Decoding Speech Comprehension from Continuous EEG recordings

Imperial College of London

Department of Bioengineering

EPSRC
Centre for Doctoral Training in
NEUROTECHNOLOGY
for life and health

Supervisor:
Tobias Reichenbach

Hugo Weissbart
hugo.weissbart12@imperial.ac.uk

Co-supervisors:
Robert Leech
Richard Wise
& Étienne Brudet

Submitted in part fulfilment of the requirement for the degree of Doctor of Philosophy in Bioengineering at Imperial College London

December 2019

Declaration of originality

This thesis is a presentation of my original research work. It is submitted to the university of Imperial College in support of my application for the degree of Doctor in Philosophy in neurotechnology. It has been composed by myself and has not been submitted in any previous application for any degree. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature. This work was done under the guidance and supervision of Tobias Reichenbach, at Imperial College London.

Copyright Declaration

Acknowledgements

A very special thanks goes to my other group members, Katerina Kandylaki for all the help she gave me during the EEG recording session, Octave Etard for the discussions on EEG processing and analysis, and the rest of the group with whom I shared experiences that built into friendship beyond the walls of our office: Marina Saiz Alia, Antonio Ellia Forte, Laura Sumner, Nikola Ciganovic, Shabnam Kadir, Mikolaj Kegler. I owe a lot to all my friends that I met in London, with who I shared a roof, a but of life, and a lot of wisdom.

Throughout the years of PhD, I have learned much more than I could ever expect, and mostly on non technical and non scientific aspects. The most beautiful times unfolded onto painful but teaching moments, I rejuvenated with that encounter and grew older with the departure. Nevertheless I am bounded, and spiritually, and morally, I discovered new perspectives on both life and work, and for this my gratitude goes to Diana Bicazan.

I would like to thank my supervisor Tobias Reichenbach, who let me set my pace into the project and allowed me the freedom to chose a direction of my own for this thesis.

I received generous support from friends who took the time to read part of the thesis and bring helpful comments and corrections. In particular, I wish to thank Phoebe Shaw Stewart, Clément Chardon and my brother Gauthier Weissbart for reading through some parts and for having the patience and heart to help during their holidays.

The patience and support from my mother and father were priceless. They stayed comprehensive and gave me the time and support needed throughout all these years of PhD. I hope they will be relieved to see this work coming to completion inasmuch as I am grateful to them.

Finally, the most precious support I received while I wrote this thesis came from Nora Aurrekoetxea. She made the seemingly unbearable look lighter, and motivated me to write when I needed it the most. I thank her from the bottom of my heart.

Abstract

Human language is a remarkable manifestation of our cognitive abilities which is unique to our species. It is key to communication, but also to our faculty of generating complex thoughts. We organise, conceptualise, and share ideas through language. Neuroscience has shed insightful lights on our understanding of how language is processed by the brain although the exact neural organisation, structural or functional, underpinning this processing remains poorly known. This project aims to employ new methodology to understand speech comprehension during naturalistic listening condition. One achievement of this thesis lies in bringing evidence towards putative predictive processing mechanisms for language comprehension and confront those with rule-based grammar processing. Namely, we looked on the one hand at cortical responses to information-theoretic measures that are relevant for predictive coding in the context of language processing and on the other hand to the response to syntactic tree structures. We successfully recorded responses to linguistic features from continuous EEG recordings during naturalistic speech listening. The use of ecologically valid stimuli allowed us to embed neural response in the context in which they naturally occur when hearing speech. This fostered the development of new analysis tools adapted for such experimental designs. Finally, we demonstrate the ability to decode comprehension from the EEG signals of participants with above-chance accuracy. This could be used as a better indicator of the severity and specificity of language disorders, and also to assess if a patient in a vegetative state understands speech without the need for any behavioural response. Hence a primary outcome is our contribution to the neurobiology of language comprehension. Furthermore, our results pave the way to the development of a new range of diagnostic tools to measure speech comprehension of patients with language impairment.

Contents

List of Figures

List of Tables

List of Abbreviations

BMI brain-machine interface. 2

CCA Canonical Components Analysis. 69, 75–78, 91

CFG context-free grammar. 65, 94

CTI Cortical Tracking Index. xi, 86–88

ECoG electro-corticography. 14, 95, 96

EEG electroencephalogram. 16, 17, 27–29, 37, 38, 40–42, 46, 50, 51, 56, 57, 66–70, 75–77, 84

ELAN early left anterior negativity. 17, 50, 63, 64

ERP Event Related Potentials. 12, 16, 17, 19, 50, 53, 63, 79, 82

fMRI functional Magnetic Resonance Imaging. 17, 95

LAN left anterior negativity. 17, 50, 65

LSTM Long Short-Term Memory. 14, 98

MEG magento-encephalogram. 16, 50

NLP Natural Language Processing. 13, 19

OLS Ordinary Least Square. 86, 87

PCFG probabilistic context-free grammar. 56, 59

PSG Phrase Structure Grammar. 52

RNN Recurrent Neural Network. 13, 14, 52, 98

TRF Temporal Response Functions. 40–42, 46, 57, 58, 63, 68, 69

Chapter I

Introduction

> “ *The builders of prison make miserable poets compared to the architects of freedom.* ”
>
> Stig Daggerman, *Our need for consolation is impossible to satisfy*, 1952

One of the most fascinating faculties of human cognition, evidently the forerunner of our excellent reasoning ability, but also a fundamental ingredient to the development of our society and culture, is the language faculty. We employed language to pass on knowledge that survives across generations but also to create knowledge. The organisation and structure of language allows to build complex train of thoughts, mental narratives, and plans of actions of the kind that might have helped our ancestors to develop more advanced strategies. Homo Sapiens could use it to foster ideation within themselves as well as to exchange those ideas among themselves. Language plays a crucial role in the high reasoning capacity of modern human. It surpasses simple forms of communication observed in other animals and seems to coincide with the development of our species on evolutionary terms.

As early as philosophy began in the Western world, in Ancient Greece, we find questioning about language. It was clear to many that living creatures were given or endowed with a *voice*, thought to emerge from their souls. But for humans this voice seems to glue meaning to sound and speaks with a richer palette of sounds, obviously different from animals. Why is that so? Why do human struggle to learn and acquire language at birth while other animals quickly pick their vocalization? What rules govern the structure employed in language so that we can understand each other, persuade each other? From sophists of the pre-Socratic era to Aristotle, Greek philosophers were already asking those questions, pointing at that key human faculty and opening the very first disciplines related to language such as the art of rhetoric, poetry, dramaturgy, and logic.

In modern terms, the disciplines revolving around the study of language changed designation but the same motivation and passion is seen within the current, more scholarly, work about language. Various research programs, born in different disciplines, are nowadays dedicated to the study of language. Linguistics is the science dealing with language in its many forms, deciphering rules of syntax, of semantics, and looking at its evolution across civilisations. Psycholinguists link psychological analysis with linguistic approaches, and more generally cognitive neuroscientists link it to more fundamental questions that emerged from the field of neurosciences. These disciplines evolve sometimes on the their own, but also by bridging ideas between them, and nowadays we are witnessing the development of an entire research program for the neurobiology of language. The need for interdisciplinary research is clearer now than ever, when we observe convergence between works on the genetic account of language, on bird songs development in evolutionary biology, or on statistical language modelling in artificial intelligence.

On a clinical perspective, there is growing pressure to *repair* hearing and in particular speech perception (as this is the first form of complaints in the elderly) in our ever older society. Hearing aids are evolving and slowly merge with brain-machine interfaces (BMIs) technologies. As a basic example, let us imagine a brain machine interface that from electrical activity recorded with electrodes placed around the ear and head would detect in real time how well you understood the incoming speech stream, adapting the noise cancellation and sound level according to the running "comprehension index".

Language impairment can be devastating. The inability to communicate one's own thoughts and emotions or to comprehend other's results in the isolation of oneself from friends and family. Among the elderly there is a greater likelihood of strokes occurring as a consequence of various conditions, such as hypertension or trauma, that can lead to brain damage in regions that are key for language processing. Aphasia is one type of language impairment where comprehension, or production, or even both can be impacted. Today, it is still diagnosed by using battery of questionnaires (Bruce and Edmundson, 2010; Azuar et al., 2013) rather than neuroimaging assessment. The route to rehabilitation is thus unclear for stroke patients diagnosed with aphasia. It is crucial to understand better the neurobiology of language processing and how to measure the mechanisms at play with current technology in order to comprehensively explain and treat speech impairments. This thesis is focussed primarily on the basic scientific knowledge of language and how we can measure the underlying processing mechanisms with EEG. However, from an engineering perspective this work could also be applied through the use of BMI devices to assess comprehension in aphasic patients for instance. We hope to see the work presented in this thesis standing as a fundamental toolkit or approach to the design of such interfaces.

In this first chapter, we will point to the development of ideas that emerge from studies on language. A first section will combine elements of cognitive science, linguistics and evolutionary neurolinguistics to lay out the foundations and basic scientific motivation behind the study of language processing. The second section will describe current knowledge on how it is implemented in the brain to further evolve into a section about how we observe, given our neuroimaging technologies, natural language processing as it occurs in the brain. Finally, after a recap on the different methods that will be used throughout the thesis, we will conclude the introductory chapter with a section about the organisation of the thesis.

I.1 Studying language, understanding human cognition

What is the need to understand language from a neuroscientific perspective? Traditionally, linguists and psychologists have worked through research programs on language, analysing its structure and development in many ways but discarding the path and methodology borrowed from cognitive neuroscience. A major obstacle for a long time has been the lack of appropriate tools to bridge these two fields. However, with the recent bloom in development of neuro-imaging technologies, we have witnessed a unification. Many aspects of how the brain is studied are unavailable for studying the faculty of language, for example the luxury of a good comprehensive animal model. Furthermore, decades of linguistics study demonstrate the complexity of human language. It is one thing to agree that the essence of language is represented and controlled by the brain, but to underpin the many mechanisms at play is another. We know for example that several levels of representations must co-exist to synthesise a linguistic construct, or percept. This goes from a representation of sound level or visual input to the syntax and semantic aspects of language. Moreover, cognitive studies in psychology revealed neural systems, as for working memory for instance, that we ought to link up with linguistic constraint. The question of domain specificity of cortical networks is key to the understanding of how language, among other cognitive function, is processed by the brain (Collins and Hagoort, 2000). However this statement should be approached with caution as we observe an accumulation of evidence suggesting a new shift in the perspective for the interpretation of neurocognitive model of language comprehension (Hasson et al., 2018).

This ability comes as highly specific to the human species, and as a greatly evolved trait. It is sometimes even thought to be the pillar of human cognition. Many questions arise naturally when we start questioning the origin, evolution and implementation of human language.

Perhaps the most obvious is the question regarding its origin. How did it evolve? Surely, gaining knowledge on the biological evolution from a genetic and functional perspective should cast some light on how it is implemented. Especially given that evolution works over very long time-scales, mostly gradually with small genetic mutations at a time (which can have massive phenotypical changes). For the language faculty this is however a very debatable topic. The origin and necessity of language is not clear in the current state of research on that matter. Indeed, the evident fact that we observe such a radical change in the hominid lineal branch with respect to cognitive abilities is somehow in opposition to the classical view of natural selection on random walks of genetic mutations for evolution. This enormous gap is left unexplained. Yet, evolutionary biologists have found even a small unlikely change that gives a a strong competitive advantage can be crucial in the development of a species thus justifying the existence of human language within the Darwinist framework.

Naturally, this requires an explanation of what kind of advantage language was giving to our ancestors that they did not already have with, say, former forms of communications. Berwick and Chomsky, (2016) and Berwick et al., (2013) argue that language is the cornerstone to advanced reasoning. It allows individuals to elaborate more complex trains of thoughts, plans and strategies. This stands against the theories of language as a product of the pressure for a more developed mean of communication, or the necessity for gossip. The latter is also a current hypothesis for the birth of the language faculty. Mainly those investigations revolve around the question of knowing which come first between language and higher cognitive abilities. The neurobiology of language asks those questions and continues with the analysis of neural processes responsible for comprehending and producing human language. Any attempt at answering these questions still generates many debates across a variety of fields in sciences. But those attempts are primordial for the quest of understanding our origin, and grasping the gist of what makes us stand apart from the rest of the animal realm.

With those first ideas laid out, we see an emerging concept that language stands apart from animal communication. It is important to narrow down our definition of language here, as we take it as a unique human faculty as well as means of communication and source for abstract and symbolic thinking and reasoning. It will not matter which of those two streams human language stems from for the rest of the present thesis, however it is crucial that we define properly the bound and scope of the object under study here. As said, language paradoxically can carry many different meanings, from programming *language* in computing science to animal *communication*. We take the following definitions from Berwick and Chomsky, 2016, p. 1:

Definition 1 ***Language:*** *the ability to construct a digitally infinite array of hierarchically structured expressions with determinate interpretations at the interfaces with other organic systems. Hence it is a finite computational system yielding an infinity of expressions, each of which has a definite interpretation in semantic-pragmatic and sensorimotor systems.*

In this thesis, we will focus on *human language* ability, and on the neurobiological mechanism underpinning this capacity. More precisely, we will look at the *neurobiology of naturalistic speech comprehension*. The term *naturalistic* will be explained later in section I.3.2 on page 18 of this introductory chapter. Moreover we are narrowing down the study to speech perception, that is language perceived through auditory modality and also to the comprehension side of language processing rather than production or discourse, pragmatic analysis.

I.2 Speech and language in the brain

> *Aristotle's classic dictum that language is sound with meaning should be reversed. Language is meaning with sound.*
>
> R. C. Berwick & N. Chomsky, *Why Only Us*, 2017

One great advance in language studies, neurolinguistics and probably in neuroscience in general started from insights about language processing. Indeed, during the past centuries, several discoveries shed new light on major aspects of the functioning of our brain. It started with studies on brain injuries responsible for aphasias. These studies gave birth to the standard view of specialized brain areas (Broca, 1865; Wernicke, 1874), that are not only specialized and localized but also interconnected and working concurrently in a distributed and parallel fashion. Nowadays this is the most modern view for neuroscientists, nevertheless old theories about models of language stood untouched through the history of neuroscience for unclear reasons since it was these same models which opened the mind of researchers about the brain. Broca claimed we *talk with the left hemisphere* and we can still hear that phrase echoing in people's minds today. This is one aspect of the theory of language processing which has raised debates among neuroscientists (Fedorenko, Nieto-Castañón, and Kanwisher, 2012). The last decade has seen a rapid development of new functional brain imaging techniques such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI), electroencephalography (EEG) and magnetoencephalography (MEG) and they allowed scientist to examine further the circuits and networks involved in language processing (Turken and Dronkers, 2011).

The thesis focuses on speech comprehension, hence we are casting light with respect to language processing research from two distinct perspectives, namely speech and comprehension. Each of those narrows down the scope of research and limits the study to some particular brain processes. Studying *speech* implies that we are looking at language processing in the brain while information entered through the auditory modality. Hence this relates to neural mechanisms for hearing (section I.2.1). The *comprehension* side on the other hand implies that we are sweeping aside speech production mechanisms. Our research focuses on understanding language as someone else's message from which the listener has to extract meaning. Even though production of speech is probably highly tied to comprehension in term of encoding principles in the cortex, the methodology applied and the set of hypotheses raised in our study are specific to the receiver's viewpoint.

After a gentle introduction to the peripheral auditory system, we will present a more in depth overview of the different language theories of the brain, from past to present. And we will finish this introduction with an overview of a non language-centric framework, namely predictive coding, that can explain results from many language studies and on which the work of the present thesis is framed.

I.2.1 Auditory pathway: from sound to speech

Hearing, one of the primary mammalian senses, starts with sound entering the outer ear. Sound itself is an acoustic pressure wave, a propagation of mechanical energy, consisting of an alternation of compression and rarefaction of molecules in an elastic medium. When travelling through air, the wave propagates at approximately 340 m/s. The velocity is determined by the coefficient of elasticity and the density of the medium in which the wave propagates. The visible part of the human ear acts as an antenna to focus the acoustic energy into the ear canal. The pressure wave is thus redirected onto the tympanum, or ear drum, a thin membrane that is put in motion by the incoming wave. This motion is then transferred to a system of three small ossicles: the malleus, incus and stapes. Those bones act like a lever and a piston (the piston being the stapes) and are a key component in the transmission of the airborne sound waveform into a wave in a fluid inside the cochlea. As we are matching a wave travelling in air, to a wave in a denser fluid, we need an impedance matching mechanism. In order to transmit a wave (i.e. avoiding pure reflection) at the interface between two media with different impedances (the resistive force against movement of a pressure wave in the given medium), a device must sit at the interface to efficiently transfer the energy. That is exactly the role of these ossicles. When the ear drum is put into motion, and hence the stapes, the oval window—to which the stapes are attached—is pushed back and forth. This in turn will provoke a pressure wave to propagate in the fluid filling the cochlea compartments called *scalae*. The cochlea is a snail shaped structure, enclosing two communicating *scalae*, the *scala vesitbuli* and the *scala tympani*. Without entering into further details, we will explain the basic principle underlying the machinery of the inner ear. Those fluid filled compartments are separated by a thin flexible membrane, the *basilar membrane*. The propagation of the sound wave in the fluid of the *scalae* create a movement of the basilar membrane. A first important stage of sound analysis occurs at this level. Indeed, the basilar membrane has a varying stiffness and width, such that close to the base it is wider and more elastic than at the apex. These physical properties allow for different locations on the membrane along the cochlea to respond preferably, through resonance, to different frequencies. In other words, the basilar membrane acts already as a spectral analyser and spreads mechanical energy along its length according to the frequency content of the sound. The final stage consists of a transduction mechanism where the mechanical energy will be transformed into electrical signals to be sent to the central nervous system. This transduction is carried out by inner hair cells in the so-called organ of Corti. Inner Hair cells are sensory neurons located all along the basilar membrane. When tilted, so when a mechanical force, or a displacement, is applied to the cilia of the cell, its membrane potential gets depolarised. Enough depolarisation leads to an action potential to be generated and sent through the axon of the inner hair cell which are then depolarising neurons constituting the auditory nerve (Hudspeth, 2013).

The spiking activity is then forwarded through the auditory pathway up to the cortex. This pathway is a rich deployment of hierarchically organised stages of processing with feedforward, feedback and recurrent connections at each level. From the inner hair cells, the pathway consists of nuclei in the brainstem and finishes in the primary auditory cortices, located in Brodmann area 42 and 41.

The brainstem is responsible for several accounts in hearing perception. Sound localisation, pitch determination and even auditory stream selectivity is observed at the brainstem level. Across the different nuclei composing the auditory pathway in the brainstem up to the auditory cortices, we observe a *tonotopical* organisation of neural populations. The spectral decomposition stemming from the spatial organisation along the basilar membrane in the inner ear is preserved at least through to the primary auditory cortices (Oertel and Doupe, 2013).

I.2.2 Speech perception: from the Classical Model to modern theories of language

The "Broca–Wernicke–Lichtheim–Geschwind" model

> “*Nous parlons avec l’hémisphère gauche!*”
>
> Paul Broca, *1864*

The "Broca–Wernicke–Lichtheim–Geschwind" model, often referred to as the Wernicke-Lichtheim model or more simply the Classic model, is one of the oldest and probably the most cited cognitive model. It stands as a neuroanatomic model for both language production and comprehension. In the late 19$^{\text{th}}$ century, neurologists such as Paul Broca and Carl Wernicke developed this model for language processing using deficit-lesion analysis. They were treating patients who had speech deficits after suffering from a stroke, and by observing location and spread of brain lesions from post-mortem analysis, Broca and Wernicke could infer how some specific areas of the brain were playing a functional role in language production or comprehension. The original idea that language function is impaired when lesion occurred in the left hemisphere can even be traced back earlier, in 1836, from notes of the neurologist Marc Dax (Dax, 1865). As depicted in figure I.1, the Classical model consists of three components: Broca’s area, Wernicke’s area (together with the angular gyrus) and the arcuate fasciculus. The "modern" version of the model, that we find in most neuropsychology and medical textbooks, is essentially the model as updated by Geschwind and extends the previous model from Wernicke by linking Broca’s area, responsible for speech production, and Wernicke’s area, responsible for comprehension, with a single fiber tract of white matter called the arcuate fascilucus. Back in the 19$^{\text{th}}$ century, this was the very first anatomical description of a language model in the brain and shifted the view of neuroscience towards localists theories of cognition. These theories assumed that brain areas are specialised: distinct brain areas are responsible for different cognitive functions. This pioneering work in neurobiology of language remained the governing concept for more than a century of research in the field. As of today, we still use the idea behind this model to classify clinical syndromes such as Broca’s and Wernicke’s aphasia (which are also known as respectively non-fluent and fluent aphasia).

However, several problems stem from this original work establishing the foundation of the Classical Model. First of all, the spatial accuracy of the model is too limited to test specific hypothesis about brain and behaviour relationships; secondly it is focused on two "language-specific regions" while more regions are known to be involved and those regions might have a domain-general role, thirdly it focuses on cortical structures, and for the most part leaves out subcortical structure and rele-

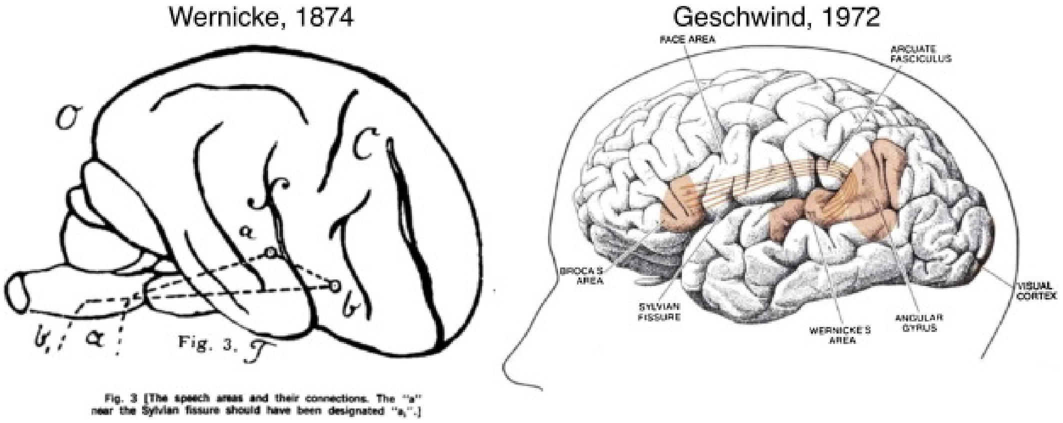


Figure I.1: Left: The original model from Wernicke, (1874). For unknown reasons, the model is represented on the right hemisphere. Right: An update of the Classic model from Geschwind, (1970). Reproduced, with permission, from Tremblay and Dick, (2016) ©(2016) Elsevier.

vant connections (there are for instance people with aphasia having only subcortical lesions, and growing amount of evidence about the implication of sub-cortical structures in speech processing, see Turken and Dronkers, 2011; Hasson and Tremblay, 2016). Even the areas described so widely in the literature are poorly defined. Broca and Wernicke did not work on many patients, Broca drew its conclusion from a single brain which he did not dissect. The brain has been recently imaged with modern techniques and revealed much wider lesions that dive into subcortical structures below the inferior frontal gyrus (Hasson et al., 2018). An example of those outdated areas are still ill defined as of today is depicted in figure I.2, where we see how Wernicke's area's boundaries vary significantly across author definitions throughout the 20th century as well as across contemporary agreed definitions from academics.

The need to modernise the neuroanatomical model, on both a structural and functional perspective, is enhanced by recent evidence of deviations from the classical model brought up with new imaging technology. Notably with fMRI researchers could analyse with unprecedented accuracy the loci of neural activity in response to a task. It is relatively fair to admit that the Classical Model is nowadays refuted as new ones emerge from evidence obtained with fMRI or EEG studies. Moreover, recent theories on language processing in the brain tackle the question with a mechanistic approach to explain *how* and *when* it is processed in more elaborate way.

Current theories

Neuroanatomical models of language aim at characterizing the brain network dedicated to language processing, as well as its functional parts. They define, in the same way the Classical Model did, the organisation of brain areas in the processing of language. To exemplify the shift in perspective for plausible neuroanatomical model of language processing, we will present two models that gained popularity in

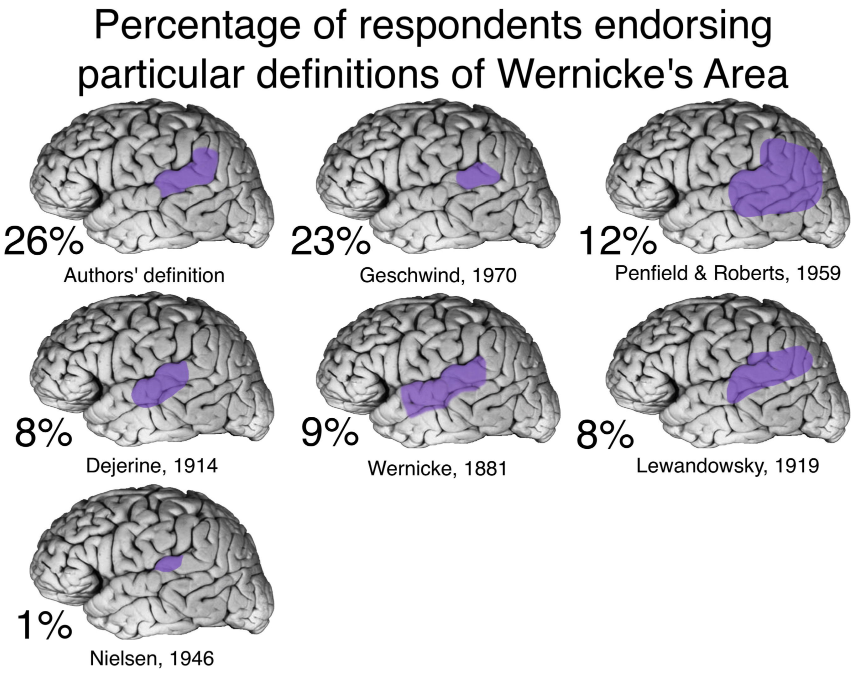


Figure I.2: Ambiguous anatomical description of Wernicke's area. The original author questioned students and professional academics in neuroscience about the definition of the Wernicke's area. Reproduced, with permission, from Tremblay and Dick, (2016) ©(2016) Elsevier.

the last decade. The first is the dual-stream model, initially described by Hickok and Poeppel, and the second will be the memory-unification-control, instanced by Hagoort. Most theories of language in the brain until recent years have focused on the formation of syntax and semantic, taking different stages of processing as separated modules that would occur in distinct brain areas. Current neuroanatomical models were built principally from meta-analysis of fMRI studies Hickok and Poeppel, 2007; Friederici, 2011, but also lesion studies and electrophysiological data. A more nuanced picture emerged where all language processing components are not necessarily lateralised.

For instance, as presented in the *dual-stream model* (Hickok and Poeppel, 2007) shown in figure I.3, partly inspired by the *what* and *where* streams found in visual processing, the mapping from sound to lexical-level meaning seem to be operated bilaterally in superior temporal lobes of both hemispheres. In contrast with the Classical model, one finds in the dorsal stream of that model connection with motor areas. Structures involved in speech production are also involved in speech perception, underpinning a sensory-motor mapping.

Hagoort, (2016) proposed another model, the Memory-Unification-Control model, abbreviated as MUC, which segregate three distinct functional components of language processing. The memory component is a language-specific component which refers to the linguistic knowledge acquired during language learning, encoded as a

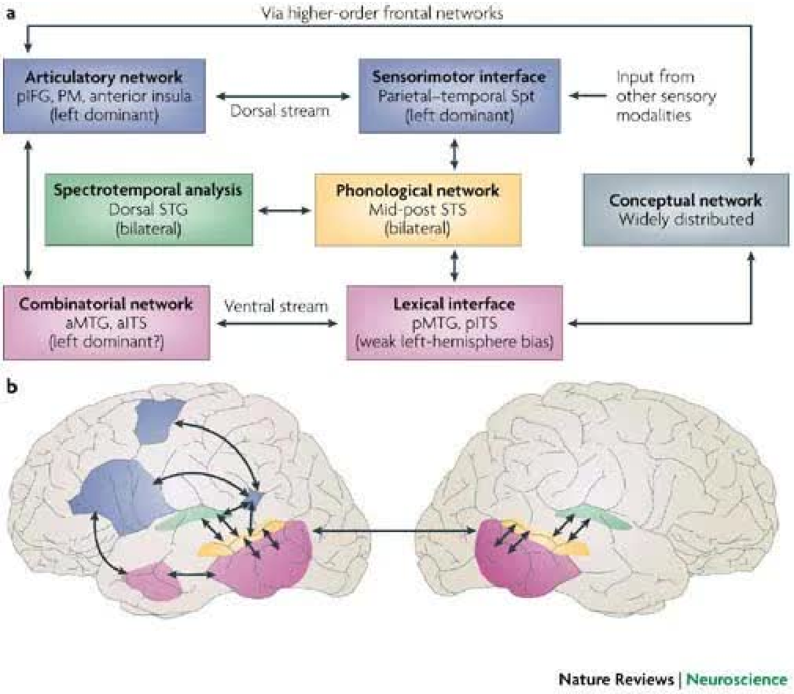


Figure I.3: Dual-stream model from Hickok and Poeppel, (2007). The dorsal stream deals with articulo-motor encoding of speech representations. Spectro-temporal and phonological processing as well as lexical access occur bilaterally. Reproduced with permission from ©2007, Springer Nature.

mapping between e.g. phonemic, spectral, syntactic representations. Then Hagoort argues that language processing is also evidently more than just a memory retrieval system that concatenates lexical entries. The combination, or composition of lexical elements makes up the richness of human language. He refers to this process as *unification* (Bastiaansen, Magyari, and Hagoort, 2010, see also), which is somehow similar to the *merge* operation defined by Chomsky, 2013 or Hasson et al., 2018. Although, here unification would not only refer to syntactic processing but would operate across all levels of representation, that is also at the semantic and phonological levels. Finally the control component relates to language as an action. This system operates the articulo-motor control for speech production, but also controls how to use language in social interactions context and at the discourse level. The brain areas potentially involved in his model are also broadly spread across both hemisphere and several non language specific areas.

Another shift in perspective from the original classical model occurred thanks to the field of linguistic which developed independently of neurosciences. In the mid 20th century, computational linguistic flourished, and slowly turned out to be an integrative part of cognitive sciences in general given at the same time the opportunity to integrate ideas from computational language model to fit in the context of neurobiological grounding.

Neuro-computational models on the other hand, define a theoretical framework on *how* linguistic inputs are processed. Poeppel, (2014) reviews anatomical organisation for speech processing but also points out at recent work characterising the *temporal* organisation and the potential role of cortical oscillations for speech processing. This aspect of neurobiology of language is core to the hypothesis and results described in this thesis.

In the dual-stream model proposed by Hickok and Poeppel, (2007), parallel routes in the mapping of acoustic input to lexical phonological representations are hypothesised. Each hemisphere is said to process sound with asymmetric sampling frequency, the left hemisphere would be mostly working in the gamma range (40-80Hz), processing segmental information while the right hemisphere would operate at the slower frequencies characterised by the theta range (4-7Hz) and would produce syllable-level representations. Again this highlights the bilateral processing of speech input.

Other models emerged that focused on simulating speech perception to match results found in the Event Related Potentials (ERP) literature (see section I.3.1). TRACE is one of such models that gained a lot of popularity. It can recognize word from the sound input by processing the spectrograms of speech in a hierarchical manner, parsing sound chunks and decoding possible phonemes first, then lexical items. The dynamic aspects incorporated in TRACE allows the model to present the time-course of word recognition as the input is being parsed. It has showed a lot of compelling result with empirical data (McClelland and Elman, 1986). Such a model was originally described from a psycholinguist perspective, but it shares similarity with modern deep neural network for speech recognition (developed independently by engineers). We briefly review below, some of the influential models developed in the field of *natural language processing* and computational linguistic that inspired some of the work of this thesis.

Statistical modelling of language stem from computational linguistic, but in the modern view of perception by inference, they now play a revived role in our appreciation of language processing in the brain.

Originally, Chomsky evoked the *poverty of stimulus* argument in *Language and problems of knowledge: the Managua lectures* (Chomsky, 1988) against the possibility of a statistical, empiricist, learning of language in human. This has shaped decades of studies in linguistic and cognitive sciences. It is known as the *rationalist* perspective. The argument consist in that language acquisition can *not* be solely due to empirical, or phenomenological, exposure to language. The development of language should rely instead on innate structure or representation of a *universal grammar* so that a child can pick quickly on generating sentences never heard before. The *empiricist* view states that indeed, through exposure and induction (or statistical inference), an agent is capable of learning language. Now the debate is still on, although the field of artificial intelligence and in particular language engineering, nowadays known as *natural language processing*, presented models capable of generalizing well on unseen sentences as well as learning grammatical rules (Manning and Schütze, 1999, see Chapter I, p.5).

This argument was stated far before the rise of more efficient computing methods for statistical learning. Nowadays, it is clear that a statistical model, if trained properly, can inject probability mass to events it has not encountered during the training phase. In the field of *natural language processing*, or NLP, modelling language through a probabilistic generative model is a key operation, that is currently broadly used in all our mobile phone that do speech recognition. The idea is to estimate the probability of occurring for a given word sequence $w_1, \ldots, w_T$. A first order estimation in such implementation would only look at the probability of occurrence of a word regardless of the previous items in the sequence. If we take the assumption that words can co-occur independently, the probability of a full sequence becomes the product of the probability of each word individually. However, words are organised in sentences following some grammatical rules, making up the syntax of a language. Early language models were based on such co-occurrence statistics were the probability of a word conditioned on the $N-1$ words preceding it is estimated simply by counting such occurrences in a large corpus of text. This method is called N-gram language modelling (Manning and Schütze, 1999). The main limitation of such models is that they have a limited window of context (N words).

