were presented with six 960-items test sequences obtained from the community structure graph comprising New Within and Between community transitions with low transition probabilities of 4% (light blue and pink colors on the graph). C) Table summarizing the local and community properties for the transitions for each condition and the 2-by-2 design with local and high order properties.*

Participants

29 healthy adults came to the lab. Due to technical issues during the MEG recording of 6 subjects, 23 recordings (16 females, mean age = 26.58, sd = 6.1) were kept for the analysis. All participants gave written informed consent prior to enrollment and received 90€ as compensation.

MEG recordings and preprocessing

Participants performed the tasks while sitting inside an electromagnetically shielded room. The magnetic component of their brain activity was recorded with a 306-channel, whole-head MEG by Elekta Neuromag® (Helsinki, Finland). The MEG helmet is composed of 102 triplets, each comprising one magnetometer and two orthogonal planar gradiometers. Brain signal was acquired at a sampling rate of 1000 Hz with a hardware high-pass filter at 0.03Hz. The data were then resampled at 250Hz to reduce computational load. Eye movements were monitored with vertical and horizontal EOGs and heartbeats with ECGs. Subjects' head position inside the helmet was measured for realignment at the beginning of each run with an isotrack Polhemus Inc. system from the location of four coils placed over the frontal and mastoids.

MEG signal was then preprocessed using MNE python pipeline with classical steps following recommendations from ***(Jas et al., 2018; Niso et al., 2018)***. We first applied Maxfilter algorithm to remove ambient noise and signal was band-pass filtered ([0.1-30]Hz). Eye movements and heartbeats were identified and removed using PCA components correlation with EOG and ECG measures.

To decode if a transition was within or between community, data was epoched in trials from 100 ms before to 300ms after the sound onset. To determine how sustained was the

neural representation of each sound across time, we also segmented the data in 2.6 seconds long epochs, from 100 ms before to 2500 ms after the sound onset. Bad data, channels and epochs were detected and removed with autoreject toolbox ***(Jas et al., 2017)***.

Within vs Between Decoding analysis

To examine whether the brain encoded the sequence structure, we employed a logistic regression decoder trained to predict whether a transition occurred *Within* a community (*Familiar Within* and *New Within*) or *Between* communities (*Familiar Between* and *New Between*). The decoder was trained on smoothed data (sliding window +- 20ms) to enhance the signal-to-noise ratio. Given that the number of trials differed between the two classes, we used the area under the ROC curve as a metric of success (ROC AUC). This analysis was conducted for each time-point of the epochs (fig 4.2). We also computed the decoding performance when the decoder was trained at time t and tested at time t', to reveal the generalization across time (GAT) of the decoder, and thus the stability of the mental representation. This analysis is presented in the GAT Matrix plot in fig 4.2. By design, the diagonal of the GAT matrix corresponds to the previously described time-by-time decoding performances.

To replicate the decoding accuracy with a different metric, we performed a whole trial decoding analysis at the subject level. This decoder used all time points across all recording channels simultaneously for training and testing. Consequently, it provided a single accuracy value for the entire epoch. Unlike time-by-time decoding, this approach can exploit the temporal dynamic of the signal to differentiate conditions

The previous analysis pulls together data from both familiar and new transitions, but we know from our previous behavioral work that these two transitions elicit different familiarity judgment ***(Benjamin et al., 2023a)***. We investigated how the neural representation of these two types of transitions differed. Specifically, we investigated whether decoding

success remained when local transition probabilities were similar in the two conditions. Therefore, we replicated the previous decoding analysis but limited to *Familiar* transitions only, which had identical local transition probabilities of 0.23 (*Familiar Within* Vs *Familiar Between*). We conducted the same analysis with *New* transitions (*New Within* Vs *New Between*), which both had a transition probability of 0.04.

Statistical Analysis

Statistical significance in the Generalization Across Time (GAT) matrix was assessed using a temporal-temporal cluster-based permutation (MNE python ***(Gramfort et al., 2013)***) for times between 0 and 300ms. For the time-by-time decoder, we performed a temporal cluster permutation test in [0, 300]ms time window. Note that these two statistical tests are not independent as the time-by-time decoding corresponds to the diagonal of the GAT Matrix. The whole trial decoding gives a single decoding value per subject, we thus performed a one-way t-test across subject to test if the performance subjects was greater than chance.

Hebbian learning estimation and linear regression

To assess the duration of the representation of a sequence item in the brain signals, we used epochs containing 10 tones (2.5s). We trained a 12-class decoder (for the 12 tones) with balanced accuracy to decode the identity of the first tone of the epoch throughout the whole epoch. To ensure that we were decoding the sustained activity related to the first tone and not a subsequent repetition of the same tone, we removed from the analysis all epochs in which the first tone was repeated during the test window (fig 4.4 A). We averaged the above chance decoding performance over the time-windows which corresponded to the interval between two consecutive items to estimate the amount of superposition of the mental states of the different elements of the sequence. We then

estimated the strength of the Hebbian association for each pair, which corresponds to the sum of the transition probability matrix between the tones at all orders (A^t), weighted by the overlap between item representations (fig 4B).

We later used the associated surprise, defined as the negative log of the Hebbian association strength, as a regressor for the MEG signal (fig 4D). We performed spatio-temporal cluster analysis on the beta value associated with this linear regression to extract electrodes and times where this Hebbian learning estimation might significantly explain the dynamic of the brain data. We also computed the average Hebbian association strength of each type of transition (fig 4C).

Results

In this experiment our aim was to find neural correlates that correspond to the encoding of the community structure and to test whether this learning results from a low-level associative process or corresponds to a late abstract and explicit discovery ***(Ren et al., 2022; Stiso et al., 2022)***. To assess the encoding of the community structure, we first decoded Within vs Between transitions type and searched for the timing of the effect. We subsequently estimated how sustained was the representation of each element of the sequence in order to estimate what an associative mechanism would predict and tested this prediction on our data.

Decoding Within Vs Between

We first tested whether participants' mental model of the sequence encoded the community structure despite uniform transition probabilities. We thus trained and tested decoders on all *Within* transitions (*Familiar Within* & *New Within*) vs all *Between* transitions (*Familiar Between* & *New Between*). We obtained a significant cluster ($p<0.05$) in the GAT matrix accuracy. Temporal cluster analysis on the Time-by-Time decoding

accuracy revealed a significant cluster between 88 and 252ms ($p<0.001$) peaking at 156ms. Finally, the trial-based decoding was significantly above chance ($p<0.01$) (see fig 4.2 *Within* Vs *Between*).

We then restricted this analysis to the *Familiar* transitions (*Familiar Within* vs *Familiar Between*, which corresponds to 92% of the trials). Since *Familiar Within* and *Familiar Between* transitions had the same transition probabilities (0.23), a significant difference could not be explained by statistical learning on consecutive items, but had to be due to a higher-order representation of the community structure. Here again, a significant cluster ($p<0.01$) was found in the GAT matrix. A temporal cluster between 76 and 280ms was found in the time-by-time decoding ($p<0.001$) with a peak at 152ms. Trial-based decoding was also significantly above chance ($p<0.001$).

Symmetrically, we have restricted the analysis to New transitions only (*New Within* Vs *New Between,* which corresponds to 8% of the trials). By design, both *New Within* and *New Between* transitions had transition probabilities of 4%, so learning only local transition probabilities would predict equal unfamiliarity with both types of transition. In line with the previous results, we found a significant temporal-temporal cluster in the generalization matrix ($p<0.05$), a significant temporal cluster in the time by time decoding ($p<0.05$, significant time = [132, 172] ms, peaking at 160ms). Trial based decoding was also significant ($p<0.05$). Due to the much smaller number of trials, the results were noisier.

We performed a series of control analyses to eliminate putative low level confounds, such as the identity of the current or of the previous tone or the identity of the pair of tones. To control for tone identity decoding, we ran the decoding analysis but restricted to trials ending in one of the four nodes connected to node of the other community (i.e. involved in a Familiar Between transition, darker nodes in fig 4.3). Depending on the previous tone, these trials could be either *Familiar Within*, *Familiar Between, New Within* or *New Between*. Thus, decoding within vs between community transitions on those trials cannot

be driven by the current tone identity. The same was done for trials starting in one of these four nodes to control for decoding the identity of the previous tone. We also controlled for the pair (previous and current tone identity simultaneously): like for the current tone control, we restricted the analysis to trials ending in one of the four nodes connected to each type of transition, and cross-validated the previous tone. To do so, we trained and tested on different previous nodes (training on two per community and testing on the two others, see batches in fig 4.3). This strategy was also used for the *Familiar Within* Vs *Familiar Between* GAT decoder. By design of the experiment, *New Within* vs. *New Between* decoders were already balanced for current and previous tone (each tone is attached to one transition of each type). We then only controlled for the pair by cross-validating on the previous node. Overall, the control analyses qualitatively and quantitatively confirmed previous results. Only the *New Within* vs. *New Between* control for pair (i.e. controlling both previous and current tone identity simultaneously) analysis did not reach significance probably due to the small number of trials in this contrast (only 8% of the data is used).

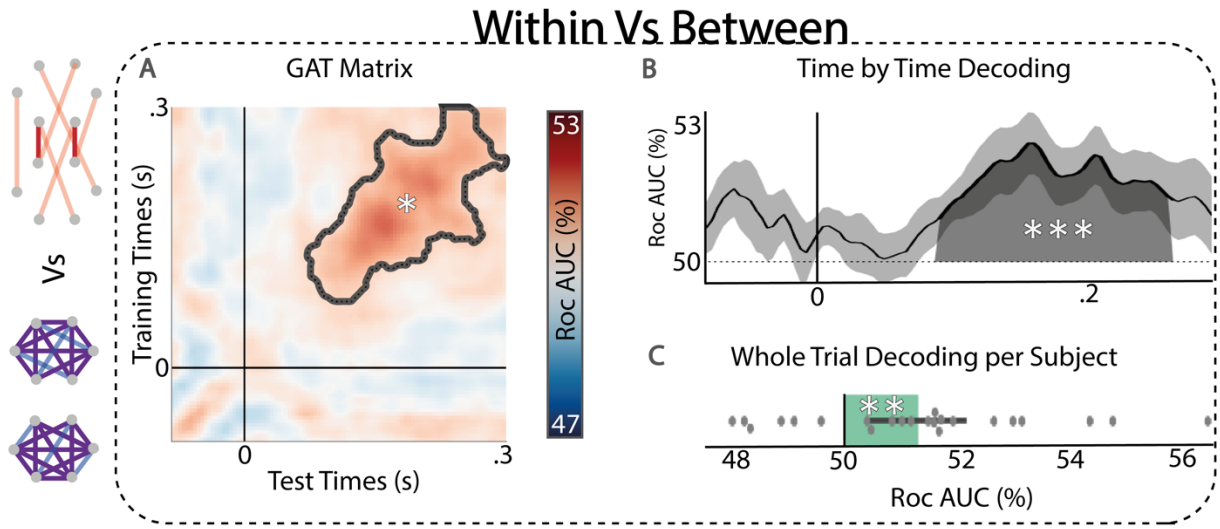
An alternative explanation for *Within* vs *Between* community transitions decoding is that the difference is related to a habituation effect since the previous presentation of the same tone is probably closer when the random walk remained within the same community than across communities. Therefore, we restricted the analysis to the first appearance of each tone after a community change. Thus, close repetitions of tones of the same community is avoided in the data used for this decoder. Despite a decrease in the number of trials, the decoding accuracy of those controls was still significant for all conditions. All generalization matrices are shown in fig 4.3.

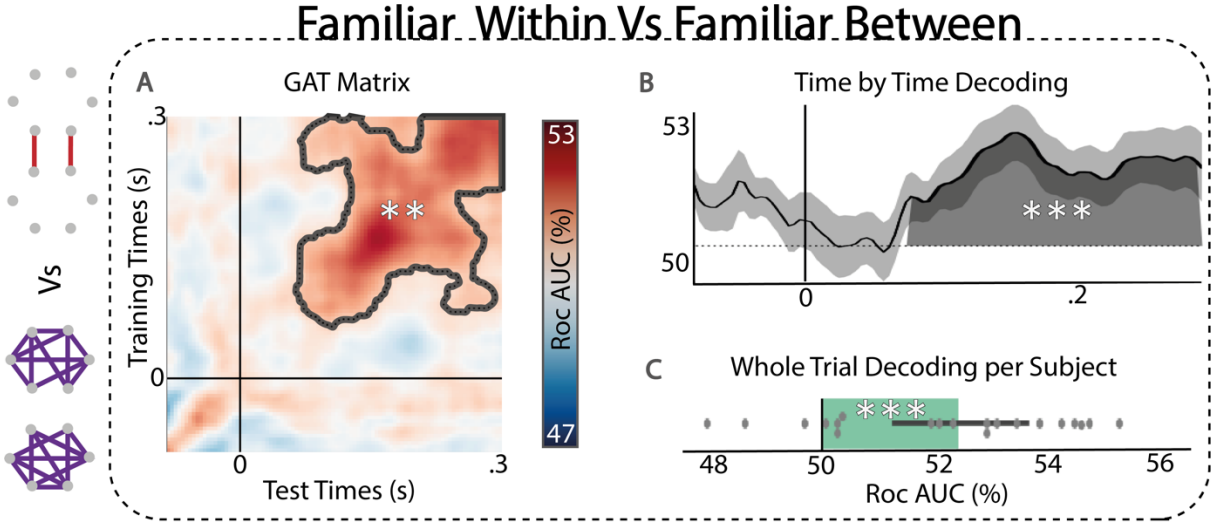
Hebbian learning estimation

We test here the hypothesis that Hebbian associative learning, commonly referred to as "fire together, wire together" can support the encoding of network structure. In our

experiment, this would imply that the mental representation of each element is sustained for a sufficient duration to allow multiple elements to overlap ***(Endress, 2010)*** and extend the distance over which associations can be formed. According to this model, it is predicted that the representation of each tone should decrease following an exponential profile. To test this hypothesis, we quantified the amount of overlap between the representations of item *n* and item *n+i.* In fact, this provides a good estimator of the weight of the non-adjacent transition probability of order i in the overall familiarity of the transition. To estimate the overlap between brain representations of different items of the sequence, we determined how long the representation of each item is seen in brain activity. To do so, we split the data into 10-items long sequences (i.e. 2.5 seconds) with no repetition of the first tone in the sequence. We train a 12-class decoder on each time-point to predict the identity of the first tone. Decoding performance is shown in fig 4.4. We averaged the above chance decoding performance over the time-windows which corresponded to the interval between two consecutive items. We observed an exponential-like decrease in performance that reached 0 after ~ 8 sequence items (fig 4.4 A).

We then estimated the Hebbian strength of each pair of tones from associative learning. To do so, we computed the sum of the different transitional probability orders weighted by the overlap between item representations as estimated from the decoding performances (fig 4.4 B). This gave us a 12x12 symmetrical matrix of Hebbian learning familiarity for each pair (fig 4.4 C). Finally, we averaged this measure of Familiarity for each condition type (fig 4.4 C) and obtained a result that is consistent with the pruning effect (difference between *Familiar Within* vs *Familiar Between* transitions) and the completion effect (difference between *New Within* and *New Between* transitions) as discussed in ***(Benjamin et al., 2023a)***.


Within Vs Between
Vs
A
GAT Matrix
Training Times (s)
Test Times (s)
.3
0
53
47
Roc AUC (%)
B
Time by Time Decoding
Roc AUC (%)
53
50
0
.2
C
Whole Trial Decoding per Subject
48
50
52
54
56
Roc AUC (%)


Familiar Within Vs Familiar Between
Vs
A
GAT Matrix
Training Times (s)
Test Times (s)
.3
0
53
47
Roc AUC (%)
B
Time by Time Decoding
Roc AUC (%)
53
50
0
.2
C
Whole Trial Decoding per Subject
48
50
52
54
56
Roc AUC (%)

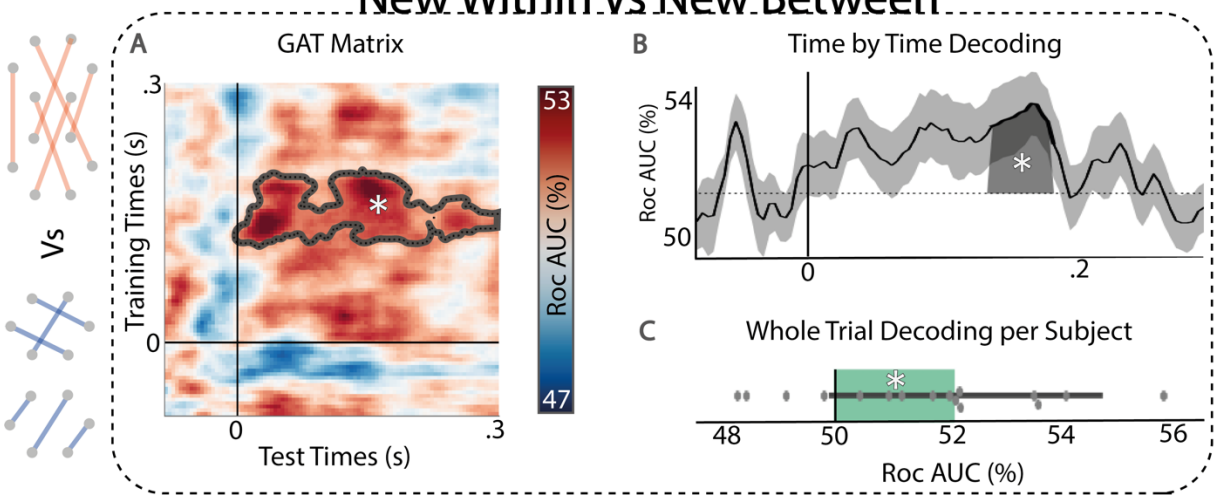

New Within Vs New Between
Vs
A
GAT Matrix
Training Times (s)
Test Times (s)
.3
0
53
47
Roc AUC (%)
B
Time by Time Decoding
Roc AUC (%)
54
50
0
.2
C
Whole Trial Decoding per Subject
48
50
52
54
56
Roc AUC (%)

Figure 4.2 : Within vs Between community decoders on the MEG signal. *Top Panel : Decoders with all Within community (Familiar & New) vs. all Between communities (Familiar & New) transitions. A) Generalization Across Time (GAT) matrix with significant cluster delineated in black. B) Time by time decoding. The shaded area indicates a significant temporal cluster. C. Individual performances based on whole trial decoding: Mean Decoding accuracy across subjects (green bar, one dot per subject). Those three analyses have been replicated with Familiar only transitions (middle panel) and New only transitions (bottom panel). Community structure was encoded in each case despite the flat local transition probability. Stars represent significant of the statistical tests (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).*

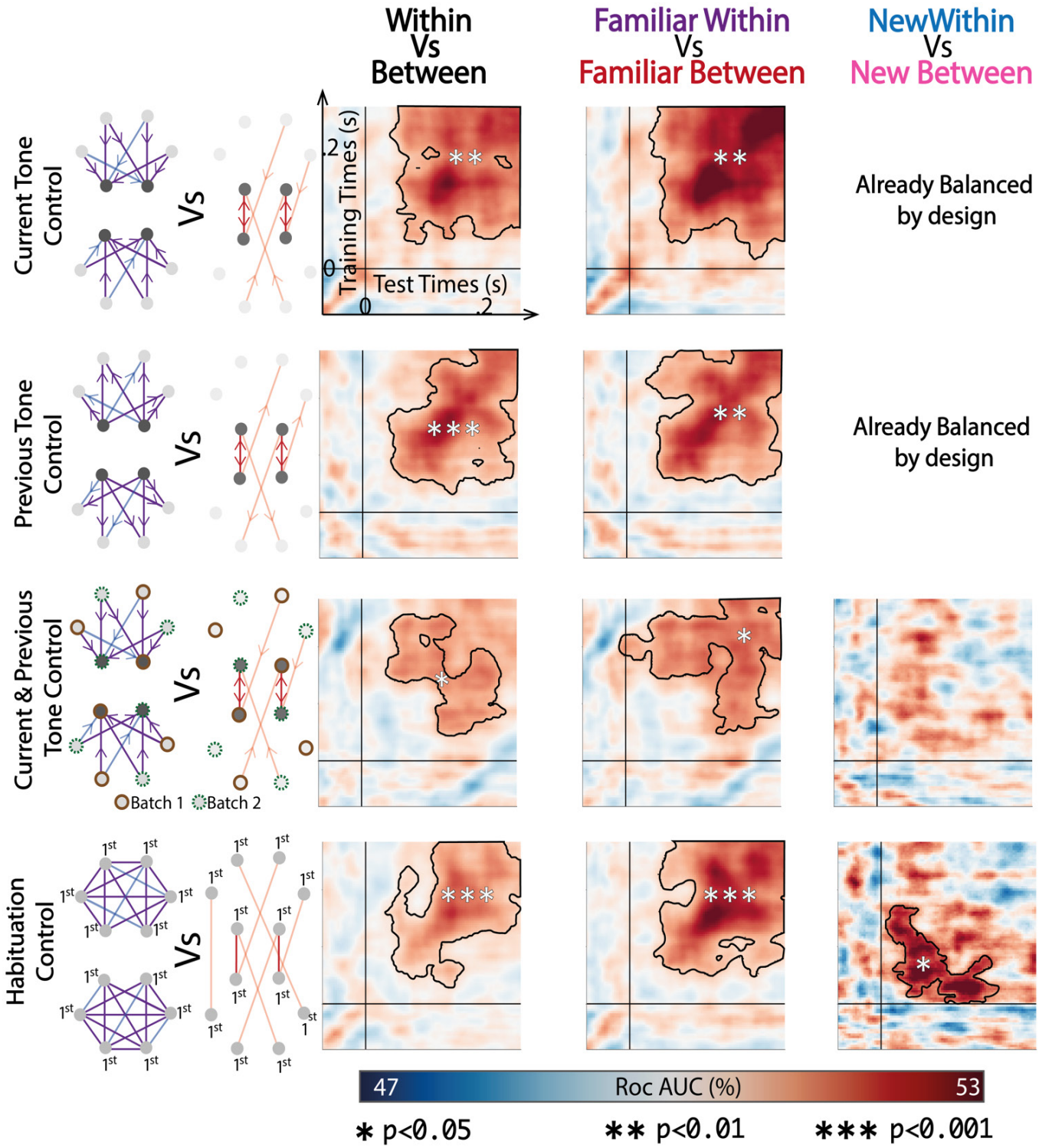

Within
Vs
Between
Familiar Within
Vs
Familiar Between
NewWithin
Vs
New Between
Current Tone Control
Vs
Training Times (s)
Test Times (s)
0
.2
Already Balanced by design
Previous Tone Control
Vs
Already Balanced by design
Current & Previous Tone Control
Vs
Batch 1
Batch 2
Habituation Control
Vs
1st
47
Roc AUC (%)
53
* p<0.05
** p<0.01
*** p<0.001

Figure 4.3 : Control analyses for the results presented in figure 4.2*. For each decoder, we controlled for the current tone, for the previous tone and for the pair (both current and previous tone simultaneously). We also controlled for temporal proximity habituation. All the analyses qualitatively and quantitatively confirmed previous results except the New Within vs. New Between control analysis that did not reach significance probably because of insufficient data.*

Hebbian learning accounts for trial variability

To further investigate the neural predictions of Hebbian learning, we correlated brain signals with the estimated Hebbian strength of each transition (fig 4.4 D). We performed a linear regression between the brain signal after each tone and the surprise elicited by each transition. Unlike most studies of sequence learning, where the surprise response is calculated solely from local transition probabilities, we computed it here as the negative log of the Hebbian learning strength. This takes into account several orders of adjacent and non-adjacent transition probabilities whose weights have been computed on the basis of the overlap of brain representations estimated by our tone decoder. A spatio-temporal cluster permutation test revealed a significant cluster in the magnetometers (right centro-occipital, time = [150 ; 290] ms, pval < 0.01) that was replicated in the gradiometers (right centro-occipital, time = [140 ; 300] ms, pval < 0.05). Furthermore, the observed clusters were still significant when the negative log of the adjacent transition probabilities was introduced as a supplementary regressor (ps < 0.05 for both magnetometers and gradiometers clusters).

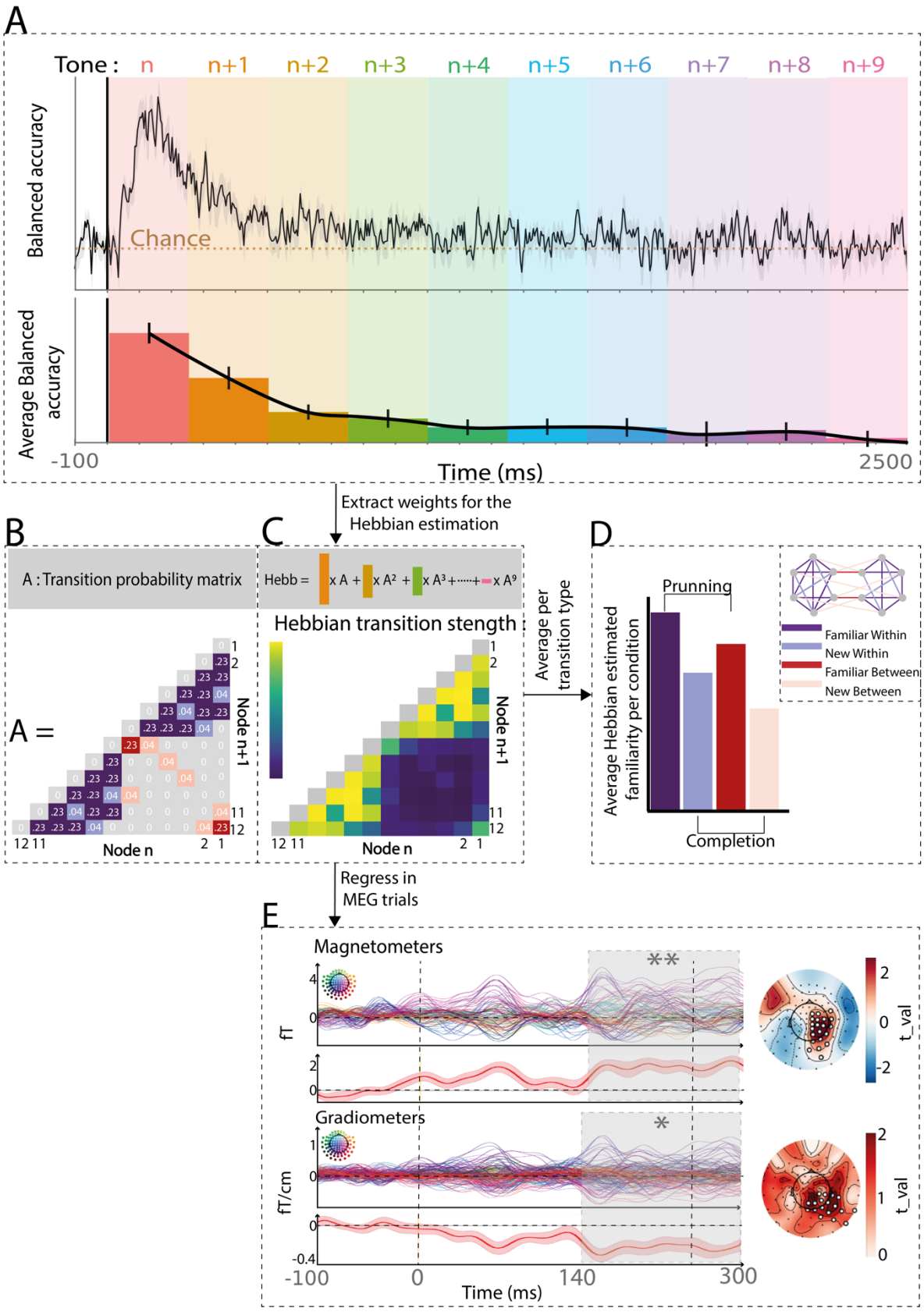

A
Tone : n n+1 n+2 n+3 n+4 n+5 n+6 n+7 n+8 n+9
Balanced accuracy
Chance
Average Balanced accuracy
-100
Time (ms)
2500
Extract weights for the Hebbian estimation
B
A : Transition probability matrix
A =
Node n
Node n+1
C
Hebb = x A + x A² + x A³ +·····+ x A⁹
Hebbian transition stength
Average per transition type
D
Prunning
Completion
Average Hebbian estimated familiarity per condition
Familiar Within
New Within
Familiar Between
New Between
Regress in MEG trials
E
Magnetometers
fT
Gradiometers
fT/cm
t_val
-100
0
140
300
Time (ms)

Figure 4.4 : Associative learning estimation and fit on MEG data. *A) Top: Decoding performance of the first item of the sequence across time (2.5s window). Shaded colors represent the SOA between each tone of the sequence. The dotted line shows the chance level. Bottom: Decoding performance averaged over the SOA between consecutive sequence items. Error bars present the standard error across subjects. It takes ~8 items for the decoder of the first tone to converge to chance level. B) Matrix of exact transition probabilities (A) associated with the graph underlying the sequence. Familiar transitions are associated with 23% transition probabilities and New with 4% see fig4.1. Impossible transitions have a null transition probability. C) Estimation of the Hebbian learning strength for each transition. Based on the decoder (panel A), we estimated the overlap between non adjacent elements of the sequence (average decoder accuracy during SOA of item n+i). We then computed the Hebbian strength (Hebb matrix) for each pair of elements as the sum of the different transitional probability orders (A^t), weighted by the overlap between item representations. D) Average of the Hebbian learning strength per condition. Pruning (Familiar Within > Familiar Between) and completion (New Within > New Between) effects are consistent with behavioral results (Benjamin et al., 2022a)) and with the decoding performance obtained in fig 4.2. E) Regression coefficient for the estimated Hebbian surprise (-log(Hebb)) for each MEG sensor. Significant time-windows are shown in shaded areas and significant sensors are indicated on the t-map topographies by the white dots. These were obtained with a spatio-temporal cluster-based analysis. The red line below the sensors value represent the time course of the average regression value on the significant sensors*

Discussion

In this study, our objective was to investigate whether local statistical learning and structure learning in sequences are driven by distinct cognitive processes or if they stem from a shared associative learning mechanism. Indeed, while learning of local statistics is often described as an associative process, network learning is usually seen as an abstract map representation. Previous studies exploring network learning have used explicit paradigms, revealing late brain signatures consistent with top-down or frontal activity ***(Ren et al., 2022; Stiso et al., 2022)***. However, based on a modeling approach, we proposed in our previous behavioral study that both local and high-order statistical scales might be supported by low-level associative learning strategies ***(Benjamin et al., 2023a)***. Thus, this hypothesis predicts that learning sequence structure does not necessitate an explicit representation and may instead rely on similar automatic and rapid (~150ms) mismatch responses observed in violation of local transition probabilities.

Network learning results from a low-level bottom-up computation

To test these predictions, we presented participants with a passive learning task using fast auditory sequences. We showed that the structure properties of the sequence were rapidly decodable from their brain recordings (~[100-250] ms after tone onset). The timing of this response aligns with the rapid mismatch responses (MMN in EEG) observed in learning paradigms based on transitional probabilities ***(Maheu et al., 2019; Todorovic and de Lange, 2012)***. Since the transition probabilities between tones were uniform and the walk within the network was random, no expectation could be built on the basis of previous tones, implying that the difference between within and between community transitions was bottom-up, triggered 150ms after the information became available. This result challenges the notion of abstract and explicit calculations as prerequisites for learning such structures. In addition, our analyses revealed a similar effect when the decoding

analysis was restricted to new transitions (*New Within* Vs *New Between*) and familiar transitions (*Familiar Within* Vs *Familiar Between),* suggesting an automatic generalization of the community structure beyond sensory evidence. This result provides neural underpinnings for the behavioral observation we previously reported, indicating that participants accurately assess the familiarity of transitions on the basis of their congruence with network structure, even when these transitions were not encountered during training.

Hebbian learning as a plausible implementation

In our previous study, we put forward the hypothesis that adult behavioral performance could be effectively explained by the free energy minimization model (FEMM). This model aggregates the different orders of statistical regularities (adjacent and non-adjacents) into a single quantity. In this study, we showed that this model can be readily implemented through a simple associative learning mechanism anchored in Hebb's rule, commonly known as "fire together, wire together" ***(Benjamin et al., 2023a; Hebb, 1949)***. In the context of structure learning, this rule implies that the mental representation of each tone needs to be sustained for a sufficient duration to enable the overlapping of several elements and the extension of associations over longer distances. According to this model, we predicted the representation of each tone to exhibit an exponential decay profile. A rapid decay of tone information would limit associations to shorter distances, while a slower decay would facilitate the formation of long-range dependencies and the generalization of the underlying structure. Thus, this exponential decay acts as a balance between generalization (slow decay) or local statistical significance (fast decay).

To test for that, here, we estimated the duration of the representation of each tone by performing a time-by-time decoding analysis of tone identity. We were able to successfully decode the identity of a tone during the presentation of the subsequent eight

tones, with a decoding performance that exhibited an exponential decline. This decay profile provided an estimation of the extent of overlap between elements in the sequence, reflecting the number of elements that "fire together." Consequently, it allowed us to quantitatively assess the strength of their associative connections or the extent to which they "wire together." Therefore, we computed the Hebbian learning strength for each transition in the network (fig 4.3 C) and found that these weights accurately accounted for the results of the Within vs. Between decoders, encompassing both familiar and novel transitions (fig. 4.3 D). Moreover, the Hebbian strength estimate significantly correlates with neural activity, aligning with the timing of the mismatch response ***(Maheu et al., 2019; Todorovic and de Lange, 2012)***. This result provides compelling evidence for the rapid encoding of structure through bottom-up processes, compatible with associative learning strategies.

It is important to acknowledge that the associative learning mechanism discussed earlier may not be the sole factor contributing to network structure learning, particularly in cases where explicit detection is required from participants. Hippocampus ***(Constantinescu et al., 2016)*** or frontal ***(Stiso et al., 2022)*** representations of abstract maps might also play a role in such tasks ***(Garvert et al., 2017; Schapiro et al., 2017, 2016)***. Recent intra-cranial recordings conducted during local statistical learning paradigms have revealed that multiple brain regions, including both cortical areas and hippocampus, can simultaneously represent the same structure while carrying different feature information ***(Henin et al., 2021)***.

Difference between implicit passive listening and explicit structure learning

Converging results provide evidence that associative learning supports the perception of the community structure in the present experiment. The Hebbian learning strength significantly accounted for the variance in brain signals (fig 4.4 E). Moreover, the pruning (*Familiar Within* vs *Familiar Between*) and completion (*New Within* vs *New Between*)

effects found with decoders (fig 4.2) can easily be explained by the same mechanism (see fig 4.3 D). However, it is worth noting that the results from our previous behavioral study do not entirely align with the current findings. Specifically, in the present experiment, the strength of associations between tones exhibited a more rapid decrease (exponential decrease factor 0.52) compared to the previous behavioral study (factor 0.058). This discrepancy suggests that participants in the current experiment might be less inclined to generalize the underlying structure.