Artificial neural networks, and in particular *recurrent* neural networks (RNN) go beyond this limitation (Elman, 1990; Bengio et al., 2000; Mikolov et al., 2010). RNN-based language models encode past input in their hidden layer through a recurrent connection. Theoretically, they can encode infinitely long sequences although they are in reality limited by the implementation used and the training method employed. It is actually extremely difficult to train such an architecture of neural network with the classic stochastic gradient descent by back-propagation algorithm. The main issue is the *vanishing* gradient. Long Short-Term Memory (LSTM) are a novel architecture that aims at avoiding this problem by gating input to a memory cell (Graves, 2013). For both architectures, we train a recurrent neural network for language modelling by teaching it to predict the next word based on the previous words it encountered. The last layer is implemented as a *softmax* layer, which is a layer of artificial neurons of which the sum of activity is normalised to one. By construction, this layer is therefore interpretable as a probability distribution. For a language model, this distribution would often be defined over the entire, or a limited portion of the vocabulary of the training corpus. To reduce the numbers of parameters to be fitted and also to enhance the performance of such networks, the first layer used is an *embedding* layer. Mikolov et al., (2013) found that it is possible to embed one-hot encoded vectors representing words, that is a vector of the size of the vocabulary with every component values at zero except a one (hence "one-hot") at the word index, into a lower dimensional space that carry some lexical information such as plurality, or gender. For instance, a vocabulary set of 35000 words could be embedded into a lower vectorial representation of 300 dimensions of dense real valued vector space. Mikolov et al., (2010) called this representation "Word2Vec". Amazingly, algebraic calculus are possible in such space such that the following sum is often taken as a canonical example: $\texttt{queen} - \texttt{king} + \texttt{man} \approx \texttt{woman}$. In this example, the word vector corresponding to "queen" minus the sum of the word vectors representing respectively "man" and "king" is close to the word vector mapped to "woman". This demonstrates that the embedding learnt a meaningful manifold to represent the vocabulary in few, but dense (as opposed to one-hot encoded), dimensions. A toy example of a RNN for language modelling is illustrated in the diagrams of figure I.4.

I.3 EEG studies in speech processing

In comparison with other neural processes such as those involved in decision making, motor control or sensory perception, studying the neural basis of language models poses unique chalenges. Inherently, the faculty of language as we know it and defined earlier is unique to human. In other words, there is no true animal model in the field which greatly limits neuroscientists. Cell recordings, ablation studies, optogenetics, are unavailable when studying speech. Nevertheless, there are numerous results found in related research programs such as in animal communication research, specifically with studies on song learning in birds or communication in primates, that bring evidence regarding the evolution of language faculty for instance. Those studies may inform on the putative underlying precursors neural systems allowing human language ability to blossom.

The most common neuroimaging techniques for studying language processing in the brain are fMRI, PET and M/EEG. Teams have rarely access to electro-

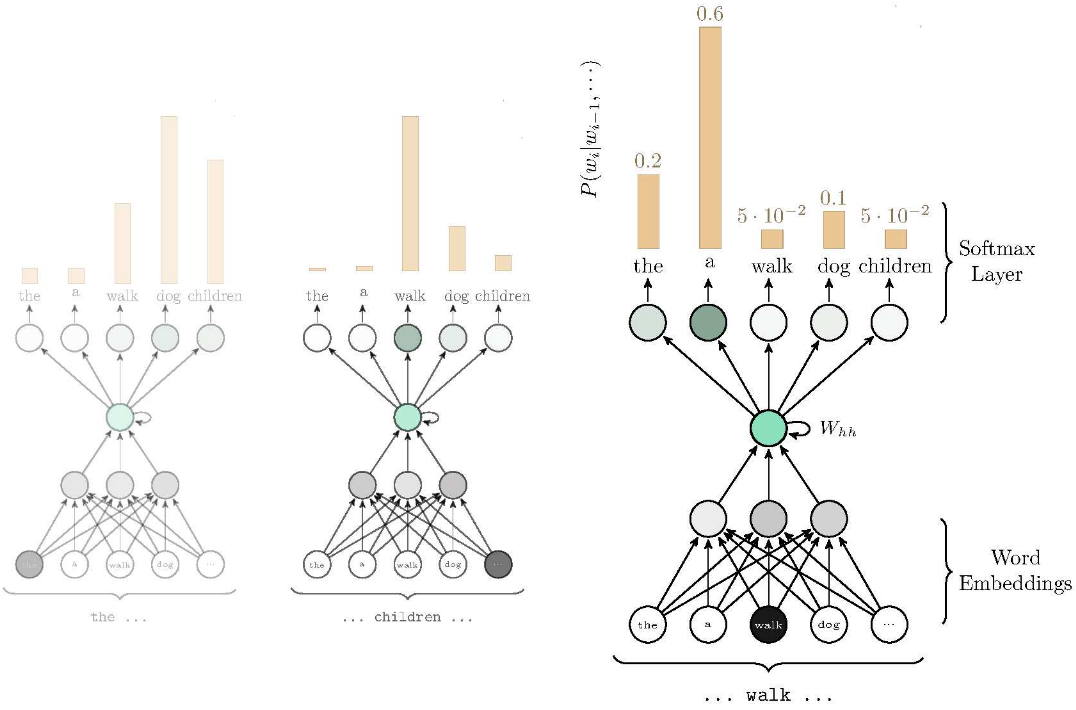


Figure I.4: Example diagram of a Recurrent Neural Network for Language modelling. In this very basic example, the vocabulary contains five words, and the embedding dimension is three. The probability mass obtained from the softmax layer is shown in a histogram above each diagram. The recurrent neural network here is evaluating the following sentence, word by word, `"the children walk a dog"`. While processing "walk", we see that the most likely next word is "a" (arbitrary values are chosen for illustration purposes).

corticography (ECoG) data because it requires setting up electrodes invasively during sub-dural surgical operations. PET/fMRI are based on the haemodynamic activity. The haemodynamic response stems from blood flow activity, the latter being partially correlated with the surrounding neuronal activity as neurons require more oxygen and nutrients when solicited. Nevertheless, haemodynamic response and neural activity are non-linearly related and are coupled through slow dynamic. As a result, fMRI or PET is better at measuring sustained brain activity but too slow to capture fast transient activity. Most studies involving PET or fMRI focused on single word responses to avoid the problem of fast dynamic integration of sentences. The biggest advantage of those imaging techniques are their spatial resolution. In fact they allow for precise mapping of language functions to brain structures. Many studies focused on the role of Broca's area for instance, and could conclude with new insights on its role and localisation thanks to such imaging techniques.

If fMRI is the tool to answer *where* the neural foci of language processing are, then electrophysiological recording technology such as electroencephalogram (EEG) or magento-encephalogram (MEG) answer the *when* question. Naturalistic speech contains information at many different time scales. Pitch and timber, as many other spectro-temporal cues are rapidly processed in the auditory cortex while in parallel phonological decoding, syntactic integration and lexical access occurs all within a short time. Moreover, information at those different timescales must be unified to form the conceptual percept representing the meaning of speech input. For instance, considering lexical phonological representations, evidence points towards a model integrating segmental information on the acoustic input at a fast rate while a parallel processing stream would parse the input on a shorter time-scale and map to syllable-level representations (dual-stream model as described in Hickok and Poeppel, (2007), see figure I.3 on page 11). Those two views of the input need to rapidly get unified to form the lexical phonological encoding of the input. A similar idea can be extrapolated to word level processing, or syntactic integration of larger constituents. In conclusion, speech input is highly dynamic, and the extraction of meaning occurs at near real-time. The neural machinery encoding speech must also reflect fast dynamic that are difficult or impossible to capture with fMRI or PET. In order to capture the time-course of brain activity to linguistic features, researchers relied mostly on the analysis of Event Related Potentials (ERP) or ERF (MEG counterpart) as we will discuss in the following section. Recently, new analysis methods have been developed and successfully used with electrophysiological recording to track natural speech processing, those new tools will be described in section I.3.3.

I.3.1 ERP studies

Most of the work looking at neural correlates of language processing with EEG relied on the computation and analysis of Event Related Potentialss (ERPs) to interpret neural activity. An ERP is the time aligned average EEG response to an aspect of the stimulus, for instance relative to the onset of a visual stimulus. It contains rich spatio-temporal dynamics usually regarded as different "ERP components". It informs whether the brain, and by inference, cognitive processing, is specifically responding to experimental manipulations, for instance semantic or syntactic violations (see Kutas and Hillyard, 1980; Kutas and Federmeier, 2011, for a review). Peak of activity in ERPs are thus pinned to local processing of corresponding brain areas. The interpretation of location of underlying electric dipole is fuzzy although methods exist to infer such information. The main advantage stems from the temporal resolution of EEG. It allows to analyse the time course of the response to a stimulus on a millisecond timescale, which behavioural paradigms or fMRI analysis cannot offer. By careful design of experimental paradigms, researchers could map the different components of word-related responses to distinct linguistic processes. Such careful designs imply almost in every case to insert controlled violations of specific aspect of speech as semantic, or morpho-syntactic, or also pragmatic/thematic violations and even prosodic violations as in Friederici, (2002). Interpreting the role of each ERP component is difficult as experimental designs eliciting those ERP share some confounds. Here is a short overview of the most seen interpretations for the word-level ERP components (for a review see Friederici and Weissenborn, (2007) and Kuperberg, (2007)):

- early left anterior negativity (ELAN): This component is associated to local phrase structure binding, the word category starts to be integrated with the current syntactic structure.

- left anterior negativity (LAN): The amplitude of the LAN is modulated by manipulation syntactic and thematic relations

- N400: The most studied ERP component for language. It has a centro-parietal negativity peaking around 300-500ms. This is often linked to semantic violation paradigms. It is hypothesised to be a signature of the process of integrating the current word to the semantic context (an example sentence is "He took coffee with sugar and dog", see Kutas and Hillyard, (1980)). Previously it was attributed to lexical retrieval difficulty, but findings from (Kutas and Federmeier, 2011) reject this view in favour of an anticipatory mechanism. It can be noted that the N400 has been widely used as a marker of semantic processing in clinical population (Kielar, Meltzer-Asscher, and Thompson, 2012).

- P600: The P600 has a similar topography as the N400 but with inverted polarity, and with the peak of amplitude at a latency of 500-800ms. It has been reported principally in response to syntactic or morpho-syntactic violations (Friederici and Kotz, 2003), or even for animacy violation (Kuperberg et al., 2003) (as in *For breakfast, the eggs would only eat toast*), but it is also found in fully grammatical but ambiguous sentences indicating that it might reflect processing difficulty in integrating complex sentences (Osterhout, McLaughlin, and Bersick, 1997).

Several functions have been attributed to these ERP components, and sometimes with confronting interpretations. One major issue concerns syntactic-effects.

The presence or absence and the latency of the ELAN/LAN, as well as their spatial distribution (bilateral or not) is still being debated (Friederici and Weissenborn, 2007). Researchers aimed at differentiating different ERP components for each stage of preocessing, although it is clear that these linguistic functions are not independent. Hagoort, (2003) showed that the amplitude of the N400 was boosted when an additional syntactic violation was completing an already present semantic violation. However, he reported that the size of the P600 was not influenced by additional semantic violation. This points towards an asymmetrical interplay between syntax and semantic processing. Even further, Fedorenko et al., (2016) argues that lexical, semantic and syntactic processing always occurs congruently and in an unified manner. Altogether, these limitations highlight the difficulty of dissociating language functions. Moreover, as we will see in the next section, working with artificial violations in single sentences is probably hindering some of the mechanism at play during natural language understanding.

I.3.2 On the use of naturalistic stimuli for speech studies

An inherent problem to trial-based experiments when studying speech is that the stimulus might lack of what characterizes language in the first place. Speech is used to communicate, or receive information. Listening to repetition of sentence or speech segments, with no narrative, appears very artificial with respect to how speech is naturally used. ERP studies suffer from this limitation.

Language comprehension happens in our daily life with naturalistic speech, where the brain is subject to multi-modal information to begin with and where speech also present long term statistics. Structures in language, entity reference, or semantic information, can span several sentences and so do cortical representations of speech. To an extreme, it has been shown as well that discourse level information modulates brain responses to similar stimulus (Kandylaki et al., 2016). Non-linguistic information such as the perceived age of the speaker combines with semantic representation of speech (such as to elicit an N400 to the word *wine* in "I like drinking wine" if the sentence is said by a child speaker, Hasson et al., (2018)). Those are forms of context that can modulate brain response to perceived speech. We see that context is broadly defined, across modalities, social cues, and not only textual or semantic. That the brain operates naturally within those contexts is clear.

One other benefit of using naturalistic experiment designs resides in that it produces results more readily generalisable to everyday language use. Various degree of ecological credibility can be achieved by choosing different type of stories (Kandylaki and Bornkessel-Schlesewsky, 2019). In Hamilton and Huth, (2018), the authors argue in favour of using more naturalistic stimuli for linguistic studies. The underline the fact that in the past it was actually more difficult or unfeasible to quantify stimulus statistics if it was not tightly controlled. However as of today, those downsides of using natural stimuli are overcome by modern statistical and computational models.

This thesis is an opportunity to explore new approaches and methodology in order to foster the use of naturalistic and complex stimuli. We build up on recent studies in the field of auditory processing and speech perception where continuous complex stimuli were used to gain knowledge on how the brain process those. Through estimating spectro-temporal receptive fields from continuous speech presentation (Mesgarani and Chang, 2012), or similarly, extracting canonical response to envelope (Ding and Simon, 2014; Crosse and Lalor, 2014) or semantic dissimilarity (Broderick et al., 2018), researchers have managed to take advantage of the richness of the stimulus to answer meaningful questions about underlying neural mechanisms. Besides, advances in statistical learning such as deep learning and NLP allows to build statistic representing linguistic information as it unfolds through a narrative. By merging the recent analysis methods from neuroscientists studying continuous speech and modern language models we hope to offer a new way standard for the study of naturalistic language perception.

I.3.3 Cortical tracking of speech features

The shift towards more naturalistic stimulus puts new constraints on the type of analysis and methods used on the EEG data. Indeed, with continuous speech and no repetition at all, it becomes obsolete to think about extracting ERPs. Instead, researchers must employ novel techniques and make use of the entire recording in an efficient way. A growing amount of studies has been using continuous recording in the last decade, with most of those studies stemming from auditory neuroscience and especially focused on speech processing. Researchers, such as Lakatos et al.; Ding and Simon; Peelle and Davis; Poeppel, paved the way with early results on cortical tracking, or *envelope entrainment*.

The term entrainment is debatable, as it implies an intrinsic oscillatory source that will adapt its frequency towards the one from the stimulus if in range. This precise mechanism has still to be proven with further electrophysiological studies. Nevertheless a phase locked neural activity to the low-frequency content of the stimulus is clear. Recent studies (Kadir et al., 2019; Zoefel and VanRullen, 2015b; Etard and Reichenbach, 2019) explicitly show how this observed alignment is not solely a passive by-product of the periodicity from the input stimulus. Indeed, the low-frequency tracking is thought to be an anticipatory mechanism, or pro-active, in the sense that cortical oscillations seek to align their "best" phase (for example the phase at which connected neuronal population can synchronously decode the parsed input) to the ongoing phase of sound envelope. This is illustrated in the diagram in figure I.5 reproduced from Giraud and Poeppel, (2012).

Speech itself is a dynamic input, therefore it is stipulated that cortical rhythms may be involved in the *parsing* of auditory input (Ghitza, 2011; Hyafil et al., 2015). Sound chunks that fall within the time scale of syllables, so roughly 150–300ms which corresponds to the theta frequency range, are continuously being parsed in order to map them to phonemic representations and so forth. Features that unfolds

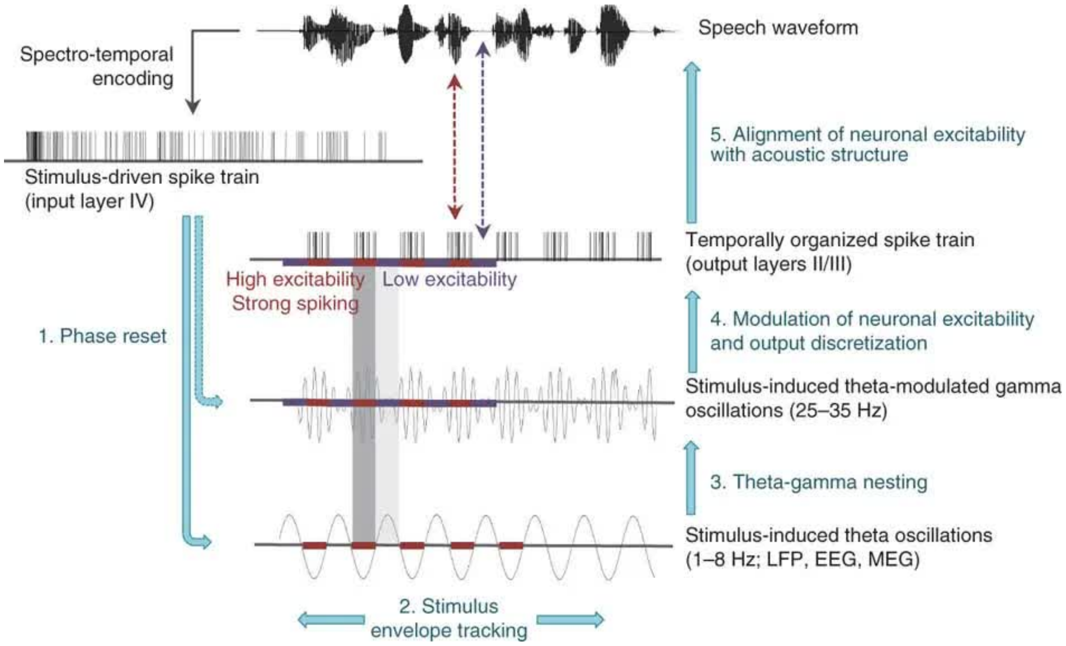

Speech waveform
Spectro-temporal encoding
Stimulus-driven spike train (input layer IV)
5. Alignment of neuronal excitability with acoustic structure
Temporally organized spike train (output layers II/III)
High excitability
Strong spiking
Low excitability
1. Phase reset
4. Modulation of neuronal excitability and output discretization
Stimulus-induced theta-modulated gamma oscillations (25–35 Hz)
3. Theta-gamma nesting
Stimulus-induced theta oscillations (1–8 Hz; LFP, EEG, MEG)
2. Stimulus envelope tracking

Figure I.5: Oscillation for Speech processing Giraud and Poeppel, 2012. Slow cortical oscillations in the theta range align their phase to the incoming speech envelope in order to match the period of higher excitability of neuronal population that encode sound chunks into phonological representations. Reproduced with permission ©(2012), Springer Nature.

at other timescales in speech like words, phrases or pragmatic discourse rhythms, can also be tracked by neural activity. The idea that there are endogenous or intrinsic oscillations that reset their phase according to the stimulus phase or instead time-locked potentials responses emerging from synchronous activity is still debatable. However it is now clear that we observe cortical activity with strong coherent oscillatory power in relationship with some speech features. Most of findings in the current literature concern processing of speech envelope, or more generally of spectro-temporal features of sounds.

Neural tracking of the envelope in the auditory cortices is probably involved in speech sounds too. It is thought to reflect the chunking of input and computations of summary statistics regarding regularities in the stimulus. A popular view, is that ongoing phase of cortical rhythm is reset by salient aspect in stimulus so that we align an optimal phase of the ongoing oscillations with the stimulus in order to process the speech segment efficiently, as detailed in figure I.5. It was argued that changes in neural tracking reflect stimulus intelligibility (Poeppel, 2014), however it has been reported that neural tracking to envelope can be as strong for reversed speech. Recent work pointed towards the contribution of high level features to the cortical tracking of envelope (Zoefel and VanRullen, 2015a; Etard and Reichenbach, 2019). Those features only exist for understandable speech. The mechanistic interpretation of how neural coding at the microscopic scale generate this tracking is still discussed. Obleser and Kayser, (2019) gives a review on the possible mechanisms such as neuronal phase coding of the stimulus to explain those results, highlighting the importance of understanding what *entrainment* is in the narrow sense.

The processing of higher-level linguistic information in speech may employ cortical tracking as well. Recent findings showed that cortical activity in the delta band and the power time course in beta band synchronized to the rhythm of phrases and sentences in continuous speech (Ding et al., 2016; Keitel, Gross, and Kayser, 2018). Another study by Broderick et al. demonstrates that semantic dissimilarity between individual words is significantly modulating amplitude of low-frequency cortical oscillations (Broderick et al., 2018). Finally, tracking of phonemic features seems to bring a strong and reliable way of encoding neural dynamics as shown by Brodbeck, Presacco, and Simon, (2018) and Di Liberto et al., (2019).

All these studies employed similar methods to analyse the relationship between speech features and electrophysiological signals. A review of this type of analysis can be found in Sassenhagen, 2019. Mainly, one aims at using time-resolved multivariate regression analysis to infer brain responses to speech features under naturalistic listening conditions. Depending on the end goal of the task and analysis, the model predicts the brain signals from the stimulus features (Mesgarani and Chang, 2012; Di Liberto et al., 2015, known as *forward models* as in), or conversely, decode the input from brain data (Etard et al., 2019; Cheveigné et al., 2018b).

I.4 Predictive Processing as a key mechanism for Language Comprehension

> “ *Uste-gabea* ”
>
> from Basque:
> literally *Without belie*f. Translate as
> *Unpredictably, suddenly,*

In sections I.2.2 and I.3.3 we already mentioned the concept of predictive processing, notably with temporal predictions of stimulus spectro-temporal structure with model of entrainments as in figure I.5 where phase alignment allow neuronal population to anticipate incoming salient part of the stimulus. The notion of predictions in neuroscience is a widespread idea and not only restricted to speech or auditory processing. Actually, the brain is often view as a predictive organ, an inference machine, capable of actively predicting its sensory inputs through a generative model. The idea that the brain generates its own perception and that it is not merely applying a series of transformations to its sensory input is not new. It refers to the original concept advanced by Hermann Von Helmoltz about unconscious inference in Helmholtz, (1866). The theory has been largely refined since and predictive coding principles are now used to explain perception and action, hence behaviour. Some researchers describe it as a fundamental and unifying principle, and argue that it can explain the theory of mind or the sense of agency (Clark, 2013; Friston, 2012).

I.4.1 Predictive Coding: a short review

Rao and Ballard were pioneers of the modern account of predictive coding. They developed a computational model of visual processing in the primary visual cortex entirely based on predictive coding by proposing that feedback projections were modulating receptive fields of neurons by predicting bottom-up activity. Their result suggested that visual processing is not an exclusively feedforward mechanism but that an efficient hierarchical strategy based on feedback connections could explain some data observed in physiological studies. At that time, the idea of efficient coding such as sparse coding and redundancy reduction was strongly present in the mind of neuroscientists working on neuronal mode of visual processing. It appears as obvious that the brain is strongly constraint by energy consumption. It uses approximately 20% of the body's energy while accounting for only 2% of its mass. Therefore the encoding schemes must capture maximal information about the stimulus in the sparsest representation. Physiological results were also bringing evidence to this energy-constraint as we observe irregular and sparse firing patterns, and a globally balanced excitatory to inhibitory activity (Yang, Zhou, and Zhou, 2017). Results from Rao and Ballard, (1999) were emerging from another perspective but still accounted for efficiency and sparsity. Recently, Chalk, Marre, and Tkačik, (2018) showed that predictive coding can be unified to efficient coding by using an information theoretic approach.

We can look at this framework, or theory, through each of the different levels from Marr's systemic representation (Marr, Poggio, and Brenner, 1979). On the computational level, we lay out the core principles of the theory. The information coming at a level of hierarchical representation is constantly predicted by higher levels, such that only a *prediction error* is passed through the levels above. The algorithmic level which describes in mathematical terms possible way to implement such principles, for instance using non-linear dynamic systems (Friston, 2005), and finally an implementational level which gives the details of biological implementation of afore mentioned descriptions. Only the first computational level is described here. The framework models how neuron populations encode the environment in an agent's brain. Perception of the environment, along with the sense of our own body and motion, the decisions for actions, all emerge from inferences made by the brain about causes of sensory inputs (Friston, 2010). This results in a generative model of the world, that is continuously confronted to actual sensory bottom-up information. The neural code for perception could thus be reduced to the difference between predictions made by the generative model and the upcoming inputs, referred to as *prediction errors*.

Perception is therefore an emerging phenomenon resulting from the interplay between predictions and sensory inputs, or more generally, between top-down predictions and bottom-up information, encoded as prediction errors. This computational element can be generalized at different levels of representation in the brain as a hierarchy of increasing complexity and abstraction. Such that lower level (closer to sensory input in the processing hierarchy) pass on the message, encoded in prediction errors to higher levels which themselves pass on predictions or beliefs to the level below. Each level tries to suppress the error generated on the incoming input based on the top-down predictions.

Friston and Kiebel, 2009 described a biologically plausible implementation of predictive coding with physiological grounding in the form of hierarchical dynamical system between activity of neurons in different cortical layers. They test their model in the context of bird songs generation and recognition with remarkable success (Yildiz and Kiebel, 2011, see also). Their model is based on an internal generative model of how songs are produced by a hierarchical system. This same model has then been ported to speech production and recognition in Yildiz, Kriegstein, and Kiebel, (2013).

Other mechanistic descriptions of the neural circuitry involved in predictive coding have been studied (Bastos et al., 2012, for the "cortical microcircuit of predictive coding"). One important finding relates to the asymmetry of frequency content between population of neurons sending predictions compared to those forwarding prediction errors. Indeed, Bastos et al. show that as the deep pyramidal cells of cortical columns is processing and propagating *predictions* in his model, they must

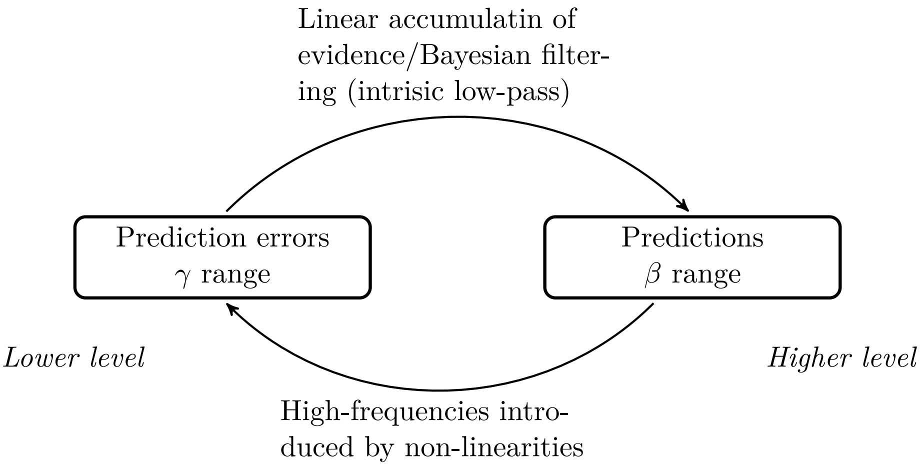

Linear accumulatin of evidence/Bayesian filtering (intrisic low-pass)
Prediction errors
γ range
Predictions
β range
Lower level
Higher level
High-frequencies introduced by non-linearities

Figure I.6: This diagram illustrates the asymmetry in oscillatory activity for predictive coding as described in Bastos et al., (2012) and Giraud and Poeppel, (2012). Note that the left node comes from the level below in the cortical hierarchy (hence they represent different neuronal populations, and/or different cortical columns).

accumulate prediction errors in order to generate their conditional expectations. Hence they produce a smooth estimate of hidden causes and suppress high-frequency fluctuations from their input. This is inherent in Bayesian filtering. On the other hand, the bottom up signal, generated in superficial layers, propagates prediction errors at higher frequencies (gamma range), see figure I.6 for an illustration. The authors predicted that deep layers should express more beta relative to gamma and conversely in the superficial layers. This was observed experimentally with extracellular recording in primates (Fries, 2015) and in human for auditory tone sequences (Sedley et al., 2016; Giraud and Arnal, 2018).

Building on those results, predictive coding for auditory processing has been linked to cortical oscillations (see Giraud and Arnal, 2018, for a review), and the asymmetric oscillatory activity mapped up to different signals in the hierarchy of predictions and prediction errors, as seen in figure I.7.

I.4.2 Predictions in Language Processing

In the context of language processing, predictive coding gained a lot of attention. It could explain many previous experimental results and provides with a framework to justify the rapidity of processing of spoken language and the robustness of comprehension in noisy environments. Although it is not a new concept for the neurobiology of language, but it was often attributed to different mechanisms than predictive coding per se. Originally, behavioural studies have indeed shown that the brain makes predictions about upcoming speech segments: words can be better distinguished from noise when transition probabilities between words are high rather than low (Miller et al., 1951), and a highly-predictable word can

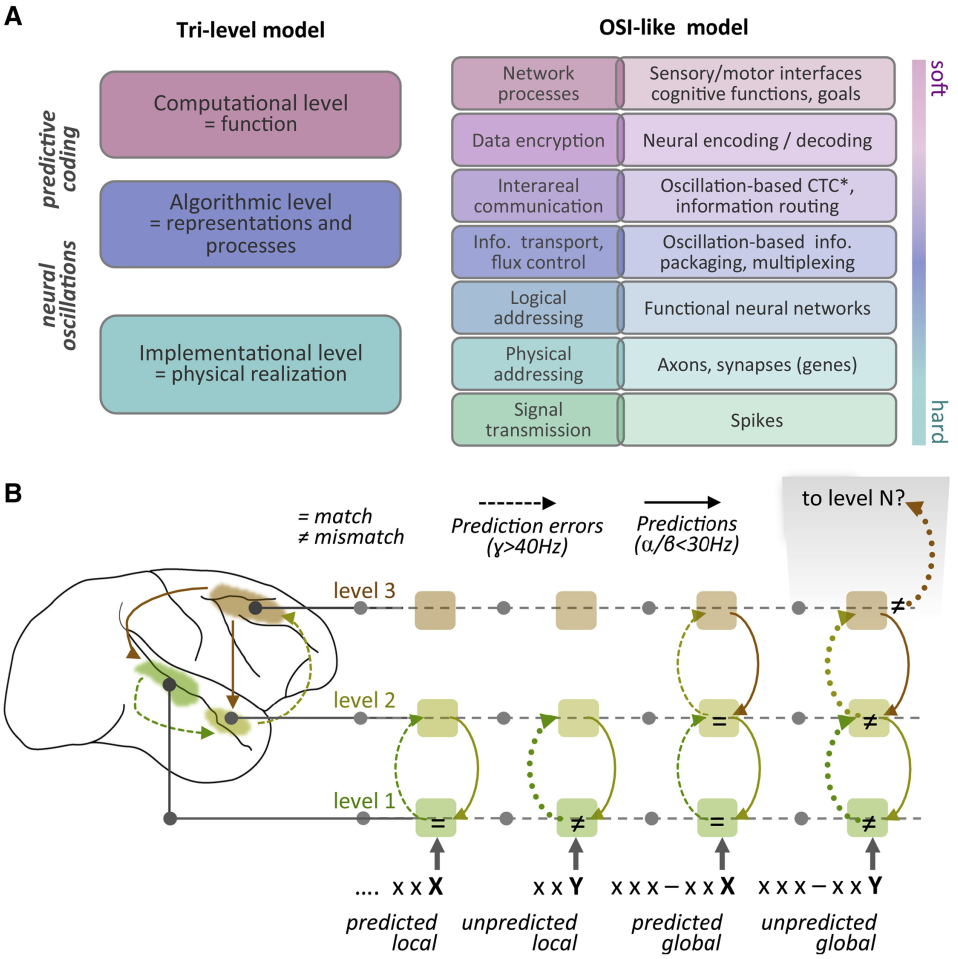


Figure I.7: Predictive coding and different level of computation. The spectral asymmetry between predictions, in the beta band, and predictions errors, in the gamma band, is also depicted. Reproduced from Giraud and Arnal, (2018), with permission, ©2018 Elsevier Inc.

be perceived as heard even when obscured by noise (Miller and Isard, 1963). Many psycholinguists have therefore proposed that anticipatory mechanisms are at play during language processing. We can read in Federmeier and Kutas, (1999) that "in the course of processing a sentence, the comprehension system is involved in some process tantamount to prediction". Predictions play a key role in language processing, but it is still unclear whether predictive coding is the mechanism implementing those. The hierarchy of representation during speech processing follows quite naturally from the encoding of sound and its characteristic acoustic features such as pitch to language-specific representations such as phonemes, lexical information, syntactic constituents and phrases. Under the framework of predictive processing, we thus expect that higher-level such as the representation of semantic content will predict the upcoming sequence of phonemes which in turn will predict spectral content from ongoing acoustic input (Levy, 2008; Lewis and Bastiaansen, 2015; Willems et al., 2016). Gagnepain, Henson, and Davis, 2012 showed that partially matching words (such as "formu…" predicting "formula") are used to predict upcoming speech segments more precisely. In their study, they show that their data are better explained by such a predictive mechanism rather than by some form of lexical competition.

In the literature there are also examples of computational models applying predictive coding principles to showcase their utility for perceptual systems as mentioned in the previous section but also to present the usability of those principles for speech processing. Using a version from the cortical microcircuit for predictive coding proposed by Friston and Kiebel, 2009, Yildiz, Kriegstein, and Kiebel, 2013 extended the model from Yildiz and Kiebel, 2011 to apply predictive coding directly to speech recognition. Keeping to two level in the hierarchy of hidden states after cochlear input, they managed to build a system of speech recognition that after being trained on a single speaker could still perform well in a competing speaker environment. This is an example of predictive coding for the speech recognition although they applied to the lowest level, where the hidden causes predicted by their system were at most equivalent to phonemic encoding.