Several factors might explain this difference. Firstly, the estimation of the generalization factor in the MEG experiment may be more prone to noise due to the small number of participants (23 vs. several hundred in the behavioral study). Since the trade-off between generalization and accuracy may vary among individuals *(Lynn et al., 2020a)***, on one side group level estimation with 23 subjects is limited and on the other side it is difficult to measure this trade-off at the individual level due to the MEG data variability. To thoroughly investigate this question, a larger sample size with multiple sessions per subject would be necessary to obtain a reliable estimation of the generalization factor at the individual level. Secondly, it is possible that associative learning represents the implicit component of this task** *(Andringa and Rebuschat, 2015)***, followed subsequently by an explicit decision-making processes involving higher level prefrontal regions. This second step might facilitate the abstraction of the structure by labeling each community as distinct** *(Koechlin et al., 2003; Koechlin and Jubault, 2006)***. This dual process might explain why explicit behavioral tasks** *(Benjamin et al., 2023a; Lynn et al., 2020a)* **exhibit a better generalization factor compared to our implicit MEG task. The same explanation might account for the late signatures of top-down activity reported by** *(Ren et al., 2022)* **who used a slow and explicit version of a community paradigm. To further explore this hypothesis, a direct comparison of passive and active learning of such networks while monitoring the representations in the auditory cortex, the hippocampus, and the lateral prefrontal cortex would be necessary.**

Conclusion

The aim of the present study was to uncover the neural mechanism underlying network learning. We proposed the sparse community paradigm as a way of combining local

statistical learning and network learning in a single sequence. Previous behavioral studies have showed that a mathematical model (FEMM) accurately captures human learning. Here we add that Hebbian learning, a plausible neural implementation of the FEMM, might account for such learning. Indeed, thanks to time-by-time decoding of the brain state associated with a tone, we observed an exponential decline in the tone representation across 7-9 elements. Using this estimate of mental representations dynamics, we estimated the strength of each transition of the network and significantly correlated this estimate with our data. The present study provides novel insights into the mechanism underlying network learning and highlights the importance of considering the role of brain dynamics in understanding sequence learning. Further investigation would be useful to better understand the role of the generalization/accuracy tradeoff offered by the speed of the exponential decay. Investigating the sparse community paradigm in different experimental conditions (explicit vs implicit), over different tone and ISI durations (if the sequence is slower, what is the overlap between tones?), different populations (non-human primates) or during early development would allow to better characterize the properties of this generalization/accuracy tradeoff.

Appendix 3 : Exploring the bi-partite graph

This work is original and a follow-up of the previous studies. It will not be published as it is more an exploration and pilots for possible future research questions than a proper study.

In the previous studies, we presented the sparse and high sparse community networks and showed that they were completed by subjects. We presented these results with evidence for bias in the computation of transition probabilities following the Free Energy Minimization Model. Another possible explanation for those results, although contradictory with our MEG results, would be that subjects learnt the actual local transition probabilities and in parallel formed an abstract representation of the structure of the network using another un-described mechanism. Based on the extraction of the structure, they could a posteriori reconstruct the network and complete it.

Thanks to the MEG analysis, we showed brain correlates favoring FEMM computation over a-posteriori reconstruction of the network. We looked for another way to disentangle between those two models and thus, looked for another type of network where computing FEMM and completing the structure of the network would make different behavioral predictions. We decided to test a sparse bi-partite network. The full bi-partite network is almost the opposite to the community network structure : it is divided in two groups of nodes where nodes from the same group are not connected to each other but every node of each group is connected to every node of the opposite group. In the sparse version, we removed one connection per node (see fig A6 A). Then we tested three types of transitions: Familiar Between (purple, locally congruent with 0.33 transition probability and congruent with the structure), New Between (blue, locally incongruent with 0

transition probability but in accordance with the structure) and New Within (pink : locally and globally deviant). Both hypothesis (FEMM and a posteriori reconstruction of the structure) postulate that the locally and globally deviant transition (New Within, pink) should be rejected by participants. However, unlike in the sparse community network, the missing transitions to complete the network structure (here the New Between, blue) are not expected to be hallucinated by participants if they compute the FEMM. However, if they have found the structure and a posteriori completed it, they should report being familiar with these missing transitions.

To test that, we used the exact same procedure as for the community designs by testing participants (N=99) familiarity with the different type of transition using a 2-forced choice task. They had to judge the most familiar pair between a Familiar Between transition (locally and globally congruent, purple) and one of the two New transitions. They significantly rejected both the New Within and the New Between transition types, and we found no difference between New Within and New Between, as predicted by the Free Energy Minimization Model. The absence of completion of this structure gives us greater confidence in the claim that participants are learning biased statistics and do not perform a posteriori reconstruction of the structure afterwards.

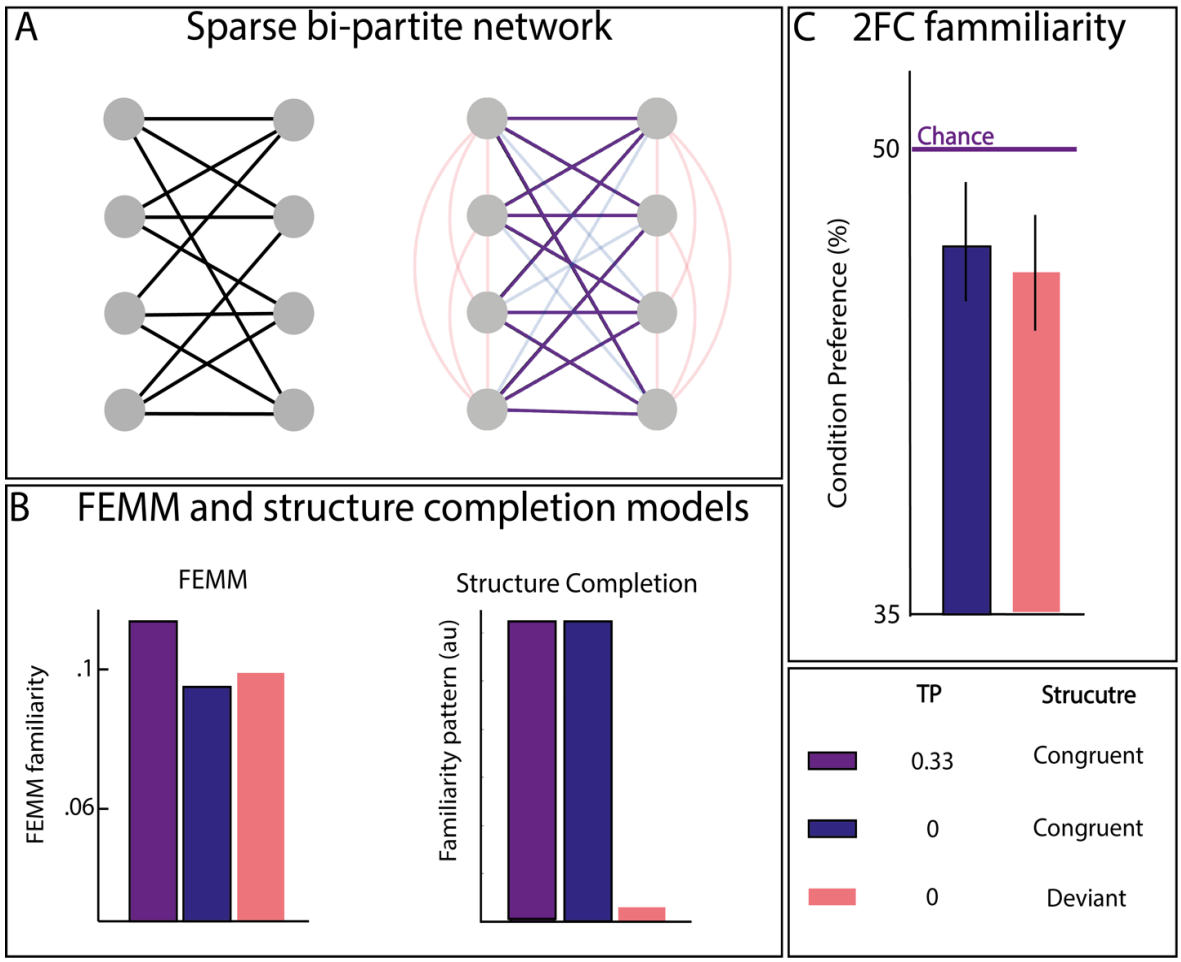


Figure A6 *: **A**. Presentation of the sparse bi-partite network. In this network we separated the pool of stimuli in two groups of four elements. Each element of a group is connected to three elements of the other group (Purple transitions= Familiar Between). Two elements of the same group cannot be connected to each other. We called this graph the sparse bi-partite design because the structure is incomplete : from each node, there is a missing link to one node of the other group.* ***B****. We proposed this design because unlike the community paradigm, the FEMM does not predict the completion of the structure (i.e. the false memory of transitions that never occurred but respect the graph structure: New Between:light blue lines). We can thus disentangle whether the completion of the structure is driven by FEMM or by a posteriori reconstruction of the missing elements of the structure.* ***C****. The two forced choice task showed that participants did not complete this structure and equally rejected New Within and New Between transitions, ruling out the hypothesis of a posteriori completion and supporting the FEMM Hypothesis.*

Discussion : are statistical learning scales different mechanisms ?

Chapter 5 : A unified mechanism for statistical sequence learning at different scales

This work is original to this PhD Thesis .

Abstract : Sensory inputs exhibit intricate temporal dependencies that often follow an underlying hidden structure that the brain tries to capture. In the case of learning temporal dependencies between items in sequences, it ranges across different scales, from local statistics between consecutive items **(Saffran et al., 1996a)** to local and global statistical dependencies across sequences of notes **(Basirat et al., 2014; Bekinschtein et al., 2009)** or more high order and abstract relationships such as pattern repetitions **(Barascud et al., 2016; Southwell and Chait, 2018; Zhao et al., 2019)**, hierarchical patterns and nested structures **(Dehaene et al., 2015),** networks **(Garvert et al., 2017; Schapiro et al., 2017)** or even rules and mental programs **(Al Roumi et al., 2021; Dehaene et al., 2022; Maheu et al., 2020; Planton et al., 2021)**. While learning rules and statistics probably depends on different brain processes and circuits **(Maheu et al., 2020*) –*** for a counterpoint see **(Fiser and Lengyel, 2022)-** the different scales of statistical learning (adjacent and non-adjacent transitions, network learning) however might result from a common mechanism. Here, we have attempted to apply the Free-Energy Minimization Model (FEMM), previously proposed to explain the learning of network structures of visual sequences **(Lynn et al., 2020a)**, to different results of the literature in a goal to provide an unifying model spanning different statistical scales. Our goal is to provide a more parsimonious and biologically plausible explanation than the coexistence of several statistical learning metrics, each used for a specific scale and sometimes in contradiction with others.

Introduction

In the first two chapters of this thesis, we presented local statistical learning paradigms with different populations. Later, we introduced network learning paradigms, investigating high-order dependencies which at first glance do not seem to rely on the same computational capacity of our brain as local statistical learning. However, in the second section we investigated the possibility of a common mechanism for statistical learning at the local scale (adjacent dependencies between consecutive elements) and at a higher order scale (network properties). For that, we introduced a mixed paradigm: the *sparse community paradigm*. The main purpose of it was to present a sequence allowing to test both local and high order congruency/deviance **(Benjamin et al., 2023a)**. Using the behavioral results, we compared different models proposed in the literature and showed that the Free Energy Minimization Model (FEMM) might account for these results, including the overgeneralization of the structure. Then, thanks to the MEG temporal resolution, we observed an early effect, at the same latency that the MMN described after the violation of local transitions. Moreover, the duration of tones representation pointed to the possibility of encoding the FEMM with Hebbian learning principles. We thus consider FEMM as a potential explanation for statistical learning across different scales.

However, in order to achieve our goal of unification, we need to test our candidate model on other results reported in the literature. When we reviewed the literature in the introduction of the thesis, we proposed a classification of statistical learning abilities into three scales of temporal dependencies (fig 0.3). The local scale encompasses all paradigms in which subjects have to learn adjacent dependencies between consecutive elements. At the intermediate scale, dependencies between non-adjacent elements must be learned in order to succeed in the task, and for the higher scale we considered paradigms that study the learning of network properties. During our literature review, we were struck by

the proliferation of results and explanatory models that were often limited to paradigms within a specific statistical learning scale. As a result, the number of required statistical calculations seems to be expanding with each study. These findings have been attributed to learning local transition probabilities, backward transition probabilities, non-adjacent transition probabilities, communicability, cosine similarity, and the free energy minimization model (FEMM) (fig 0.3). Given that humans are initially unaware of the underlying structure of a sequence, they should compute and maintain multiple statistical regularities simultaneously. However, the memory requirements for tracking all these computations make the accumulation of so many models biologically implausible. Additionally, many of these paradigms have been proposed in the context of language acquisition, which necessitates considering the limitations of infant's memory.

All these considerations make this multi-computational approach costly and not very parsimonious for explaining the many observed statistical effects of the literature. Therefore, we will explore FEMM as a possible unifying model for bringing together different statistical learning scales under a common mechanism. To test this hypothesis, we systematically reviewed statistical learning capacities at different scales reported in the literature. Our goal is not to show that this model fits the data better than the models described in the respective papers, but to point out that there is a possibility of a single model to account for this variety of results.

Formal description of the candidate model : FEMM

The Free Energy Minimization Model has been proposed by Lynn and colleagues **(Lynn et al., 2020a)** to account for high order statistical learning in visual sequences following network structures, where adjacent transition probabilities alone were deemed insufficient to account for human performance. It has been initially described as a trade-

off between computational accuracy and complexity. Intuitively, performances can be explained as memory errors in the process of computing adjacent transition probabilities in sequences. Indeed, participants exposed to a sequence of inputs reinforce the association between element i and i-1 (local transition probabilities). However, errors in this process may lead participants to sometimes bind current element i with element i-2, i-3, i-4…. with a decreasing probability. This intuitive idea can be analytically formalized, and the optimal decreasing probability distribution can be computed by minimizing the free energy function. It results in a probability distribution following the Boltzmann function (exponential decay). Therefore, the estimated mental model of transition probability is biased compared to the sequence objective transition probabilities, enabling participants to encode high-order structure. In more detail, the mental model is a linear combination of the transition probability matrix (A) and non-adjacent transition probabilities of every order $(A^{\Delta t})$ with a weight of $P(\Delta t)$ following the Boltzmann distribution were Δt is the order of non-adjacency (or size of the memory error, ie $\Delta t = n$ corresponds to $P(E_t|X.....XE_{t-n})$). The estimated model can then be written as:

$$\hat{A} = \sum_{\Delta t=0}^{+\infty} P(\Delta t) A^{\Delta t+1}$$

With

$$P(\Delta t) = \frac{1}{Z} e^{-\beta \Delta t}$$

where A is the transition probability matrix of the graph. β parameter was previously estimated to 0.06 in network learning tasks with human adults **(Benjamin et al., 2022a; Lynn et al., 2020a).** β plays the role of a generalization parameter: a low value implies very high generalization (the long-distance dependencies are strongly taken into account in the computation). At the contrary, a high β value means a lower generalization, as only short

scale dependencies truly influence the FEMM value. To the limit, $\beta = \infty$ corresponds to a total absence of generalization, meaning that only adjacent transition probabilities are learned.

Importantly, this mathematical model appears biologically relevant through a simple associative learning mechanism anchored in Hebb's rule, commonly known as "fire together, wire together". In fact, if the representation of each element of the sequence is hold for a sufficient amount of time in the brain, a significant overlap between several successive element representations will enable simultaneous short and long-distance dependency learning. To match the predictions of the FEMM, the sustained representation of sequence element in the brain should follow the same Boltzmann function.

An equivalent model has already been proposed from the reinforcement learning point of view. Indeed, in this literature, the hippocampal place cells have been proposed to represent maps of probabilistic future states and reward by encoding *successor representation* instead of positional cognitive maps **(Dayan, 1993; Momennejad, 2020; Stachenfeld et al., 2017)**. As possible future states are not limited to the next one, but can encompass several steps forward, the theoretical question is similar to the one of integrating different order of transition probability. Indeed, successor representation has then been formally defined as a pondered sum of probabilistic future states, and can then be written SR = $\sum_{\Delta t} \gamma^{\Delta t} A^{\Delta t}$, with A the transition probability matrix between successive states. This approach is very similar to FEMM with an infinite sum of all powers of the transition matrix, pondered by an exponentially decreasing factor. Here the factor is $\gamma^{\Delta t}$ with $0 < \gamma < 1$ and generally $\gamma = \frac{0.85}{\lambda max}$ with λmax the largest eigenvalue of the transition matrix **(Garvert et al., 2017)**. It is important to note that even if those models have been proposed from different backgrounds (statistical learning vs. cognitive map

formation), they are mathematically equivalent (up to a constant) if the parameters of both models follow the relation $\gamma = e^{-\beta}$ suggesting the possible implication of the hippocampus in the FEMM computation.

Computation of the FEMM for each paradigm

To examine the applicability of FEMM across various experimental contexts, we carefully curated landmark studies that encompass different scales of statistical learning. Generally, these studies comprise a training period during which participants are exposed to sequences of stimuli, followed by a test period in which items corresponding more or less to the training are compared. Learning is assumed when there is a difference in familiarity/deviance rating of the different types of items. Therefore, we conducted simulations to estimate the strength of associations between elements given a specific training, computed an estimation of the familiarity associated with each tested condition and finally qualitatively fit our estimation with the measured data from the literature. When possible, we also made a quantitative comparison between our estimations and the participants' performances.

More specifically, we estimated the mental model predicted by FEMM in each paradigm, but to avoid overfitting of each paradigm and keep the most parsimonious model possible, we fixed the βparameter for all paradigms $\beta = 0.06$, i.e. the value proposed by Lynn et al **(2020)**. Then, for each of these paradigms, we used the following approach:

- When the TP matrix was available and the training sequence was a random walk into the associated network, we used the analytical formula of the FEMM to compute the exact FEMM matrix from the TP matrix .
- In the case of AXC or artificial grammar paradigms, the TP matrix was not sufficient to describe the training material as the input followed non-Markovian relationships. Therefore, we estimated the FEMM from the corpus of training

sequences, by computing and summing (with the associated decreasing exponential factor) all TP orders within each training sequence. The final FEMM matrix was obtained by averaging each matrix of FEMM estimation for each training sequence that a subject experienced.

Once the participants' mental model was computed based on the training material for each paradigm, we estimated the familiarity of each test condition.

- When experimental conditions corresponded to pairs, the estimated familiarity referred directly to the corresponding association on the FEMM matrix previously obtained. This is the case for all network paradigms (fig 5.1 E, F, G & I)
- For longer test conditions, we computed a pondered sum of each pair of elements within the test sequence. For example, the familiarity of a test sequence with three elements ABC is estimated following the equation : $FEMM(AB) + FEMM(BC) + e^{-\beta}FEMM(AC))$. This formula was used for all paradigms with test sequences of more than 2 elements (fig 5.1 A, B, C &D).
- In their study, Lynn et al (2020a) also compared whether learning was different for two types of networks: the community and the Lattice network. We thus followed their method and computed the average FEMM familiarity of all transitions of the network for both networks.

This method enabled us to obtain quantitative predictions of estimated FEMM familiarity for each condition in each paradigm, which we could then compare with experimental measures like behavioral judgment, looking time, reaction time or fMRI activity.

Re-considering previous results at different scales

At the local scale

Saffran paradigm

At the local scale, learning transition probabilities has been proposed to account for several results in the literature, among which the important study on speech segmentation by 8 months-old infant by Saffran and colleagues **(Saffran et al., 1996a)**. In this study and its many variations, subjects heard structured sequences of syllables (or other auditory, visual of haptic stimuli) with embedded trisyllabic pseudo-words (ABC), so that local transitions between consecutive syllables within a word are predictable (TP=1), while between two words the transition probability drops at 1/3. After learning the structure, subjects are typically exposed to isolated triplets corresponding either to a word ($A_iB_iC_i$) or to a so-called part-word ($B_iC_iA_k$ for example) straddling a word boundary. A difference between Word and Part-word conditions (in looking time, familiarity rating, EEG responses…) is usually interpreted as a correct learning segmentation of the sequence based on the learning of transition probabilities between syllables. We confirmed these results in the studies presented in the first chapters. Although infants were not able to segment the speech stream when quadrisyllabic words were used, they did learn the transition probabilities as shown by the difference between non-words and the other two conditions (words and part-words). Moreover, comatose patients showed neural correlates of segmentation in a similar task suggesting a partially preserved computation of local statistics.

Comparison with FEMM: To account for those results with the FEMM, we computed the FEMM mental model from the TP matrix of the training sequence and estimated the familiarity with both Words and Part-words conditions. Word familiarity was estimated up to 19% higher compared to Part-word (see estimated FEMM familiarity in **figure 5.1A**). As adjacent transition probability is the first order (and most prominent)

component of the FEMM, it is not surprising that a TP difference induces a FEMM difference. FEMM can therefore perfectly explain the effects described above.

Statistical learning vs chunking

While the Word vs Part-word effect is consensual, the mechanism involved is the subject of debate, with two main proposals: a statistical approach based on the transition probabilities between elements of the sequence **(Endress and Johnson, 2021; Endress and Mehler, 2009; Fiser and Aslin, 2005; Saffran and Wilson, 2003)** vs. a chunking approach where repeated series are recognized and stored into memory **(French et al., 2011; Isbilen et al., 2020; Perruchet, 2019; Perruchet and Poulin-Charronnat, 2012; Perruchet and Vinter, 1998; Slone and Johnson, 2018)**. To disentangle these two mechanisms, authors have proposed the study of phantom words, i.e. triplets that never occur as such but within which local transition probabilities are correct. **(Endress and Mehler, 2009; Polyanskaya, 2022; Slone and Johnson, 2018).** For example, in a training sequence where the pseudo-words /ra/ /ti/ /fu/ and /bo/ /ti/ /ma/ were presented, the corresponding phantom word would be /ra/ /ti/ /ma/. Indeed, high familiarity with phantom words would support the statistical learning mechanism only, while rejection of these words would argue in favor of a chunking mechanism. Unfortunately, both results have been found in the literature, leaving the problem unsolved.

Comparison with FEMM: Our current FEMM model cannot disentangle between the two mechanisms, however it is worth noting that because it integrates adjacent and non-adjacent transition probabilities in a single familiarity measure, FEMM predicts a difference between a word and a phantom word ($FEMM_{ra,fu} > FEMM_{ra,ma}$) while being a statistical model. We thus believe that the use of phantom words to demonstrate chunking over statistics is insufficient to conclude. Moreover, the expected FEMM

difference between Words and Phantom Words being more subtle than Words vs Part-Words, it might explain the results instability for this type of test in the literature.

At the intermediate scale

Grammar learning

Some studies investigating grammar learning, such as Milne et al's work **(2018)**, can be seen as testing adjacent and non-adjacent dependencies. We chose this specific study because exhaustive training and test sequences were explicitly stated in the article, allowing us to precisely run our model. After being exposed to 8 different training sequences following the artificial grammar (each sequence repeated 6 times), subjects were presented with 5 items sequences and had to judge if those were violating or not the training grammar. The authors showed that the average percent of violation judgment for a test sequence was correlated with local transition probabilities in both visual and auditory modalities.

Comparison with FEMM: We extracted the percentage of time subjects judged the sequence as a violation of the grammar from this study. To account for those results with our model, we first estimated the FEMM mental model based on the 8 training sequences that subjects experienced. Then we computed the predicted FEMM familiarity for each test sequence. Correlations between those predictions and the behavioral responses of human adults showed an overall good agreement in both auditory and visual modalities (Visual $R=-0.65$ $p<0.05$, Auditory $R=-0.8$ $p<0.01$). Thus, FEMM correctly accounts for those results.

AXC paradigm

At the intermediate level, sensitivity to non-adjacent dependencies has been described as essential to language learning. Indeed, to account for grammatical rules such

as tense agreement in English ("*is ... ing*", "*has … ed*"), one needs to learn non-adjacent associations, independently of the central element. It has then been shown that participants could learn non-adjacent transition probabilities of the form P(C|XA), in AXC triplets where the syllable A predicts C with variable X syllables in between. For instance, Peña et al **(2002)** showed that participants could learn such dependencies when presented with AXC triplets with subliminal pauses between each triplet. Specifically, the authors showed that after listening to a sequence including AXC triplets with pauses between each word, participants preferred test words that respected the A-C dependency (Rule Word AX'C where X' has never been presented in AXC triplet) over Part-Words (XC_iA_k) participants have heard.

A surprising property of this paradigm is that by increasing the number of possible Xs to be inserted in the middle of an A-C relationship, the familiarity of the A-C relationship improves. This has been formally measured by Gomez **(2017)**. In this study, she compared the performances of a two forced choice task between pseudo-words respecting or not the A-C relationship (Words - A_iXC_i- vs non-Words -A_iXC_k-). She tested this performance for training sequences with 2, 6, 12, and 24 possible Xs in the middle of the A-C relationship, and showed that the subjects' performances increased with the number of possible Xs. As the A-C relationship does not change from one condition to the next, learning only the non-adjacent dependency or memorizing the A-C rule cannot explain this variation in familiarity with the number of X.

Comparison with FEMM: We first estimated the FEMM familiarity of Words, Part-words and Rule-Words in the classical case where three Xs can be presented in the middle of each pair **(Buiatti et al., 2009)**. We showed that Word (A_iXC_i) was rated higher than Rule-Words ($A_iX'C_i$), itself higher than Part-Words (XC_1A_2) (**Fig 5.1C**), in agreement with results presented in the literature **(Buiatti et al., 2009; Newport and Aslin, 2004; Pena, 2002).** In agreement with Peña et al, we also found that in the absence of pauses, when estimating

the FEMM on a continuous sequence with embedded AXC, Rule-Words were no longer estimated more familiar compared to the Part-Words. This correctly accounts for the necessity of the pause between words in such paradigm. We also compared familiarity for every number of possible X between 2 and 29 elements. We showed that the predicted Words vs non-Words differences in FEMM familiarity increased with the number of Xs. This qualitatively fits the behavioral results observed in Gomez' study **(2017)**. In fig 5.1D, the left panel represents the FEMM prediction for Words vs non-Words for different number of possible X ranging from 2 to 29. On the right panel is the behaviorally measured familiarity from Gomez et al study. However, a detailed comparison between the model predictions and the results showed only a partial agreement, qualitatively accounting for the increased familiarity with number of Xs, but quantitatively only partially accounting for the behavioral familiarity. Indeed, the model seems to predict a logarithmic increase of the Word vs non-Word dissimilarity with the number of possible X while the empirical results seem more compatible with a linear trend (although differentiation between linear and logarithmic trend with only four measured values is unreliable). A more careful exploration of this question, with more conditions, would be interesting to validate or contradict this model.

At the higher order scale

Garvet et al network

At a higher-order scale, description of the structure is often based on a network formalism, and a variety of networks and their properties have been explored. For instance, Garvert and colleagues **(2017)** scanned participants with fMRI while they were traversing an abstract network composed of 12 objects. Among other results, the authors reported a correlation between the brain signal and Communicability, another metric of distance between nodes in networks. The conceptual idea behind Communicability and

FEMM is similar, but the penalization of long-distance dependencies is greater in Communicability than FEMM. Using a searchlight method, they found a cluster in the hippocampus and entorhinal cortex correlating with communicability.

Comparison with FEMM: Based on the network structure, we computed the FEMM and thus estimated a predicted familiarity for each transition of the network. We could then re-analyze the average activation of the cluster that they evidenced and found that its mean activation per transition also significantly correlated with FEMM predictions of the transitions ($p<0.05$) (**Fig 5.1E**).

Community network

Schapiro and colleagues **(2013)** proposed the community network to demonstrate the sensitivity of humans to clusters in network structure, despite the absence of information at the local scale. Indeed, this network is composed of three communities (groups of nodes almost entirely connected to each other but weakly connected to other groups) and all the nodes in the network have exactly four neighbors. The resulting sequence from such network is then completely flat in terms of adjacent transition probabilities, with each pair of elements in the sequence having exactly ¼ transition probability. Despite this lack of local transition probability information, Schapiro and colleagues **(2013)** as well as many others later on **(Benjamin et al., 2023a; Kahn et al., 2018; Karuza et al., 2019, 2017; Lynn et al., 2020a)** showed that humans could discriminate transitions staying within communities from transitions switching between communities. They also compared the community network with lattice network structures. Like the community network, it is composed of twelves nodes with four neighbors per nodes. They found that participants trained on a lattice network were more familiar with the structure (faster reaction time) compared to participants trained on a community network.

Comparison with FEMM: The FEMM was initially proposed in this study to specifically account for those results. In particular, they showed that the drop in familiarity when switching community was correctly predicted by the model (see **fig 5.1F**). Moreover, they showed that the learning difference between the community and the lattice network was also correctly accounted by the FEMM. Indeed, the average of the FEMM familiarity on every transition of the lattice network is higher than the one on the community network, which is congruent with faster reaction time when learning the lattice compared to the community network (see **fig 5.1H**).

Ring graph

In this important study, another type of network was also investigated by Lynn and colleagues: the ring graph. In this 15 nodes network, each node is connected to four neighbors (+1 / +2 / -1 / -2). Then, from a node, four are directedly reachable by a single transition (Distance 1), four others are reachable in two steps (Distance 2) and the six others can be reached in three transitions (Distance 3). Measuring the familiarity of each pair of this network using reaction time revealed a preference for connected pairs (Distance 1) over unconnected pairs of distance 2, itself preferred to unconnected pairs of distance 3.

Comparison with FEMM: This also matches the FEMM predictions as shown by Lynn and colleagues **(2020a)** see **Fig 5.1G.** Indeed, participants' reaction time to Distance 3 pairs was higher than for Distance 2 pairs, itself higher than for connected pairs.

Sparse and high sparse community networks

As reported in earlier chapters of this thesis, we introduced the sparse community design **(Benjamin et al., 2023a)**, in which one (sparse community) or two (high sparse community) possible transitions per node inside each community in the network were

removed. Thus, it creates a 2-by-2 design where a transition between two sounds can be locally congruent (Familiar vs New transitions i.e. these two sounds are connected or not) or respect higher-order community membership (Within vs Between transitions: these two sounds belong to the same community or not). We measured familiarity judgments of subjects and showed that they were sensitive to the community structure. Their pattern of errors showed that they even hallucinated non existing transitions between two elements belonging to the same community, despite having never been heard together (New Within transitions). To explain this behavior, we compared many models proposed in the statistical learning and network learning literature and we have shown that it is the FEMM that provides the best fit to the behavior (see Fig 5.1I).

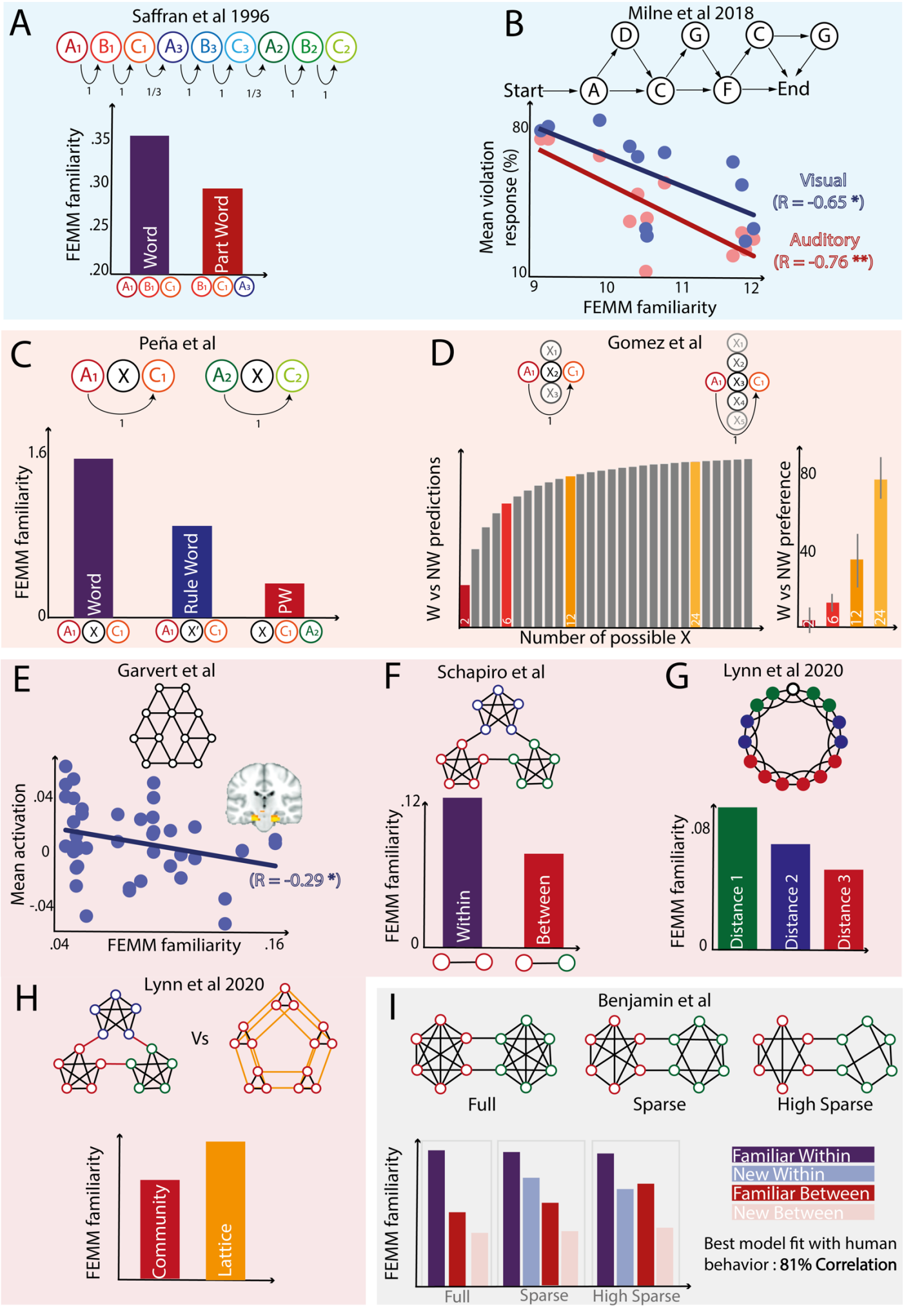

A
Saffran et al 1996
A1 B1 C1 A3 B3 C3 A2 B2 C2
1 1 1/3 1 1 1/3 1 1
FEMM familiarity
.35
.30
.25
.20
Word
Part Word
A1 B1 C1
B1 C1 A3
B
Milne et al 2018
D G C G
Start A C F End
Mean violation response (%)
80
10
9 10 11 12
FEMM familiarity
Visual
(R = -0.65 *)
Auditory
(R = -0.76 **)
C
Peña et al
A1 X C1
A2 X C2
1
FEMM familiarity
1.6
0
Word
Rule Word
PW
A1 X C1
A1 X' C1
X C1 A2
D
Gomez et al
X1 A1 X2 C1 X3
X1 X2 A1 X3 C1 X4 X5
1
W vs NW predictions
2 6 12 24
Number of possible X
W vs NW preference
80
40
2 6 12 24
E
Garvert et al
Mean activation
.04
0
-.04
(R = -0.29 *)
.04
FEMM familiarity
.16
F
Schapiro et al
FEMM familiarity
.12
0
Within
Between
G
Lynn et al 2020
FEMM familiarity
.08
0
Distance 1
Distance 2
Distance 3
H
Lynn et al 2020
Vs
FEMM familiarity
Community
Lattice
I
Benjamin et al
Full
Sparse
High Sparse
FEMM familiarity
Full
Sparse
High Sparse
Familiar Within
New Within
Familiar Between
New Between
Best model fit with human behavior : 81% Correlation