The asymmetry described in the previous section, and found in the auditory domain by Sedley et al., 2016 are also being hypothesised for language processing. In a review, Lewis and Bastiaansen, 2015 present how results from language studies converge towards a predictive coding account of sentence-level language comprehension that involves different cortical oscillations too. Lewis and Bastiaansen propose to reframe the classical findings on P600 into a predictive coding framework in which this ERP component would be a marker of syntactic prediction update or prediction error. Moreover, following Bastos et al.'s theory, they advocate that beta oscillations carry predictions to lower-level of linguistic representations, while gamma oscillations would convey prediction errors to higher levels. Those hypothesised involvement of oscillatory activity for predictive processing during language comprehension found empircal data bringing evidence and conclusive results for such a theory. Beta power has been found to be reduced by semantic and syntactic violations (Bastiaansen, Magyari, and Hagoort, 2010; Kielar et al., 2014) and gamma power has been observed to increase when a word is highly predictable but not when its predictability is low (Molinaro, Barraza, and Carreiras, 2013; Wang, Hagoort, and Jensen, 2017). Besides the contribution of oscillatory activity to predictive mechanisms for language processing, others have focused on quantifying prediction in order to relate those to brain activity. This has been done mainly with fMRI, hence they could not asses which frequency band where supporting the measured processes. Willems et al., (2016) recorded the fMRI of subjects while they where exposed to ongoing speech. He found areas that were significantly responding to the *entropy* and to the *surprisal* of a word as defined by a language model that predicts the next word based on the previous context. With a similar experimental design, Frank and Willems, (2017) observed that the predictability of a word activated distinct brain areas than the semantic dissimilarity, and that it correlated with the amplitude.

We aim at unifying the views on predictive coding for language processing, by analysing electrophysiological data during naturalistic story comprehension. The nature of EEG provides us an access to high frequency brain activity in response to the stimulus and instead of using artificial sentences, with semantic or syntactic violation, we will quantify key metrics for predictive coding, like surprisal and entropy, in order to represent the ongoing speech stimulus and to model the stimulus-response relationship.

I.5 Recap on Linear Models and Regression Analysis

One of the strongest and boldest assumption we will make to study the link between recorded EEG signals and the stimulus is that there exists a linear relationship between those. In other words, we are looking for a linear map between a vectorial representation of the stimulus and the EEG time series, or vice-versa. Although this assumption might appear quite restrictive at first sight, it is in its simplicity that we get the most of its value. Through such linear approaches, we are effectively trying to determine a linear, and also time-invariant system that takes the stimulus as an input signal and outputs the multi-channel EEG activity. That is, if we can find such a system that significantly explains the data observed, we would have explained and modelled at best that much of the relationship between stimulus and response. It can be seen as a first order relation, where all the non explained noise, or residual, contains task-irrelevant signals, speech-independent responses or high order activity. The most basic form of equation for such a system, with a one channel signal output and one dimensional input is:

$$y(n) = \beta_0 + \sum_{k=-l_1}^{l_2} \beta_k x(n-k) + \epsilon \tag{I.1}$$

Where $y(n)$ is the n^{th} sample of the EEG single channel data $\mathbf{y} \in \mathbb{R}^{T\times 1}$, with T the total number of samples (hence the duration of the data in sample unit). The stimulus is represented in the signal x. The intercept is estimated through β_0 and the coefficient β_k are multiplicative factor of the input signal x at different lags. This operation is a discrete convolution between the input x and the kernel formed by the coefficients β. We can extend this model to multi-dimensional features $\mathbf{x} \in \mathbb{R}^{T \times N_{feat}}$, this is often referred to as multivariate linear regression:

$$y(n) = \beta_0 + \sum_{i=1}^{N_{feat}} \sum_{k=-l_1}^{l_2} \beta_k^i x^i(n-k) + \epsilon \tag{I.2}$$

The dummy indices i runs across feature dimensions, so there is a linear coefficient for each feature i and lag k: β_k^i. Note that equation (I.2) can be completely reversed, such that we reconstruct a one-dimensional feature x from the multi-channel EEG signals. This would thus be referred to as a *backward model* (see below). Finally we can simplify further equation (I.2) by writing it in its matrix form. First of all, let us consider the multi-channel, or multi-dimensional matrix of input $\mathbf{x} \in \mathbb{R}^{T \times N_{feat}}$, where each row represents a sample of the multi-dimensional input $\mathbf{x(n)} = \begin{pmatrix} x^1(n) & \cdots & x^{N_{feat}}(n) \end{pmatrix}$. Then for each lag, we define a vector of coefficients $\beta_\mathbf{k} \in \mathbb{R}^{N_{feat} \times 1}$ such that we can get the summation over feature dimension through the dot product:

$$\mathbf{x(n-k)} \cdot \beta_\mathbf{k} = \sum_{i=1}^{N_{feat}} \beta_k^i x^i(n-k)$$

We can now define the lag-matrix $\mathbf{X} \in \mathbb{R}^{\tilde{T} \times N_{lags} \cdot N_{feat}}$ of the input data. This will be a Tœplitz matrix formed of shifted version of $\mathbf{x}$:

$$\begin{aligned}
\mathbf{X} &= \begin{pmatrix} & & \vdots & & & & \vdots \\ x^1(n+l_1) & \cdots & x^{N_{feat}}(n+l_1) & \cdots & x^1(n-l_2) & \cdots & x^{N_{feat}}(n-l_2) \\ & & \vdots & & & & \iddots \end{pmatrix} \\
&= \begin{pmatrix} \vdots & \vdots & \vdots \\ \mathbf{x(n+l_1)} & \mathbf{x(n)} & \mathbf{x(n-l_2)} \\ \vdots & \vdots & \vdots \end{pmatrix}
\end{aligned}$$

And by doing so, we can vectorize the summation over lag indices to end up with:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X} \cdot \beta \\ \hat{\mathbf{Y}} &= \mathbf{X} \cdot \underline{\beta}\end{aligned} \tag{I.3}$$

Note the ˆ denoting an *estimated* value (therefore rendering this equation exact so we can drop the noise term ϵ). Hence in equation (I.3), $\hat{\mathbf{y}}$ refers to the estimated EEG time series of one channel and $\hat{\mathbf{Y}}$ to the estimated matrix of EEG data, the concatenation of all channel time series.

Now we can isolate β in equation (I.3), such that we obtain from the pseudo inverse of $\mathbf{X}$ the estimate $\hat{\beta}$:

$$\underline{\hat{\beta}} = \mathbf{X}^{\dagger} \cdot \mathbf{Y} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathsf{T}}\mathbf{Y} \tag{I.4}$$

$\mathbf{X}^{\dagger}$ is the Moose-Penrose pseudo-inverse of the matrix $\mathbf{X}$.

The covariance matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ can be non invertible if it is rank deficient. In that case, one on more of its eigenvalues will be zero. Numerically, this does not usually happen as the machine will reach numerical precision for near zero eigenvalues. Practically, the inversion of the matrix is stable if it is well *conditioned*, that is, if the range between the highest and the smallest eigenvalues is within the dynamical range of the computing machine. To assure numerical stability of the inversion in equation (I.4), one can shift all eigenvalues away from zero by adding a constant λ to all diagonal element of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ so that equation (I.4) becomes $\underline{\hat{\beta}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1} \mathbf{X}^{\mathsf{T}}\mathbf{Y}$. This is referred to as Tikhonov, or Ridge regularization.

Family of Linear models

Several linear modelling approaches are available to establish the relationship between stimulus and response, but they all fall into three major classes:

- Forward models: The model predicts the response from the stimulus. Namely the input of a forward model is the stimulus representation and its output is the neural response. It allows to look into how the brain *encodes* features from stimulus space (as measured through EEG signals as observables).
- Backward models: For this class of methods, one *decodes* stimulus features from brain responses. Therefore, the weight learn are not directly interpretable as brain sources, but rather as spatio-temporal filters over brain responses that map onto stimulus space.
- Hybrid models: Those models are a form of encoding/decoding, where both stimulus and responses are projected into a subspace where they are, for instance, maximally correlated. This can be useful to quantify the strength of

the relationship between a feature set and the EEG response. See methods in section IV.3.1.

Each family of methods comes with mutually exclusive advantages that strongly depend on the end goal of the analysis. Forward models are more directly interpretable than backward ones. Indeed, the weights β_k in equation (I.1) act as a scaling factors of the input features x and therefore are in the unit of y when the input is made adimensional (by z-scoring or whitening for instance). They represent the convolution kernel of the model and characterize *fully* the linear time-invariant system. By looking at the coefficient series $\beta_{1:k}$ as the discretization of the signal $\beta(\tau)$, we are considering the canonical response of the system to an impulse (dirac) from the stimulus space. Observing that canonical response gives us insights on the underlying brain mechanisms at play (as of where and when in the brain the stimulus feature is being processed). However weight from backward models are not directly interpretable as such but provides a direct projection into stimulus space, optimal in the least-square sense, which can be thus used as brain computer interface in decoding tasks. Backward models are principally used in the literature for decoding selective attention in competing speaker experiments. Finally, hybrid methods are better to inform on the quality of cortical tracking and to assess the degree of linear relationship existing between two datasets, in a symmetrical fashion (Cheveigné et al., 2018b; Di Liberto et al., 2019).

In all three following chapters, we will be using *forward* models to capture the time-course of the linear relationship between EEG and linguistic predictors derived from the stimulus. This is actually the only model that is readily interpretable as mentioned above. The linear coefficients computed during fitting such model are alike brain potentials. This will allow a direct transcription into meaningful brain response to each linguistic feature. Moreover, only the *forward* and *backward* model take in consideration the covariance between different speech (linguistic) features. It can be seen in equation (I.4), where the computation of β coefficients imply to know X^TX, the covariance matrix of the input. Our goal will be to measure brain response to a set of features describing linguistic information at once, and it is therefore important that we take the full covariance of the stimulus space in account. These two points make up the main reasons for choosing forward modelling approach. In the last chapter though, in one analysis we will be using a *hybrid* model to better characterize the strength of the linear relationship between stimulus and response, but again this does not allow to interpret time-course of coefficients as forward models do.

I.6 Goals and organisation of the thesis

Having in mind the different models for language processing and available methodology, we aim at bringing new tools for analysing EEG data in the context of naturalistic speech comprehension. We designed an experiment allowing the participant to be exposed to language with a full narrative context, and developed analysis methods to extract information from continuous EEG signals. We wanted to explore how statistical features derived from a generative language model are encoded in EEG signals. To extend to other theories of language processing, we looked also at syntactic features, derived from rule-based algorithms. Both set of features could give significant result, and we finally tested how each would benefit a decoding system to predict comprehension state from EEG. Hence we are asking the following questions:

How are different word level linguistic features contributing to comprehension of language? Can we decode whether speech is being understood from EEG?

Chapter II focuses on statistical models for language, taking inspiration from predictive coding models established for auditory perception. Mainly, inspired by the results obtained by Sedley et al., 2016 in the auditory domain using *surprisal* and *precision* but applied to the linguistic domain. Moreover, we wanted to investigate the relative contribution of different frequency bands to the response to those features following hypothesis from Arnal and Giraud, 2012; Lewis and Bastiaansen, 2015

In chapter III, we thought of exploiting recent results from Ding et al., 2016 to study how the brain respond to syntactic structures during naturalistic language comprehension. Moreover, this allowed us to define linguistic features that could relate to the *merge/unification* operation described by Bastiaansen and Hagoort, 2006; Chomsky, 2013. It has been shown with non natural speech, that sentence building and syntactic unification presented neural correlate in theta and beta band activity. We hope to reproduce this result in a naturalistic experiment.

Finally the last chapter will try to characterize in more depth the dynamics of those different stage of processing and to establish their relative roles for language comprehension. We will reflect on the use of envelope acoustic tracking such as described in section I.3.3 to understand how linguistic features can modulate low-level acoustic tracking and conclude with a decoding analysis where the task is to effectively decode comprehension from the EEG signals.

Chapter II

Cortical tracking of surprisal and precision entropy during continuous speech comprehension

BEsides the intrinsic complexity of speech, there is some meaningful redundancy and structure within its linguistic constituents. We demonstrate in this chapter how we can record a response to data-driven statistical features extracted from speech.

The brain continuously decodes the message being transmitted by speech to extract meaning from it. It can do so in a variety of environments, where the speech signal can sometimes be mixed with many other acoustic sources, other speech sounds, or environmental noise. More variability also comes from the variation of acoustic properties stemming from physical differences between speakers (for instance, different pitch corresponding to differences in vocal tract shapes), as well as different accent and so on. Though our brain can robustly encode information despite this variability and still extract the original meaning of the message remarkably fast. One putative mechanism to be able to process speech in unpredictable, noisy environment and in real-time, is for the neural population responsible in decoding language to predictively process speech signals.

This chapter is a study on the predictive processes occurring in the brain during speech comprehension. The work hereafter has been published in Weissbart, Kandylaki, and Reichenbach, (2019a), what follows is the manuscript of the actual article modified slightly for formatting and adjusting content.

II.1 Introduction

To understand spoken language, a listener must rapidly process information that unfolds over several timescales, including the duration of syllables at around 150 ms, words of about 300 ms, and phrases of 1 s (Giraud and Poeppel, 2012). Recent studies have shown that cortical activity in the delta, theta and gamma frequency bands tracks acoustic features of speech such as the speech envelope as well as phonemic features (Lakatos et al., 2007; Zion Golumbic et al., 2013; Ding and Simon, 2014; Di Liberto et al., 2015; Ding et al., 2018). This cortical tracking of speech features has accordingly been proposed to reflect neural mechanisms of speech processing, for instance an online segmentation of speech into acoustic speech tokens, such as phonemes that occur on the time scale of a few hundreds of milliseconds (Giraud and Poeppel, 2012; Hyafil et al., 2015).

The processing of higher-level linguistic information in speech may employ cortical tracking as well. Recent findings showed that cortical activity in the delta and beta frequency bands synchronized to sequential cues such as the rhythm of phrases and sentences in continuous speech (Ding et al., 2016; Keitel, Gross, and Kayser, 2018), to hierarchical cues such as context-free grammar structure (Brennan and Hale, 2019), as well as to the semantic dissimilarity between successive words (Broderick et al., 2018).

An important property of word sequences is that they can allow the prediction of an upcoming word, resulting in a word expectation. The degree to which a word can be predicted is referred to as *precision* and reflects the certainty with which a neural population generates its prediction. Predictions and precision are both closely related to putative implementations of predictive processing mechanisms (Friston, 2005; Feldman and Friston, 2010; Heilbron, 2018). Behavioral studies have indeed corroborated that the brain makes predictions about upcoming speech segments: words can be better distinguished from noise when transition probabilities between words are high rather than low (Miller et al., 1951), and a highly-expected word can be perceived as heard even when obscured by noise (Miller and Isard, 1963).

Neurophysiological research on event-related potentials elicited by a word in a sentence has shown that the brain response to a word reflects the word expectancy through modulation of the N400 response (Kutas and Hillyard, 1980; Kutas and Federmeier, 2011, for a review). Although this response has not been found to be further modulated by the precision of the prediction (Federmeier et al., 2007), precision can influence the neural power in the alpha and theta band (Rommers et al., 2017; Sedley et al., 2016) The power in the beta frequency band has been found to be reduced by semantic and syntactic violations, and may therefore relate to word expectation as well (Bastiaansen, Magyari, and Hagoort, 2010; Lewis and Bastiaansen, 2015; Kielar et al., 2014). Gamma power has been observed to increase when a word is highly predictable but not when its predictability is low (Molinaro, Barraza, and Carreiras, 2013; Wang, Hagoort, and Jensen, 2017). However, these prior studies on neural correlates of word expectancy and precision have focused on specific words in single sentences, contrasting words with high and low expectancy as well as with high and low precision. But natural speech often consists of many sentences, and the expectancy and the corresponding precision of successive words take a range of values that do not fall in only two classes of 'high' and 'low'. It therefore remains unclear how neural responses to word expectancy and precision correlate with this graded variability. An analysis taking in account the time course of such variability as it occurs in natural speech could potentially better characterize the neural mechanisms underlying speech processing.

Furthermore, assessing the cortical responses to these linguistic features of successive words using naturalistic stories as stimuli, instead of only focusing on a particular word in single sentences presentation, allows to quantify the cortical tracking of these features. Recent investigation on word predictability and hierarchical structure in naturalistic speech used such an approach to show cortical tracking of word surprisal (Brennan and Hale, 2019; Frank and Willems, 2017), but did not investigate the influence of precision entropy nor the power modulation in higher frequency bands. Here we therefore set out to investigate cortical tracking, including through power modulation in higher frequency bands, of word surprisal and the precision of word prediction in naturalistic stories. The surprisal of a word denotes the log-transformed conditional probability of a word based on the preceding context, has been argued to relate to processing load (Levy, 2008) and predicts reading time (Smith and Levy, 2013; Frank et al., 2015). Precision has been computed as the inverse of the entropy of the conditional probability distribution over a close vocabulary set. We quantified word surprisal and precision from naturalistic stories using language modelling as estimated by a recurrent deep neural network, and then related the obtained word features to electroencephalographic (EEG) responses of volunteers who listened to the stories.

II.2 Materials and Methods

II.2.1 Experimental design

Participants. 13 subjects (aged 25 ± 3 years, 6 females) participated in the experiment. The volunteers were all right-handed native English speakers. They had no history of hearing or neurological impairment. All participants provided written informed consent. The experimental procedures were approved by the Imperial College Research Ethics Committee.

We employed naturalistic speech narratives in the native language of our subjects (English). The experiment consisted of one session in which we measured electroencephalographic (EEG) responses to the short stories 'Gilray's flower pot' and 'My brother Henry' by J.M. Barrie as well as 'An undergraduate's aunt' by F. Anstey (Patten, 1910). The stimuli were sourced from the public domain `librivox.org` and were spoken by a male voice. The corresponding text was obtained from the project

Gutenberg[1]. The audio material was presented in 15 parts, each of which were 2.6 ± 0.43 min long. The total length of the stories was 40 min. After each part of a story, participants answered comprehension questions about what they just heard in the form of multiple choice questions, where they had to select their answer among four possible answers. Participants were asked 30 questions in total; the questions were presented and answered on a monitor.

II.2.2 Statistical Model of Language

We used computational linguistics methods to quantify linguistic features in the employed stories (Graves, 2013). Specifically, we employed statistical language modelling to compute word frequency, entropy and surprisal from the text of the stories. Word frequency is a property of each individual word out of any context, which was computed from Google N-grams by using only the *unigram* values. This word feature is an estimate of the unconditional probability of occurrence of a word, $p(w_i)$. We used the negative logarithm of this probability such that all our information theoretic word features are expressed in the same unit. Both entropy and surprisal follow from conditional probabilities of a particular word given the preceding words. We denote by $p(w_i|w_{i-1}\dots w_1)$ the conditional probability of the i$^{\text{th}}$ word in the sequence, w_i, given the previous $w_{i-1}\dots w_1$ words. Taking the negative logarithm of this probability yields the surprisal value for that word:

$$S(w_i) = -\log(p(w_i|w_{i-1}\dots w_1)) \tag{II.1}$$

The surprisal, also referred to as self-information or information content, quantifies the information gain that an upcoming word generates with respect to the sequence of words formed with its context. It can be related to how unexpected a word is given the previous words in the sentence. Inasmuch as surprisal informs about expected words, precision relates to the confidence about the predictions made (Koelsch, Vuust, and Friston, 2019). We implement this degree of certainty by taking the inverse of entropy. A high precision translates in a high confidence about a word expectation, meaning that the word is predictable. The entropy $E(i)$ at word i, that is, the uncertainty for predicting the word from the context $w_{i-1}\dots w_1$, is given by the sum of the conditional probabilities for each possible word w_i, weighted by the logarithm of this probability. In other words, the entropy is the expected surprisal:

$$E(i) = \mathop{\mathbb{E}}_{w\in V}[S] = -\sum_{w} p(w|w_{i-1}\dots w_1)\log(p(w|w_{i-1}\dots w_1)) \tag{II.2}$$

The *precision entropy* of the m$^{\text{th}}$ word follows as the inverse of entropy.

[1] http://www.gutenberg.org/ebooks/32846

The conditional probabilities for the different words in the sequence, given the preceding words, were computed through a recurrent neural network language model (Bengio et al., 2003; Graves, 2013). The network had a hidden layer with recurrent connections to encode previous input. Such networks are particularly useful for processing sequences and have previously been successfully applied to language modelling (Bengio et al., 2003; Graves, 2013). In particular, a recurrent neural network can capture long-term dependencies, of variable length, by encoding preceding words through its recurrent connection into the state of the hidden neurons. This is enabled by a careful balance between short- and long-term memory and means that there is in principle no limit on the number of preceding words that such a network can thereby take into account (Bengio et al., 2003; Pascanu, Mikolov, and Bengio, 2012). This contrasts with N-gram language models, for instance, that are limited to a context window of N-1 words (Brown et al., 1992).

The network was implemented using the feature-augmented recurrent neural network language modelling toolkit (Mikolov et al., 2010; Mikolov et al., 2011). To decrease the computational time required for training, this toolbox assigns words to classes and factorizes the output layer into a part that describes the probability of each class given the previous words, as well as another part that describes the probability of each word within a class given the previous words. This factorization yields a significant decrease in training time at a small cost to accuracy; importantly, the network still computes the probability of individual words following the previous words. We employed 300 classes. As an embedding layer we used the pre-trained global vectors for word representation trained on the Wikipedia 2014 and the Gigaword 5 datasets (Mikolov et al., 2013, for the original paper on word embeddings; and see Pennington, Socher, and Manning, 2014, for the embeddings used here). The recurrent layer encompassed 350 hidden units. The source code was customized to compute the entropy of each word, a feat that the original code did not allow. The neural network was then trained on the text8 dataset that consists of 100 MB of data from Wikipedia[2], using back propagation through time, truncated to five words with a starting learning rate of 0.1. The data was cleaned to remove punctuation, html tags, capitalisation and numbers before training. Since the network can only train well on words that appear frequently enough in the training data to allow meaningful training, we limited the vocabulary to the 35,000 most common words in the training dataset. The remaining words were mapped to an "unknown" token. Infrequent words in the stories, such as compound nouns used for style, that appeared repeatedly throughout the stories did therefore not obscure the results.

The output of the recurrent neural network was obtained from a softmax function, and could therefore be interpreted as the probability distribution for an upcoming word given the preceding words in the input sequence, as shown in the diagram of figure I.4 on page 15. The network was therefore trained to predict the next word, that is, to compute an output that was as close as possible to a probability distribution that was one for the actual upcoming word and zero for all remaining ones. The trained network was then run on the stories that the participants heard.

[2] retrieved from http://www.mattmahoney.net/dc/textdata.html

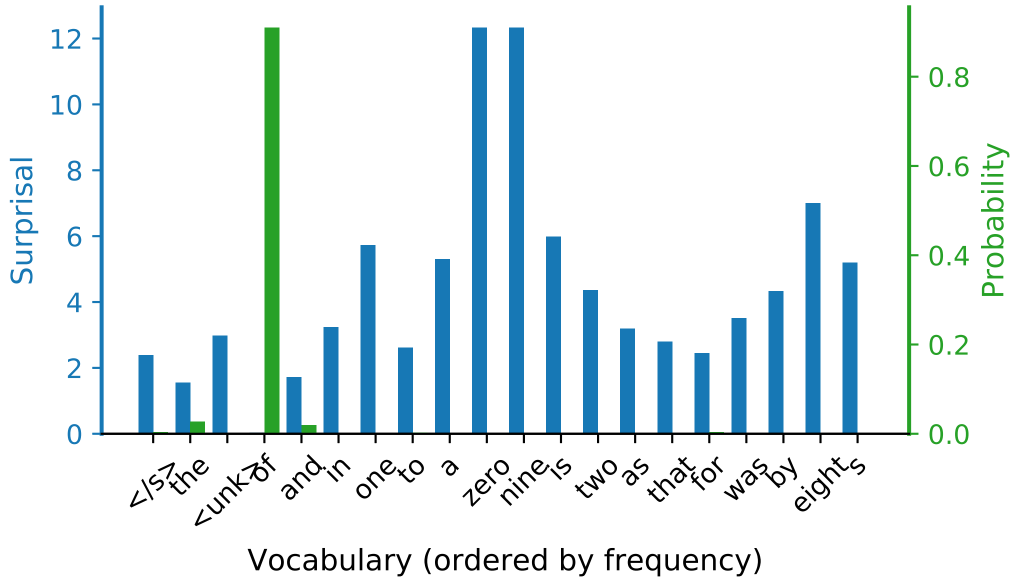


Figure II.1: Probability distributions (green) obtained after evaluating the network on a sequence of words. The corresponding surprisal values in base 10 are also shown in blue.

Entropy (and hence precision) and surprisal of each word were determined from the network's computed probability distribution at the corresponding word through equations (II.1) and (II.2). An example of the actual values obtained by a forward sweep through the neural network once trained is shown on figure II.1.

II.2.3 Speech Representation: Linguistic-modulated comb

To relate surprisal and entropy to the EEG data, we constructed a time series for each linguistic feature. We first aligned each word of the speech to the acoustic signal through forced alignment using the Prosodylab-Aligner software from Gorman, Howell, and Wagner, (2011). We thereby obtained the time at which each word began. To construct features for surprisal and for precision that were aligned with the speech stimuli, we assigned each of the time points where a new word started a spike of a magnitude that corresponded to the surprisal respectively precision of that word (figure II.2). A similar procedure has been employed recently for assessing neural responses to the semantic dissimilarity of consecutive words (Broderick et al., 2018).

Because surprisal and precision are high-level linguistic features of speech, we sought to ascertain that any putative cortical tracking of them could not be explained by lower-level features. To this end we added three low-level speech features. First, cortical activity can track the onset of words, which can partly be based on changes

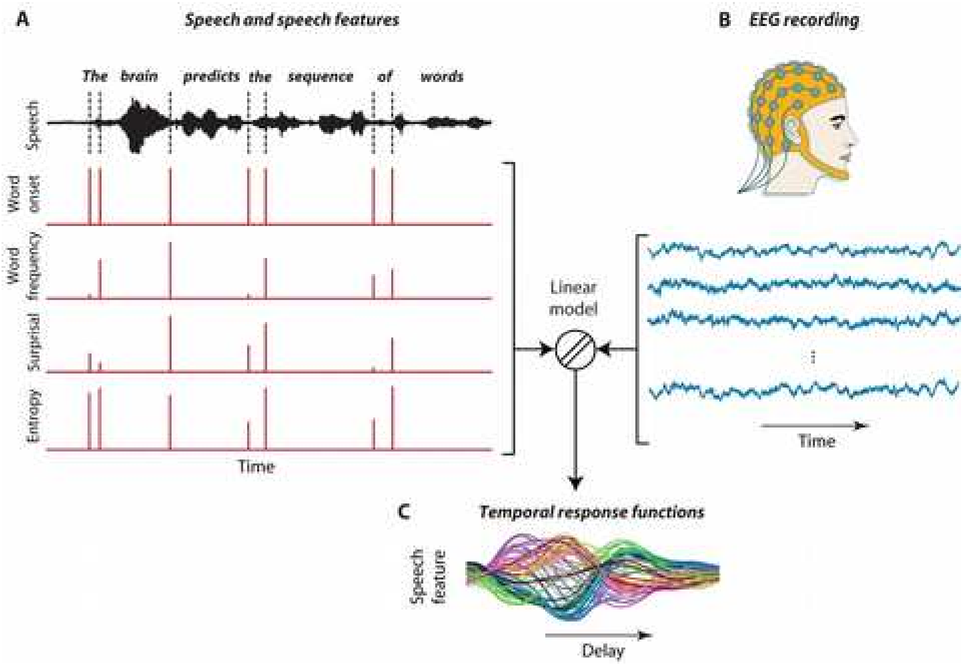


Figure II.2: Experimental overview. (A), We employ continuous speech narratives and utilize speech processing as well as language modelling to extract acoustic and linguistic features, namely word onset, word frequency, precision and surprisal. (B), The participant's neural activity is recorded through EEG while they listen to the stories. (C), We extract temporal response functions for each of the four speech features through computing a linear model that estimates the EEG recordings from the speech features.

in the acoustics at word boundaries and partly result from the brain's parsing of the acoustic signal to form discrete linguistic units (Ding and Simon, 2014; Brodbeck, Presacco, and Simon, 2018). To account for this onset response, we constructed a word onset feature as a series of spikes, each of which had unit amplitude and was located at the onset of a word. This feature also takes the role of capturing the variance in EEG, elicited by any word, which cannot be explained by the variance from surprisal and precision values. Second, we computed the word position within a sentence and used it as another word feature. The latter can be correlated with precision, as the entropy tends to decrease across words within the sentence. Hence the word position feature is for us a control to ensure that the neural response to precision is not solely due to the incremental processing occurring throughout a sentence. Third, the frequency of a word in a given language, outside its context, is a linguistic feature that acts as a prior probability for computing the probability of a word in a sequence (Brodbeck, Presacco, and Simon, 2018). Word frequency can also interfere with surprisal: less frequent words may indeed often be more surprising. To capture the share of the neural response that could be explained away by word frequency, we included the latter as a third linguistic feature. This feature was computed by scaling the amplitude of the spike at each word onset by the negative logarithm of the frequency of the corresponding word. The logarithm was used such that word frequency and surprisal were expressed in the same units.

Finally, to investigate the weighting effect that precision may have on surprisal, we added an interaction term "Surprisal * Precision". Indeed, in the framework of predictive processing for instance, precision comes as a top-down signal that modulates the degree of integration of predictions against bottom up information. Inaccurate predictions (associated with a low precision entropy) will let bottom up speech representations prevail against the prediction itself and vice-versa. This was computed by multiplying precision values with surprisal such that the interaction feature effectively stands as a confidence-weighted version of surprisal. In summary, we computed five speech features: one acoustic feature, word onset, and four linguistic features, word position in its sentence, word frequency, precision, and surprisal. To those, we add the interaction term between surprisal and precision. Each feature was a time series of spikes, which each spike being located at the onset of a word. The amplitude of the spike was constant for the word onset feature, and for each other feature it was scaled the corresponding value for each respective linguistic feature. All values of the different linguistic features have been standardized to have unit variance and zero mean across all words of our speech stimulus.

II.2.4 EEG acquisition and pre-processing

We recorded brain activity using 64 active electrodes (actiCAP, BrainProducts, Germany) and a multi-channel EEG amplifier (actiCHamp, BrainProducts, Germany). The presented sound was recorded simultaneously through an acoustic adapter (Acoustical Stimulator Adapter and StimTrak, BrainProducts, Germany) and was used for aligning the EEG recordings to the audio signals. Both the EEG and the audio data were acquired at a sampling rate of 1 kHz. The ear lobe was used as a reference for the EEG. The EEG data was processed by first applying an anti-aliasing filter (Kaiser window, FIR filter, cutoff -6 dB at 125 Hz, transition bandwidth 50 Hz, order 130) and by downsampling the data to 250 Hz to reduce the computation time of subsequent operations. A high-pass filter (Hanning window, sinc type I linear phase FIR filter, cutoff -6 dB at 0.3Hz, transition bandwidth 0.15 Hz, order 5168), was then applied to every channel to remove non-stationary trends such as slow drifts and offsets. Bad channels were identified using the procedure ’clean_rawdata’ from the EEGLAB plugin ASR (Artefact Subspace Reconstruction); they were then removed and interpolated with spherical interpolation. All channels were then average referenced. We subsequently ran an ICA decomposition and removed artifacts from eye blink, eyes movement as well as muscle motion by visual inspection of the ICA components. The cleaned data were low-pass filtered (Hamming window, linear phase FIR filter, cutoff -6 dB at 62 Hz, transition bandwidth 10 Hz, order 138) and further down-sampled to 125 Hz. The filtered EEG data therefore contained the broad frequency range from 0.3 Hz to 62 Hz.

We computed Temporal Response Functionss (TRFs) from EEG data in several frequency bands. The TRFs followed from a linear forward model that expressed the EEG signal at each electrode as a linear combination of the speech features shifted by different latencies, as detailed in section I.5 (Ding and Simon, 2012; Di Liberto et al., 2015). Every filter used were FIR type I filters, designed with synced windowed method, and using a hamming window. We filtered the EEG data in several frequency bands of interest: delta band (low-pass filter, cutoff at 4.5 Hz, filter order 132), theta band (band-pass filter, cutoff frequencies at 4 Hz and 8 Hz, order 206), alpha band (band-pass filter, cutoff frequencies at 8 and 12 Hz, order 206), beta band (band-pass filter, cutoff 20 Hz and 30 Hz, order 82) and gamma band (cutoff at 30 and 60 Hz, order 164). For every frequency band other than delta, we computed the power modulation by taking the absolute value of the Hilbert transform of the band passed data and further band passed it between 0.5 Hz and 20 Hz (filter order 824) to remove DC offset and higher frequencies that do not occur in our speech representations.

II.2.5 EEG data analysis

To relate the speech features to the EEG data, we used a linear spatio-temporal forward model that reconstructed the EEG recordings from the acoustic feature and the three linguistic features, shifted by different delays (see figure II.2). A detailed mathematical description of such is given in the introduction, section I.5 on page 27. Such an approach has recently been used successfully for assessing the cortical tracking of the speech envelope, phonemic information as well as semantic dis similarity of words in speech (Ding and Simon, 2012; Di Liberto et al., 2015; Broderick et al., 2018). The coefficients resulting from this regression constitute the TRFs that inform on the brain’s response to each feature at different latencies. In particular, the forward model sought to express the pre-processed EEG recordings

$\{x_i(t_n)\}_{i=1}^{N_{chan}}$ of the $N_{chan} = 64$ channels at each time instance t_n through the time series $\{y_j(t_n - \tau_k)\}_{j=1}^{N_{feat}}$ of the $N_{feat} = 6$ speech features (word onset, word frequency, word position, word precision, word surprisal, and the product of surprisal and entropy), shifted by N_{lags} different delays $\{\tau_k\}_{k=1}^{N_{lags}}$. This actually corresponds to a discrete convolution between features and TRF kernels, which is linear in the kernel coefficients. The model is thus:

$$\hat{x}_i(t_n) = \sum_{j=1}^{N_{feat}} \sum_{k=1}^{N_{lags}} \beta_{ij}(\tau_k) y_j(t_n - \tau_k) \qquad \text{(II.3)}$$
$$\forall n \in \{1, \ldots, T\}\,; \quad \forall i \in \{1, \ldots, N_{chan}\}$$

This equation is the application of equation (I.1) on page 27. We hereby considered equally spaced delays $\{\tau_k\}_{k=1}^{N_{lags}}$ that ranged from -400 ms to 1,100 ms. At the sampling rate of 125 Hz this yielded a number of $N_{lags} = 188$ lags. T is the total number of samples in the sampled recording of the speech aligned EEG. The obtained *estimate* for the EEG channel i is denoted by $\hat{x}_i$. The coefficient $\beta_{ij}(\tau_k)$ is the TRF for the ith EEG channel and speech feature j at the latency τ_k. The preprocessed EEG recording $\{x_i(t_n)\}_{i=1}^{N_{chan}}$ was either the EEG signal in the delta band or the power of higher frequency activity. We computed those coefficients for each participant separately, leading to a set of TRFs on which we could apply our group-level statistical analysis. The different speech features that we employed were partly correlated. The largest correlation emerged between surprisal and the interaction term "surprisal * precision", at a value of 0.61. We wondered if these correlations would hinder the EEG analysis, and in particular if they would obscure the neural responses to the individual speech features through the linear regression analysis, an issue known as multicollinearity (Kumar, 1975). A high multicollinearity between features could result in higher variance or leakage between the coefficient $\beta_{ij}(\tau_k)$. However, the Frisch–Waugh–Lovell theorem from econometrics states that linear regression based on correlated features yields the same results as when the features are first orthogonalized, that is, decorrelated (Frisch and Waugh, 1933; Lovell, 2011).