Figure 5.1 *: Re-interpeting results from different statistical learning scales with FEMM. Comparision between FEMM simulation and results from the litterature.*

A) *Segmentation task of embedded words in a continuous sequence have been proposed as a demonstration of sensitivity to adjacent transition probabilities. Bar plots displays the FEMM estimated familiarity for Word and Part-Word conditions.*

B) *Artificial grammars have also been used to investigate local transition probability learning. The mean violation response measured in Milne et al 2018 significantly correlates with FEMM familiarity in both visual (blue) and auditory (red) modalities. .*

C) *At the intermediate scale, sensitivity to non-adjacent dependencies have been demonstrated using AXC paradigms. The reported preference for RuleWord over PartWord is also correctly accounted by the FEMM.*

D) *Varying the number of possible X in the AxC paradigm causes a variation of participant's' performances in Word vs NonWord discrimination despite a similar non-adjacent transition probability P(C|XA) = 1 in all cases .The left barplot represents the FEMM estimation for number of possible Xs ranging from 2 to 29 showed an increased Word vs NonWord discrimination performance, similar to Gomez et al behavioral results (right barplot).*

E) *Using fMRI, Garvert and colleagues showed that when exposed to a sequence deriving from a network structure, a cluster in the entorhinal cortex significantly correlated with Communicability. Here, we show that the average activation within this cluster also significantly correlated with FEMM.*

F) *In the community paradigm, Schapiro and colleagues showed that participants could learn the cluster structure and differentiate transitions Within from transitions Between communities despite no local transition probability difference. This difference is well modeled by the FEMM as shown in the barplot comparing within from between transitions' familiarity .*

G) *The ring graph, has also been investigated with this framework. Authors that two connected nodes (white to green) where jugged more familiar than unconnected nodes two transitions apart (white to blue), themselves more familiar than unconnected nodes three transitions apart (white to red). The estimated FEMM familiarity displayed in the barplot correctly predicts this behavior.*

H) *They also showed that, in accordance to the FEMM predictions, learning a community structure was easier (faster reaction time) than a lattice structure. Indeed, the average FEMM familiarity is higher for lattice (orange) than for community (red) network.*

I) *Finally, in a previous experiment we introduced the sparse and high sparse community designs. We modified the community paradigm by removing possible transitions within each community, thus mixing local and high order statistical properties. The FEMM predicted a pattern of familiarity including the generalization of the structure where non-existant transitions within communities would be judged much more familiar than those between communities. The behavioral and MEG results experimentally confirmed this prediction. Critically, we compared many statistical models on this paradigm and showed that FEMM significantly outperformed the other statistical models of the literature.*

Limitations of our hypothesis

We have shown above how FEMM might account for a large variety of results from the literature, spanning different scales of statistical learning in sequences. However, we should point to the limitations of the model and to why we think that this model does not explain all participants' performances.

This model is a first approach to characterize how low-level and simple properties of the sensory system can already explain a large number of performances, but it is not envisaged to cover the entire hierarchy of processing until decision-making. For example, several studies have shown that prior expectation on the structure of the sequence can greatly influence statistical learning performances **(Elazar et al., 2022a; Lew-Williams and Saffran, 2012; Onnis and Thiessen, 2013a; Potter et al., 2017; Siegelman et al., 2018).** The FEMM does not take into account prior knowledge or expectations, and only describe the learning of the statistical structure assuming no prior knowledge. However, it might be possible to add this information, but it would require specifically designed experiments to correctly tune it. We summarize below other limitations:

Learning backward transition probabilities:

Some studies have also examined backward transition probabilities: For example, if subjects are exposed to bi-syllabic words in sentences like /fu/ /ga/ and /me/ /lo/ in sentences, but the syllables /fu/ and /me/ were also the first syllable of two other words, so that the forward transition probability within word (from /fu/ to /ga/ or from /me/ to /lo/) was 0.33, the same as the fromward transition probability between words (from /ga/ to /me/ for example). However, the backward transition probability within words (from /ga/ to /fu/ and from /lo/ to /me/) was 1 because these syllables were only present in the two words /fuga/ and /melo/ while between words (from /me/ to /ga/) was 1/9. Then the authors argued that the only available cue for segmentation are the backward transition

probabilities in that case. 8-month-old infants, as adults **(Perruchet and Desaulty, 2008b)**, were sensitive to this backward probability **(Pelucchi et al., 2009b)**. With this particular design the FEMM does predict a difference between Word and Part-Word, however the Part-Words are expected to be more familiar than the Words, unlike what have been found in the literature. However, in addition to backward transition probabilities, by design the sequence also follows a rule where every two syllables is either X, Y or Z. In particular, some parts of the sequence contain alternation, like 'X A X C X B X C…' for example. This repetition of an element every two items could serve as an anchor for segmentation of the pairs and is similar to some language sequence with alternations of pronouns followed by nouns. Participants could then rely on this alternance rule between (X,Y,Z) and the other set of syllables (A,B,C,D,E,F,G,H,I) to extract the words rather than statistical information.

Statistical learning and segmentation

We previously reported that successful segmentation of words into a continuous sequence based on transition probabilities was restricted to bi-syllabic and tri-syllabic words but failed with quadri-syllabic words **(Benjamin et al., 2023b)** without the help of a pause (**chapter 2**) or other contextual cues (**appendix 1**). Nor the local transition probability model or the FEMM can explain this result. As we wrote in this paper (**chapter 2**), we believe that segmentation of a sequence and tracking statistical regularities are different processes and, that for longer words, only statistical regularities are extracted without allowing segmentation process to build upon. As the FEMM does not describe segmentation of sequences but extraction statistical regularities, it does not help to better describe the critical difference that we observe between tri and quadri-syllabic segmentation paradigm in adults and neonates.

Statistical and rule learning

Statistical regularities are not the only type of regularities that can be extracted from sequences of input. Indeed, rule learning in sequence has also been extensively studied and refers to structures that are not optimally described by their statistics. For example, we can consider alternating sequences (ABABABAB….) or more complex rules such as alternation of repetitions (AABBAABB….), symmetry (ABCDDCBA)… In order to memorize these sequences, knowing the strength of association between pairs is not enough, and one must encode a condensed representation of the sequence. Recent studies argues for a compressed memorization of such sequences in a language like manner **(Al Roumi et al., 2021; Dehaene et al., 2022; Planton et al., 2021)**, recycling the idea of languages of thought **(Fodor, 1975; Quilty-Dunn et al., 2022)**. Yet, it is still debated whether rule and statistical learnings share a common **(Fiser and Lengyel, 2022)** or distinct **(Maheu et al., 2020)** cognitive mechanism. From the point of view we adopted in this discussion, FEMM is only capable of calculating an associative strength for each pair of elements and therefore cannot take into account rules such as repetitions, symmetries, embedding, etc. It thus goes along well with the hypothesis of separate mechanisms for statistical and rule learning in sequences, which would need more empirical support. FEMM is thus not a general model for sequence learning, but is restricted to a general model of statistical learning in sequences, and in the case of evidence for similar cognitive mechanism for rule and statistical learning this would contradict this approach and a more general model would need to be found.

Following the same line, another well studied paradigm in sequence learning is the local global paradigm **(Bekinschtein et al., 2009; King et al., 2014)**. In this paradigm, subject hear small sequences of 5 elements where the last can be the same (XXXX : locally congruent) or different (XXXY : locally deviant). But crucially those sequences are in blocks which creates a global regularity. Thus, a XXXX sequence is globally congruent in a block of XXXX

sequences, but globally deviant if presented in a block of XXXY sequences. Local and global regularities have been shown to elicit different brain responses (early localized response for the local mismatch and later more distributed one for the global mismatch) with different impact of consciousness (local mismatch is preserved in comatose or sleeping subjects while the global mismatch seem to need conscious attention). In this design, the local aspect is likely the result of a statistical computation (and thus predicted by FEMM) while the global one, in awake healthy adults, might rely on a later more abstract rule mechanism and is thus not accounted by the FEMM. Although we should remain very careful to the equivalence between a global effect and consciousness or rule learning. Indeed, the probability of XX and XY transitions are different in XXXX blocks and XXXY blocks. For example, in Basirat et al **(2014)**, the total TP over the whole block was 93.75% for XX transition in XXXX block, but only 81.25% for XX transition in XXXY block. Similarly, TP was 6.25% for XY transition in XXXX block and 18.75% for XY transition inside XXXY block. This leads to an early global effect, observed in babies and in preterm infants that very likely to be driven only by statistical effect over the whole block, rather than the rule property. As this effect is supported by TP, it fully transfers to FEMM and the correlation between the FEMM predictions and the TP predictions in this case exceed 90% and correctly accounts for the results reported in Basirat et al (2014). Here again, FEMM is compatible with this data although it does not directly improve the prediction made by adjacent TP only.

Finally, many studies described networks learning as more closely related to rule-learning than statistical learning and signature of top-down activities have been found in accordance to this hypothesis **(Ren et al., 2022).** While the statistical hypothesis for community learning that we proposed here, and rule-learning hypothesis seems in contradiction, we do not believe that they must be mutually exclusive. Indeed, our work here highlights that abstraction of the structure is not always needed to learn networks

properties, however when presented with explicit tasks that require such abstraction, subjects might form a subsequent rule-based representation of the network. Indeed, recent intra-cranial recordings conducted during local statistical learning paradigms have revealed that multiple brain regions, including both cortical areas and hippocampus, can simultaneously represent the same structure while carrying different feature information **(Henin et al., 2021)**. Similarly, it is likely that several encodings of the community structure could be maintained in parallel and in particular an abstract representation of the structure could be derived from its FEMM representation.

Variation of the generalization parameter

In all the results discussed above, we fixed the β parameter to a single value in order to avoid overfitting the model to every piece of data available and to assess the real prediction value of this model. However, the value of this parameter, which is the single free parameter of the model, drastically changes the model predictions by modulating the degree of integration of the different statistical scales. Therefore, it might be a relevant tool to study changes in this cognitive process between different contexts, ages or populations.

Modulation of the model generalization

At the limits, when $\beta \rightarrow \infty$, the model converges asymptotically to the true transition probability structure. Conversely, for $\beta \rightarrow 0$, all statistical scales are equally weighted and thus all stimulus pairs will be uniformly familiar. For all values between these two limits, the beta parameters act as a generalization factor. The lower β is, the more the memorized structure will be an overgeneralization of the real structure. The higher β is, the more the memorized structure will match the real transition probabilities. While in this paper we kept a fixed value of 0.06 based on previous studies **(Benjamin et al.,**

2023a; Lynn et al., 2020a), this parameter might in reality be variable depending on population, context or even vary at the subject level. Thus, we believe that this parameter is an appropriate modeling tool to account for possible differences in structure perceptions over individual subjects, populations, development stages or even species. Indeed, even if estimating generalization per subject is difficult, Lynn and colleagues **(2020a)** showed that an individually fitted β parameter correlated with subjects memory capacity in a n-back task, providing first evidence that not all individuals have the same generalization/accuracy trade-off.

Variation of the β parameter across different populations

Building upon this line, we speculate that this parameter could be a useful modelling approach to account for interindividual differences in network representations. For example, several studies pointed out the great generalization capacities on neonates at the expense of precision that tend to decrease with age **(Newport, 1990)**. Moreover, statistical learning capacities are known to change with development **(Forest et al., 2023)**, and thus characterizing the β parameter along development could help better investigating this question. Similarly, investigating statistical learning capacities in different species through this framework and measuring the associated generalization factor (β) of different species could help understand if learning high order structures rely on a specific human capacity or if different species have different degree of integration of statistical scales corresponding different generalization/accuracy tradeoff.

Variation of the β parameter with attention

Moreover, the generalization parameter could also depend on attention and/or active or passive involvement in the task. This hypothesis has not been formally tested yet. However, using the sparse community paradigm in an active behavioral task, we

estimated the average β of the tested population at 0.06. Using the same paradigm with completely passive MEG analysis, we estimated it at 0.5. While this difference could be due to very different measure method (behavior and MEG), active involvement in the task might also lead to greater generalization by holding long distance element longer in memory increasing the weight of long dependencies in sequences.

Variation of the β parameter with cortical area

Finally, the generalization parameter could also vary across brain regions enabling a gradient of temporal integration within cortical areas similar to what have been proposed for successor representation **(Brunec and Momennejad, 2022; Momennejad, 2020)** and that would be consistent with autocorrelation patterns measured in primate cortex **(Chaudhuri et al., 2015; Chen et al., 2015)**. This organization would enable multi scale temporal representation and flexibility in sensory integration over time. However, measuring such properties is challenging and would necessitate precise spatial and temporal resolution. In the following paragraph we discuss a possible future eCog experiment that could help investigating this question among other predictions of this model.

Predictions of our model and future work

More than accounting for a large variety of results from the existing literature, our aim was to define a new theoretical framework to allow new predictions. Indeed, this model comes with several predictions that remain to be tested. We detail some possible future explorations that might help better understand brain representations of statistical sequences.

Spatial exploration

First, if the nature of learning local statistics and network properties is indeed similar, network learning should naturally inherit from many of the local statistical learning properties. In particular, similar brain regions should be involved in paradigms studying the different scales of statistical learning. In particular, spatially defined imaging, such as fMRI or intracranial recordings, are thus necessary to explore this issue. To explore this question, we would like to re-analyze and possibly collect new intra-cranial data on patients exposed to sequence of inputs to have recordings of the brain activity with a fine-grained localization. The pattern of electrode implantation often includes the auditory cortex and hippocampus, which makes this technique adapted for the exploration of auditory sequence learning. A strong prediction of our hypothesis of similar brain mechanism for statistical and network learning, is that we should observe similar brain circuits when performing both types of tasks. This is still to be explored as our current experiments in EEG and MEG lack spatial resolution. Moreover, based on the MEG results described in **chapter 4**, we proposed that the representation of elements in a sequence would be maintained for a time sufficient to overlap several successive items. However, the brain implementation of this long-lasting profile is still unclear. In particular, thanks to eCog recordings, we would like to investigate if the generalizing factor (and so the decreasing profile of representation of the elements) is the same across different areas of the brain or if we could observe a gradient of integration within the cortex. Another question is whether this decreasing profile should be measured in absolute time or number of elements. Changing the presentation rate in sequence learning would help to tackle this question. There is also a converging literature on the role of the hippocampus in network learning but the integration with the model that we propose here is unclear. Could the hippocampus store a condensed version of the model, compressing the clusters

(see fig 3.5) ? Or does it perform a similar associative learning algorithm, possibly with a different integration decay ?

Community structure and neonates

Another prediction of our unifying hypothesis is that statistical learning skills observed in neonates should transfer to learning higher order structure modulo the duration of the overlap between elements. Crucially, the statistical learning capacities described in the first section of this PhD in sleeping neonates (**Chapter 2**) should predict their sensitivity to the community structures. To our knowledge, it has never been tested and we plan to use EEG to assess sleeping neonate's learning of the sparse community paradigm. Indeed, this would give more insight on the developmental trajectory of the cognitive capacity. In the first section of this PhD, we showed how similar adults and neonates' responses to local dependencies were. I would like to explore if this similarity extends to community structure learning and if the generalization effect is stable across development or if the great learning capacities of neonates could be explained by a greater generalization of underlying structures. For this purpose, measuring the generalization factor at different ages could provide insightful information about the change in statistical sequence representations during development.

Similarly, non-human primates and other species that have shown sensitivity to local statistics should also be tested with higher-order statistical regularities to confirm or infirm our hypothesis of a unified model. This aspect might be experimentally difficult given that the sparse community experiments requires a correct discrimination of twelve different stimuli and to maintain and update a 12x12 matrix of familiarity at all time.

Quantitative predictions about new test word in Saffran paradigm

In the Saffran et al's **(1996a)** paradigm, a different familiarity effect for certain types of test words that are not differentiated by their local statistical properties is predicted by FEMM. For instance, $A_iC_iB_{i.}$ (Shuffling of the syllables of a word) and $A_iC_kB_{j.}$ (Each syllable belonging to a different word) are both associated with null adjacent transition probabilities (TP). However, temporal proximity should favor $A_iC_iB_{i.}$ over $A_iC_kB_j$ and FEMM, contrary to adjacent TP only, would predict different familiarity with both test words. If subjects reports a familiarity difference between those two conditions, this would confirm the hypothesis that even in local statistical learning paradigms, familiarity is based on more than just adjacent Transition Probabilities (Otherwise both test words should be equally rejected). Moreover, as the chunking hypothesis postulates the extraction and memorization of chunks rather than computing statistics, it does not predict any difference between those two conditions. Thus, a difference between those two test words would also be in favor of a statistical learning approach rather than chunking.