In addition, in our implementation of the multiple linear regression we used a singular value decomposition of the design matrix of time-lagged features, resulting in transformed features that were mutually uncorrelated. The correlation of the features was therefore not problematic. The only issue that multicolinearity can cause is significantly increased variance for each $\beta_{ij}(\tau_k)$ estimate, which typically emerges when the variance inflation factor (VIF) is above 5. For our speech features we obtained VIFs between 1.22-2.25, indicating that increased noise due to correlated features is not an issue.

We fitted the TRF models on N-1 subjects and evaluated each models on the left-out subject so that the model has never seen those data before. This was repeated for each subject. The evaluation of a model was done by computing the *reconstruction* *accuracy*, as measured by the correlation between predicted EEG signals and the true EEG in a given frequency band:

$$\rho_i = \text{corr}(\mathbf{\hat{y}_i}, \mathbf{y_i}) \qquad i \in \{1..64\} \tag{II.4}$$

Where y_i and $\hat{y}_i$ are the EEG time-series of channel i and the time series predicted by the TRF respectively. This gives a total of 64×13 scores as it is evaluated on each subject. The evaluation of those scores was systematically done using a cross-validation procedure. Here, on the subject level, the cross-validation occurs within subjects. Namely, we used a 5-folds cross-validation where the TRF model is trained on 4 folds, and evaluated on the fifth. This is then repeated for each fold and the average score across folds is used as the subject score. One quick approach to evaluate the quality of a given model was to take the average score across electrodes for each participants as in figure II.3.

As an additional control that our TRFs did not contain leakage from responses to different features, we developed a null model that was employed to assess the statistical significance of the actual TRFs (see below). The null model was constructed such that a potential leakage between features would appear similarly both in the actual model and in the null model, and therefore would not result in statistical significant results. It follows that any statistically-significant part in the TRFs that we obtained did not result from leakage between the features.

II.2.6 Statistical significance

In order to determine the statistical significance of the estimated TRFs , we computed chance-level TRFs as a null model. The chance-level TRFs were generated by constructing unrelated speech features, and by regressing these to the EEG recordings in the same way as for the computation of the actual TRFs . To establish chance-level linguistic TRFs , only the linguistic information of interest contained in the spike amplitude of the speech features but not the acoustic information in the spike timing needed to be unrelated to the EEG. We therefore created unrelated speech features by keeping the timing of the spikes identical to those in the true model. The speech feature that described word onsets was therefore not altered. However, we changed the amplitude of the spikes for the other linguistic speech features by taking their values from an unrelated story (that is a story not aligned with EEG during the experiment). To obtain a large number of null models, we then considered permutations of all our 15 story parts. Through permutation of the entire story parts, and not the order of individual words, the statistical relationship between the linguistic features of successive words was conserved. However, because we kept the timing of the spikes in the null model as in the actual stories, the obtained null model could only be used to determine the significance of the neural responses to the linguistic features, but not for those to the acoustic word onset.

The subject-level TRFs were then analysed for statistical significance at the group level by comparing them to the 1,000 null models. The comparison was obtained from a permutation test together with cluster-based correction for multiple comparison (Oostenveld et al., 2011), where only clusters of at least four electrodes were kept. Specifically, we used the function spatio_temporal_cluster_test from MNE python library (Gramfort et al., 2013).

The procedure consists in determining whether the global null hypothesis can be rejected, i.e. that there is at least one cluster (here across space, or channels, and time lags) which makes data non-exchangeable between experimental conditions (our two conditions being the true TRF and the null ones). The calculation of cluster statistics is derived first from a t-test applied to every sample (a (channel, time)-pairs). The computed t-value is *not* the cluster-based test statistic. Then, samples whose t-values are larger than 0.01 are selected as a possible candidate member of a cluster. If the candidate samples are connected, based on temporal and spatial adjacency, they will be assigned to a *cluster*. From here, one can compute the *cluster-based statistics* by taking the sum of the t-values of each samples within the clusters. Finally, a reference distribution is created by permuting samples between the two conditions (the true TRF and the null models) and re-calculating cluster statistics on this random partition. This permutation is done 1000 times thus providing us with a distribution of cluster statistics values forming the reference distribution. We reject or accept the null hypothesis by computing the proportion of permutation statistics (from the reference distribution) that are greater than each cluster-based statistic. This corresponds to the p-value of each cluster. If the p-value is smaller than the critical level we can conclude that the two experimental conditions are significantly different, which in our case results in having significant samples at the group-level for our TRF that differ from the null models. We considered only clusters with a p value greater than 0.05/10 (hence correcting for the multiple tests done, as we carried this analysis for 4 frequency bands of interest with only a subset of linguistic features for each band, yielding a total of ten such analysis). This permutation test effectively controls for multiple comparison across channels and time.

II.3 Results

II.3.1 Behavioural assessment

We first assessed to which degree the participants understood the stories through asking them comprehension questions. These questions were answered with an average of 96% accuracy, evidencing that the volunteers consistently understood the speech and paid attention.

II.3.2 Cortical tracking of acoustic and linguistic speech features

The cortical tracking of the speech features can be found in different frequency bands. First, because all four features relate to words, the frequency range of the features is similar to the rate of words in speech. The latter is about 1 – 4 Hz and corresponds to the delta frequency range. Cortical activity at low frequencies, including the delta frequency band, can therefore be evoked by or entrain to the rhythm set by the acoustic and linguistic word features. Second, the amplitude of the neural activity in higher frequency bands can be modulated by the speech features. In order to obtain results comparable to previous work in this domain we opt to focus on activity in the classic EEG frequency bands. This is a relatively arbitrary decomposition but it reflects main band of activity as observed in raw EEG data and provides us with comparable results. This allows to directly map our contributions to similar work such as Bastiaansen et al., (2008) where power increase in beta and gamma bands have been studied or as in Etard et al., (2019), which showed how activity in theta and delta bands differentially decode comprehension and attention. Our stimulus residing mainly in the delta band, we ought to take the power of higher frequency band in order to observe any response to the stimuli at those higher frequencies in the EEG using a linear model. The classic frequency bands for EEG are the theta band (4 - 8 Hz), the alpha band (8 – 12 Hz), the mid-high beta frequency band (20 – 30 Hz, and gamma band (30 – 100 Hz), the power of which can be modulated by prediction in sentence comprehension (Bastiaansen and Hagoort, 2006; Bastiaansen, Magyari, and Hagoort, 2010; Weiss and Mueller, 2012; Wang, Hagoort, and Jensen, 2017).

Figure II.3 presents the reconstruction accuracy, measured as the correlation between predicted and true EEG and averaged across all channels. Averaging over the whole head is misleading, as some electrodes might not benefit form certain word features in the reconstruction, but this is a crude way to measure the predictive power of each feature to encode EEG data. We trained a different model for each set of individual features (so not for interaction) in the delta band, that is the band where phase locked activity to word onset is plausible. We observed significant increase ($t = 5.4$, degree of freedom = 12, p-value $< 1e\text{-}4$) for word frequency and surprisal only with respect ro reconstruction accuracy achieved with a model containing only word onsets. For reference, we also show the topography of reconstruction accuracy of a model with word onset only in panel **B** of figure II.3 while panel **C** present the difference in reconstruction accuracy between the full model and the word onset only model. We observe an improvement in parietal electrodes and also in the left anterior temporal ones.

We started by quantifying the neural tracking of the word features at low frequencies. In figure II.4, we observe in (A), the responses to the word onset appear as insignificant due to the construction of the null model. In panels (B,C), we obtain significant neural responses to word frequency as well as surprisal for delays around 400 ms (D), Significant neural responses to precision arise around delays of 100 ms as well as at 500 ms. In figure II.4 panel (E), the interaction between surprisal and

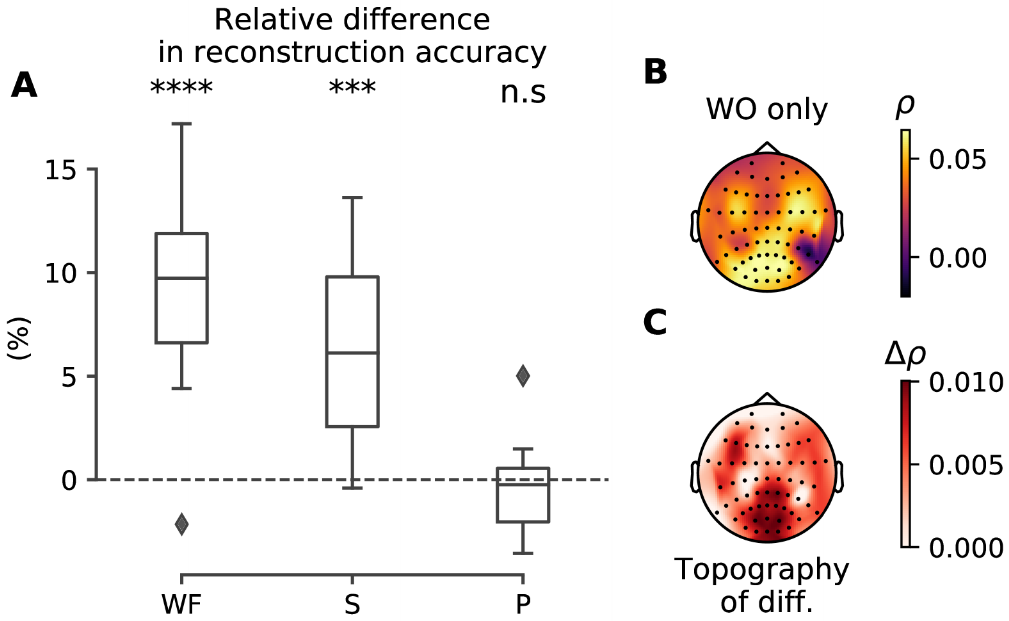


Figure II.3: A: Relative increment in reconstruction accuracy (see equation (II.4)) with respect to a model with only word onset feature, averaged across electrodes. B: Topography of reconstruction accuracy ρ at each electrode, average across all participants, for a model with only word onset feature. C: Difference in accuracy with the full model.
WO: word onsets only, S: word onsets and surprisal, WF: word onset and word frequency, P: word onset and precision.

precision leads to a neural response at a delay of 400 ms as well as at a long delay of 1,000 ms. The topographic plots of the responses show large differences between the temporal scalp areas on the one hand and the parietal and occipital areas on the other hand.

Importantly, we found significant responses to the word surprisal around a delay of 450 ms (figure II.4). These responses emerged predominantly in the temporal and occipital scalp areas and was lateralised on the left hemisphere. Precision was tracked by cortical activity at delays of around 100 ms and around 500 ms. Moreover, we observed a significant neural response to the interaction of surprisal and precision, at an earlier latency of around 400 ms and at a longer latency of around 1,000 ms. We also computed the modulation of the power in the theta band, the alpha band, the lower and higher beta band as well as in the gamma band by the acoustic and linguistic features (figures II.6 and II.7). While the power in the alpha band and in the lower beta band was not significantly related to the linguistic features, the power in the theta band was shaped by word frequency at delays of around 300 ms and around 1,000 ms (figure II.5). Furthermore, the power in the theta band was significantly decreased by precision at delays of about 700 ms.

The power in the higher beta band correlated positively with surprisal at delays of around 700 ms and 1,000 ms (figure II.6). At the latter delay, the influence of surprisal was strongest at the left temporal channels. Moreover, the power in the higher beta band was modulated by precision at a delay of about 700ms, with the main contributions coming from the occipital channels The power in the gamma band was shaped by surprisal around the early delays of about 0 ms, with pronounced modulation of the gamma power at the electrodes from the left temporal area (figure II.7). The gamma power was also increased by word with higher surprisal at the long latency of around 1,000 ms, again mainly for the left temporal channels. The interaction of surprisal and precision shaped the gamma power as well, at the early delay of about 0 ms.

Even though we carefully designed the analysis in a conservative way, evaluating each score on held-out data and using multiple comparison corrections with permutation based cluster statistics to establish significance at the population level. We were intrigued how much single participants were biasing the grand average. Therefore we computed several averages, each time leaving a single participant out, in order to verify any participant was driving the effects observed. The figure II.8 show those results.

Figure II.8 represents one channel (randomly picked, for illustrative purposes). For this channel and a given TRF model (e.g. Word Frequency feature), we compute the grand average of this signal but discarding *one* participant each time, and then repeat this for all participants. Thus we have 12 different grand averages, always with a different subject being left out. In red we can observe the actual grand average. The same data are then also presented in the right panel, but we subtracted the mean (red) and show now for, each participant, the spread of samples (taking each lags as independent sample) around that mean. We see already on the figure how

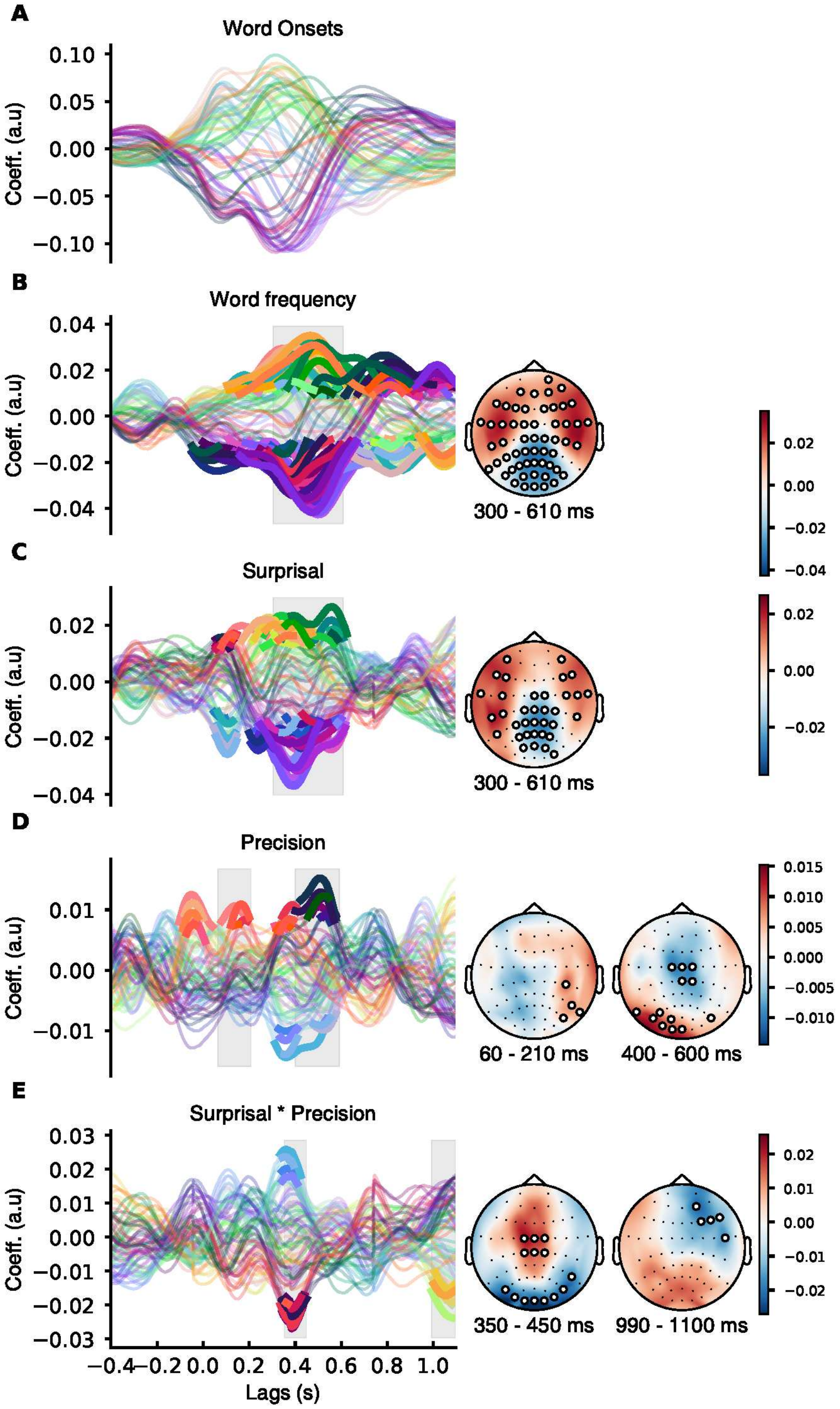


Figure II.4: Temporal response functions (TRFs) for acoustic and linguistic speech features. The temporal response functions for each electrode are shown in bold at time instances where they are significant compared to a null model that is based on shuffled data. EEG channels that yield a significant response within a particular range of delays, highlighted in grey, are indicated in white in the topographic plots. Each rows represent the TRF coefficient $\beta(\tau)$ for each feature used respectively. Individual lines correspond to TRF coefficient amplitude for different electrodes.

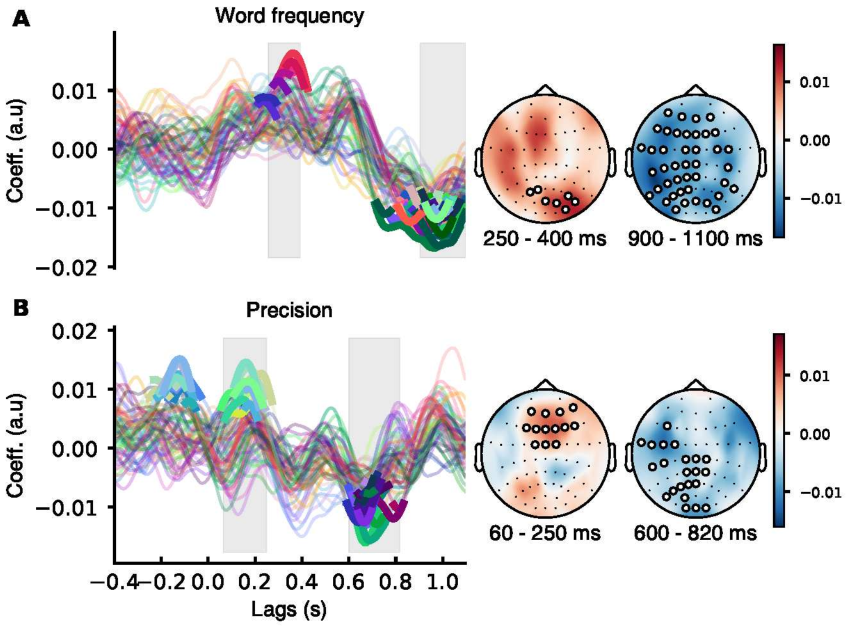


Figure II.5: Neural responses in the theta frequency band. (A), Word frequency is positively correlated to theta power at a delay of 300 ms, and is negatively correlated at a delay of 1,000 ms. (B), Words that can be predicted with higher precision lead to a decreased theta power at a latency of 700 ms.

little spread, for each “left-out" subject, there is around the grand average value.

We then used the Kruskal-Wallis H-test which is a non-parametric test that tests the null hypothesis that a population median of all groups are equal. It is the non-parametric version of ANOVA. Our groups are the subject (more precisely each grand averages with one subject out only) and the samples are the coefficients values at each lags. We could not reject the null hypothesis (statistic=6.96, p-value=1.), asserting that all “reduced" grand averages have the same population median.

II.4 Discussion

We have shown that cortical activity tracks the surprisal of words in speech comprehension. Such cortical tracking has emerged at low frequencies, that is, within the delta band that encompasses a similar frequency range as the rate of words in speech. Importantly, we found that the neural activity in the faster theta, beta and gamma frequency bands tracks the surprisal as well. These frequency bands have previously been suggested to be involved in the bottom-up and top-down propagation of predictions and prediction errors (Lewis and Bastiaansen, 2015). We have further demonstrated that the cortical tracking of word surprisal is modulated by the precision: the interaction between surprisal and precision lead to responses both in the slow delta band as well as in the power of the faster gamma band. In par-

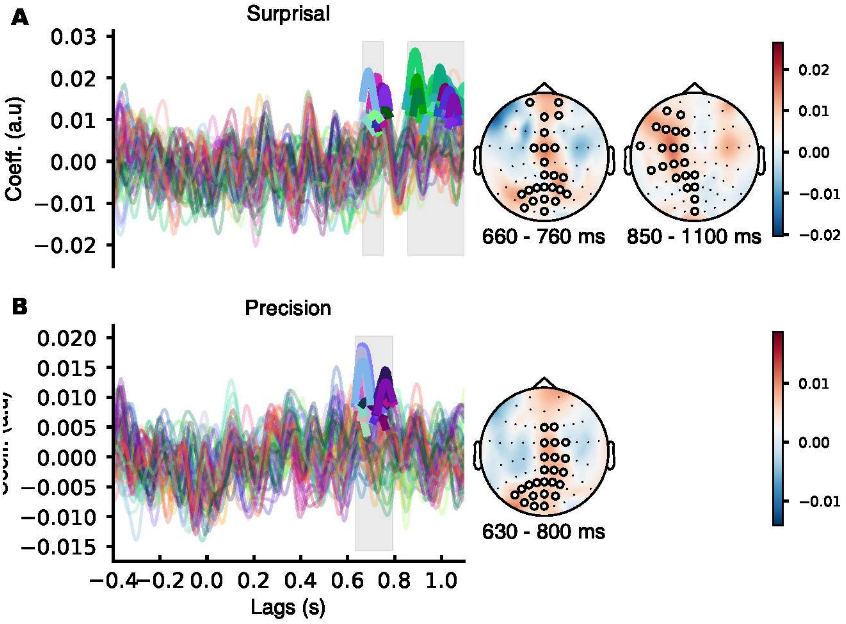


Figure II.6: Neural responses in the higher beta frequency band. (A), There are significant neural responses to surprisal, emerging at delays of 700 ms and 1,000 ms. (B), Precision causes an increased power in the higher beta band activity around a delay of 700 ms but also at early onset spreading around 0ms. This early significant rise may seem anti-causal, however we note that to compute prediction entropy at a given word in the sentence one needs only information up to the previous word and hence such temporal locus for prediction is still a causal effect. This is further discussed in

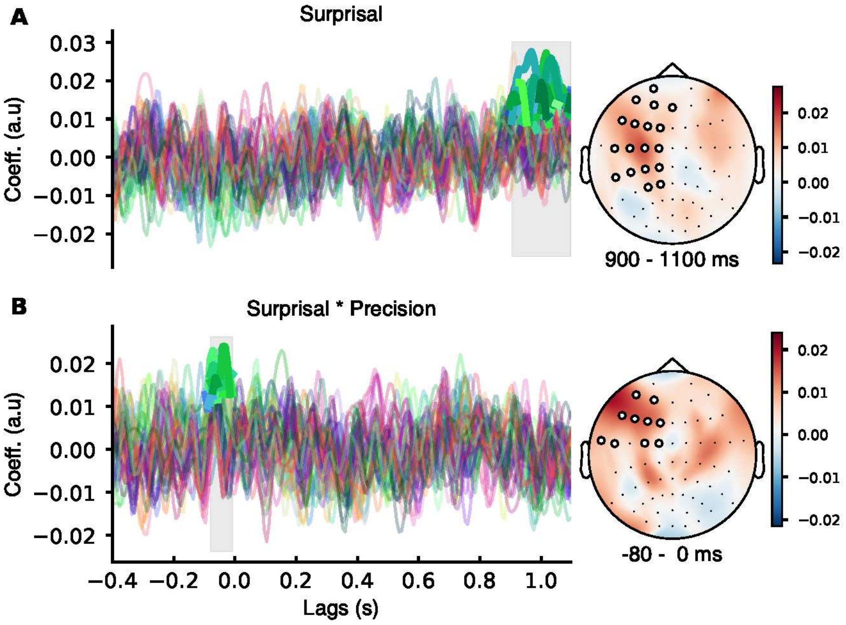


Figure II.7: Tracking of surprisal by gamma-band activity. (A), The gamma activity is increased in response to words with high surprisal at a delay of 0 ms, and is decreased at around 1,000 ms, both times mostly in the left temporal and frontal scalp areas. (B), The interaction between precision and surprisal leads to a modulation of the gamma power at the latency of around 0 ms. This modulation occurs predominantly for left temporal and frontal channels as well. As in figure II.5, we note that prediction does not need to "know" the current word to be computed, which allows for early effects to still be causally accounted for.

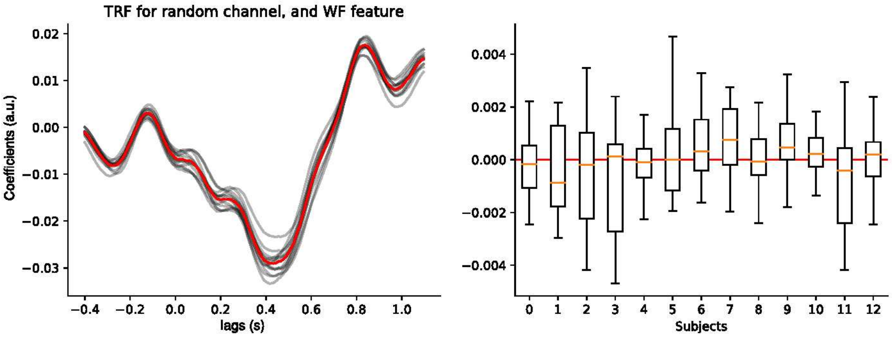


Figure II.8: Left panel: Each grey line correspond to the TRF coefficients for Word Frequency at one randomly selected channel location for twelve different “grand averages", namely each grand average misses one different subject. The red line indicate the true grand average. On the right panel we show those data as box plots. Each box corresponds to the samples, for a given left out subject, across all lags and from which we subtracted the true grand average (red horizontal line at zero).

ticular, word predictions that are made with high precision but then lead to large surprisal cause an increased gamma power at zero lag. However, as opposed to a previous study on event-related potentials, we did not observe a significant effect in the theta or alpha bands (Rommers et al., 2017). This difference may be due to our use of naturalistic stimuli, and the inclusion of all words in the analysis, while the previous study used specialized sentences with final words that had either high or low surprisal, and either high or low precision.

II.4.1 Evidence for predictive coding account of language processing

The cortical tracking of surprisal may indicate predictive processing by the brain. Predictive processing is a framework for perception in which it is assumed that the brain infers hypotheses about a sensory input, that the hypotheses are constantly updated as new sensory information becomes available (Friston and Kiebel, 2009; Friston, 2010; Kanai et al., 2015). In particular, the surprisal of a word reflects a prediction error, a key quantity in the framework of predictive coding (Friston, 2010; Sedley et al., 2016). However, the expectancy of a word based on previous words also correlates with the plausibility of a word in a particular context (DeLong, Urbach, and Kutas, 2005; DeLong, Quante, and Kutas, 2014; Nieuwland et al., 2020). Further studies are therefore required to disentangle neural correlates of actual word prediction from those that do not require predictive processing, such as word plausibility. The surprisal of a word can reflect both its semantic as well as syntactic information, and previous investigations into the neurobiological mechanisms of language comprehension have manipulated both independently (Henderson et al., 2016; Humphries et al., 2006). In contrast, our approach has taken a naturalistic and holistic approach to surprisal; we employed natural speech without manipulations combined with statistical learning of a rich variety of natural language cues through a recurrent neural network. Because the neural network infers both syntactic rules as well as semantic information from the training of the speech material, the reported neural response to word surprisal can reflect both semantic as well as syntactic information (Elman, 1990; Collobert et al., 2011).

We observed a significant response to precision entropy around 0 ms (onset time of a word), with peaks before and after the word onset for the theta band (see figure II.5) or right at 0 ms for the gamma band (figure II.7). This might appear as an anti-causal response. However we shall carefully examine what quantity is proposed for the precision entropy. Indeed, as shown in equation (II.2), precision at some position in a sentence only depends on previous words and *not* on the current word (as opposed to surprisal or word frequency). No information on the word currently heard is needed. In other words, the precision entropy can be computed fully before the word onset and such early latency does not reflect an anti-causal response. Interestingly, the effect is seen exactly around 0 ms. This highlights the possible prediction of the latency at which words are perceived. As the cortical response to precision is already acting on neural populations to acquire the newly incoming word by weighting neural activity with its prior knowledge. This furthermore emphasizes the role of precision entropy as a Bayesian prior in the context of word prediction.

II.4.2 Link with other language related ERPs

It is instructive to compare the reported neural responses to surprisal to the well-characterized event-related responses that can be elicited by violations of semantics, syntax or morphology in sentences. In particular, semantic violations can cause the N400 response, a negativity at 200 − 500 ms at the central and parietal scalp areas (Kutas and Hillyard, 1980; Kutas and Federmeier, 2011). Syntactic anomalies due to ungrammaticality or temporary misanalysis elicit the P600, a broad positive potential that is located at the posterior scalp area and arises around 600 ms after the anomaly (Friederici and Kotz, 2003; Hagoort, 2003). More specific syntactic anomalies can lead to negative potentials that occur anteriorly and that can be left lateralised, either occurring at 300 − 500 ms (LAN) or earlier, at 125 − 150 ms (ELAN) (Rösler et al., 1998; Van Den Brink, Brown, and Hagoort, 2001; Friederici, 2002; Steinhauer and Drury, 2012).

These Event Related Potentialss (ERPs) do presumably not reflect the activation of single static neural sources, but rather waves of neural activity that propagate in time across different brain areas (Maess et al., 2006; Tse et al., 2007; Kutas and Federmeier, 2011). In the case of the N400, for instance, this wave of activity starts at about 250 ms in the left superior temporal gyrus, and then propagates to the left temporal lobe by 365 ms as well as to both frontal lobes by 500 ms (Helenius, 1998; Halgren et al., 2002; Van Petten and Luka, 2006). A recent theory suggests that this wave of activity reflects reverberating activity within the inferior, middle and superior temporal gyri that corresponds to the activation of lexical information, the formation of context and the unification of an upcoming word with the context (Baggio and Hagoort, 2011). The spatio-temporal characteristics of the responses to surprisal that we have measured here share certain similarities with these ERPs . In particular, we have found neural responses to surprisal at latencies between 300 ms and 600 ms. These responses show a central-parietal negativity that is reminiscent of the N400. However, other features of the neural responses that we describe here appear distinct from these ERPs . The neural response to surprisal in the delta band at the latency of 600 ms does, for instance, not display the posterior positivity of the P600. Moreover, we have identified late responses around 700 ms and 1,000 ms. We have also shown that neural responses to surprisal arise in various frequency bands, beyond the delta band that matters for the ERPs . However, a further comparison of the neural response to surprisal to the related ERPs is hindered by the lack of spatial resolution offered by EEG recordings. Future neuroimaging studies using intracranial recordings or magento-encephalogram (MEG) may localize the sources of the neural response to surprisal that we have measured here and quantify potential shared sources with these ERPs .

The difference of the cortical tracking of surprisal to the well-known neural correlates of semantic, syntactic or morphological anomalies, and in particular the late responses at a delay of around one second, may come as a result of our use of natural speech that differs from the artificially constructed and tightly controlled stimuli used to measure ERPs . First, in our experiment the subjects encountered no violations of semantics, syntax and morphology, but instead heard naturalistic speech, within which the words occurred in context. Second, our stimuli did not contain artificial manipulations of word surprisal or entropy. Instead of altering the stimuli, we focused on quantifying surprisal and precision, or confidence, as it varied naturally in the presented stories. Third, we assessed the responses to surprisal and precision at each word in the story, and hence for words in every sentence position, rather than for words at a particular position within each sentence. Adding the word position within sentences, we can thus avoid the possibility of sentence position having an effect on the results (Bastiaansen, Magyari, and Hagoort, 2010). Fourth, we did not employ isolated sentences but continuous stories so that information of integration occurred over time scales exceeding a few seconds.

II.4.3 Possible source localisation

While our EEG recordings showed the cortical tracking of surprisal in different frequency bands, they did not allow us to precisely localize the sources of the activity in the cortex. Pairing EEG with functional magnetic resonance imaging (fMRI) or employing magnetoencephalographic (MEG) may allow to add spatial information to the temporal tracking that we have assessed here. A recent fMRI study, for instance, found that the left inferior temporal sulcus, the bilateral posterior superior temporal gyri, and the right amygdala responded to surprisal during natural language comprehension, while the left ventral premotor cortex and the left inferior parietal lobule responded to entropy (Willems et al., 2016). Another recent magnetoencephalographic (MEG) measurement of the brain's natural speech processing found that entropy and surprisal play a role in the assembly of phonemes into words, and involves brain areas such as core auditory cortex and the superior temporal sulcus (Brodbeck, Presacco, and Simon, 2018). Combining the temporal precision of EEG with the spatial precision of fMRI, or harnessing the ability of MEG to locate neural sources temporally and spatially, will allow to further clarify the spatio-temporal mechanisms of natural language comprehension in the brain.

Conclusion

In summary, we showed that neural responses to word surprisal can be measured from EEG responses to naturalistic stories. Our results demonstrate that both the slow delta band as well as the power in higher frequency bands, in particular the theta and higher beta band, are shaped by surprisal. Moreover, we also showed that the neural response to surprisal is modulated by the precision of a prediction. In particular, predictions made with high precision which lead to high surprisal modulate gamma power in the left temporal and frontal scalp areas. In addition, we also demonstrated that neural activity in the delta, theta and beta frequency bands is shaped by the precision of word prediction directly. These responses arise at different latencies and at different scalp areas, suggesting a rich spatio-temporal dynamics of neural activity related to word prediction.