Long distance non-adjacent dependencies

Similarly, to the AXC paradigm, non-adjacent dependencies at longer distances should also be learnable under some circumstances. Indeed, AXXB or more Generally AX*B are theoretically grasped by the FEMM. However, because of the exponential decay in the model, the difficulty of the task should exponentially increase with the distance of the dependency and conversely the associated predicted effect size should exponentially decrease. This makes a quantitative prediction that remains to be tested with an experiment specifically designed to compare effect size in learning dependencies at different distances (AB vs AXB vs AXXB vs AXXXB…). Based on the shape of the decoding profile that we observed in the MEG experiment (**chapter 4** fig 4.4A) and the overlap of tone representations lasting around 7 elements, we could speculate that those

dependencies could be captured by our brain up to five central elements (AXXXXXB). Nevertheless, the magnitude of the effect may be so small that it would be very difficult to measure it experimentally.

Hierarchical fractal structure in networks

Another prediction we can draw from this model is the sensitivity to the hierarchical fractal structure in networks. Indeed, modelling the estimated familiarity with the FEMM on a network made up of "communities of communities" predicts sensitivity to this hierarchical organization. This hypothesis is difficult to test given the number of different stimuli required to build such a network and the small expected effect size. However, it is worth noting that in such situation, properties like hierarchy or recursion can spontaneously emerge from a simple associative learning mechanism.

The role of sleep

Finally, I would like to explore the role of sleep in such learning. In particular, in the discussion of **chapter 3**, we speculated that the bias in transition probability learning that we described could later be used to form a condensed representation of the structure (see fig 3.5). However, we haven't shown the existence of such representations. Memory consolidation and forgetting during sleep have been described as possible mechanisms to re-modelled events that have been encoded during wakefulness **(Diekelmann and Born, 2010; Feld et al., 2022, 2016)**. By exposing participants to sparse community structures and testing their mental representation after sleep or wakefulness period, we would like to explore whether sleep enables to move from bias transition probability learning to a more abstract condensed representation of a two-cluster structure.

Conclusion

The exploration of the brain statistical capacities is a rich and fertile domain to understand how humans, as probably other animals, structure the environment. In this PhD we explored different statistical paradigms and tried to better defined the cognitive mechanisms at stakes. First studying neonates, we showed that learning adjacent transition probabilities and segmenting sequences were two separated mechanisms. We postulated that extracting statistical regularities was a very basic mechanism upon which other processes can build (segmentation, compression, memorization of the first element…). We then tried to extend this basic statistical learning mechanism to more complex types of probabilistic dependencies. For that, we moved to adults and showed that network learning tasks, considered as separated from statistical learning, could in fact rely on the same process. We thus proposed a unifying framework and detailed a series of predictions that gives a direction for future exploration of statistical learning in sequences.

APPENDIX

List of studies and collaborations I have been involved in during the four years spent at UNICOG.

As a 1st Author :

- Benjamin L, Dehaene-Lambertz G, Fló A.
  Remarks on the analysis of steady-state responses: spurious artifacts introduced by overlapping epochs. ***Cortex 2021***. **(Appendix 1)**
- Benjamin L, Fló A, Palu M, Naik S, Melloni L, Dehaene-Lambertz G.
  Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates. ***Developmental Science 2023***. **(Chapter 2)**
- Benjamin L, Fló A, Al Roumi F, Dehaene-Lambertz G.
  Humans parsimoniously represent auditory sequences by pruning and completing the underlying network structure. ***eLife 2023*** **(Chapter 3)**
- Benjamin L, Sablé-Meyer M, Fló A, Dehaene-Lambertz G*, Al Roumi F*
  Associative learning explains human sensitivity statistical and network structures in auditory sequences ***in prep*** **(Chapter 4)**
- Benjamin L, Dehaene-Lambertz G
  A unified mechanism for statistical sequence learning at different scales ***in prep*** **(Intro + Chapter 5)**
- Benjamin L, Di Z, Fló A, Zengxin Q, Liping W, Xuehai W, Peng G, Dehaene-Lambertz G
  Statistical learning is partially preserved in minimally conscious patients ***in prep*** **(Chapter 1)**

As a collaborator :

- Fló A, Benjamin L, Palu M, Dehaene-Lambertz G.

  Sleeping neonates track transitional probabilities in speech but only retain the first syllable of words. ***Scientific Reports* 2022**

- Fló A, Gennari G, Benjamin L, Dehaene-Lambertz G.

  Automated Pipeline for Infants Continuous EEG (APICE): A flexible pipeline for developmental cognitive studies. ***Developmental Cognitive Neuroscience* 2022 (2022b)**

- Fló A, Benjamin L, Palu M, Dehaene-Lambertz G.

  Statistical learning across multiples features of speech at birth ***in prep.***

- Sablé-Meyer M, Benjamin L, Potier Watkins C, He C, Al Roumi F, Dehaene S

  The neural mechanisms of geometric shape perception ***in prep***

References

1. *Al Roumi F, Marti S, Wang L, Amalric M, Dehaene S. 2021. Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. Neuron **109**:2627-2639.e4. doi:10.1016/j.neuron.2021.06.009*

2. *Ambrus GG, Vékony T, Janacsek K, Trimborn ABC, Kovács G, Nemeth D. 2020. When less is more: Enhanced statistical learning of non-adjacent dependencies after disruption of bilateral DLPFC. Journal of Memory and Language **114**:104144. doi:10.1016/j.jml.2020.104144*

3. *Andrillon T, Kouider S. 2016. Implicit memory for words heard during sleep. Neuroscience of Consciousness **2016**:niw014. doi:10.1093/nc/niw014*

4. *Andrillon T, Poulsen AT, Hansen LK, L É Ger D, Kouider S. 2016. Neural markers of responsiveness to the environment in human sleep. Journal of Neuroscience **36**:6583–6596. doi:10.1523/JNEUROSCI.0902-16.2016*

5. *Andringa S, Rebuschat P. 2015. NEW DIRECTIONS IN THE STUDY OF IMPLICIT AND EXPLICIT LEARNING: An Introduction. Stud Second Lang Acquis **37**:185–196. doi:10.1017/S027226311500008X*

6. *Avarguès-Weber A, Finke V, Nagy M, Szabó T, d'Amaro D, Dyer AG, Fiser J. 2020. Different mechanisms underlie implicit visual statistical learning in honey bees and humans. Proc Natl Acad Sci USA **117**:25923–25934. doi:10.1073/pnas.1919387117*

7. *Bagou O, Frauenfelder UH. 2018. Lexical Segmentation in Artificial Word Learning: The Effects of Converging Sublexical Cues. Language and Speech **61**:3–30. doi:10.1177/0023830917694664*

8. *Bakker A, Kirwan CB, Miller M, Stark CEL. 2008. Pattern separation in the human hippocampal CA3 and dentate gyrus. Science **319**:1640–1642. doi:10.1126/science.1152882*

9. *Barascud N, Pearce MT, Griffiths TD, Friston KJ, Chait M. 2016. Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. Proceedings of the National Academy of Sciences of the United States of America **113**:E616–E625. doi:10.1073/pnas.1508523113*

10. *Basirat A, Dehaene S, Dehaene-Lambertz G. 2014. A hierarchy of cortical responses to sequence violations in three-month-old infants. Cognition* ***132****:137–150. doi:10.1016/j.cognition.2014.03.013*

11. *Batterink LJ, Choi D. 2021. Optimizing steady-state responses to index statistical learning: Response to Benjamin and colleagues. Cortex. doi:10.1016/j.cortex.2021.06.008*

12. *Batterink LJ, Paller KA. 2019. Statistical learning of speech regularities can occur outside the focus of attention. Cortex* ***115****:56–71. doi:10.1016/j.cortex.2019.01.013*

13. *Batterink LJ, Paller KA. 2017. Online neural monitoring of statistical learning. Cortex* ***90****:31–45. doi:10.1016/j.cortex.2017.02.004*

14. *Batterink LJ, Zhang S. 2022. Neuropsychologia Simple statistical regularities presented during sleep are detected but not retained. Neuropsychologia* ***164****:108106. doi:10.1016/j.neuropsychologia.2021.108106*

15. *Bekinschtein T, Dehaene S, Rohaut B, Tadel F, Cohen L, Naccache L. 2009. Neural signature of the conscious processing of auditory regularities. Proceedings of the National Academy of Sciences of the United States of America* ***106****:1672–1677. doi:10.1073/pnas.0809667106*

16. *Benjamin L, Dehaene-Lambertz G, Fló A. 2021. Remarks on the analysis of steady-state responses: spurious artifacts introduced by overlapping epochs. Cortex. doi:10.1016/j.cortex.2021.05.023*

17. *Benjamin L, Fló A, Al Roumi F, Dehaene-Lambertz G. 2023a. Humans parsimoniously represent auditory sequences by pruning and completing the underlying network structure. eLife* ***12****:e86430. doi:10.7554/eLife.86430*

18. *Benjamin L, Fló A, Al Roumi F, Dehaene-Lambertz G. 2022a. Humans parsimoniously represent auditory sequences by pruning and completing the underlying network structure. bioRxiv. doi:10.1101/2022.05.19.492659*

19. *Benjamin L, Fló A, Palu M, Naik S, Melloni L, Dehaene-Lambertz G. 2023b. Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates. Developmental Science* ***26****:e13300. doi:10.1111/desc.13300*

20. *Benjamin L, Fló A, Palu M, Naik S, Melloni L, Dehaene-Lambertz G. 2022b. Tracking transitional probabilities and segmenting auditory sequences are dissociable processes in adults and neonates. Developmental Science. doi:10.1111/desc.13300*

21. Bergelson E, Swingley D. 2012. At 6–9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences **109**:3253–3258. doi:10.1073/pnas.1113380109

22. Black A, Bergmann C. 2017. Quantifying infants' statistical word segmentation: A meta-analysis. Proceedings of the 39th Annual Conference of the Cognitive Science Society 124–129.

23. Boersma P, Weenink D. 2020. Praat: doing phonetics by computer [Computer program] retrieved October 2020.

24. Boros M, Magyari L, Török D, Bozsik A, Deme A, Andics A. 2021. Neural processes underlying statistical learning for speech segmentation in dogs. Current Biology **31**:5512-5521.e5. doi:10.1016/j.cub.2021.10.017

25. Brunec IK, Momennejad I. 2022. Predictive Representations in Hippocampal and Prefrontal Hierarchies. J Neurosci **42**:299–312. doi:10.1523/JNEUROSCI.1327-21.2021

26. Bruno M-A, Vanhaudenhuyse A, Thibaut A, Moonen G, Laureys S. 2011. From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes: recent advances in our understanding of disorders of consciousness. J Neurol **258**:1373–1384. doi:10.1007/s00415-011-6114-x

27. Buiatti M, Peña M, Dehaene-Lambertz G. 2009. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. NeuroImage **44**:509–519. doi:10.1016/j.neuroimage.2008.09.015

28. Bulf H, Johnson SP, Valenza E. 2011. Visual statistical learning in the newborn infant. Cognition **121**:127–132. doi:10.1016/j.cognition.2011.06.010

29. Buzsáki G. 2006. Rhythms of the brain, Rhythms of the brain. New York, NY, US: Oxford University Press. doi:10.1093/acprof:oso/9780195301069.001.0001

30. Capilla A, Pazo-Alvarez P, Darriba A, Campo P, Gross J. 2011. Steady-State Visual Evoked Potentials Can Be Explained by Temporal Superposition of Transient Event-Related Responses. PLOS ONE **6**:e14543. doi:10.1371/journal.pone.0014543

31. Chaudhuri R, Knoblauch K, Gariel M-A, Kennedy H, Wang X-J. 2015. A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. Neuron **88**:419–431. doi:10.1016/j.neuron.2015.09.008

32. Chen J, Hasson U, Honey CJ. 2015. Processing Timescales as an Organizing Principle for Primate Cortex. Neuron **88**:244–246. doi:10.1016/j.neuron.2015.10.010

33. Choi D, Batterink LJ, Black AK, Paller KA, Werker JF. 2020. Preverbal Infants Discover Statistical Word Patterns at Similar Rates as Adults: Evidence From Neural Entrainment. Psychol Sci **31**:1161–1173. doi:10.1177/0956797620933237

34. Chomsky N. 1986. Knowledge of language: Its nature, origin and use.

35. Christophe A, Dupoux E, Bertoncini J, Mehler J. 1994. Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. Journal of the Acoustical Society of America **95**:1570–1580. doi:10.1121/1.408544

36. Constantinescu AO, Jill O, Behrens TEJ. 2016. Organizing conceptual knowledge in humans with a gridlike code **352**.

37. Conway CM, Christiansen MH. 2006. Statistical Learning Within and Between Modalities: Pitting Abstract Against Stimulus-Specific Representations. Psychol Sci **17**:905–912. doi:10.1111/j.1467-9280.2006.01801.x

38. Conway CM, Christiansen MH. 2005. Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition **31**:24–39. doi:10.1037/0278-7393.31.1.24

39. Conway CM, Christiansen MH. 2001. Sequential learning in non-human primates. Trends in Cognitive Sciences. doi:10.1016/s1364-6613(00)01800-3

40. Cowan N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences **24**:87–114. doi:10.1017/S0140525X01003922

41. Cunillera T, Càmara E, Toro JM, Marco-Pallares J, Sebastián-Galles N, Ortiz H, Pujol J, Rodríguez-Fornells A. 2009. Time course and functional neuroanatomy of speech segmentation in adults. NeuroImage **48**:541–553. doi:10.1016/j.neuroimage.2009.06.069

42. Dayan P. 1993. Improving Generalisation for Temporal Difference Learning: The Successor Representation 14.

43. de Heering A, Rossion B. 2015. Rapid categorization of natural face images in the infant right hemisphere. eLife **4**:e06564. doi:10.7554/eLife.06564

44. de Leeuw JR. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behav Res **47**:1–12. doi:10.3758/s13428-014-0458-y

45. DeCasper A, Fifer W. 1980. Of Human Bonding : Newborns prefer their mother's voice. Science.

46. DeCasper AJ, Lecanuet J-P, Busnel M-C, Granier-Deferre C, Maugeais R. 1994. Fetal reactions to recurrent maternal speech. Infant Behavior and Development **17**:159–164. doi:10.1016/0163-6383(94)90051-5

47. Dehaene S, Al Roumi F, Lakretz Y, Planton S, Sablé-Meyer M. 2022. Symbols and mental programs: a hypothesis about human singularity. Trends in Cognitive Sciences **26**:751–766. doi:10.1016/j.tics.2022.06.010

48. Dehaene S, Charles L, King JR, Marti S. 2014. Toward a computational theory of conscious processing. Current Opinion in Neurobiology **25**:76–84. doi:10.1016/j.conb.2013.12.005

49. Dehaene S, Meyniel F, Wacongne C, Wang L, Pallier C. 2015. The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. Neuron **88**:2–19. doi:10.1016/j.neuron.2015.09.019

50. Dehaene-Lambertz G, Dehaene S, Hertz-Pannier L. 2002. Functional neuroimaging of speech perception in infants. Science **298**:2013–2015. doi:10.1126/science.1077066

51. Delorme A, Makeig S. 2004. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods **134**:9–21. doi:10.1016/j.jneumeth.2003.10.009

52. Diekelmann S, Born J. 2010. The memory function of sleep. Nat Rev Neurosci **11**:114–126. doi:10.1038/nrn2762

53. Ding N, Melloni L, Yang A, Wang Y, Zhang W, Poeppel D. 2017. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Frontiers in Human Neuroscience **11**:1–9. doi:10.3389/fnhum.2017.00481

54. Ding N, Melloni L, Zhang H, Tian X, Poeppel D. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci **19**:158–164. doi:10.1038/nn.4186

55. *Ding N, Melloni L, Zhang H, Tian X, Poeppel D. 2014. Cortical Tracking of Hierarchical Linguistic Structures in Connected Speech. Physiology & behavior* ***63****:1–18. doi:10.1038/nn.4186.Cortical*

56. *Doelling KB, Florencia Assaneo M, Bevilacqua D, Pesaran B, Poeppel D. 2019. An oscillator model better predicts cortical entrainment to music. Proceedings of the National Academy of Sciences of the United States of America* ***116****:10113–10121. doi:10.1073/pnas.1816414116*

57. *Dutoit T, Pagel C, Pierret N, Baraille E, der Vrecken van. 1996. The Mbrola project : Towards a set of high quality speech synthesizers free of use for non commercial purposes.*

58. *Elazar A, Alhama RG, Bogaerts L, Siegelman N, Baus C, Frost R. 2022a. When the "Tabula" is Anything but "Rasa:" What Determines Performance in the Auditory Statistical Learning Task? Cognitive Science* ***46****:e13102. doi:10.1111/cogs.13102*

59. *Elazar A, Alhama RG, Bogaerts L, Siegelman N, Baus C, Frost R. 2022b. When the "Tabula" is Anything but "Rasa:" What Determines Performance in the Auditory Statistical Learning Task? Cogn Sci* ***46****:e13102. doi:10.1111/cogs.13102*

60. *Ellis CT, Skalaban LJ, Yates TS, Bejjanki VR, Córdova NI, Turk-Browne NB. 2021. Evidence of hippocampal learning in human infants. Current Biology* ***31****:3358-3364.e4. doi:10.1016/j.cub.2021.04.072*

61. *Endress AD. 2010. Learning melodies from non-adjacent tones. Acta Psychologica* ***135****:182–190. doi:10.1016/j.actpsy.2010.06.005*

62. *Endress AD, Johnson SP. 2021. When forgetting fosters learning: A neural network model for statistical learning. Cognition 104621. doi:10.1016/j.cognition.2021.104621*

63. *Endress AD, Mehler J. 2009. The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. Journal of Memory and Language* ***60****:351–367. doi:10.1016/j.jml.2008.10.003*

64. *Fama ME, Schuler KD, Newport EL, Turkeltaub PE. 2022. Effects of healthy aging and left hemisphere stroke on statistical language learning. Lang Cogn Neurosci* ***37****:984–999. doi:10.1080/23273798.2022.2030481*

65. *Farthouat J, Atas A, Wens V, De Tiege X, Peigneux P. 2018. Lack of frequency-tagged magnetic responses suggests statistical regularities remain undetected during NREM sleep. Scientific Reports* ***8****:1–16. doi:10.1038/s41598-018-30105-5*

66. Feld GB, Bernard M, Rawson AB, Spiers HJ. 2022. Sleep targets highly connected global and local nodes to aid consolidation of learned graph networks. Sci Rep **12**:15086. doi:10.1038/s41598-022-17747-2

67. Feld GB, Weis PP, Born J. 2016. The Limited Capacity of Sleep-Dependent Memory Consolidation. Frontiers in Psychology **7**.