Chapter III

Cortical responses to hierarchical syntactic features

SPeech signal is complex in nature. Its information content span many different time scales, from a few tens of milliseconds in which phonemic information is contained within spectro-temporal energy fluctuations, to several seconds in the form of syntactic structures that need to be encoded appropriately to extract the corresponding meaning of the speech utterance. However the complexity of speech signal is not only in that it unfolds across time scales of different order of magnitude but also in that it is constructed in a hierarchical way, where syntactic structures can be represented as tree structure of grammatical functions and items.

III.1 Introduction

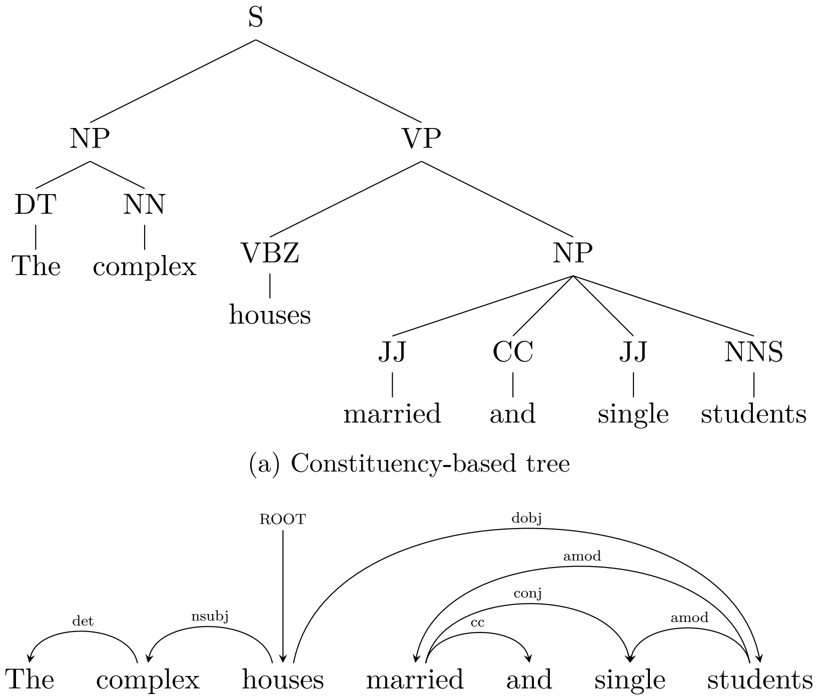
Whether the brain continuously performs prediction during language processing, it must embed the information in a compact and efficient representation to encode meaning of sentences. Indeed, human language is by nature organised in recursive and hierarchical fashion. Its content is not a pure linear sequence of items (words) but rather a complex organisation of those. Structure of words, hence syntax, can endow further meaning to the utterance beyond individual lexical information at the word level. In other words, grammar is an important aspect of language, and it allows to combine group of words and phrases such that one can alter the meaning of its subcomponents.

One important aspect of language modelling via RNN is that they do not rely on any particular linguistic assumptions. For instance, there is no information about syntactic categories or hierarchical structures, as they are simply trained on sequences of words with little lexical information contained in the embedding layer. The latter is itself derived from co-occurrence statistics in large corpus of text data. Conversely, Phrase Structure Grammars (PSGs) are built from prior linguistic knowledge, and are the foundations on which are constructed hierarchical syntactic structures.

Syntactic structures are categorical attributes given to a word or a group of words (phrase). They take place within a theoretical framework in which a finite set of rules suffices to generate an infinite amount of sentences (Chomsky, 2015; Chomsky, 2007; Hauser, Chomsky, and Fitch, 2002). This set of rules is called a *grammar*. Human language is uniquely shaped such that meaningful units can be arranged in a recursive and hierarchical way to form complex sentences. This distinguishes human language from any known form of animal communication such as bird songs as described by Hauser, Chomsky, and Fitch, (2002) and Friederici, (2011).

There are several ways to apply a grammatical rules in order to *parse* structure from an utterance. We can, for instance, describe a sentence by its syntactic constituent, which are nothing less than phrase structures, sub-parts of the sentence itself. Those groupings are pre-defined from grammatical rules such that a *noun phrase*, NP (another often seen constituent is the *verbal phrase*, VP), can be formed by a determinant (DET) and a noun (N), or a determinant, and adjective (JJ) and noun, or proper noun, and so on. From such a description, a tree representing the hierarchical organisation of such constituents is constructed. This is referred to as a **constituency-based tree**. Another approach is to relates words in the sentence based on their dependency relationships with each other. Each word is the *child* of a *parent* node, or word, with the verb as the root for the whole sentence. Such graph is called a **dependency-based tree**. We can see an example in figure III.1.

The parsing of a sentence can be ambiguous when several categories can be attributed to a word. Generally, the most frequent one will be chosen first leading to a *garden-path* effect (Osterhout, Holcomb, and Swinney, 1994) if the actual syntactic category was different (see also figure III.1). Researches have looked at brain responses, using electroencephalography, to those ambiguity resolution and word category violation. Studies showed that a systematic response, often linked to *syntactic repair*, emerges after 600ms in posterior parts of the scalp, the P600 ERP component (Osterhout, McLaughlin, and Bersick, 1997; Molinaro, Barber, and Carreiras, 2011a). Many ERP studies focused on the processing of grammar by looking at syntactic violations, and one hypothesis for the origin and mechanism in play behind the P600 component is related to the integration of a word into larger phrasal content. Integrating a word into a phrase structure that stand higher in the syntactic hierarchy boils down to merging its representation with the incremental build up of the phrase structure encoding. One key computation for such processingis the so-called *merge* operation (Berwick et al., 2013). It consists in merging two or more meaningful units, words or phrases. For instance, X and Y (e.g. $X=$"predicts" and $Y=$"the sequence"), into one new unified unit, $Z \sim X + Y$ ("predicts the sequence"), which carries a synthesised meaning. The compound Z might represent more than the sum of its parts. In that sense, *merging* is an essential computational process for syntactic processing and language comprehension.



(a) Constituency-based tree

(b) Dependency-based tree

Figure III.1: Example of Phrase Structure Grammars, here on a grammatically ambiguous sentence (this is effect is referred to as *garden-path* in linguistics). The word complex is more frequently used as an adjective (adjectives are commonly tagged as JJ in such representations), and the word houses as a plural noun (NNS) whereas in this example *complex* is a noun (NN) and *houses* a verb (VB) at the third person singular (VBZ). Determinant are tagged as DT, coordinators as CC, noun phrases as NP, verbal phrases as VP and a whole sentence as S. (b) represents a parse linking word by their relationship with each other: "The" is the determinant of "complex" which is itself the subject (noted as *nsubj*) of the verb "houses". "Students" is the direct object of the verb (*dobj* and is modified (*amod*)) by the adjectives "married" and "single". The initial parse naturally made by the brain is led as in a *garden-path* to take those first words according to their more common categories until the ambiguity arises.

More recently, Ding et al., (2016) and Ding et al., (2017) have shown results that they argue were direct evidence for cortical tracking of hierarchical syntactic structures. However, the structures employed in their stimulus were always similar from one sentence to another. This was a design choice so they could measure the frequency of supra-word level (phrasal, and sentence-level) tracking directly via electrophysiology recording although it also plays as a major issue in the paradigm. Indeed Frank and Yang, (2018), demonstrated in response to Ding et al.'s study that they could elicit the same pattern of response with a simulated model that was taking only lexical information into account (and hence could not rely on hierarchical syntactic structures at all). By using natural speech, we can explore a broad range of syntactic tree structures, such that the response observed in cortical activity, if any, will not be only a by-product of rhythmic presentation of specific word categories. However, in another study, Frank and Christiansen shows that a form of *syntactic surprisal*, computed as the amount of information gain obtained at each word given a certain tree structure, correlates with reading time and moreover is predictive of behavioural outputs above and beyond a model relying on non-hierarchical (hence sequential) linguistic information (Frank and Christiansen, 2018).

We propose to measure the cortical tracking to syntactic complexity during natural story comprehension. To describe syntax, we used features deriving from constituency-based trees themselves obtained from a context-free grammar. Attributes of those tree structures will give us insight on the interplay of different mechanisms occurring in the brain during processing of syntax. As a proxy for the *merge* operation for instance, we chose to look at the brain response to the number of branches in the hierarchy being closed by each word in a sentence. The aim is to extract brain responses to syntactic hierarchy, and to address the more general question of whether and how the cortex builds nested phrase structure representations.

A similar method of analysis that the one used in the first chapter to measure correlates of mechanisms involved in predictive processes during speech comprehension has been applied here. This methods allows to track brain responses using continuous recording and hence to take the advantages of using a more naturalistic and ecologically valid stimulus as explained in I.3.2. It also gives a method to track continuous response to hierarchical tree structures as they unfold on natural speech stimuli, without the need of artificially built syntactic mismatch or violations. This is restrictive to one type of tree parsing, but the goal here is not to underpin the exact mechanism used by the brain but rather to get a proxy of such computations, and a reliable measure of syntactic processing such as merging operation and tracking of phrasal structures at the individual subject level.

III.2 Methods

III.2.1 Material

We used the same material as in chapter II, also accessible in Weissbart, Kandylaki, and Reichenbach, (2019a), where English native speakers were asked to attentively listen to audiobook stories while their EEG was recorded ($N = 13$).

III.2.2 Syntax representations

The method used to analyse the response to syntax was also adopted from chapter II, yet here the speech representations were different. Instead of *surprisal* and *precision-entropy* derived from sequences of word we used linguistic informed features obtained from rule-based phrase structures. To extract the constituency trees for each sentence of our stimuli material we used a probabilistic context-free grammar (PCFG) parser form Stanford natural language processing software [1] (Klein and Manning, 2003). The parser outputs constituency-based trees on which were extracted features that describe the hierarchical structure present in the stimulus. It was important for us to be able to obtain a value representing the syntactic hierarchy for each word in the sentence, so that we could compare those results with our previous study.

A first level of representation of the hierarchy resides in the depth at which each word is situated within the tree structure. A second feature was then built to extract possible neural correlate to the *merge* operation. Having in mind that such an operation occurs after the last word of a constituent phrase is encountered. We modelled this with a feature that is zero for every words except those that act as closing nodes in the syntactic hierarchy. Moreover, the non-zero value was chosen to scale up with the amount of words *merged* into a constituent so as to represent the increased relaxation offered by freeing up working memory and also increased effort in unifying larger structures. This feature was labelled *closing branch* as it is given by the number of "branches" the current word is closing in the phrase structure tree hierarchy. Similarly, we represented the number of branches being *opened* by each word to account for possible prediction in cognitive loading to forthcoming words. The diagram in figure III.2 shows the afore described syntactic features for a given sentence.

Our control features, namely word onsets and word order within a sentence, were kept the same to account for any acoustic and linguistic responses not captured by our present syntactic features and avoid possible confound with sentence length, such controls were also used in Brennan and Hale, (2019).

As in chapter II, *reconstruction accuracy* is evaluated by measuring the cor-

[1] open source software available at https://nlp.stanford.edu/software/lex-parser.html

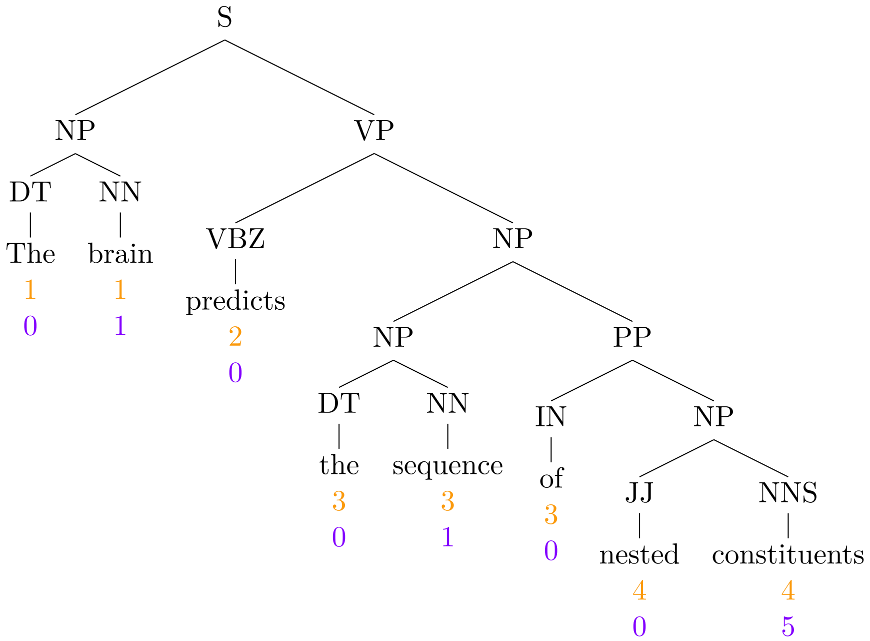


Figure III.2: Example of syntactic features used. In orange, the *syntactic depth*, and in purple the number of *closing branches*. Abbreviations stand for: sentence (S), noun phrase (NP), verbal phrase (VP), prepositional phrase (PP), determinant (DT), noun (NN), verb at third person (VBZ), preposition (IN), adjective (JJ), and plural noun (NNS).

relation between predicted EEG signals and true EEG (see equation (II.4)). The evaluation was carried out on held out data, using a 5-fold cross validation procedure and repeating the procedure for every subject.

III.3 Results

A significant response can be observed for all syntactic features used. That is, an increased EEG potential in specific scalp regions correlated with the depth in the syntactic hierarchy along with responses to either words that open new phrasal structures or words that are closing them. Importantly, the latter responses scale with the number of subtrees being opened or closed.

From the TRF model, we can also predict EEG activity from stimulus representation. We looked at the cross-validated score (correlation between true and predicted EEG) differences between a model with word onsets alone and a model with both word onsets and syntactic depth. The resulting topography shows where, in sensor space, the reconstruction benefits from adding the depth feature into the model. This analysis is also a way to verify that we are not observing spurious amplitudes in the TRF coefficients arising from multicolinearity. If the model were

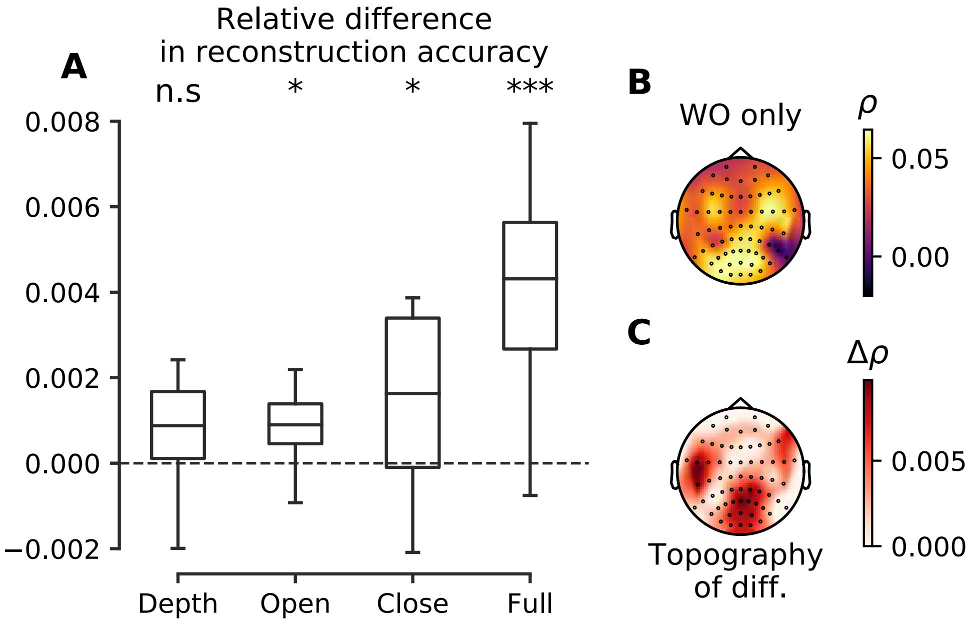


Figure III.3: A: Relative increment in reconstruction accuracy obtained with respect to a model with only word onset (WO) feature. B: Topography of reconstruction accuracy ρ at each electrode, average across all participants, for a model with only word onset feature. C: Difference in accuracy with the full model.

overfitting we would not observe any improvement from adding new features. Figure III.3 shows the overall (average across all channels) gain in reconstruction accuracy for each features independently. For those we added to word onsets only one other feature and computed the corresponding TRFs in the delta band. The greatest benefit occurred for the model containing all features. The feature that had the strongest increase was the number of closing nodes. Syntactic depth did not provide a benefit strong enough to be significant across all electrodes, although across participants a small group of electrodes had a significantly higher reconstruction accuracy than with a model with word onset only (see figure III.4) where we observed a left lateralised response at occipital, and left posterior temporal electrodes as well as at centro-parietal sensors.

We found a significant difference with our null models in the delta band. Full TRF are shown in figure III.5. *Syntactic depth* and *closing nodes* gave the greater responses in amplitude. The peak latencies are relatively early compared to usual syntactic violation studies. We report a first peak of significant response between 100ms and 200ms for those two features. The TRF for *closing nodes* is slightly left lateralised with a stronger posterior negativity at parieto-occipital electrodes. The *syntactic depth* shows a positivity at occipital locations and then a later centro-parietal positive peak of amplitude between 600ms and 700ms. As in chapter II, there is a sustained high amplitude response at later latencies beyond 800ms, but

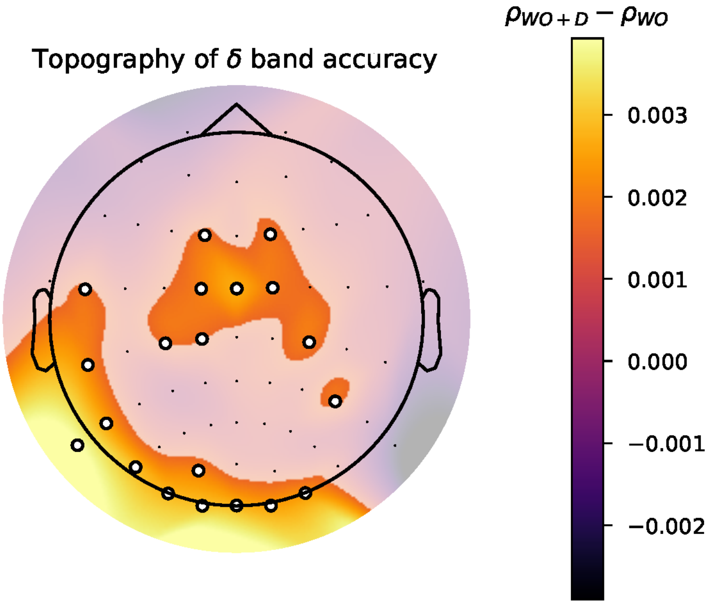


Figure III.4: Difference in accuracy between model with and without syntactic depth feature. Scores were computed using correlation between predicted and true activity at each electrode location. Non-significant are masked with transparency.

only for the response to *closing nodes*.

For higher frequency band, we found only significant TRF amplitude for the number of closing branches feature and only for theta and beta bands (see figures III.6 and III.7). None of our syntactic features TRF achieved significance in the alpha or gamma band. We expected to see syntactic depth to give a theta or gamma response as it was reported that theta power monotically increased with syntactic constituent (Bastiaansen and Hagoort, 2006). However it seems that the TRF obtained from the number of closing nodes explained enough of the theta power modulation to shadow any effect from syntactic depth.

III.4 Discussion

If the brains maintains a unified representation for each constituent, it has to do so for each level of nested structures and possibly with an increased cognitive load. This must happen incrementally as the utterance unfolds and will call for higher working memory demand as more nested phrases are parsed. The first goal of our study was to measure brain activity correlating with such cognitive mechanism. Hence the choice of syntactic depth as a representative feature. This number characterizing the depth of each constituent in the hierarchy was obtained from a PCFG parser. In other words, we are giving prior linguistic knowledge, specifically

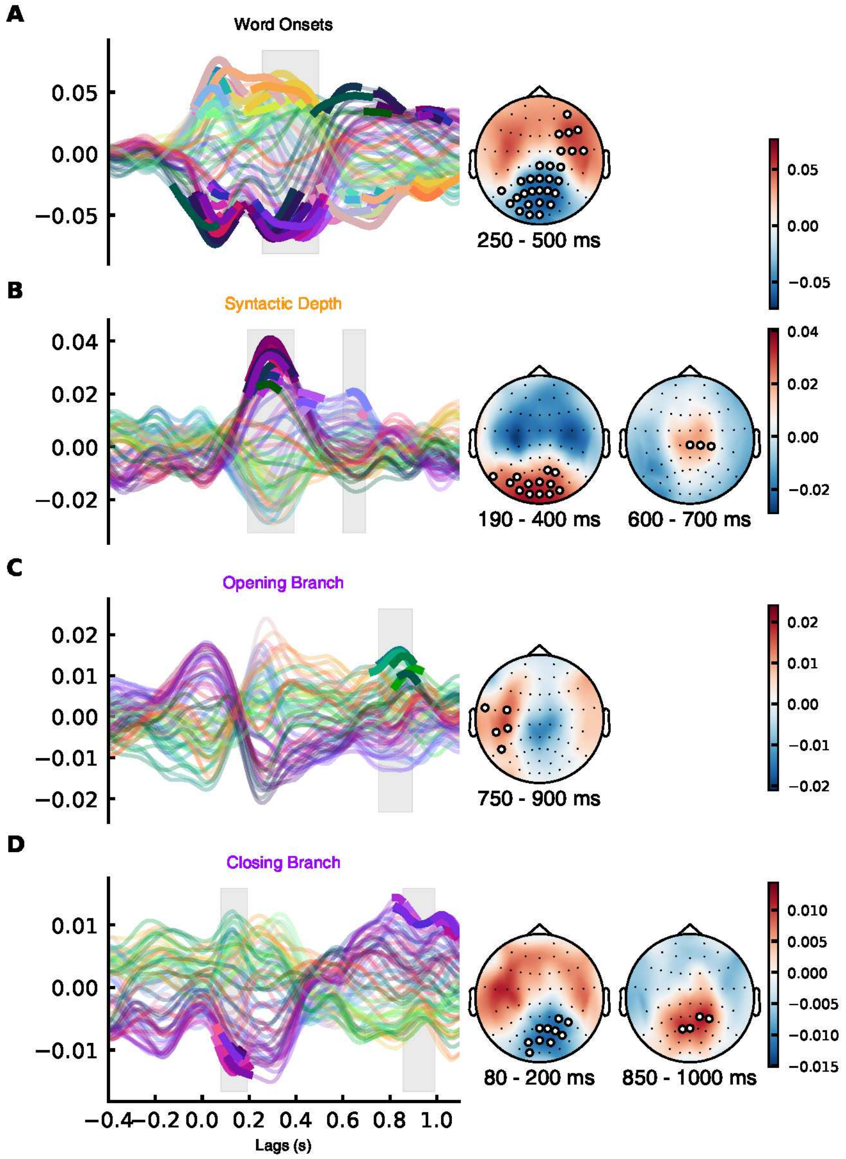


Figure III.5: TRF for syntactic features in the delta band. The temporal response functions for each electrode are shown in bold at time instances where they are significant compared to a null model that is based on shuffled data. EEG channels that yield a significant response within a particular range of delays, highlighted in grey, are indicated in white in the topographic plots. Each rows represent the TRF coefficient $\beta(\tau)$ for each feature used respectively. Individual lines correspond to TRF coefficient amplitude for different electrodes.

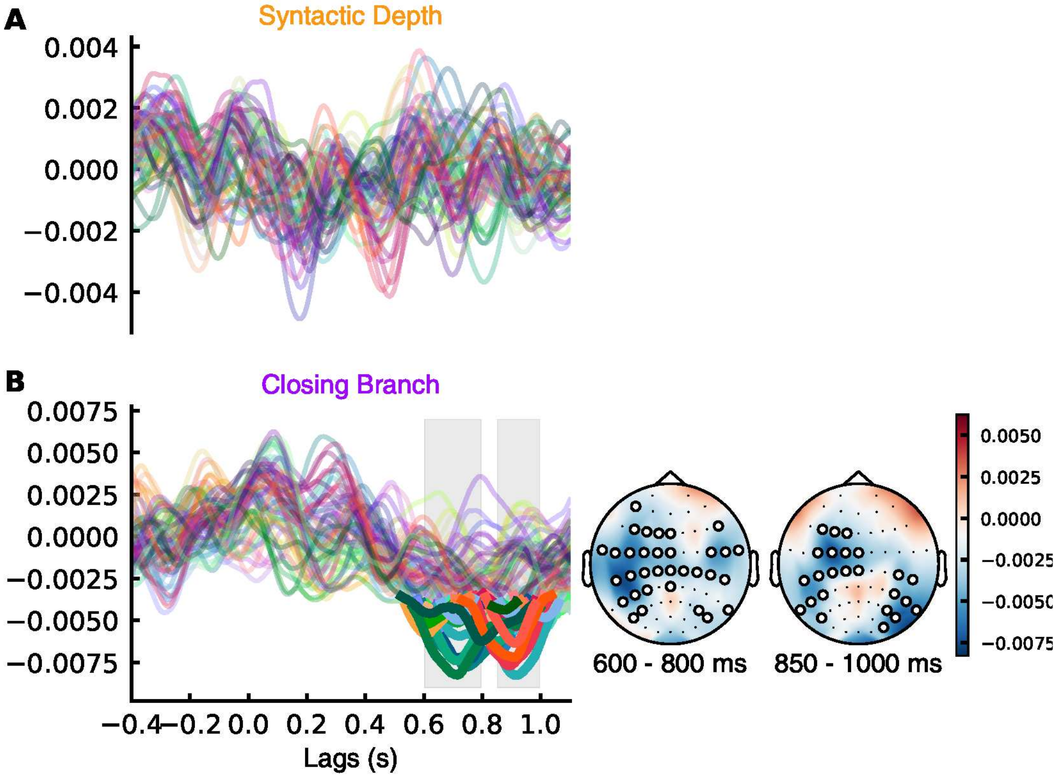


Figure III.6: TRF for syntactic features in the theta band

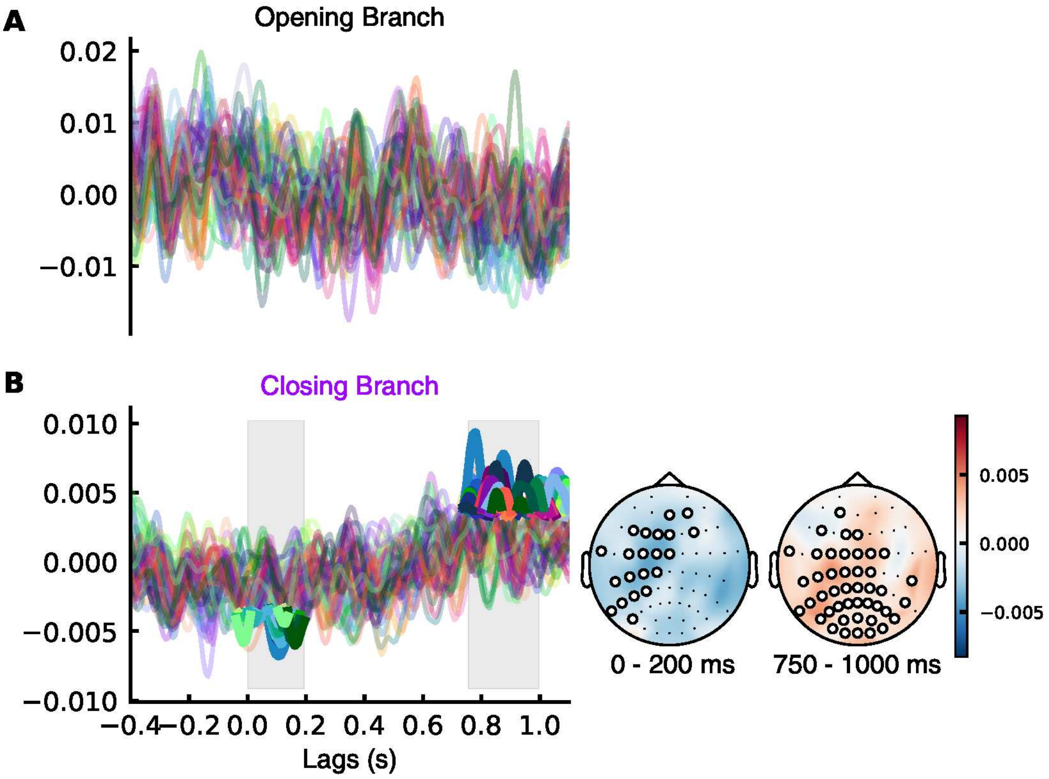


Figure III.7: TRF for syntactic features in the beta band

about grammar structure, to our model and extract brain responses in line with this information. A second objective in the study was to identify if the putative operation *merge* would elicit a strong enough response to be measured at the scalp level of individuals. "*Merge*" occurs when a list of words get reduced into a single hierarchically higher node, i.e. one phrase. A neural correlate of such operation might then be captured shortly after the last word of each syntactic constituent is processed by the brain. At that point, the brain can release the words held in memory as they enter a *merged* representation. Generally speaking, we were able to measure reliably cortical activity correlated with the number of syntactic branches a word closes.

III.4.1 A brain correlate to syntax complexity

The observed response for *depth* happens relatively early compared to syntactic violations studies (Left Anterior Negativity - LAN, 300-450 msec and P600 after 500 msec Molinaro, Barber, and Carreiras, (2011b)). This could be explained by the fact that we are using a continuous and more ecological stimuli. Therefore the response observed is the canonical brain activity in relation rather than in response to violation of grammar. The latter obviously incorporate a confound with a *repair* operations that is triggered when a violation occurs. However, earlier components have been linked to complexity of sentences and ambiguity resolution. Those aspects of syntax processing are not violation per se and can occur in natural speech. I argue that the depth TRF is tightly linked to processing of complex sentences. The deeper a word in the syntactic structures, due for instance to nested constituents and generally this happens for complex sentences, the stronger the cortical response to that word will be (positive or negative) at those significant electrode locations (see second row, right panel of figure III.5). However, the underlying mechanism is not necessarily solely related to syntax but could also be explained by the maintenance of syntactic structure in memory.

III.4.2 Closing nodes and Merge operation

The interaction between depth and closure however, should inform us on brain processes closely related to the *merge* operation. When enough embedding occurs such that the brain must maintained a deeply intricate structure in memory, the closure should be even more of a relief for the system to relax by chunking the whole structure in a condensed representation and move on to the new incoming input.

Interestingly, these data do not reflect a typical syntactic response as observed in ERP analysis for grammatical processing. The main difference between the present study and those ERP studies arise from our experimental design where we opted on using naturalistic stimulus. This obviously as to be taken in consideration when compared to data from syntactic violation studies which will involve processes that are not necessarily occurring during normal comprehension. The P600 component usually measured by violation studies has been linked by Molinaro, Barber, and Carreiras, (2011a) and Friederici, (2011), to *merge*. However, our *closing nodes* TRF is rather weak compared to reported P600 effect and happens at earlier latencies, with a negativity in left, centro-parietal areas rather than left anterior positivity. Actually the region left-anterior area shows an early positive peak, being the exact opposite in sign to the ELAN ERP component. However, the ELAN is usually not linked to the integration of new input into larger syntactic structure. Observing such an early latency for closing nodes could be indicative of predictive processes underlying the syntactic integration.

In a very good review about oscillatory processes in language processing, Bastiaansen and Hagoort, 2006 postulate that syntactic unification is reflected by high-frequency band synchronisation such as in the beta and gamma band. We did observe a decrease for the number of closing branches, but failed to see a linear increase following the locally monotonically increasing values of syntactic depth as we advance within a syntactic sub-tree. The syntactic depth did present a small resynchronisation in the theta band around 600ms after word onset though but it did not came out significant from the cluster-based statistic (see figure III.6 on page 64). Bastiaansen and Hagoort also pinned theta oscillations to memory retrieval. They observed a linear increase in theta power for each word added to the sentence. Although this is not captured by our syntactic depth feature, indicating that this role might not be dependant on syntactic complexity. But we recall the significant theta power modulation by word frequency in chapter II that could relate to memory retrieval.

In term of localisation, we do not have the data to correctly compute sources from those EEG responses. Notwithstanding the topographies we obtained for syntactic depth and the number of closing branches in the delta band could potentially fit the network of structures involved in syntactic structure building observed by Brennan et al., 2012. They also used naturalistic story to measure responses to languages coupled with fMRI imaging to localise the neural sources. They could associate word by word syntactic structure building with the posterior superior temporal gyrus bilaterally, which we argue to correspond with the response to closing branches, and the left anterior temporal lobe as a locus of neural correlate to syntactic complexity. The latter could be associated to our syntactic depth feature, which also present a left anterior scalp distribution of reconstruction accuracy (figure III.4).

III.4.3 Relation to ELAN and P600 ERP components

In the perspective of well studied ERP components related to syntactic processing, the present study shows several similarities. We observe an early component in both *depth* and *closure* TRFs. The latencies of that first part matches with the "early left anterior negativity" (ELAN) ERP component at around 200ms (figure III.5). This response is observed for morphosyntactic violation or for unexpected word-categories. Our topographies show a somehow less anterior activation. But the

TRF for closing nodes in the hierarchy is also left-lateralised although with the sign flipped when compared to the ELAN. One hypothesis to explain those differences with ERP components of syntactic processing is that the latter are stemming from violation of syntax whereas our TRF is measured using natural, grammatical speech, thus with no syntactic violations. As violations entrain a strong mismatch between top-down information and bottom up inputs. The same process is likely less important during comprehension of grammatically correct sentences.

The early latency is for processing closure could be the reflection of predictive mechanism where the neural population anticipate syntactic category that predict a larger merge operation. That is, as the sentence unfolds and longer and intricate phrase structure are met, the constraint put on the system to fuse word representation into grouped representation is increased too. That can be explained with our *closure* feature as it scales with the number of nodes, and hence words, closed by the given words. Our TRFs for both syntactic depth and closure are also more parietal or temporal, but given the difficulty of localizing neural sources with EEG we can not give a conclusive result on the exact source of our component.