68. Fenson L, Dale PS, Reznick JS, Bate E, Thal DJ, Pethick SJ. 1994. Variability in early communicative development. Monographs of the Society for Research in Child Development **59**:1–4. doi:10.1111/j.1540-5834.1994.tb00169.x

69. Fernandes T, Kolinsky R, Ventura P. 2010. The impact of attention load on the use of statistical information and coarticulation as speech segmentation cues. Attention, Perception, and Psychophysics **72**:1522–1532. doi:10.3758/APP.72.6.1522

70. Fernandes T, Ventura P, Kolinsky R. 2007. Statistical information and coarticulation as cues to word boundaries: a matter of signal quality.

71. Ferry AL, Fló A, Brusini P, Cattarossi L, Macagno F, Nespor M, Mehler J. 2016. On the edge of language acquisition: Inherent constraints on encoding multisyllabic sequences in the neonate brain. Developmental Science **19**:488–503. doi:10.1111/desc.12323

72. Fiser J, Aslin RN. 2005. Encoding Multielement Scenes: Statistical Learning of Visual Feature Hierarchies. Journal of Experimental Psychology: General **134**:521–537. doi:10.1037/0096-3445.134.4.521

73. Fiser J, Aslin RN. 2002. Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences of the United States of America **99**:15822–15826. doi:10.1073/pnas.232472899

74. Fiser J, Lengyel G. 2022. Statistical Learning in Vision. Annu Rev Vis Sci **8**:265–290. doi:10.1146/annurev-vision-100720-103343

75. Fló A, Benjamin L, Palu M, Dehaene-Lambertz G. 2022a. Sleeping neonates track transitional probabilities in speech but only retain the first syllable of words. Scientific Reports **12**:4391. doi:10.1038/s41598-022-08411-w

76. Fló A, Brusini P, Macagno F, Nespor M, Mehler J, Ferry AL. 2019. Newborns are sensitive to multiple cues for word segmentation in continuous speech. Developmental Science **22**. doi:10.1111/desc.12802

77. Fló A, Gennari G, Benjamin L, Dehaene-Lambertz G. 2022b. Automated Pipeline for Infants Continuous EEG (APICE): A flexible pipeline for developmental cognitive studies. Developmental Cognitive Neuroscience **54**:101077. doi:10.1016/j.dcn.2022.101077

78. Fodor J. 1975. The Language of Thought.

79. Forest TA, Schlichting ML, Duncan KD, Finn AS. 2023. Changes in statistical learning across development. Nat Rev Psychol. doi:10.1038/s44159-023-00157-0

80. French RM, Addyman C, Mareschal D. 2011. TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. Psychological Review **118**:614–636. doi:10.1037/a0025255

81. Garvert MM, Dolan RJ, Behrens TEJ. 2017. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. eLife **6**:1–20. doi:10.7554/eLife.17086

82. Geschwind N. 1970. The Organization of Language and the Brain. Science.

83. Giacino JT, Ashwal S, Childs N, Cranford R, Jennett B, Katz DI, Kelly JP, Rosenberg JH, Whyte J, Zafonte RD, Zasler ND. 2002. CME The minimally conscious state.

84. Giacino JT, Kalmar K, Whyte J. 2004. The JFK Coma Recovery Scale-Revised: Measurement characteristics and diagnostic utility1. Archives of Physical Medicine and Rehabilitation **85**:2020–2029. doi:10.1016/j.apmr.2004.02.033

85. Giraud AL, Poeppel D. 2012. Cortical oscillations and speech processing: Emerging computational principles and operations. Nature Neuroscience **15**:511–517. doi:10.1038/nn.3063

86. Girvan M, Newman MEJ. 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America **99**:7821–7826. doi:10.1073/pnas.122653799

87. Goldfine AM, Victor JD, Conte MM, Bardin JC, Schiff ND. 2011. Determination of awareness in patients with severe brain injury using EEG power spectral analysis. Clinical Neurophysiology **122**:2157–2168. doi:10.1016/j.clinph.2011.03.022

88. Gomez R. 2017. Variability and detection of invariant structure. SciFed Journal of AIDS & HIV Research **1**:431–436. doi:10.23959/sfahrj-1000006

*89. Gramfort A, Luessi M, Larson E, Engemann D, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen M. 2013. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience **7**.*

*90. Green C. 2017. Usage-based linguistics and the magic number four. Cognitive Linguistics **28**:209–237. doi:10.1515/cog-2015-0112*

*91. Gui P, Jiang Y, Zang D, Qi Z, Tan J, Tanigawa H, Jiang J, Wen Y, Xu L, Zhao J, Mao Y, Poo M ming, Ding N, Dehaene S, Wu X, Wang L. 2020. Assessing the depth of language processing in patients with disorders of consciousness. Nature Neuroscience **23**:761–770. doi:10.1038/s41593-020-0639-1*

*92. Harris ZS. 1955. From Phoneme to Morpheme. Language **31**:190–222. doi:10.2307/411036*

*93. Hassin RR, Bargh JA, Engell AD, McCulloch KC. 2009. Implicit working memory. Consciousness and Cognition **18**:665–678. doi:10.1016/j.concog.2009.04.003*

*94. Hauser MD, Newport EL, Aslin RN. 2001. Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. Cognition **78**:53–64. doi:10.1016/S0010-0277(00)00132-3*

*95. Hay JF, Pelucchi B, Estes KG, Saffran JR. 2011. Linking sounds to meanings: Infant statistical learning in a natural language. Cognitive Psychology **63**:93–106. doi:10.1016/j.cogpsych.2011.06.002*

96. Hebb DO. 1949. The organization of behavior: a neuropsychological theory. Mahwah, N.J: L. Erlbaum Associates.

*97. Henin S, Turk-Browne NB, Friedman D, Liu A, Dugan P, Flinker A, Doyle W, Devinsky O, Melloni L. 2021. Learning hierarchical sequence representations across human cortex and hippocampus. Science Advances **7**:1–13. doi:10.1126/sciadv.abc4530*

*98. Hochmann JR, Endress AD, Mehler J. 2010. Word frequency as a cue for identifying function words in infancy. Cognition **115**:444–457. doi:10.1016/j.cognition.2010.03.006*

99. Hublin J-J, Seytre B. 2011. Quand d'autres hommes peuplaient la Terre : Nouveaux regards sur nos origines.

*100. Isbilen ES, McCauley SM, Kidd E, Christiansen MH. 2020. Statistically Induced Chunking Recall: A Memory-Based Approach to Statistical Learning. Cognitive Science **44**:e12848. doi:10.1111/cogs.12848*

*101. James LS, Sun H, Wada K, Sakata JT. 2020. Statistical learning for vocal sequence acquisition in a songbird. Scientific Reports **10**:1–18. doi:10.1038/s41598-020-58983-8*

*102. Jas M, Engemann DA, Bekhti Y, Raimondo F, Gramfort A. 2017. Autoreject: Automated artifact rejection for MEG and EEG data. NeuroImage **159**:417–429. doi:10.1016/j.neuroimage.2017.06.030*

*103. Jas M, Larson E, Engemann DA, Leppäkangas J, Taulu S, Hämäläinen M, Gramfort A. 2018. A Reproducible MEG/EEG Group Study With the MNE Software: Recommendations, Quality Assessments, and Good Practices. Front Neurosci **12**:530. doi:10.3389/fnins.2018.00530*

*104. Johnson EK, Tyler MD. 2010. Testing the limits of statistical learning for word segmentation. Developmental Science **13**:339–345. doi:10.1111/j.1467-7687.2009.00886.x*

105. Jonathan D, Reece D. 2020. eSpeaker, [Computer program] retrieved October 2020.

*106. Kabdebon C, Pena M, Buiatti M, Dehaene-Lambertz G. 2015. Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. Brain and Language **148**:25–36. doi:10.1016/j.bandl.2015.03.005*

*107. Kahn AE, Karuza EA, Vettel JM, Bassett DS. 2018. Network constraints on learnability of probabilistic motor sequences. Nature Human Behaviour **2**:936–947. doi:10.1038/s41562-018-0463-8*

*108. Kakaei E, Aleshin S, Braun J. 2021. Visual object recognition is facilitated by temporal community structure. Learn Mem **28**:148–152. doi:10.1101/lm.053306.120*

*109. Karuza EA, Kahn AE, Bassett DS. 2019. Human sensitivity to community structure is robust to topological variation. Complexity **2019**. doi:10.1155/2019/8379321*

*110. Karuza EA, Kahn AE, Thompson-Schill SL, Bassett DS. 2017. Process reveals structure: How a network is traversed mediates expectations about its architecture. Scientific Reports **7**:1–9. doi:10.1038/s41598-017-12876-5*

111. Karuza EA, Thompson-Schill SL, Bassett DS. 2016. Local Patterns to Global Architectures: Influences of Network Topology on Human Learning. Trends in Cognitive Sciences **20**:629–640. doi:10.1016/j.tics.2016.06.003

112. King JR, Gramfort A, Schurger A, Naccache L, Dehaene S. 2014. Two distinct dynamic modes subtend the detection of unexpected sounds. PLoS ONE **9**. doi:10.1371/journal.pone.0085791

113. Kirkham NZ, Slemmer JA, Johnson SP. 2002. Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition **83**:B35–B42. doi:10.1016/S0010-0277(02)00004-5

114. Koechlin E, Jubault T. 2006. Broca's Area and the Hierarchical Organization of Human Behavior. Neuron **50**:963–974. doi:10.1016/j.neuron.2006.05.017

115. Koechlin E, Ody C, Kouneiher F. 2003. The Architecture of Cognitive Control in the Human Prefrontal Cortex. Science **302**:1181–1185. doi:10.1126/science.1088545

116. Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin J-J, Hänni C, Fortea J, Rasilla M de la, Bertranpetit J, Rosas A, Pääbo S. 2007. The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals. Current Biology **17**:1908–1912. doi:10.1016/j.cub.2007.10.008

117. Kuhl PK. 2004. Early language acquisition: cracking the speech code. Nat Rev Neurosci **5**:831–843. doi:10.1038/nrn1533

118. Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. Science **255**:606–608. doi:10.1126/science.1736364

119. Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature **413**:519–523. doi:10.1038/35097076

120. Laureys S, Schiff ND. 2012. Coma and consciousness: Paradigms (re)framed by neuroimaging. NeuroImage **61**:478–491. doi:10.1016/j.neuroimage.2011.12.041

121. Lecanuet JP, Granier-Deferre C, Jacquet AY, Busnel MC. 1992. Decelerative cardiac responsiveness to acoustical stimulation in the near term fetus. Q J Exp Psychol B **44**:279–303. doi:10.1080/02724999208250616

122. Lew-Williams C, Pelucchi B, Saffran JR. 2011. Isolated words enhance statistical language learning in infancy. Developmental Science **14**:1323–1329. doi:10.1111/j.1467-7687.2011.01079.x

123. Lew-Williams C, Saffran JR. 2012. All words are not created equal: Expectations about word length guide infant statistical learning. Cognition **122**:241–246. doi:10.1016/j.cognition.2011.10.007

124. Liu KY, Gould RL, Coulson MC, Ward EV, Howard RJ. 2016. Tests of pattern separation and pattern completion in humans-A systematic review. Hippocampus **26**:705–717. doi:10.1002/hipo.22561

125. Locke J. 1690. Essai sur l'entendement humain.

126. López-Barroso D, Catani M, Ripollés P, Dell'Acqua F, Rodríguez-Fornells A, de Diego-Balaguer R. 2013. Word learning is mediated by the left arcuate fasciculus. Proc Natl Acad Sci U S A **110**:13168–13173. doi:10.1073/pnas.1301696110

127. Lynn CW, Bassett DS. 2020. How humans learn and represent networks. Proceedings of the National Academy of Sciences of the United States of America **117**. doi:10.1073/pnas.1912328117

128. Lynn CW, Kahn AE, Nyema N, Bassett DS. 2020a. Abstract representations of events arise from mental errors in learning and memory. Nature Communications **11**. doi:10.1038/s41467-020-15146-7

129. Lynn CW, Papadopoulos L, Kahn AE, Bassett DS. 2020b. Human information processing in complex networks. Nature Physics **16**:965–973. doi:10.1038/s41567-020-0924-7

130. MacDermot KD, Bonora E, Sykes N, Coupe A-M, Lai CSL, Vernes SC, Vargha-Khadem F, McKenzie F, Smith RL, Monaco AP, Fisher SE. 2005. Identification of FOXP2 Truncation as a Novel Cause of Developmental Speech and Language Deficits. The American Journal of Human Genetics **76**:1074–1080. doi:10.1086/430841

131. Maheu M, Dehaene S, Meyniel F. 2019. Brain signatures of a multiscale process of sequence learning in humans. eLife **8**:1–24. doi:10.7554/eLife.41541

132. Maheu M, Meyniel F, Dehaene S. 2020. Rational arbitration between statistics and rules in human sequence learning. doi:10.1101/2020.02.06.937706

133. Mandel DR, Jusczyk PW, Kemler Nelson DG. 1994. Does sentential prosody help infants organize and remember speech information? Cognition **53**:155–180. doi:10.1016/0010-0277(94)90069-8

134. Marchetto E, Bonatti LL. 2015. Finding words and word structure in artificial speech: The development of infants' sensitivity to morphosyntactic regularities. Journal of Child Language **42**:873–902. doi:10.1017/S0305000914000452

135. Marcus GF, Vijayan S, Bandi Rao S, Vishton PM. 1999. Rule learning by seven-month-old infants. Science **283**:77–80. doi:10.1126/science.283.5398.77

136. Mark S, Moran R, Parr T, Kennerley SW, Behrens TEJ. 2020. Transferring structural knowledge across cognitive maps in humans and models. Nature Communications **11**:1–12. doi:10.1038/s41467-020-18254-6

137. Marr D. 1982. Vision: a computational investigation into the human representation and processing of visual information. Cambridge, Mass: MIT Press.

138. Mattys SL, White L, Melhorn JF. 2005. Integration of multiple speech segmentation cues: A hierarchical framework. Journal of Experimental Psychology: General **134**:477–500. doi:10.1037/0096-3445.134.4.477

139. McNealy K, Mazziotta JC, Dapretto M. 2006. Cracking the Language Code: Neural Mechanisms Underlying Speech Parsing. J Neurosci **26**:7629–7639. doi:10.1523/JNEUROSCI.5501-05.2006

140. Mehler J, Bertoncini J, Barriere M. 1978. Infant recognition of mother's voice. Perception **7**:491–497. doi:10.1068/p070491

141. Mehler J, Jusczyk P, Lambertz G, Halsted N, Bertoncini J, Amiel-Tison C. 1988. A precursor of language acquisition in young infants. Cognition **29**:143–178. doi:10.1016/0010-0277(88)90035-2

142. Menyhart O, Kolodny O, Goldstein MH, DeVoogd TJ, Edelman S. 2015. Juvenile zebra finches learn the underlying structural regularities of their fathers' song. Front Psychol **6**. doi:10.3389/fpsyg.2015.00571

143. Milne AE, Petkov CI, Wilson B. 2018. Auditory and Visual Sequence Learning in Humans and Monkeys using an Artificial Grammar Learning Paradigm. Neuroscience **389**:104–117. doi:10.1016/j.neuroscience.2017.06.059

144. Momennejad I. 2020. Learning Structures: Predictive Representations, Replay, and Generalization. Current Opinion in Behavioral Sciences **32**:155–166. doi:10.1016/j.cobeha.2020.02.017

145. Nemeth D, Janacsek K, Polner B, Kovacs ZA. 2013. Boosting Human Learning by Hypnosis. Cerebral Cortex **23**:801–805. doi:10.1093/cercor/bhs068

146. Nespor M, Vogel I. 2006. Prosodic phonology, The Dialects of Italy. doi:10.4324/9780203993880-15

147. Newman MEJ. 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America **103**:8577–8582. doi:10.1073/pnas.0601602103

148. Newman MEJ. 2003. The structure and function of complex networks. SIAM Review **45**:167–256. doi:10.1137/S003614450342480

149. Newport EL. 1990. Maturational constraints on language learning. Cognitive Science **14**:11–28. doi:10.1016/0364-0213(90)90024-Q

150. Newport EL, Aslin RN. 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. Cognitive Psychology **48**:127–162. doi:10.1016/S0010-0285(03)00128-2

151. Newport EL, Hauser MD, Spaepen G, Aslin RN. 2004. Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. Cognitive Psychology **49**:85–117. doi:10.1016/j.cogpsych.2003.12.002

152. Nicholls EK, Hempel de Ibarra N. 2014. Bees associate colour cues with differences in pollen rewards. Journal of Experimental Biology jeb.106120. doi:10.1242/jeb.106120

153. Niso G, Gorgolewski KJ, Bock E, Brooks TL, Flandin G, Gramfort A, Henson RN, Jas M, Litvak V, T. Moreau J, Oostenveld R, Schoffelen J-M, Tadel F, Wexler J, Baillet S. 2018. MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. Sci Data **5**:180110. doi:10.1038/sdata.2018.110

154. Norcia AM, Appelbaum LG, Ales JM, Cottereau BR, Rossion B. 2015. The steady-state visual evoked potential in vision research: A review. J Vis **15**:4. doi:10.1167/15.6.4

155. Norman KA, O'Reilly RC. 2003. Modeling Hippocampal and Neocortical Contributions to Recognition Memory: A Complementary-Learning-Systems Approach. Psychological Review **110**:611–646. doi:10.1037/0033-295X.110.4.611

156. Onnis L, Thiessen E. 2013a. Language experience changes subsequent learning. Cognition **126**:268–284. doi:10.1016/j.cognition.2012.10.008

157. Onnis L, Thiessen E. 2013b. Language experience changes subsequent learning. Cognition **126**:268–284. doi:10.1016/j.cognition.2012.10.008

158. Ordin M, Polyanskaya L, Laka I, Nespor M. 2017. Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. Memory and Cognition **45**:863–876. doi:10.3758/s13421-017-0700-9

159. Palmer SD, Mattys SL. 2016. Speech segmentation by statistical learning is supported by domain-general processes within working memory. Quarterly Journal of Experimental Psychology **69**:2390–2401. doi:10.1080/17470218.2015.1112825

160. Pelucchi B, Hay JF, Saffran JR. 2009a. Learning in reverse: Eight-month-old infants track backward transitional probabilities. Cognition **113**:244–247. doi:10.1016/j.cognition.2009.07.011

161. Pelucchi B, Hay JF, Saffran JR. 2009b. Learning in reverse: Eight-month-old infants track backward transitional probabilities. Cognition **113**:244–247. doi:10.1016/j.cognition.2009.07.011

162. Pena M. 2002. Rôle du calcul statistique dans l'acquisition du langage.

163. Peña M, Bonatti LL, Nespor M, Mehler J. 2002. Signal-driven computations in speech processing. Science **298**:604–607. doi:10.1126/science.1072901

164. Perruchet P. 2019. What Mechanisms Underlie Implicit Statistical Learning? Transitional Probabilities Versus Chunks in Language Learning. Top Cogn Sci **11**:520–535. doi:10.1111/tops.12403

165. Perruchet P, Desaulty S. 2008a. A role for backward transitional probabilities in word segmentation? Memory & Cognition **36**:1299–1305. doi:10.3758/MC.36.7.1299

166. Perruchet P, Desaulty S. 2008b. A role for backward transitional probabilities in word segmentation? Memory & Cognition **36**:1299–1305. doi:10.3758/MC.36.7.1299

*167. Perruchet P, Poulin-Charronnat B. 2012. Beyond transitional probability computations: Extracting word-like units when only statistical information is available. Journal of Memory and Language **66**:807–818. doi:10.1016/j.jml.2012.02.010*

*168. Perruchet P, Vinter A. 1998. PARSER: A Model for Word Segmentation. Journal of Memory and Language **39**:246–263. doi:10.1006/jmla.1998.2576*

*169. Peykarjou S, Hoehl S, Pauen S, Rossion B. 2017. Rapid Categorization of Human and Ape Faces in 9-Month-Old Infants Revealed by Fast Periodic Visual Stimulation. Sci Rep **7**:12526. doi:10.1038/s41598-017-12760-2*

170. Piaget J. 1964. Six études de psychologie.

171. Picq P, Sagart L, Dehaene-Lambertz G, Lestienne C. 2008. La plus belle histoire du langage.

*172. Picton TW, John MS, Dimitrijevic A, Purcell D. 2003. Human auditory steady-state responses: Respuestas auditivas de estado estable en humanos. International Journal of Audiology **42**:177–219. doi:10.3109/14992020309101316*

173. Planton S, van Kerkoerle T, Abbih L, Maheu M, Meyniel F, Sigman M, Wang L, Figueira S, Romano S, Dehaene S. 2021. A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. PLoS Computational Biology. doi:10.1371/journal.pcbi.1008598

*174. Polyanskaya L. 2022. Cognitive mechanisms of statistical learning and segmentation of continuous sensory input. Mem Cogn **50**:979–996. doi:10.3758/s13421-021-01264-0*

*175. Pothos EM, Juola P. 2007. Characterizing linguistic structure with mutual information. British Journal of Psychology **98**:291–304. doi:10.1348/000712606X122760*

*176. Potter CE, Wang T, Saffran JR. 2017. Second Language Experience Facilitates Statistical Learning of Novel Linguistic Materials. Cognitive Science **41**:913–927. doi:10.1111/cogs.12473*

177. Pudhiyidath A, Morton NW, Duran RV, Schapiro AC, Hinojosa-Rowland DM, Molitor RJ, Preston AR. 2022. Representations of temporal community structure in hippocampus and precuneus predict inductive reasoning decisions 51.

*178. Pudhiyidath A, Roome HE, Coughlin C, Nguyen KV, Preston AR. 2020. Developmental differences in temporal schema acquisition impact reasoning decisions. Cognitive Neuropsychology **37**:25–45. doi:10.1080/02643294.2019.1667316*

179. Querleu D, Lefebvre C, Titran M, Renard X, Morillion M, Crepin G. 1984. Reaction of the newborn infant less than 2 hours after birth to the maternal voice. J Gynecol Obstet Biol Reprod (Paris) **13**:125–134.

180. Quilty-Dunn J, Porot N, Mandelbaum E. 2022. The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences. Behavioral and Brain Sciences 1–55. doi:10.1017/S0140525X22002849

181. Reber AS. 1967. Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior **6**:855–863. doi:10.1016/S0022-5371(67)80149-X

182. Regan D. 1977. Steady-state evoked potentials. J Opt Soc Am, JOSA **67**:1475–1489. doi:10.1364/JOSA.67.001475

183. Ren X, Zhang H, Luo H. 2022. Dynamic emergence of relational structure network in human brains. Progress in Neurobiology **219**:102373. doi:10.1016/j.pneurobio.2022.102373

184. Rosenblum LD, Schmuckler MA, Johnson JA. 1997. The McGurk effect in infants. Percept Psychophys **59**:347–357. doi:10.3758/bf03211902

185. Sablé-Meyer M, Ellis K, Tenenbaum J, Dehaene S. 2022. A language of thought for the mental representation of geometric shapes. PsyArXiv. doi:10.31234/osf.io/28mg4

186. Sablé-Meyer M, Fagot J, Caparos S, van Kerkoerle T, Amalric M, Dehaene S. 2021. Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. Proceedings of the National Academy of Sciences of the United States of America **118**:1–10. doi:10.1073/pnas.2023123118

187. Saffran JR, Aslin RN, Newport EL. 1996a. Statistical Learning by 8-Month-Old Infants. Science **274**:1926–1928.

188. Saffran JR, Johnson EK, Aslin RN, Newport EL. 1999. Statistical learning of tone sequences by human infants and adults. Cognition **70**:27–52. doi:10.1016/S0010-0277(98)00075-4

189. Saffran JR, Kirkham NZ. 2018. Infant Statistical Learning. Annu Rev Psychol **69**:181–203. doi:10.1146/annurev-psych-122216-011805

190. Saffran JR, Newport EL, Aslin RN. 1996b. Word segmentation: The role of distributional cues. Journal of Memory and Language **35**:606–621. doi:10.1006/jmla.1996.0032

*191. Saffran JR, Wilson DP. 2003. From Syllables to Syntax: Multilevel Statistical Learning by 12-Month-Old Infants. Infancy **4**:273–284. doi:10.1207/S15327078IN0402_07*

*192. Santolin C, Rosa-Salva O, Regolin L, Vallortigara G. 2016. Generalization of visual regularities in newly hatched chicks (Gallus gallus). Anim Cogn **19**:1007–1017. doi:10.1007/s10071-016-1005-2*

*193. Schapiro AC, Gregory E, Landau B, McCloskey M, Turk-Browne NB. 2014. The necessity of the medial temporal lobe for statistical learning. J Cogn Neurosci **26**:1736–1747. doi:10.1162/jocn_a_00578*

*194. Schapiro AC, Kustner LV, Turk-Browne NB. 2012. Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. Current Biology **22**:1622–1627. doi:10.1016/j.cub.2012.06.056*

*195. Schapiro AC, Rogers TT, Cordova NI, Turk- NB, Botvinick MM. 2013. Neural representations of events arise from temporal community structure **16**:486–492. doi:10.1038/nn.3331.Neural*

*196. Schapiro AC, Turk-Browne NB, Botvinick MM, Norman KA. 2017. Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. Philosophical Transactions of the Royal Society B: Biological Sciences **372**. doi:10.1098/rstb.2016.0049*

*197. Schapiro AC, Turk-Browne NB, Norman KA, Botvinick MM. 2016. Statistical learning of temporal community structure in the hippocampus. Hippocampus **26**:3–8. doi:10.1002/hipo.22523*

*198. Scher MS. 2008. Ontogeny of EEG-sleep from neonatal through infancy periods. Sleep Medicine **9**:615–636. doi:10.1016/j.sleep.2007.08.014*

*199. Schön D, Boyer M, Moreno S, Besson M, Peretz I, Kolinsky R. 2008. Songs as an aid for language acquisition. Cognition **106**:975–983. doi:10.1016/j.cognition.2007.03.005*

*200. Shukla M, Nespor M, Mehler J. 2007. An interaction between prosody and statistics in the segmentation of fluent speech. Cognitive Psychology **54**:1–32. doi:10.1016/j.cogpsych.2006.04.002*

*201. Shukla M, White KS, Aslin RN. 2011. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. Proceedings of the National Academy of Sciences of the United States of America **108**:6038–6043. doi:10.1073/pnas.1017617108*

*202. Siegelman N, Bogaerts L, Elazar A, Arciuli J, Frost R. 2018. Linguistic entrenchment: Prior knowledge impacts statistical learning performance. Cognition **177**:198–213. doi:10.1016/j.cognition.2018.04.011*

*203. Siew CSQ. 2013. Community structure in the phonological network. Front Psychol **4**. doi:10.3389/fpsyg.2013.00553*

204. Sigurd B, Van De Weijer J. 2004. Word length, sentence length and frequency – Zipf revisited.

*205. Simon HA. 1962. The Architecture of Complexity. Source: Proceedings of the American Philosophical Society **106**:467–482.*

*206. Slone LK, Johnson SP. 2018. When learning goes beyond statistics: Infants represent visual sequences in terms of chunks. Cognition **178**:92–102. doi:10.1016/j.cognition.2018.05.016*

207. Smalle EHM, Daikoku T, Szmalec A, Duyck W, M€ Ott€ Onen R. n.d. Unlocking adults' implicit statistical learning by cognitive depletion. doi:10.1073/pnas.2026011119/-/DCSupplemental

*208. Sonnweber R, Ravignani A, Fitch WT. 2015. Non-adjacent visual dependency learning in chimpanzees. Anim Cogn **18**:733–745. doi:10.1007/s10071-015-0840-x*

*209. Southwell R, Chait M. 2018. Enhanced deviant responses in patterned relative to random sound sequences. Cortex **109**:92–103. doi:10.1016/j.cortex.2018.08.032*

*210. Stachenfeld KL, Botvinick MM, Gershman SJ. 2017. The hippocampus as a predictive map. Nat Neurosci **20**:1643–1653. doi:10.1038/nn.4650*

*211. Stiso J, Lynn CW, Kahn AE, Rangarajan V, Szymula KP, Archer R, Revell A, Stein JM, Litt B, Davis KA, Lucas TH, Bassett DS. 2022. Neurophysiological Evidence for Cognitive Map Formation during Sequence Learning. eNeuro **9**:ENEURO.0361-21.2022. doi:10.1523/ENEURO.0361-21.2022*

*212. Strauss M, Sitt JD, King JR, Elbaz M, Azizi L, Buiatti M, Naccache L, Van Wassenhove V, Dehaene S. 2015. Disruption of hierarchical predictive coding during sleep. Proceedings of the National Academy of Sciences of the United States of America **112**:E1353–E1362. doi:10.1073/pnas.1501026112*

*213. Swingley D. 2005. Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology **50**:86–132. doi:10.1016/j.cogpsych.2004.06.001*

*214. Takahasi M, Yamada H, Okanoya K. 2010. Statistical and Prosodic Cues for Song Segmentation Learning by Bengalese Finches (Lonchura striata var. domestica): Song segmentation learning in finches. Ethology **116**:481–489. doi:10.1111/j.1439-0310.2010.01772.x*

*215. Teinonen T, Fellman V, Näätänen R, Alku P, Huotilainen M. 2009. Statistical language learning in neonates revealed by event-related brain potentials. BMC Neuroscience **10**. doi:10.1186/1471-2202-10-21*

*216. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. 2011. How to grow a mind: Statistics, structure, and abstraction. Science **331**:1279–1285. doi:10.1126/science.1192788*

*217. Todorovic A, de Lange FP. 2012. Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields. Journal of Neuroscience **32**:13389–13395. doi:10.1523/JNEUROSCI.2227-12.2012*

*218. Toro JM, Sinnett S, Soto-Faraco S. 2005. Speech segmentation by statistical learning depends on attention. Cognition **97**:25–34. doi:10.1016/j.cognition.2005.01.006*

*219. Toro JM, Trobalón JB. 2005. Statistical computations over a speech stream in a rodent. Perception and Psychophysics **67**:867–875. doi:10.3758/BF03193539*

*220. Tummeltshammer K, Amso D, French RM, Kirkham NZ. 2017. Across space and time: infants learn from backward and forward visual statistics. Dev Sci **20**:e12474. doi:10.1111/desc.12474*

*221. Tyler MD, Cutler A. 2009. Cross-language differences in cue use for speech segmentation. The Journal of the Acoustical Society of America **126**:367–376. doi:10.1121/1.3129127*

222. Varga N, Morton N, Preston A. 2022. Schema, Inference, and Memory. doi:10.31234/osf.io/m9adb

*223. Vouloumanos A, Hauser MD, Werker JF, Martin A. 2010. The Tuning of Human Neonates' Preference for Speech. Child Development **81**:517–527. doi:10.1111/j.1467-8624.2009.01412.x*

*224. Vouloumanos A, Werker JF. 2007. Listening to language at birth: evidence for a bias for speech in neonates. Developmental Science **10**:159–164. doi:10.1111/j.1467-7687.2007.00549.x*

*225. Vouloumanos A, Werker JF. 2004. Tuned to the signal: the privileged status of speech for young infants. Developmental Science **7**:270–276. doi:10.1111/j.1467-7687.2004.00345.x*

*226. Wakefield JA, Doughtie EB, Lee Yom BH. 1974. The identification of structural components of an unknown language. Journal of Psycholinguistic Research **3**:261–269. doi:10.1007/BF01069242*

*227. Wang FH, Zevin JD, Trueswell JC, Mintz TH. 2020. Top-down grouping affects adjacent dependency learning. Psychonomic Bulletin and Review **27**:1052–1058. doi:10.3758/s13423-020-01759-y*

*228. Wang T, Saffran JR. 2014. Statistical learning of a tonal language: the influence of bilingualism and previous linguistic experience. Frontiers in Psychology **5**.*

*229. Werker JF, Gilbert JHV, Humphrey K, Tees RC. 1981. Developmental Aspects of Cross-Language Speech Perception. Child Development **52**:349–355. doi:10.2307/1129249*

*230. Werker JF, Tees RC. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development **7**:49–63. doi:10.1016/S0163-6383(84)80022-3*

*231. Werker JF, Yeung HH, Yoshida KA. 2012. How Do Infants Become Experts at Native-Speech Perception? Curr Dir Psychol Sci **21**:221–226. doi:10.1177/0963721412449459*

*232. Whittington JCR, Muller TH, Mark S, Chen G, Barry C, Burgess N, Behrens TEJ. 2020. The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. Cell **183**:1249-1263.e23. doi:10.1016/j.cell.2020.10.024*

233. Xu C, Li H, Gao Jiaxin, Li L, He F, Yu J, Ling Y, Gao Jian, Li J, Melloni L, Luo B, Ding N. 2022. Statistical learning in patients in the minimally conscious state. Cerebral Cortex bhac222. doi:10.1093/cercor/bhac222

*234. Yassa MA, Stark CEL. 2011. Pattern separation in the hippocampus. Trends in Neurosciences **34**:515–525. doi:10.1016/j.tins.2011.06.006*

*235. Zhao S, Chait M, Dick F, Dayan P, Furukawa S, Liao H-I. 2019. Pupil-linked phasic arousal evoked by violation but not emergence of regularity within rapid sound sequences. Nat Commun **10**:4030. doi:10.1038/s41467-019-12048-1*

*236. Zhou H, Melloni L, Poeppel D, Ding N. 2016a. Interpretations of Frequency Domain Analyses of Neural Entrainment: Periodicity, Fundamental Frequency, and Harmonics. Front Hum Neurosci **10**:274. doi:10.3389/fnhum.2016.00274*

*237. Zhou H, Melloni L, Poeppel D, Ding N. 2016b. Interpretations of frequency domain analyses of neural entrainment: Periodicity, fundamental frequency, and harmonics. Frontiers in Human Neuroscience **10**:1–8. doi:10.3389/fnhum.2016.00274*

238. Zipf GK. 1935. The psycho-biology of language, The psycho-biology of language. Oxford, England: Houghton, Mifflin.