Brennan and Hale, (2019) showed that they could extract ERP-like responses to context-free grammar (CFG) surprisal that differs from the N400 and share some similarity with the left anterior negativity (LAN). These results are more focused on the predictive aspect of word processing, given syntactic information, while we want to distinguish processes that are linked to syntactic representations. Their study gives a neat evidence towards a combined mechanism between word level prediction and hierarchical syntax computation although it does not directly address the possible mechanism happening at those particular "event" of the phrase structure hierarchy such as deeper structures and closing/opening of subtrees.

Molinaro, Barber, and Carreiras, (2011a) distinguishes a late and an early subcomponents to the P600. We prefer to interpret this as wave of information flow (Maess et al., 2006; Tse et al., 2007; Kutas and Federmeier, 2011) propagating from anterior towards posterior areas, potentially through the dorsal pathway as described in Friederici, (2011). This could be a modulation mechanism implemented in areas dedicated to sequences processing towards regions encoding lexical representations. More specifically, top-down information from Broca's area is projecting towards posterior superior temporal gyrus. The latencies and topographies of those early and late phases of P600 components match our result for *syntactic depth* and *closure*. We have an earlier peak in the depth TRF starting at 600ms with a central activation followed by late peaks beyond 750ms, for the *closure* only, in more parieto-occipital areas of the scalp. This striking similarity with the P600 suggest that we can decompose those early and late components into two distinct mechanism, one of which depending only on the depth, the other on the ability to relieve the load on processing input by integrating lexical information to larger syntactic structure. The former would be captured by our *syntactic depth* TRF, so the difficulty to maintain more nested structures in memory, and the latter by response to closing branches.

In contrast to semantic violations attributed to modulate the N400 component, syntactic anomalies have been linked with a left-anterior negativity (LAN, at about 400ms post-stimulus) as well as a centro-parietal positive component emerging around 600ms after word onset, called the P600 (Friederici and Männel, 2014; Kielar et al., 2015). However, both P600 and N400 have a similar centro-parietal topography although with a opposed polarity which suggests that overlapping populations of neurons are processing semantic and syntactic anomalies. Similarly, we note an analogous resemblance between the topography at peak latency for surprisal (figure II.4 on page 47) and the late scalp distribution of amplitude for the closing branch feature (figure III.5) suggesting here a relation with neuronal population involved in the N400/P600 wave. One main difference is that we capture the response from naturalistic stimuli with no violations of semantic or syntax.

III.5 Conclusion

Even though the brain is continuously predicting linguistic input. It has to build a compact representation of the complex hierarchical structures involved in language. The meaning of an entire word sequence is thus inferred from semantic *and* syntactic properties, be they predicted or not, and the network involved for such process engages temporal, parietal and frontal areas over a wide range of latencies (Fedorenko, Nieto-Castañón, and Kanwisher, 2012; Heer et al., 2017). EEG lacks of spatial resolution, and without electrode position digitisation and structural MRI, source localisation is not very precise. We have reported significant EEG responses to syntactic features as extracted from constituency-based trees as they occurred during natural speech listening. Those response function show spatio-temporal dynamics similar to ERP components related to syntactic processing. We do not claim that this responses are entirely reflecting hierarchical phrase structure tracking, but they certainly depend on several descriptive attribute of the syntactic hierarchy. More accurate imaging, possibly invasive recording at specific sites of the macro scale brain network engaged in such processing would be required to underpin the underlying encoding mechanism and exact neural dynamics that generate those hierarchical representations. Though we could verify that such linguistic features elicit a strong enough brain response that can be observed at the population level at least with the amount of data we had (40 min). Different analysis and experiment would be necessary to examine the robustness of this effect at the individual level. This could pave the way to new diagnostic tools to better assess what stage of the hierarchy of speech comprehension is affected in patients with language impairments.

Chapter IV

Characterizing the relationship between linguistic features and EEG

We saw significant EEG responses to linguistic features as extracted from speech. Cortical response to surprisal, precision and syntactic features could be reliably computed from the EEG recorded during presentation of continuous speech stimuli. Yet we have not analysed the strength of the relationship between those features and the EEG signal. If they are well represented in the EEG time series, one should be able to extract information about the language content of the stimulus. This approach is known as decoding, as we attempt to decode information represented in the cortex about the stimulus and environment from its recorded activity.

Listening to speech does not necessarily imply understanding it. When listening to foreign languages, it is natural to recognize that we hear speech sounds, we might even infer from the prosody speech structures like sentence boundaries or some of the phrasal structures. However we will not extract any meaningful content from it. This is one reason why we opted for mapping cortical responses to the linguistic content of speech stimuli rather than its spectro-acoustic features. By doing so, we are more directly targetting the neural processes involved in comprehension. Decoding the degree of comprehension from electrophysiological recordings is a challenging and interesting achievement. It provides us with new insights into what aspects of EEG are important for comprehension and also it could be ported to future applications such as brain-machine interfaces to verify that vegetative patients understood what they heard.

IV.1 Introduction

Speech comprehension involves many brain processes from lower perceptual acoustic processing to domain-general sequence learning, as well as language-specific lexical retrieval and semantic disambiguation. Each of those level of representations support each other in order to lead to comprehension. The lower levels feeding in the forward sense bottom up inputs, e.g. neural encoding of phonemic representations, to higher-levels, like ones processing words to extract individual word meanings from the sound chunks just processed; while higher level areas are modulating lower level representation and encoding via predictive information such that the syntactic category of a word can help to predictively pre-activate neural networks encoding a word which itself predicts phonemic representation (this was presented in greater details in section I.4. This information flow is present although disrupted when listening to foreign speech. The different techniques reviewed in this thesis as well as employed in the previous chapters help to disentangle some of this processing hierarchy.

IV.1.1 Aims

Mainly we want to decode comprehension using some of the higher level linguistic features and their electro-encephalographic encoding. Nevertheless we ought to go beyond the sole purpose of decoding stimuli as we also aim at deciphering mechanisms underlying the neurobiology of language processing. With this in mind, it is relevant to analyse both encoding and decoding models with different perspectives. Indeed, the computation of TRF in chapters II and III revealed the time-course and loci of neural signatures to the different speech features of interest. But to go further, we want to assess (1) the gain in reconstruction accuracy achieved by each predictor individually and their capacity to generalize to unseen data, (2) the strength of relationship between a set of features from the stimulus and the EEG when we pool frequency bands information, temporal and spatial filters, on both stimulus and EEG, and finally (3) how much those linguistic features can tell us about preceding (in cortical hierarchy) low-level acoustic tracking quality.

This characterisation of coupling between linguistic features and EEG paves the way to expand this work to online decoding of comprehension and of linguistic content. In previous chapters we focused on analysing TRF by contrasting them to a null model. Both TRF and null models were built using all subjects. This was a chosen design to emphasize any significant coefficients within each kernel that robustly stands out as different from zero above and beyond null model distribution of those same coefficient. However in this chapter we focus on the reconstruction power of linear models and their capacity to generalize *across* participants. This is more constraining than before as not only we want each predictor to get significant filter coefficients but also they need to reconstruct the EEG better than any other linguistic feature for unseen data despite the high intra-subject variance. This highlights our interest in decoding the stimulus rather than describing the EEG response to those linguistic features.

The goal is to get a more complete picture of the interaction between word-level linguistic features and EEG data. In addition to the afore mentioned techniques, we further characterize the role of linguistic features by looking at the dependency of cortical tracking time-course on those word-level prediction statistics. We seek at demonstrating how the reconstruction of low-level stimulus features such as the acoustic envelope is linked to the presence of highly unpredictable words or not. This shows how low level tracking has a top-down component and might be driven by predictions as well as by spectro-temporal features of speech.

IV.1.2 General Methods

In the analysis of the present chapter, unless mentioned otherwise, we used the same pre-processing pipeline as described in chapter II and Weissbart, Kandylaki, and Reichenbach, (2019a).

Several methods have been employed to measure the degree of relationship between speech linguistic features and EEG. First we measured the correlation between the predicted EEG using the forward encoding models computed through TRF modelling from chapter II. Then we used a more efficient algorithm in modelling linear relationships between two dataset, namely Canonical Components Analysis (CCA) (though better at establishing correlated subspace between stimuli and response, this method does not allow to accurately study the time-course of cortical responses to the different features individually). Finally we also developed a classification task in which we assess the respective contribution of the cortical tracking of each individual linguistic features in classifying listening condition (Dutch vs English).

To summarize, we use three distinct methods to measure the quality of tracking between a set of linguistic features and EEG data:

- TRF (forward models) reconstruction accuracy, following models and methods from the two previous chapters, to assess the power of generalisation to unseen participant data for each feature, and hence the quality of robustness of representation of the feature in EEG
- CCA correlations across canonical component pairs to measure the strength of representation of linguistic features and how this differs across listening condition
- Classification of comprehension condition (Dutch vs English) to test whether feature subset can reveal neural basis for comprehension, that is to ask: *are we able to know if the language heard is understood using the quality of reconstruction from our encoding models?* For this one first needs to verify that

English models give different reconstruction accuracy (better?) than the one computed from the Dutch condition.

Our word-level features are difficult to reconstruct from linear modelling as they are not smooth in time and the corresponding brain responses are generated from high-level processes, and hence less robustly measurable as for lower-level features such as the acoustic envelope. To reconstruct such non-linear features (time series of "spikes" at word onsets) one would more naturally use a non-linear model, which be harder t fit on noisy EEG data. Moreover, using linear regression models would not allow us to jointly consider each linguistic features when they are used as targets. A multiple regression will compute coefficients mapping EEG to each individual features without leveraging their possible interactions. For these reasons we argue that a backward model is less adequate than a forward model to carry on with such analysis. Finally, by using a backward model one needs to regularize strongly as the autocorrelation naturally occurring in EEG data renders the inversion of its covariance matrix unstable. However it is possible that such regularization (necessary to avoid overfitting) will also hinders performance as we privilege principal components with higher variance and those could be stimulus-irrelevant. We indeed found very little performance while trying to reconstruct our stimuli with backward models. Using cross-validation, we trained a backward model for each subject to reconstruct either the true word-level representations or a shuffle representation (null model). We did not observed significant reconstruction of word-level features using backward models (not greater than with a null model: paired t-test: t-statistics $= -0.9$, degree of freedom $= 12$, p-value $= 0.38$). Although this might be investigated in the future with different methodology.

IV.2 Generalisation of reconstruction accuracy for each feature

In both chapters II and III we aimed at underpinning the dynamics of cortical responses to specific linguistic features, namely surprisal, precision and syntactic trees descriptives.

A first analysis consist in assessing the degree at which word-level linguistic features are represented in EEG. We expect a stronger relation with models that carry a richer representation of speech (such as having both precision and surprisal as characteristic features of prediction in language). However this will not necessarily hold true for every locations of the brain and hence for every sensors. We must look at the topography of accuracies to fully apprehend the beneficial impact of adding predictors into a model. Indeed, for un-regularized models, we should observe no gain in accuracy or overfitting at electrodes that do not measure brain sources responsible for the encoding of the linguistic predictors used in the models.

IV.2.1 Methods

To evaluate the accuracy of our encoding models, we scored the model at each electrode location by taking the correlation between predicted signal and the true underlying sensor signal.

Subject-out cross-validation: The evaluation was done through a leave-one-subject-out cross-validation procedure. Each subject is in turn left out, as a left-out dataset on which the model will be evaluated. The model is trained on all the other remaining subjects. The training and evaluation is repeated for each subject, such that we end up with a correlation score ρ for each channel and subject. The procedure is illustrated in figure IV.1. This generalizes the model predictions across participant and hence across experimental conditions too. However in that case, we expect to be overfitting at some electrode locations where the predictors used can only generalise to a certain extent across participants.

This was repeated for every set of possible features, and all frequency band of interest, to assess the respective gain of adding each predictor to the model. Namely we evaluated a model for all possible combination of the following linguistic features (with the word onset feature always present): Word Frequency, Surprisal, and Precision.

Within Subject Cross-validation (for classification) The evaluation was done through a 4-folds cross-validation within subjects. For each subject, the classifier was fitted on three folds while the fourth one was kept to compute an accuracy score. Thus we end up with a classification accuracy for each (left-out) subject and fold.

Statistical Analysis

In order to evaluate the significance of reconstruction scores, we ran paired t-tests at each electrode location between samples of correlation obtained from the model with word onset only and the models with added linguistic features. The significance threshold was set to $\alpha = 0.05/64$ to control fo multiple comparison at each of the 64 electrode sites with Bonferroni correction. This kind of correction is here very conservative and puts the result on the safe side. Another way to correct for family-wise error rate, which are likely given the structure of EEG signals (sensors pick up correlated signals) would be to use a cluster-based approach. This was done in chapters II and III. Nevertheless here we use this first t-test to have a quick and robust way to evaluate topographies, and we test for location and model difference with ANOVA, as explained below. Figures IV.2 and IV.4 show topography of the difference between correlation scores of the model with word onset only and with prediction features added and non significant areas masked by transparency.

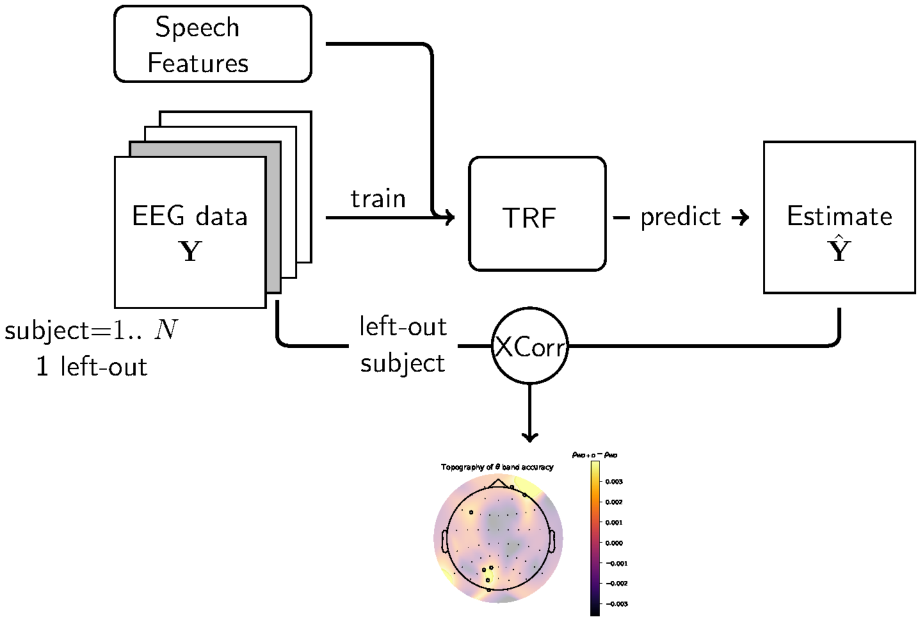


Figure IV.1: Leave-one subject-out evaluation procedure to generate reconstruction accuracies.

Aggregated values in four regions of interest were then computed by taking the electrode average scores in those regions. We employed 2-way ANOVA to determine the main effect of features and aggregated locations, or region of interests, and ran post-hoc pairwise t-tests correcting for multiple comparison to assess how much better than a model with word onset only a particular region and feature was.

IV.2.2 Results

Figure IV.2 presents improvements in reconstruction accuracy in the δ band respectively for surprisal, word frequency and precision with respect to a model with only word onset.

When adding either surprisal or word frequency, there is a significant improvement in the four regions of interests. We can see from the topography of difference in reconstruction accuracy that the surprisal feature gives a left-lateralised improvement while word frequency increases the score mostly in parietal and right anterior areas. The increase at the left anterior electrode sites for surprisal is above and beyond the one observed for word frequency. Adding precision, however, does not benefits the encoding of EEG signal with respect to word onset model in the delta band. The full model, containing all features, shows better reconstruction in at each region of interest as well as on average at all electrode sites compared to any other model.

In figure IV.3, we show a similar analysis but using features developed in chap-

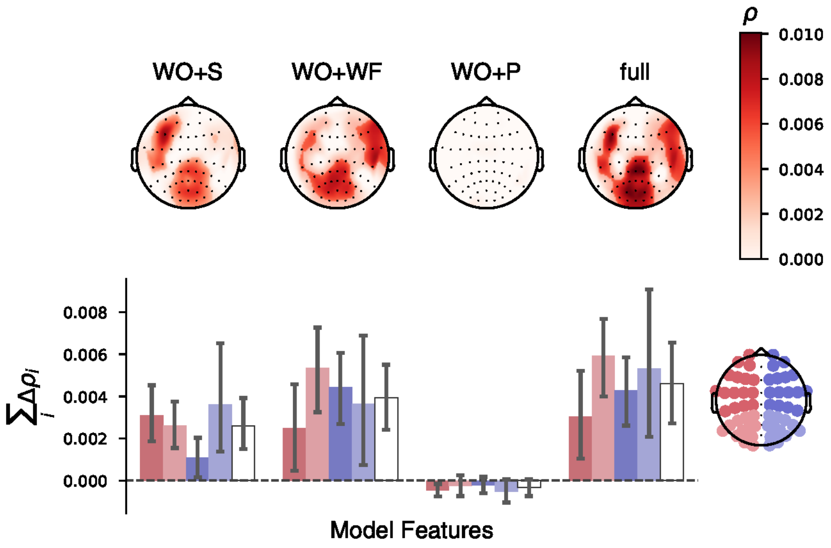


Figure IV.2: Difference in the delta band of reconstruction accuracy with respect to a model with word onset only.
Top panels: Topographies of those differences. Interpolated values that fall below the 0.05 significance level (corrected for multiple comparison using Bonferroni correction) are shown with transparency.
Bottom panel: a bar graph representing aggregated scores, the sum of differences in reconstruction accuracy $\Sigma_{i \in sensors} \Delta\rho_i$ over electrodes of a specific subset, at scalp location of interest (as shown by different colours in the bottom right corner: in red the left hemisphere, in blue the right hemisphere, more saturated colours for anterior electrodes). Error bars show 95% margin errors computed by bootstrapping. The full model refers to a model containing: Word Onset (WO), Word Frequency (WF), Surprisal (S), Precision (P) and Surprisal weighted by Precision.

ter III. We observe the strongest increase in accuracy for the response to closing nodes, with a left-lateralised parieto-temporal scalp distribution. The full model benefits strongly from each feature as we measure better accuracy in every regions of interest. The topohgraphy obtained for the full model is similar than the one for the model with closing branches with an additional right anterior area being significantly greater than the word onset model.

Figure IV.4 shows similar information as in figure IV.2 but in the theta, alpha, beta and gamma bands. Importantly, here we reconstruct the power activity rather than the actual time-course within those frequency band. Hence both phase-locked and induced activity are contributing to the results here. Except for panel D, figure IV.4 shows only topographies and bar plot of feature where significant increase where observed. The word frequency feature contributed to predicting EEG signals in the theta and alpha band. We note that it was the only feature to have a significant out-of-subject improvement in alpha band power. Precision, conversely to the results obtained in the delta band (figure IV.2), is now significantly giving better reconstruction accuracies in the left posterior and right posterior electrode sites for beta and gamma band power respectively. Reconstruction of EEG beta

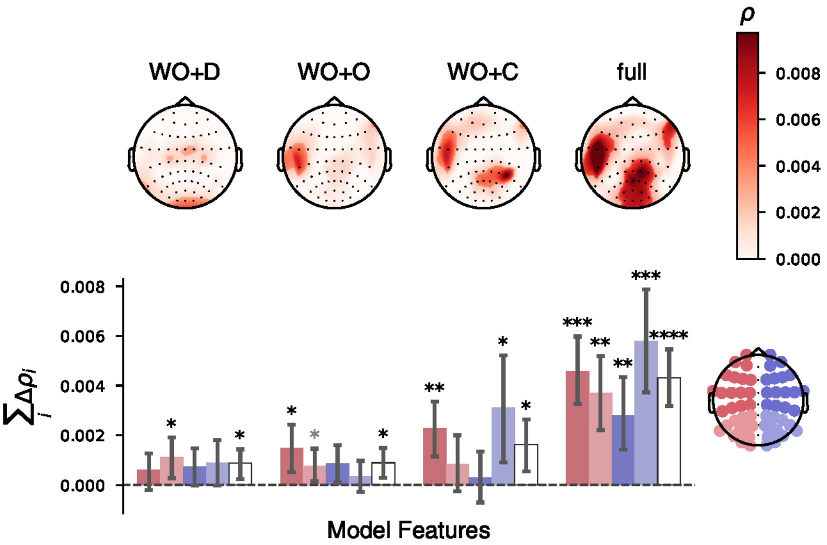


Figure IV.3: Difference in the delta band of reconstruction accuracy with respect to a model with word onset only.
Top panels: Topographies of those differences. Interpolated values that fall below the 0.05 significance level (corrected for multiple comparison using Bonferroni correction) are shown with transparency.
Bottom panel: a bar graph representing aggregated scores, the sum of differences in reconstruction accuracy $\Sigma_{i \in sensors} \Delta\rho_i$ over electrodes of a specific subset, at scalp location of interest (as shown by different colours in the bottom right corner: in red the left hemisphere, in blue the right hemisphere, more saturated colours for anterior electrodes). Error bars show 95% margin errors computed by bootstrapping from samples across subjects. The full model refers to a model containing: Word Onset, syntactic depth (D), the number of opening (O) and closing (C) sub-trees.

power using surprisal were significantly better than using only word onset for some electrodes. Interestingly, the corresponding topography is actually similar to the significant electrodes in the theta-band for word frequency (bilateral, with a slight left hemisphere preference and mostly above temporal lobes).

In contrast to those results that relied on linguistic features from chapter II, syntactic features described in chapter III did not allow for much better EEG reconstruction at least in our out-of subject evaluation procedure. So it could be that the generalisation is the issue or that those features are less predictive of high frequency power of EEG. We only observed significant improvement over a model with only word onset when adding the number of closing nodes to predict theta power of EEG signals, as shown in figure IV.5.

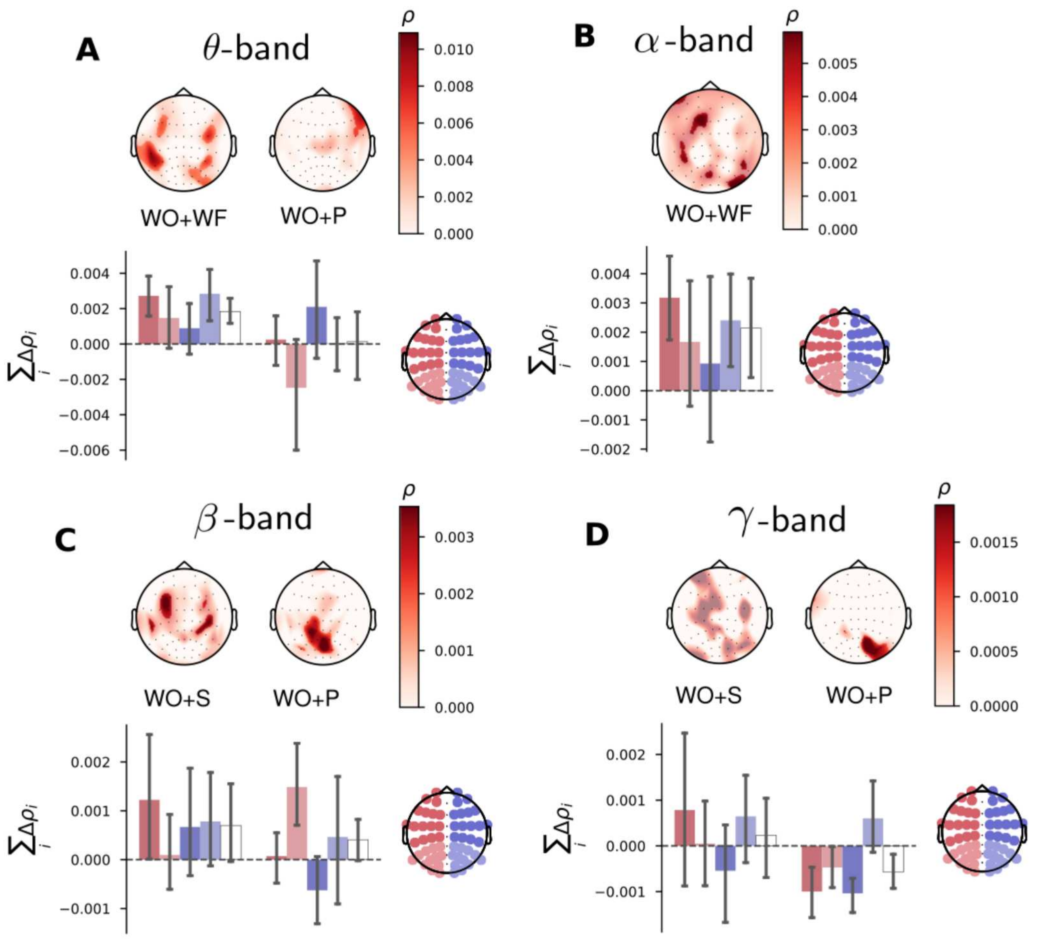


Figure IV.4: Reconstruction accuracy evaluated on left out subject. Panel **A** presents the benefits of adding word frequency and precision w.r.t to word onset to predict θ band power activity. In panel **B**, word frequency model in α band. Finally panels **C** and **D** shows surprisal and precision results in the β and γ bands respectively. Average score at electrode sites are shown in corresponding bar graphs as in figure IV.2, as well as transparency for non significant increase on topographies.

IV.3 Quantifying the strength of relationship

To assess the strength of the relationship between the stimulus representation used and the EEG activity, we employed a hybrid decoding/encoding approach using Canonical Components Analysis (CCA). This linear technique circumvents disadvantages from forward and backward modelling by forming a hybrid model. The coefficients can be interpreted as TRF and spatio-temporal filters, though here we chose to pass EEG data and stimulus through a filterbank to maximise the reconstruction accuracy. Using a filterbank allows to process data simultaneously from different time-scales and enhances the representation of data before applying the

CCA by increasing the SNR in the bands of interest. However, the coefficients become less directly interpretable. With this method, the goal is not to underpin the time-course of EEG response to each feature but rather to measure the degree of coupling existing between a specific feature set and the EEG signals, therefore we opted to maximise this by supplying to the CCA algorithm a richer representation of both stimulus and response.

IV.3.1 Method

The CCA method operates on both datasets simultaneously by projecting the speech feature representation **X** and the EEG data matrix **Y** into a new vector space in which the two transformed matrices have each pair of columns maximally correlated. In other words, CCA combines the forward and backward models into a hybrid encoding-decoding model where stimulus and response are simultaneously temporally filtered across feature dimensions as well as linearly combined (which can be interpreted as spatial filtering for the EEG side). That is, the neural activity is decoded and correlated with an encoded version of the stimulus (Dmochowski et al., 2018; Cheveigné et al., 2018a). Such a formulation allows to remove any variance that is not related to speech representation in any linear way.

$$\hat{u}(t) = (s \star \beta)(t) \tag{IV.1}$$

$$\hat{v}(t) = \sum_i r_i(t) \star w_i(t) \tag{IV.2}$$

The parameters $\beta(t)$ and w are then found by maximising the correlation between the encoded stimulus $\hat{u(t)}$ and the decoded response $\hat{v(t)}$ (equations (IV.1) and (IV.2)). This correlation can be reported as a score to quantify the quality of fit of the model (Cheveigné et al., 2018a):

$$\rho(\hat{u}, \hat{v}) = \frac{\langle \hat{u}, \hat{v} \rangle}{\|\hat{u}\| \|\hat{v}\|} \tag{IV.3}$$

The global coupling strength was measured by taking the sum of correlations between the first 25 pairs of canonical components for each language. We also show the score obtained from the first canonical component only, which is the one capturing most of the variance in both dataset (stimulus and response). In order to highlight the difference between English and Dutch, we normalised the score for each language condition by the score obtained with a word onset only model and looked at the difference of this relative score (relative to word onset model). The two bottom panels of figure IV.6 show this difference in relative scores for both the first canonical component and the sum across the 25 first pairs.

Statistical significance was assessed first with ANOVA to evaluate how each model differed with feature sets or with condition. The following design was being tested: condition×feature set, with language as *condition.* A post-hoc analysis (used in figure IV.6) was implemented with a paired t-test between score samples per language to establish significant influence of each feature set on score values. On the other hand we used an independent t-test, correcting for multiple comparison with Bonferroni ($\alpha = 0.05/4$, for 4 feature sets) to assess the difference between language conditions. The Bonferroni test was chosen because it is preferable here to avoid type I errors. These results are new and we want to stay on the most conservative side. Moreover, the number of comparisons is not so large (four) so the method is not too constraining in this case.

IV.3.2 Results

There is an increased correlation between speech representations and EEG with added linguistic features. Namely, there is a stronger coupling when the speech representations encompasses more complexed information on the stimulus such as *precision* of predictions and the *interaction* between prediction errors and precision. The strongest increment in score is observed when adding the prior probability of words, indexed by word frequency, to the model. The statistical results from a 2-way ANOVA with language condition and feature sets as grouping variable is presented in table IV.1.

Source	SS	DF	MS	F-statistic	p-value	η_p^2
Feature set	12.461	4	3.115	77.835	**3.477e-31**	**0.7389**
Language	0.842	1	0.842	21.037	**1.199e-05**	0.1605
Feature set * Language	0.077	4	0.019	0.482	7.492e-01	0.0172
Residual	4.403	110	0.040			

Table IV.1: 2-way ANOVA on the CCA scores (sum of correlations across all canonical components), with "feature set" and "language" (English or Dutch) as groups. We see a main effect of features and language condition. Interaction was not significant. The effect size computed in η_p^3 (partial η^2) shows a stronger effect size for the feature sets.

The most discriminative models to contrast comprehension state (English vs Dutch) are the models that incorporate probabilistic quantities computed over *sequences* of words instead of word frequencies alone. Indeed, we see in the bottom right panel of figure IV.6 on the next page (positive values on bottom panels of the figure favour English models) that a model with word-frequency alone (in addition to word onsets) do not have a better global score in English than in Dutch relative to the baseline model (word onset). Though the model taking in account surprisal, precision and precision weighted surprisal have incrementally stronger scores for the English condition only.

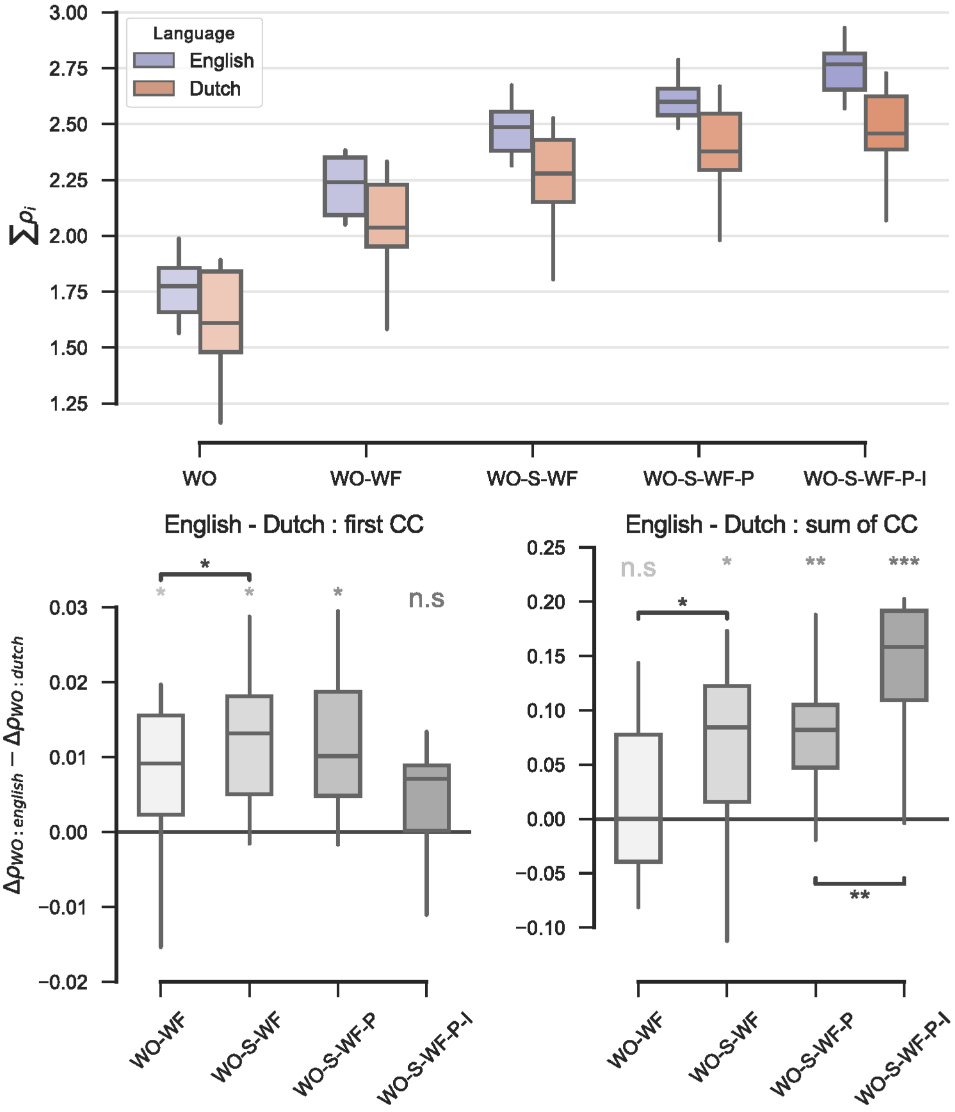


Figure IV.6: CCA correlations and scores

Yet, this is only observable when aggregating correlation values for several canonical pairs (here we used the first 25). For the first canonical component pair (bottom left panel of figure IV.6) only adding surprisal benefits English score compared to the model with only word onset and unigram probabilities. Moreover, the first canonical component do not contain discriminative (for comprehension) information on the precision weighted surprisal, as the latter is not significantly better for English than for Dutch.

IV.4 Decoding Comprehension: English vs Dutch

Some of those linguistic features can only be reliably encoded if the speech stimuli is being comprehensively processed. That is if there is an understanding of the language. For instance, the computation of uncertainty about prediction of the next word based on its preceding context can only be done if the statistics of word sequence of the given language is known, at the word level. Furthermore, grouping into syntactic phrase structure can only occur during comprehension, at least for complex sentences. With both those observations at hand, we know that only speech that is being understood will elicit reliable responses to the high-level representations such as word-level surprisal and the number of closing nodes. By shifting our perspective on the data, this means it is possible to measure the degree of comprehension a listener have given a speech stimulus by computing the strength at which those linguistic features are represented in their EEG signals. We want to verify this hypothesis by computing our models with EEG recorded while participants listen to foreign, incomprehensible speech. If there is a gain in using those high level linguistic features to encode EEG signals, one should be able to decode the English condition apart from the foreign language condition solely based on the accuracy at which the model reconstruct the signals.

After comparing the accuracy of encoding models, we will attempt to directly decode the listening condition from the EEG data. We chose Dutch as the foreign language condition. Participants were all native English speakers, with no previous knowledge of Dutch. However, the task is made relatively hard for a decoder system as both language are western Germanic languages with common syllabic structures and phonotactic constraints (Collins and Mees, 1999, p.16). Hence the decoder must rely on high-level linguistic responses that are specific to the comprehension of speech through extraction of meaning.

However most of the brain activity is captured by the response to word onset only, nearly 90% of the variance captured at all electrodes by a model using all predictors is already accounted for with just that simple onset feature in the English condition. For instance, in figure II.3 on page 45 we can see the relative increase in reconstruction accuracy of a model with word frequency and word onset with respect to the word onset only model, and it achieves in average a 10% improvement in accuracy. The latter is being measured by correlation coefficients so we loosely link this up to the amount of variance captured as all models are linear. We expect the Dutch *word onset* response to be also large, although different. Indeed, one does not need to understand speech to be able to extract word boundaries. A clear evidence for this comes from language development studies, where we can observe in babies and toddlers evidence of correct chunking of sounds into syllables and words (Saffran, Aslin, and Newport, 1996; Saffran, 2003; Mattys et al., 1999). Moreover both English and Dutch are Germanic languages, there are compelling overlap between their grammar and phoneme sets. Hence we expect to see a strong response to word onset in Dutch, but which somehow must be different to the English one as words are usually not recognizable in Dutch. Those dissimilarities can stem from divergent brain processes regarding lexical access and/or binding to other meaningful units into larger phrasal structures. We will build on top of a simple word onset model to disentangle the effect of higher level linguistic features when decoding comprehension. Altogether, we can already observe at the word response potentials a difference between conditions that would allow significant classification. This can help to investigate the time-course of comprehension relative to word onset from ERPs to words. Then we will move on more specific decoder, adding our high-level linguistic predictors to see how much one gains from those.

In order to characterise at the same time the feature of interest as well as the time course of relevant discriminatory activity we ran thus several analysis:

1. Firstly, we established an ERP-like analysis, epoching data around word onsets and classifying comprehension state at each lags
2. Second, we want to see how much a backward model on the varying linguistic features can, given our experimental conditions, provide us with a classifier on comprehension state.
3. Finally, in order to correctly incorporate the multivariate representation of speech into a decoder, we will first use an encoding step and classify comprehension based on the accuracy of the encoder.

In the following section we will often be referring to accuracy score. However we have several stage in our decoding procedure which creates some overlap in the terminology. Indeed, we are building a decoder, or classifier, based on the performance of the encoding model. That is, we directly feed the *reconstruction accuracies* of EEG from short speech segments, as obtained from a forward model, to our classifier and this result to a *classification accuracy*. To avoid further confusion, we define here for the sake of clarity the terms *reconstruction accuracy* and *classification accuracy*:

Definition 2 ***Reconstruction accuracy*** *Correlation value between true EEG signal and predicted EEG from TRF model. Its range is [-1,1].*

Definition 3 ***Classification accuracy*** *or decoding accuracy: Fraction of correctly classified condition (here English or Dutch) measured on held out segments of data. It ranges from 0 to 1.*

IV.4.1 Methods

Experimental procedure

The same set of subjects that participated in the first study, detailed in chapter II and Weissbart, Kandylaki, and Reichenbach, (2019a), underwent a second

EEG session. In this mew experiment, the task was to pay attention to a speech stimulus in a foreign language. We chose Dutch as it shares many acoustic properties with English. Hence we expect the difference in brain responses to highlight deviation in comprehension rather than in spectro-temporal processing of varying acoustic inputs. The design was very similar as the first session, participants listened to story parts extracted from an audiobook (excerpt from Arthur Conan Doyle's Sherlock Holmes, read in Dutch[1]), and had to answer comprehension question in between story parts presented on screen. In addition to the comprehension questions we added a self-report questionnaire assessing their level of attention. Hence between each story parts, for any condition (English or Dutch) they had to rate their level of attention.

Time-course of comprehension: decoding per-lag

First we want to determine the time of interest as well as the minimal decoding accuracy achievable in a simple framework. Namely, the first analysis consisted in using only EEG activity at different electrodes at a single time lag relative to word onset in order to decode comprehension state.

After preprocessing (similar as in section IV.2), EEG data were epoched around word onsets of *content words* only (N=2904), using a window spanning from -200ms to 1000ms after word onset. For each subject, we simply classified comprehension (English or Dutch condition) based on the electrodes signal amplitude at each lag so to localise and characterise the time-course of comprehension-specific activity, solely based on filtered EEG signals.

To test whether the interaction between EEG potential values and surprisal helps to decode comprehension, we developed a second decoder which were using a set of 65 features at each lag: the 64 electrode potentials plus the surprisal value of the preceding word. This allows to incorporate information on how scalp potentials and surprisal co-varying might help to decode the listening condition. The hypothesis being that if it does then not only EEG sensors signal are modulated by surprisal (as we saw already in previous chapters) but also this modulation improve our ability to decode comprehension from EEG.

A significance level was computed from the margin of error obtained above chance level using the series of Bernoulli trials, following a binomial distribution. After fixing a critical threshold α and given the total number of trials N and the fraction of English trials p, we can compute the significance level as:

[1] Audio accessible at https://www.librivox.org and text at: https://www.gutenberg.org/ebooks/30933

$$p_{\text{significant}} = \max(p, 1-p) + z_{\alpha/2} \cdot \text{sem} \tag{IV.4}$$

where:

$$z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$$

$$\text{sem} = \sqrt{\frac{p(1-p)}{N}}$$

Where Φ is the cumulative distribution function of the standard normal distribution $\mathcal{N}(0, 1)$. When referring to a *significance level* in the following text and in the context of binary classification, we will refer to equation (IV.4), which denotes the upper bound level of an interval with confidence level of $100 \times (1 - \alpha)\%$ around the chance level of the decoding task. An clear situation where this computation becomes obvious is to think of a statistical test where the alternative hypothesis would be that a coin is loaded (null hypothesis it is not). Observing 80% of heads on 5 flips only gives a significant level of 86.8%, the chance level is 50% but besides our observations there is still more than 5% chance that the outcome is observed by chance.

We corrected for multiple comparison (across lags) using a Bonferroni correction directly on the significance level. This is very conservative in this case and we are aware that it might be prone to a number of type II errors. However we are not interested in finding exactly the significant latencies and are merely pointing out at the fact that there are significant latencies at all. Thus we opted for such a conservative correction here.

Classification from forward models scores

We follow a similar procedure as in section IV.2, where trained TRF are evaluated to predict EEG data from a left-out subject, as presented in figure IV.1 on page 75. This allows to avoid overfitting as well as to test how much we can generalise across participants. Thus we are assessing the efficiency of such decoding paradigm in revealing patterns common in a population but specific to comprehension of speech from the response to linguistic features.

The reconstruction accuracy were evaluated on left out subjects for both Dutch and English condition. We segmented the data in chunks of 10 seconds, and computed the correlation between predicted and true EEG for all chunks, so to obtain one correlation value per electrode, language condition and chunk. Then a support vector machine classifier from scikit-learn python library with radial basis function was trained on the concatenation of scores for all electrodes. This gives us 128 features in total, 64 scores from the English reconstruction scores, one per electrode, and analogously 64 reconstruction features for the Dutch condition. The classification accuracy from the support vector machine classifier was obtained from a nested 4-folds cross-validation (nested inside the leave-one subject-out cross-validation procedure done to obtain TRF scores, as explained by diagram of figure IV.1).

Significance level of decoding were directly estimated using equation (IV.4). The number of trials is determined from the segment size (5 sec segment length gave 480 Dutch segments and 528 English ones), and the probability p is given by the ratio between the number of English trials and the total number of segments. This gave a significant level at roughly 53% of classification accuracy.

We repeated this procedure with different subset of linguistic or acoustic features ranging from acoustic envelope, to the features of chapter II and syntactic features but also to combinations of those. Since we are evaluating the reconstruction on left out subject the accuracies of reconstruction suffers form adding too many predictors in some cases, which can be interpreted as a sign of the difficulty to generalise across subjects. On the other hand when classification accuracy reach significant levels we can postulate a robust classification of comprehension at the individual level.

IV.4.2 Results

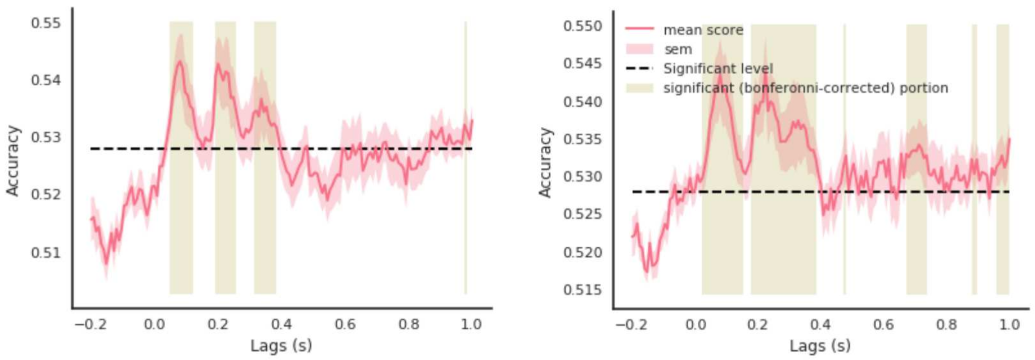
Time-course of comprehension: decoding per-lag

Figure IV.7 shows the time course of decoding accuracy. From the middle panel, we see that we can discriminate comprehension condition as early as 50ms after word onset. This early response suggests an early differentiation in the processing of words between the two listening condition. At peaks, the accuracy reached a maximum value of 0.54, just above significance level. However, when adding the surprisal as a feature to electrode potentials, the decoding accuracy increased by up to 1% at previously significant lags. Interestingly, a new time region became significant. It was possible to decode comprehension above chance from those ERP topographies around 750ms after word onset when scalp potential were augmented with surprisal value. This indicates that the covariance structure between EEG and surprisal benefits to our discriminatory task.

From encoding models

Firstly, we observed that the global reconstruction accuracy (average over all electrodes) is better in English (p-value=0.003, paired t-test, n=12), this is presented along with individual subject scores in figure IV.8. We conclude that a better representation of linguistic speech features occurs during comprehension. Importantly this indicates that we can potentially use the reconstruction accuracy to decode the listening condition.

Secondly, we looked at the topography of the reconstruction accuracies in each condition. Figure IV.9 contains the topography of reconstruction for both English and Dutch as well as of the difference. At a first glance, we highlight the strong



(a) Decoding from EEG sensors

(b) Decoding from EEG sensors and surprisal value jointly

Figure IV.7: Single lag decoding. Right panel shows time course of accuracy of the decoder when we incorporate surprisal value alongside scalp electrodes amplitudes for each lag to decode between English and Dutch condition. The significance level corresponds to the significant chance level in the binary decoding task at alpha = 0.05 (see equation (IV.4)).

similarity in the topographies. This is supposedly showing that the dominant part of the EEG are acoustically entrained, possibly generated from the response to word onset (without meaning or contextual information). In other words, we can reconstruct most of the EEG signals using the TRF to word onset only. The difference gained in adding other, more complex, linguistic features is of higher order and of lower variance. Their effect is more subtle in contrast to the bare response to acoustic onset of words. The rightmost topography in figure IV.9 shows the difference between both conditions, with significant sensors marked with a white circle. As in section IV.2 on page 73, we find an anterior and parietal network with a slight left-dominance for the anterior electrodes. We expect a model trained to decode comprehension to weight the score of each electrode differently depending on their beneficial contributions.

The decoder built from our TRF reached much higher accuracy scores than the single lag decoding even using only word onset feature (figure IV.10 on page 90). As discussed in the introduction of section IV.4, this is not surprising as we expect that words are processed differently when they are being understood. Therefore their respective EEG response is likely to differ. The goal is not to get the best accuracy in such decoding but rather to question whether richer representation of speech, which include higher-level linguistic features, will benefit the decoder. In the results presented here the decoder was applied on segments of five seconds. If we were to use longer segments, the decoding accuracy would be higher although this would be of little use in a real-time application for instance.

Figure IV.10 presents concisely the benefit of adding either syntactic or prediction based linguistic features to the model in different frequency band. We observe that in the β band, decoding is significantly better when using surprisal to represent speech, whereas in the γ band, only syntactic features representation of speech al-

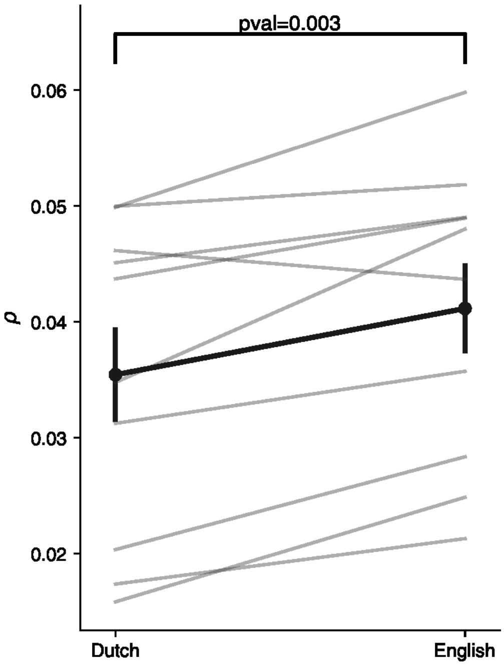


Figure IV.8: English and Dutch reconstruction accuracies measured with correlation. Grand average in black thick line, and individual subjects data are shown in light grey. We observe a significant difference between the two conditions.

lowed decoding beyond significance. We note from these results that a model using only word onset can be better than any other model. This happens only in the θ band. Finally, the highest scores are reported in the δ band where also prediction based linguistic features (surprisal, precision and their interaction) lead to a better decoding than model based solely on syntactic structure representation.

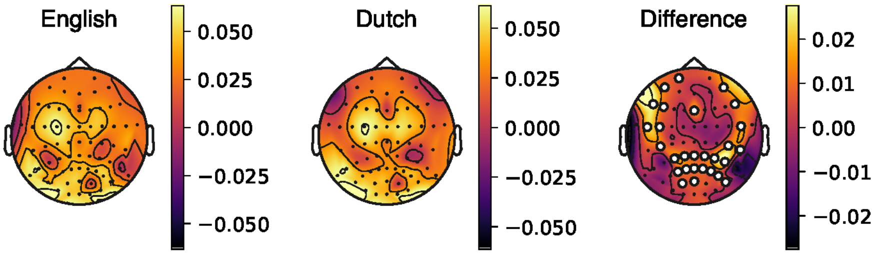


Figure IV.9: Topographies of reconstruction accuracy ρ: for English in the left panel, Dutch in the middle panel, and for the difference between both conditions in the right panel. We can see that reconstruction accuracies is quite similar for both condition although stronger in the English condition. However some electrodes still give a significantly stronger reconstruction accuracy in English as compared to Dutch, as seen in the right panel. Significant electrodes, those with higher reconstruction accuracy across all participants in English, are marked in white (t-test, $\alpha = 0.05$, corrected for multiple comparison with Bonferroni, $\alpha/64$).

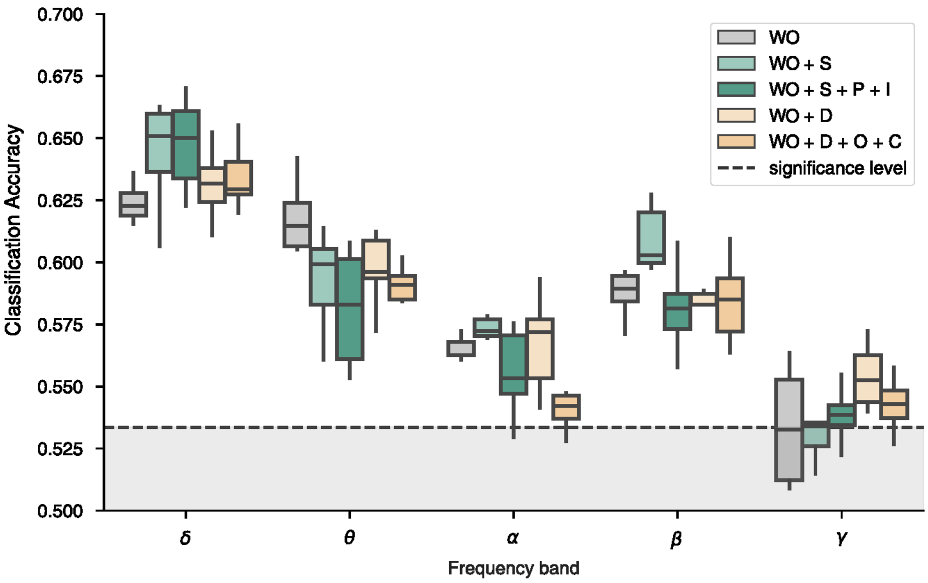


Figure IV.10: Decoding English vs Dutch. Comparing different set of features containing either prediction based statistical quantities, such as surprisal (S) and surprisal, precision and precision weighted surprisal (S, P, I), shown in shades of green, or syntactic representations such as depth (D), and the number of opening and closing branches (O, and C), in shades of yellow, for different frequency bands. Word onset feature is abbreviated WO.

IV.5 Envelope Decoding accuracy Modulated by Word Sequence Predictability features

Quality of cortical tracking can be quantified by the correlation between predicted acoustic envelope from a backward model and the actual sound envelope. We observed that this *entrainment* was similar across participants. The correlation taken on non-overlapping sliding windows were significantly high for all subjects at specific times with respect to stimuli. We suggest that this might be due to top-down linguistic modulation. Namely, part of the stimulus with words that have a high uncertainty and are not easily predictable from context will cause bottom up information to be processed with higher precision or importance. This is in line with a Bayesian perspective on sensory integration. If predictions of linguistic features (say lexical information) are unreliable, one must listen more carefully to the sound, and modulate the gain, or weight, with which acoustic information is integrated with higher level predictions.

The results below demonstrate that low-level representation of speech is modulated by probabilistic measures of predictability of word level input.

IV.5.1 Methods

In order to evaluate the influence of high level linguistic features on low level processing such as the cortical tracking of acoustic envelope we trained a backward model. The latter predicts the envelope from the EEG signals. The model was evaluated on held out data of the same subject. This cross-validation procedure was repeated for each subject to obtain a prediction of the stimulus envelope from each subject EEG activity. The backward model needs more regularisation than our forward models used so far since the lagged matrix of the EEG data is not well conditioned (its condition number, the ratio between the highest and lowest singular values, is low, indicating rank deficiency and strong correlations between columns that originate from spatial correlation as well as autocorrelation at each sensor signal) . We used another cross-validation loop on the training portion of the data to estimate the best regularization parameter via a validation set. The EEG data were processed following methods from chapter II.

Both the true and predicted envelope were then segmented into windows of five seconds with 500ms overlap. We computed the correlation score between each window pairs to get a time course of the Cortical Tracking Index (CTI). The Cortical Tracking Index (CTI) thus represents the quality of amplitude envelope encoding in the EEG. When the CTI is high, the prediction of sound envelope from the EEG data is more accurate. This allows us to obtain a measure, through time (although at a lower sampling rate as we compute the CTI on 5s window segments with little overlap), of the quality of encoding of sound envelope amplitude. The hypothesis being that low-level representation of sound is modulated by high-level features of language. In other words CTI, under this hypothesis, should be modulated by linguistic features. We hereafter will refer to the correlation between a segment of predicted envelope from EEG and the true envelope as the Cortical Tracking Index (CTI).

Finally, to test the hypothesis mentioned, a linear model was fitted between CTI across participants and linguistic features: surprisal, word frequency, precision and interaction.

Statistics

We tested the relationship between correlation scores and linguistic features by fitting a multivariate linear model, using Ordinary Least Square (OLS) regression, between the scores and the linguistic predictors. The significance of the reconstruction were established using F-statistic for the overall model being better than a model where with no dependent variable. We used the methods implemented in `statsmodels` python library (Seabold and Perktold, 2010) to extract those statistics.

Significance were shown for the model with all variables, a p-value associated with each predictors were indicative of which coefficient were significantly different from zero. We then realized a cross-validated training, to get a sample of independent coefficients (trained on non overlapping folds form the dataset) which gave us a second statistical test assuring strict non null value of the coefficient. The margin of error was computed with a Bonferroni correction to get a 95% confidence interval that takes in account the multiple comparison in the most conservative way.

IV.5.2 Results

We found a negative correlation between CTI and the word-level precision of prediction only. However, using the window segments does not provide us with a reliable estimate of linguistic features as several words are appearing within one 5sec segment. The results from table IV.2 present statistics from OLS computed on averaged linguistic feature values for each window segment. It is more likely than one word only leads the increase in CTI so we also tried with the maximum value within each window and got similar results. However, there is a strong bimodal distribution in values of precision as content and function words are clearly in contrast. Aggregating precision entropy values of words into their mean or maximum might hinder the change in values occurring between those two different word categories (function and content word). We need to separate out those two categories, and to do so we cannot simply use one value per five second segment.

Therefore we ran a new analysis, where the correlation score was obtained from windows centred around word onset (from -0.1 to 0.6 seconds around word onset). Thus we can take individual word score and represent the accuracy of cortical track-

No. Observations:	1587	Log-Likelihood:	1599.6
Df Model:	4	F-statistic:	4.250
Df Residuals:	1582	Prob (F-statistic):	**0.00200**
R-squared:	0.011	Scale:	0.0078240

	Coef.	Std.Err.	t	P> |t|	[0.025	0.975]
const	0.1231	0.0022	55.4509	0.0000	0.1188	0.1275
Surprisal	-0.2766	0.1636	-1.6911	0.0910	-0.5974	0.0442
Word Frequency	0.2687	0.1399	1.9205	0.0550	-0.0057	0.5432
Precision	-0.4878	0.1392	-3.5052	**0.0005**	-0.7607	-0.2148
Interaction	0.2045	0.1599	1.2784	0.2013	-0.1092	0.5182

Table IV.2: Results of the ordinary least squares regression of CTI with a sliding windows of 5s with 500ms overlap

ing faithfully from word-level features. We added a categorical predictor to explain away any correlation with content/function word attribute. Coefficient for linguistic features are thus now also being considered in their interaction with this new categorical variable in the OLS model. Table IV.3 present those new results. We now obtained a much stronger model (p-value of F-statistic $< 1e^{-}11$ while $\simeq 0.002$ earlier).

Those results show that there is an effect of word-level feature as well as word category (function versus content words). The linguistic features showed a significant interaction with score only for function words though. We see a negative correlation of word frequency and precision with the cortical tracking accuracy indicating a modulation of neural alignment to the envelope that becomes less strong for infrequent words or for highly predictable words. On the other hand, a positive relationship between surprisal of function words and cortical tracking was present. The predictive power of cortical tracking from linguistic features was quite low overall as the r squared value from this linear regression was only of 0.011.

IV.6 General Discussion

This chapter presented a series of analysis applied to encoding and hybrid models to assess the stimulus-response relationship and the ability to decode comprehension from our speech representation that is using only word-level linguistic features.

We could reveal several aspect of language processing by characterizing more fully the interplay between those features, in how they affect the cortical tracking in different frequency bands, and how each subset of feature helps in predicting comprehension state for instance.

No. Observations:	6337	Log-Likelihood:	3365.0
Df Model:	7	F-statistic:	9.603
Df Residuals:	6329	Prob (F-statistic):	**6.25e-12**
R-squared:	0.011	Scale:	0.020270

	Coef.	t	P> |t|	[0.025	0.975]
Intercept	0.0359	2.4204	**0.0155**	0.0068	0.0650
C(func)	0.0804	3.4060	**0.0007**	0.0341	0.1266
Surprisal	0.0012	1.5356	0.1247	-0.0003	0.0028
Surprisal:C(func)	0.0051	4.0175	**0.0001**	0.0026	0.0076
WordFrequency	0.0015	1.6059	0.1084	-0.0003	0.0034
WordFrequency:C(func)	-0.0082	-3.4614	**0.0005**	-0.0128	-0.0035
Precision	0.0111	0.9521	0.3411	-0.0118	0.0340
C(func):Precision	-0.0129	-2.4565	**0.0141**	-0.0232	-0.0026

Table IV.3: Results of the ordinary least squares regression on CTI around words with categorical variables for function and content words added as a predictor.

IV.6.1 Top down modulation of envelope entrainment

The precision is a significant predictor of how accurate is the reconstruction of the acoustic envelope from the EEG recordings. In other word, cortical entrainment to envelope and spectro-temporal structures of sound is modulated by how precisely words can be predicted. Such a precision can be computed from syntactic structure constraints or lexical ones, and more generally from any conditional variables that help building a model to predict the next word. In our experiment, the recurrent neural network used has been shown to learn long term dependencies as well as being able to create hierarchical representations of the input within the recurrent connections of its hidden layers (Elman, 1990; Graves, 2013; Frank and Christiansen, 2018).

This result is in line with suggestions from predictive coding where information from lower level processing stages must be more reliably accounted for if the predictions from higher level are uncertain (Friston, 2018). That is, when the precision about prediction of the next word is low, the brain must rely more on the bottom-up acoustic information. Now this presumes that cortical tracking does indeed represent this low-level acoustic processing. It surely gives us a glimpse on acoustic representations in the brain but it is far from the full representation of speech at its lower level. Previous studies have stipulated the role of such entrainment in parsing of input (Giraud and Poeppel, 2012; Hyafil et al., 2015; Ghitza, 2017). The fact that some portion of the stimulus are less strongly represented in the brain (since backward models will be less able to reconstruct the stimulus from brain responses) is indicative of the dynamic nature of such processing. The cortex is continuously looking for information to explain away the meaning from utterances and this information can be found in a feed forward way from the acoustic input, but also from feedback connections and predictions of linguistic content. If we assume the predictive coding model, it appears as natural that brain process relies as much as possible on those high level predictions as they are directly available in such an encoding scheme.

However we noted that only a small fraction of cortical tracking variance could be explained by linguistic features ($R^2 = 0.011$). Consequently we conclude that most of the envelope tracking stem from phase alignment to lower-level acoustic features such as speech edges and fine spectro-temporal structures although aided from top-down processes in a relatively subtle manner.

IV.6.2 Spectral and spatial distribution of language encoding

Reconstruction accuracies reported in section IV.2 depict spatial loci of cortical responses to linguistic features. It is presented here as where (which electrode locations) in the brain do we see an improvement with respect to word onset model. We found significant benefits of using informed linguistic features such as *surprisal* and *precision* in several frequency bands of interest and interestingly with different outcomes for each predictor. For instance, little information could be gained when adding *syntactic depth* to our model, at least in that case for reconstructing left out subject data. It had some little benefit in the delta band, but this is minor considering the large amplitude of TRF observed in chapter III for *depth*.

It is in the delta band (0.5 to 4.5 Hz here) that we found the largest increase in reconstruction accuracy. Notably, this is the only frequency band where the full model was significantly better than a word onset only model for both prediction-based and syntactic features. This probably indicates an incapacity for those predictors to linearly regress EEG data and generalize across participant (as we used a leave-one subject out design for evaluation of model scores, see figure IV.1) for high frequency bands. The signal to noise ration is also higher in lower frequency bands, such as delta and theta. Power of high frequency EEG activity might be more prone to underfitting from this low signal to noise ratio, as a more complex model would be needed to denoise signals (probably even non-linear models).

The main network involved is formed of anterior temporal cortices together with centro-parietal areas. *Precision* did not give better than baseline reconstruction of EEG data in the delta band. This might imply a weaker representation of precision that is not captured by our model. It is worth noting that *precision* is a second order statistic (Koelsch, Vuust, and Friston, 2019). Here it is computed across the vocabulary, we expect it to modulate activity at different level of the hierarchy although it might not be effectively visible if there are not broad areas of synchronized activity involved.

However, we observed strong improvement for both word frequency feature (we remind the reader here that word frequency is inversely proportional to the unigram probability of a word) and *surprisal*. A model with word onset and word frequency had a right lateralised increase in reconstruction accuracy and conversely, the model with surprisal was better in left temporal regions. Together with the results for syntactic features, where we also see a left lateralised benefit of reconstruction accuracy (for the opening and closing branches predictors), these data converge towards empirical observation of left-lateralisation for language specific tasks although we overall see an improvement bilaterally. The left dominance is usually reported for most neural organisation model for language processing (Fedorenko, Behr, and Kanwisher, 2011; Friederici, 2011), but it would need more investigation to localise the sources generating the scalp topographies we reported.

There is a noticeable similarity between the response observed for word frequency in the theta band and the one for surprisal in the beta band. Also, we recall that the word frequency of a word is directly linked to the probability $p(w_i)$ of occurrence of this word in an utterance. This probability is unconditional, independent from any context and is only estimated from statistics across many English corpora of text. In the classical view of Bayesian inference, to generate a prediction the likelihood of the data is weighted by a prior knowledge. Here the prior on a word occurrence would actually correspond to this context-free probability. However we did not test further the link between such prior and our surprisal. This observation though might indicate a potential overlap in neural population encoding those quantities for word processing, and that both stream of information seem to be channelled through different frequency bands in EEG activity.

In other studies, theta band activity has been linked to memory retrieval Brennan et al., 2012 while beta band power has been attributed to top-down prediction signalling (Lewis and Bastiaansen, 2015; Sedley et al., 2016). The latter hypothesis fit well our results although surprisal as such is more related to a prediction error rather than predictions per se. In order to encode the new word in the current constituent representation, or into the global meaning of the sentence, both prediction error and lexical information must co-activate in a *coherent* way in order to pass the updated information (and the next prediction) to lower levels Levy, 2008; Arnal and Giraud, 2012. Now representation of syntax involves different brain structure or neural population, and similarly prediction error could be generated in other brain structures. This would imply that distant cortical areas communicate information in a coordinated manner as the speech is being processed. Long range communication in large neural networks has been observed by Fries, 2005. He used microelectrode recordings to reveal data suggesting the existence of communication through *coherence* across different cortical areas. Neural population would fire in coherent ways such that high-frequency activity enters in synchrony.

Precision must play a role in the propagation of prediction error as the predictions received from higher levels are weighted against sensory input relative to the precision at which they were made (Friston, 2012; Koelsch, Vuust, and Friston, 2019). In line with the idea that prediction errors are signalled in higher frequency bands such as gamma (Giraud and Poeppel, 2012; Fries, 2015; Sedley et al., 2016), we observed a robust representation of precision in the gamma band in panel D of figure IV.4. Together with the gain in scores measured by CCA when adding the interaction term (figure IV.6) these results indicate the importance of precision to process prediction error in the gamma range.

IV.6.3 On the role of word level predictions and syntax in comprehension

In the delta band, all features used in our decoding task could classify comprehension better than a model with only word onsets. However we note that for the theta band power, the model with word onset only outperformed any other. This is in line with theories concerning the role of theta oscillations in the parsing of speech input (Ghitza, 2011). For attention and intelligibility decoding, Etard and Reichenbach, 2019 also showed that theta band activity was involved in predicting the clarity, or acoustic quality of the speech input, but not the comprehension. Here we show that word-level linguistic features, such as surprisal or syntactic depth, are beneficial as an encoding model feature in delta or beta but not in theta where the mere response to word onset suffices to decode which language is being listened to.

We were quite conservative in our decoding task in that we used left-out subject to build the feature space on which the decoder was trained. Hence the encoding of our linguistic features must be robust enough to be generalisable across participant in order to give significant results. Critically, we observed that beta power could be reliably encoded by *surprisal* and predict above and beyond other models the comprehension state. This is a strong evidence that prediction errors are measurable in high frequency power of EEG. Lewis and Bastiaansen, 2015; Giraud and Poeppel, 2012 hypothesised that beta synchronisation would carry predictions rather than prediction error signals. Somehow the effect observed is opposite, as surprisal is our proxy for prediction error (Sedley et al., 2016). However, these classification data should be taken with care when interpreting underlying mechanism for neurobiology as we are training noisy evaluation on left out subject which probably under-fit high frequency power such as gamma band activity (because of the low SNR of high frequency power and the variability across participant for induced activity latencies relative to word onsets).

Syntactic features could also decode comprehension above significance level, even for the gamma band. This might reflect lower-level processing in syntactic unification that do not require predictive processing and may be mapped more directly to coherent neural population firing. It is also in good agreement with results from Bastiaansen, Magyari, and Hagoort, 2010; Brennan et al., 2012 where high-frequency activity is linked to syntactic unification.

Chapter V

Conclusion

We conclude the work presented in this thesis by summarizing novel contributions of each chapters and by giving a transverse perspective on the entire work with overreaching and more general goals in the study of neurobiology of language. This is a global discussion covering the main findings across the whole thesis and confronting them, as a whole, to the existing literature.

Two main axes are presented to highlight the impact of our results and their possible applications. Section V.1 will put the focus on outcomes of our work with respect to current theories of natural language understanding in the brain. Another section will be dedicated to the horizon of applications, for instance in clinical diagnosis, that the present thesis offers. Throughout both sections, limitations and constrains as well as possible future work will be addressed.

V.1 On the neurobiology of language processing

Thanks to advanced methodology and more ecologically valid experiments, we were able to bring new evidence for a predictive coding account of neural processing of language while not ruling out any claims on syntactic processing of hierarchical structures of language.

In chapter II we presented results of the analysis aiming at isolating the cortical response to *surprisal*, *precision entropy* and the interaction of those. We found significant response to each, although with different degree in each classic frequency bands. Low-frequency cortical oscillations, or delta rhythms, show the strongest and most reliable response to linguistic features. Interestingly, it is rather in the theta band that the envelope is best tracked by cortical activity Ghitza, 2017; Etard and Reichenbach, 2019. This suggests that word-level linguistic information or supra-segmental representations, beyond phonemic or syllables, entrains cortical activity and potentially modulates faster processes such as acoustic processing of auditory input. However, our study could not underpin the source and neural basis for the estimation of surprisal. Indeed, the probability density used to compute both *surprisal* and *entropy* could be derived from any language model. We opted for a recurrent neural network to estimate the probability of upcoming words but other systems, for instance that incorporate grammatical rules could be implemented in the brain. A recent study attempted to disentangle the origin of prediction on the computational level with a behavioural task (reading time as dependant variable) (Frank and Christiansen, 2018). They advocate that a hierarchical, rule-based model, rather than sequential or *markovian* one, gave more accurate predictions of reading time. Last year, Brennan and Hale presented a study were they replicated such an investigation using EEG and found similar results (Brennan and Hale, 2019, in). Hence a grammar-based surprisal, such as CFG grammar, gave stronger correlates of brain activity than surprisal from simple recurrent neural network. However, they used a language model trained on *part-of-speech* tags (namely, word categories such as adjectives, verbs, etc. . .) which rules out entirely any semantic processing. This could introduce a bias towards grammar-based models. In our study, both syntax and semantic are taken into account by the language model. A natural follow-up study would test different implementations of surprisal but not only using part-of-speech. By doing so we can test for a better representation of brain activity and thus unveil the neural basis of prediction in language processing.

Rather than seeking a "semantic-free" *surprisal* estimation as in Frank and Christiansen, 2018, we opted to design completely new features to represent hierarchically structured constituents of phrases in chapter III. Motivated by recent results from Ding et al., 2016 and by the growing body of literature about *unification* of constituents (Shetreet, Friedmann, and Hadar, 2009; Bastiaansen, Magyari, and Hagoort, 2010; Brennan et al., 2012), also described as the MERGE operation (Fedorenko, Behr, and Kanwisher, 2011; Chomsky, 2013), we aimed at designing syntactic features that could be aligned with ongoing continuous speech stimulus. However, in our results we did not find an effect of syntactic depth, namely the level of nesting, in higher frequency bands as suggested in those studies. This occurred instead for the number of closing branches in the hierarchy. Both are related though, as the syntactic depth is decreased by that same number between each level of the syntactic hierarchy. Also we observed a left lateralised topography in the improvement of EEG reconstruction in comparison with the word onset model. This topography was qualitatively different for what was observed in chapter II, meaning that we are probably looking at distinct neural processes.

The last chapter puts in perspective findings of previous chapters by introducing encoding or decoding performances. The goal was to characterize more precisely which were the most prominent feature representations in the EEG signals, and in which frequency band. An end goal of this analysis was to provide us with a speech representation that could help us to decode comprehension, such as whether the participant is listening to native or foreign language, from the EEG recordings. With the use of CCA, we found that adding the interaction between surprisal and entropy provides an encoding of stimulus response relationship for English that supplants models without it. This increase is measured as the correlation scores of canonical components. In contrast with score obtained on the Dutch dataset, with participants listening to a foreign speech, all English models presented a greater degree of relationship between EEG and speech representations. This translated into the ability to decode comprehension from segment of EEG data based on TRF modelling. Interestingly, features played different roles in the decoding task, with surprisal being a greater predictor of comprehension in the beta band while word onset was greater in the theta band. The best decoding occurred in the delta band, were all word-level models performed better than model using only acoustic information such as the envelope. Yet, using the envelope could still give above chance level decoding accuracy. Indeed we observed a correlation between acoustic cortical tracking or entrainment to envelope and linguistic features indicative of putative top-down modulation in the neural representation of the acoustic envelope from prediction of precision entropy.

A major issue with EEG is the difficulty to localise sources of neural activity with precision, and this becomes even worse without structural MRI scan and digitised positions of electrodes. Further experimentation using different neuroimaging techniques such as invasive electro-corticography (ECoG) or fMRI are required to associate accurately spatial foci of activity to responses reported in our results.

V.1.1 Rule-based versus statistical models of language

The debate is still on. However, the work in this thesis is an attempt at unifying both views, as we can not rule out one from the other. We observed significant responses emerging from those "opposed" framework. The information theoretic framework gave us a starting point to analyse predictions in language. Although this do not let any claim transpire against the possibility of *universal grammar* being present in human brain to help develop the ability to understand and speak. However, it is clear that the mechanisms at play in the cortex to continuously track linguistic inputs rely heavily on predictive processes. We observed strong response to both surprisal of words as well as precision, and those data-driven features were positively contributing to the decoding of comprehension.

Ding et al. observed cortical tracking of hierarchical structures and concluded hastily about the possible role of such mechanism over statistical learning. In response, Frank and Yang, 2018 showed that only lexical information as produced by statistical models could actually reproduce similar results. However, in an other study Frank et al., 2015 they also argued that hierarchically structured predictions were stronger predictors than sequential ones. I think there is no debate on the question of the involvement of hierarchical structures in the neural representation of language. The debate remains around the idea that those features are learned or generated from an innate *universal grammar*, although as argued by Frank et al., they might still underlie a predictive mechanism. In other words, even if we admit the existence of a universal grammar, we suggest that the brain utilises it in a predictive way. We have developed a technique to explore through the use of naturalistic stimuli how the brain perceives different structures of utterance depending on both the *hierarchical dependencies* as well as their dynamic *prediction through context.*

V.1.2 Predictive processing during language comprehension

Most studies on predictive coding concerned low-level sensory processing and perceptual systems (Rao and Ballard, 1999; Friston, 2005; Friston, 2010). On other instances, more formally they were analysing computational models at the level of cortical micro-circuits and neuronal networks to compare with physiological data evidence and exemplify the validity of predictive coding (Bastos et al., 2012; Friston and Kiebel, 2009; Chalk, Marre, and Tkačik, 2018). Nevertheless, the predictive coding framework got a lot of attention from cognitive scientists (Clark, 2013; Levy, 2008). The idea of having predictions based on context during language comprehension is not new (Federmeier and Kutas, 1999). However it is only since recently that researchers put an effort into linking elements of computational models of predictive coding with neural activity in response to language processing. Those key quantities are information theoretic measures that qualify the probabilistic generative model at play. *Surprisal* and *precision entropy* are both important quantities in the probabilistic description of a hierarchical generative implementing predictive coding. When parametrised through a linear dynamic model with Gaussian prior and likelihood, as in a Kalman filter for instance, the *precision* is the inverse of the covariance (or determinant thereof in a multivariate case), of which the logarithm is linearly proportional to the entropy (entropy of a Gaussian probability density is $\frac{1}{2}(1 + \log(2\pi D|\Sigma|)))$. To our knowledge the work from chapter II is the first relating natural language processing to the interaction of precision entropy and surprisal, as a proxy of precision-weighted prediction error. An inspiring study carried out by Sedley et al., demonstrated the relative stronger effect of surprisal versus a linear projection of predictive error in ECoG data with a simple auditory task. Converging findings between their results and ours bring more weight to the theory of predictive coding as a broad computational mechanism across the cortex.

Precision seems to play an important role but somehow not as strong as hypothesised. In lower-level cognitive processes such as auditory perception (Sedley et al., 2016), but also during the perception of music (Koelsch, Vuust, and Friston, 2019), it has been shown to be a key element in the neural representation of sensory input (Knill and Pouget, 2004). We measured a strong effect of the interaction between precision and surprisal. This is in line with predictive coding theories where predictions get updated depending on the precision *weighted* surprisal. We understand this result as an evidence for predictive coding in semantic and syntactic processing.

Importantly, we could separate the contribution of different frequency band of interest often referred to have specific and distinct roles in both electrophysiology literature Arnal and Giraud, 2012 and in predictive coding literature Sedley et al., 2016. Even though the fine grained neural mechanism underpinning those predictions remains unclear, we could measure neural correlates of features that are key in the predictive processing framework. The cortical responses measured were different depending on the frequency band in which we filtered the data.

Yet more research is necessary to break down the predictive processes involved during language understanding. With invasive data and better source localisation, researchers could link functional activity to structural neural organisation in order to establish a better neural model of language processing.

V.2 New methodology, novel applications

Most of the analysis in the thesis relied on a particular treatment of EEG signals. We worked with naturalistic continuous presentation of speech, which were both a motivation for developing new analysis tools and a constraint in term of what technique could be used. In the last five years, many studies with EEG for speech comprehension leaned towards the use of continuous stimulus instead of trial-based experiments. Most of the analysis methods employed in those studies revolve around the idea of using a linear model to represent the stimulus-response relationship (Di Liberto et al., 2015; Ding and Simon, 2014; Ding et al., 2016; Peelle and Davis, 2012; Zoefel and VanRullen, 2015b; Keitel et al., 2016). However, those studies mostly focused on acoustic features, such as the envelope or spectrogram. Such representations are involved in non-speech auditory processing as well, or in processing of foreign speech and do not efficiently decouple neural mechanisms responsible for language comprehension from mere auditory processing. Although the attention has been brought up more recently to higher level speech representations such a semantic similarity or phoneme surprisal as in Brodbeck, Presacco, and Simon, 2018; Broderick et al., 2018, our study paved the way to word level linguistic features as a probe to assess speech comprehension through EEG.

V.2.1 A toolkit for studying naturalistic speech

One goal of this thesis was to bring new tools in the methodology for studying spoken language processing. It seemed natural to pursue the route of using more ecologically valid stimuli to foster research and applications in naturalistic environment. On one hand, this gives a view on neural processes in their context, as they occur in everyday life, and on the other hand it embeds brain activity within the natural neural context. Hasson et al., 2018

When population of neurons, sparsely connected to form complex network undergo their non-linear dynamics, it can be advantageous to analyse their activity with a fully engaged brain. Other areas, non-task specific modulation may actually be captured by richer model to provide us with better gist on what is actually at play.

On the other hand, the use of naturalistic stimuli allows to deploy experiments more easily, to the public or in the clinic, with task that seem more natural, and hence less mentally exhausting than artificial and complex cognitive tasks. When one takes a diagnostic of an aphasic patient, it may sound cumbersome to ask the patient to attentively listen to a repeated set of non-sense words.

V.2.2 Decoding comprehension: applications

We assessed comprehension by comparing cortical activity in response to exposure to native language and foreign language separately. Above chance classification was possible using segment of EEG of a few seconds only. This could potentially inspire new tools for language disorder diagnosis and also could be used in new technologies such as hearing aid devices.

The first aspect of technology concerns assessment of language impairment, such as with aphasia, where therapists would benefit from having a quantitative assessment of the degree and type of impairment. Currently, aphasia is diagnosed by speech specialists using a battery of tests such as the Comprehensive Aphasia Test (Bruce and Edmundson, 2010). Beyond the lesions analysis and observations from neurologists of MRI scans no quantitative measure of neural processing of speech are being used for such assessment.

Another potential application is in relation to brain-machine interface to detect whether an unconscious patient is hearing, and understanding speech. Some attempts have been made to detect and classify the state of patients in coma, whether they are in a locked-in or vegetative state. It is important to assess their level of consciousness and awareness as their treatment while in coma might differ.

Finally, decoding comprehension from EEG signals could also benefit hearing aids (neuro-steered hearing aid devices are attracting interests with recent findings on selective attention, see Das et al., (2020)). If a real-time decoder allows the technology to detect whether the wearer missed a part of the attended speech, one could potentially enhance acoustic aspects of the signal picked by the hearing aid microphone to increase intelligibility accordingly.

However the work presented here regarding decoding did not have an engineering perspective and was carried out mainly to answer fundamental scientific questions about speech processing. One caveat is that it requires the knowledge of textual data from the speech heard to be able to decode comprehension. Furthermore, we did not test the ability to get the technology near to real-time decoding. With this in mind, one could improve the encoding model as well as the language model that extract features from speech a straightforward improvement of the language model is to use Long Short-Term Memory (LSTM) instead of classic RNN architectures to estimate the language model (Hochreiter and Schmidhuber, 1997; Sundermeyer, Schlüter, and Ney, 2012).

Our methods have been applied in a naive fashion. Realistically, it should be tested with a real-time experiment, although further work is necessary on the speech recognition side as well, as such an application decoding comprehension in real-time using our methodology would need the speech transcript being passed through a trained language model such as Long Short-Term Memory (LSTM) or RNN.

Appendix A

Code

Most of the code, scripts, class and methods written throughout the PhD have been assembled into a python library. The library is well documented and relies on a few other scientific common python packages. It allows efficient analysis of EEG data that has been recorded along with continuous speech stimulus and especially contains method to easily align word-level features to the EEG signals.

All the package code is available at `https://github.com/Hugo-W/pyEEG`.

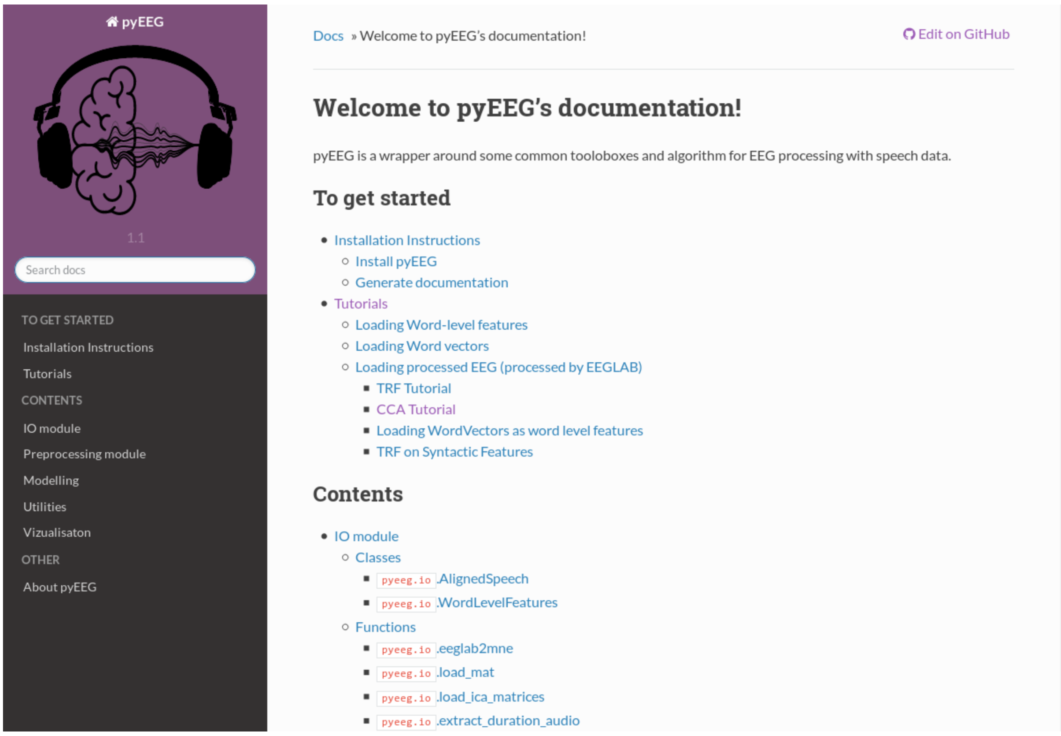
The library created heavily relies on MNE-python (Gramfort et al., 2013) to import, process and store structured EEG data. However, the rest of the custom-written code has been designed to easily handle time-aligned speech data along with EEG signals. The aligned speech segments can be continuous, and can contains different corresponding features such as acoustic ones like the envelope amplitude, or linguistic features such as word-level surprisal or categories.

The library also makes better use of algebra to efficiently compute TRFs of several subjects in a grand average model, simply by averaging the covariance matrices before inverting them. This gives a consequent gain in computation time and it is also more memory-efficient.

An example of the documentation page generated with python libraries is shown in figure A.1.

Here is a non-exhaustive list of other functions and methods implemented in the library:

- integration of word vectors representations to align those features to the EEG signal
- CCA
- multiway-CCA to preprocess group EEG data (denoising technique)



Welcome to pyEEG's documentation!

pyEEG is a wrapper around some common tooloboxes and algorithm for EEG processing with speech data.

To get started

- Installation Instructions
  - Install pyEEG
  - Generate documentation
- Tutorials
  - Loading Word-level features
  - Loading Word vectors
  - Loading processed EEG (processed by EEGLAB)
    - TRF Tutorial
    - CCA Tutorial
    - Loading WordVectors as word level features
    - TRF on Syntactic Features

Contents

- IO module
  - Classes
    - `pyeeg.io`.AlignedSpeech
    - `pyeeg.io`.WordLevelFeatures
  - Functions
    - `pyeeg.io`.eeglab2mne
    - `pyeeg.io`.load_mat
    - `pyeeg.io`.load_ica_matrices
    - `pyeeg.io`.extract_duration_audio

Figure A.1: Example of a documentation page from the EEG processing python library.

- multi-channel Wiener filter to remove artefacts from EEG
- plotting functions

Appendix B

Data Availability

The dataset has been made publicly partly available to promote reproducibility and foster more experimentations.

A first release was made along the publication of chapter II at https://doi.org/10.6084/m9.figshare.9033983.v1 (Weissbart, Kandylaki, and Reichenbach, 2019b). We released pre-processed EEG data. Namely, the signals are already band pass filtered in the delta band, and re-references to the average signal.

We supplied surprisal and precision values of words presented during the experiment along the EEG signals, and a portion of code that allows any user to regenerate figure II.4 on page 47 easily.

Appendix C

Proof of permissions

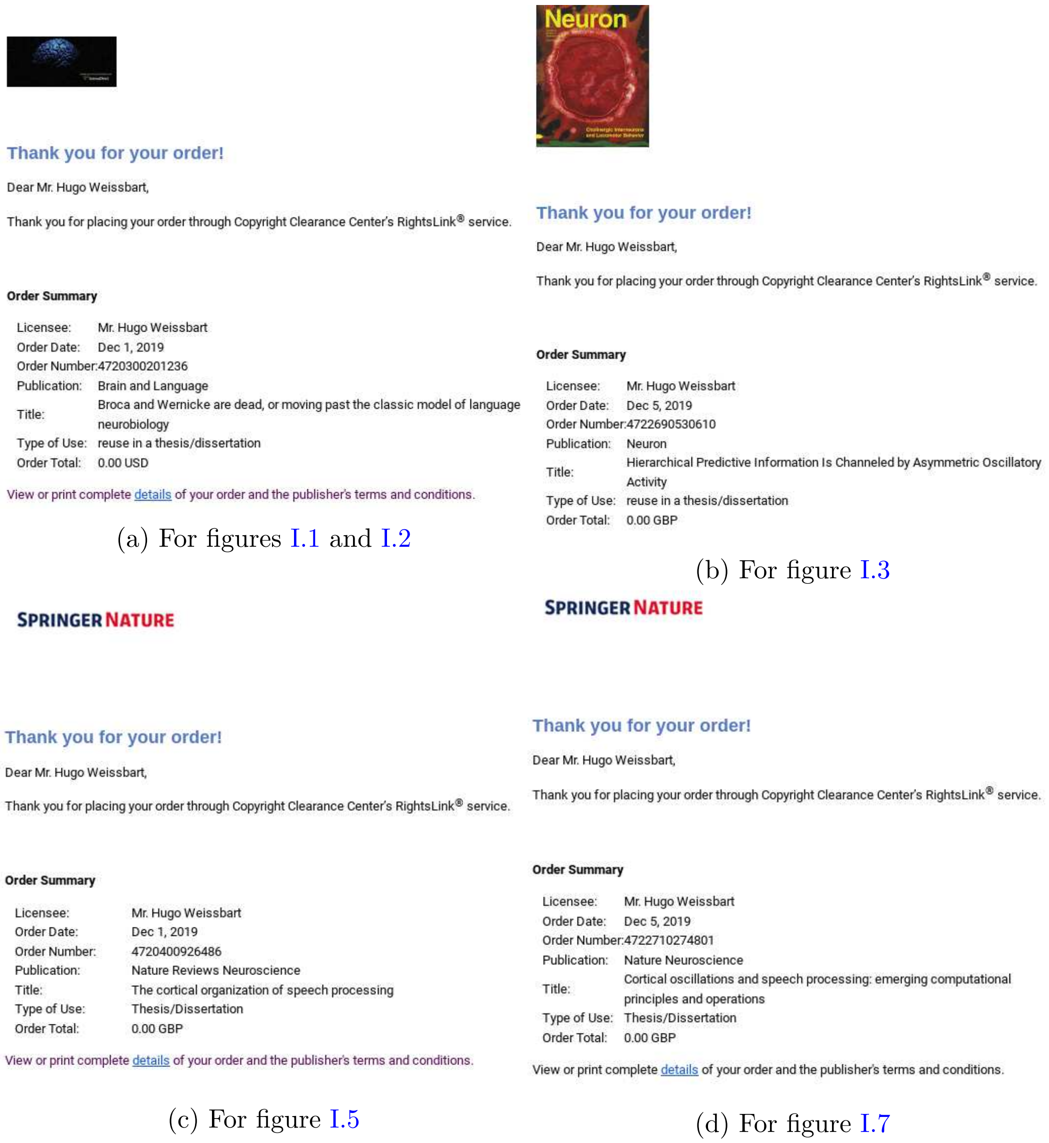
As described in the following table (table C.1 on the next page), I requested permission to reuse the following figures from different journals:

1. figure I.1 on page 9
2. figure I.2 on page 10
3. figure I.3 on page 11
4. figure I.5 on page 20
5. figure I.7 on page 25

A copy of the proof of permission received by mail is given here for each of those figures.

Page No.	Type	Source	Copyright holder	Permission requested on
9	figure	Tremblay and Dick, 2016	© 2016 Elsevier	01/12/19
10	figure	Tremblay and Dick, 2016	© 2016 Elsevier	01/12/19
10	figure	Hickok and Poeppel, 2007	© 2007 Nature	08/12/19
19	figure	Giraud and Poeppel, 2012	© 2012 Springer Nature	
24	figure	Giraud and Arnal, 2018	© 2012 Elsevier Inc.	

Table C.1: Permission summary table. All rights were obtained in written form.



Thank you for your order!

Dear Mr. Hugo Weissbart,

Thank you for placing your order through Copyright Clearance Center's RightsLink® service.

Order Summary

Licensee:	Mr. Hugo Weissbart
Order Date:	Dec 1, 2019
Order Number:	4720300201236
Publication:	Brain and Language
Title:	Broca and Wernicke are dead, or moving past the classic model of language neurobiology
Type of Use:	reuse in a thesis/dissertation
Order Total:	0.00 USD

View or print complete details of your order and the publisher's terms and conditions.

(a) For figures I.1 and I.2

Thank you for your order!

Dear Mr. Hugo Weissbart,

Thank you for placing your order through Copyright Clearance Center's RightsLink® service.

Order Summary

Licensee:	Mr. Hugo Weissbart
Order Date:	Dec 5, 2019
Order Number:	4722690530610
Publication:	Neuron
Title:	Hierarchical Predictive Information Is Channeled by Asymmetric Oscillatory Activity
Type of Use:	reuse in a thesis/dissertation
Order Total:	0.00 GBP

(b) For figure I.3

SPRINGER NATURE

Thank you for your order!

Dear Mr. Hugo Weissbart,

Thank you for placing your order through Copyright Clearance Center's RightsLink® service.

Order Summary

Licensee:	Mr. Hugo Weissbart
Order Date:	Dec 1, 2019
Order Number:	4720400926486
Publication:	Nature Reviews Neuroscience
Title:	The cortical organization of speech processing
Type of Use:	Thesis/Dissertation
Order Total:	0.00 GBP

View or print complete details of your order and the publisher's terms and conditions.

(c) For figure I.5

SPRINGER NATURE

Thank you for your order!

Dear Mr. Hugo Weissbart,

Thank you for placing your order through Copyright Clearance Center's RightsLink® service.

Order Summary

Licensee:	Mr. Hugo Weissbart
Order Date:	Dec 5, 2019
Order Number:	4722710274801
Publication:	Nature Neuroscience
Title:	Cortical oscillations and speech processing: emerging computational principles and operations
Type of Use:	Thesis/Dissertation
Order Total:	0.00 GBP

View or print complete details of your order and the publisher's terms and conditions.

(d) For figure I.7

Figure C.1: Proof of permissions to reuse figures

Bibliography

Arnal, Luc H. and Anne-Lise Giraud (2012). “Cortical oscillations and sensory predictions”. In: *Trends in Cognitive Sciences* 16.7, pp. 390–398. ISSN: 13646613. DOI: 10.1016/j.tics.2012.05.003.

Azuar, C., A. Leger, C. Arbizu, F. Henry-Amar, S. Chomel-Guillaume, and Y. Samson (2013). “The Aphasia Rapid Test: an NIHSS-like aphasia test”. In: *Journal of Neurology* 260.8, pp. 2110–2117. ISSN: 0340-5354. DOI: 10.1007/s00415-013-6943-x.

Baggio, Giosuè and Peter Hagoort (2011). “The balance between memory and unification in semantics: A dynamic account of the N400”. In: *Language and Cognitive Processes* 26.9, pp. 1338–1367. ISSN: 0169-0965. DOI: 10.1080/01690965.2010.542671.

Bastiaansen, Marcel C.M., Robert Oostenveld, Ole Jensen, and Peter Hagoort (2008). “I see what you mean: Theta power increases are involved in the retrieval of lexical semantic information”. In: *Brain and Language* 106.1, pp. 15–28. ISSN: 0093934X. DOI: 10.1016/j.bandl.2007.10.006.

Bastiaansen, Marcel and Peter Hagoort (2006). “Oscillatory neuronal dynamics during language comprehension”. In: *Progress in Brain Research* 159, pp. 179–196. ISSN: 0079-6123. DOI: 10.1016/S0079-6123(06)59012-0.

Bastiaansen, Marcel, Lilla Magyari, and Peter Hagoort (2010). “Syntactic Unification Operations Are Reflected in Oscillatory Dynamics during On-line Sentence Comprehension”. en. In: *Journal of Cognitive Neuroscience* 22.7, pp. 1333–1347. ISSN: 0898-929X. DOI: 10.1162/jocn.2009.21283.

Bastos, Andre M. A.M. M., W. Martin Martin W.M. Usrey, R.A. Rick A. A. Adams, George R. G.R. George R. Mangun, Pascal Fries, and Karl J. K.J. J. Friston (2012). “Canonical Microcircuits for Predictive Coding”. In: *Neuron* 76.4, pp. 695–711. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.10.038.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin (2000). “10.1162/153244303322533223”. In: *CrossRef Listing of Deleted DOIs* 1, pp. 1137–1155. ISSN: 0003-6951. DOI: 10.1162/153244303322533223.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin (2003). “A Neural Probabilistic Language Model”. In: *Journal of Machine Learning Research* 3.Feb, pp. 1137–1155. ISSN: ISSN 1533-7928.

Berwick, Robert C. and Noam Chomsky (2016). *Why Only Us? Evolution of Language.* MIT Press.

Berwick, Robert C., Angela D. Friederici, Noam Chomsky, and Johan J. Bolhuis (2013). "Evolution, brain, and the nature of language." In: *Trends in cognitive sciences* 17.2, pp. 89–98. ISSN: 1879-307X. DOI: 10.1016/j.tics.2012.12.002.

Brennan, Jonathan R. and John T Hale (2019). "Hierarchical structure guides rapid linguistic predictions during naturalistic listening". In: *PLOS ONE* 14.1. Ed. by Johan J Bolhuis, e0207741. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0207741.

Brennan, Jonathan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen (2012). "Syntactic structure building in the anterior temporal lobe during natural story listening". In: *Brain and Language* 120.2, pp. 163–173. ISSN: 0093934X. DOI: 10.1016/j.bandl.2010.04.002.

Broca, Paul (1865). *Sur le siège de la faculté du langage articulé.* DOI: 10.3406/bmsap.1865.9495.

Brodbeck, Christian, Alessandro Presacco, and Jonathan Z. Simon (2018). "Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension". In: *NeuroImage* 172, pp. 162–174. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2018.01.042.

Broderick, Michael P., Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor (2018). "Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech". In: *Current Biology* 28.5, 803–809.e3. ISSN: 09609822. DOI: 10.1016/j.cub.2018.01.080.

Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer (1992). "Class-Based *n*-gram Models of Natural Language". In: *Computational Linguistics* 18.4, pp. 467–480.

Bruce, Carolyn and Anne Edmundson (2010). "Letting the CAT out of the bag: A review of the Comprehensive Aphasia Test. Commentary on Howard, Swinburn, and Porter, "Putting the CAT out: What the Comprehensive Aphasia Test has to offer"". In: *Aphasiology* 24.1, pp. 79–93. ISSN: 0268-7038. DOI: 10.1080/02687030802453335.

Chalk, Matthew, Olivier Marre, and Gašper Tkačik (2018). "Toward a unified theory of efficient, predictive, and sparse coding". In: *Proceedings of the National Academy of Sciences* 115.1, pp. 186–191. ISSN: 0027-8424. DOI: 10.1073/PNAS.1711114115.

Cheveigné, Alain de, Daniel D.E. Wong, Giovanni M. Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund Lalor (2018a). "Decoding the auditory brain with canonical component analysis". In: *NeuroImage* 172, pp. 206–216. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2018.01.033.

Cheveigné, Alain de, Daniel E. Wong, Giovanni M. Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund Lalor (2018b). "Decoding the auditory brain with canonical component analysis". In: *NeuroImage* 172, pp. 206–216. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2018.01.033.

Chomsky, Noam (1988). *Language and problems of knowledge: the Managua lectures.* The MIT Press.

— (2007). *Aspects of the theory of syntax.* MIT Press.

— (2013). “Problems of projection”. In: *Lingua* 130, pp. 33–49. ISSN: 00243841. DOI: 10.1016/j.lingua.2012.12.003.

— (2015). *Syntactic structures*. Martino.

Clark, Andy (2013). “Whatever next? Predictive brains, situated agents, and the future of cognitive science.” English. In: *The Behavioral and brain sciences* 36.3, pp. 181–204. ISSN: 1469-1825. DOI: 10.1017/S0140525X12000477.

Collins, Beverley and Peter Hagoort (2000). *The Neurocognition of Language*. Oxford University Press. ISBN: 9780198507932.

Collins, Beverley and Inger Mees (1999). *The phonetics of English and Dutch*. Brill.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). “Natural Language Processing (Almost) from Scratch”. In: *J. Mach. Learn. Res.* 12.null, pp. 2493–2537. ISSN: 1532-4435.

Crosse, Michael J and Edmund C Lalor (2014). “The Cortical Representation of the Speech Envelope is Earlier for Audiovisual Speech than Audio Speech.” In: *Journal of neurophysiology*, pp. 1–31. ISSN: 1522-1598. DOI: 10.1152/jn.00690.2013.

Das, Neetha, Jeroen Zegers, Hugo Van hamme, Tom Francart, and Alexander Bertrand (2020). “EEG-informed speaker extraction from noisy recordings in neuro-steered hearing aids: linear versus deep learning methods”. In: *bioRxiv*. DOI: 10.1101/2020.01.22.915181. eprint: https://www.biorxiv.org/content/early/2020/01/23/2020.01.22.915181.full.pdf.

Dax, Marc (1865). “Lesions de la moitie gauche de l’encephale coincident avec l’ouble des signes de la pensee”. In: *Gazette Hebdomadaire de Medecine et de Chirurgie* 2.2, pp. 259–260.

DeLong, Katherine A., Laura Quante, and Marta Kutas (2014). “Predictability, plausibility, and two late ERP positivities during written sentence comprehension”. In: *Neuropsychologia* 61.1, pp. 150–162. ISSN: 18733514. DOI: 10.1016/j.neuropsychologia.2014.06.016.

DeLong, Katherine A, Thomas P Urbach, and Marta Kutas (2005). “Probabilistic word pre-activation during language comprehension inferred from electrical brain activity”. In: *Nature Neuroscience* 8.8, pp. 1117–1121. ISSN: 1097-6256. DOI: 10.1038/nn1504.

Di Liberto, Giovanni M., Daniel Wong, Gerda Ana Melnik, and Alain de Cheveigné (2019). “Low-frequency cortical responses to natural speech reflect probabilistic phonotactics”. In: *NeuroImage* 196, pp. 237–247. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2019.04.037.

Di Liberto, Giovanni M. et al. (2015). “Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing”. In: *Current Biology* 25.19, pp. 2457–2465. ISSN: 09609822. DOI: 10.1016/j.cub.2015.08.030.

Ding, N. and J. Z. Simon (2012). *Emergence of neural encoding of auditory objects while listening to competing speakers*. DOI: 10.1073/pnas.1205381109.

Ding, Nai, Lucia Melloni, Xing Tian, and David Poeppel (2017). “Rule-based and word-level statistics-based processing of language: insights from neuroscience”. In: *Language, Cognition and Neuroscience* 32.5, pp. 570–575. ISSN: 2327-3798. DOI: 10.1080/23273798.2016.1215477. arXiv: 15334406.

Ding, Nai, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel (2016). "Cortical tracking of hierarchical linguistic structures in connected speech". In: *Nature Neuroscience* 19.1, pp. 158–164. ISSN: 1097-6256. DOI: 10.1038/nn.4186.

Ding, Nai, Xunyi Pan, Cheng Luo, Naifei Su, Wen Zhang, and Jianfeng Zhang (2018). "Attention Is Required for Knowledge-Based Sequential Grouping: Insights from the Integration of Syllables into Words". In: *Journal of Neuroscience* 38.5, pp. 1178–1188. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.2606-17.2017. eprint: https://www.jneurosci.org/content/38/5/1178.full.pdf.

Ding, Nai and Jonathan Z. Simon (2014). "Cortical entrainment to continuous speech: functional roles and interpretations." In: *Frontiers in human neuroscience* 8.May, p. 311. ISSN: 1662-5161. DOI: 10.3389/fnhum.2014.00311.

Dmochowski, Jacek P., Jason J. Ki, Paul DeGuzman, Paul Sajda, and Lucas C. Parra (2018). "Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity". In: *NeuroImage* 180. New advances in encoding and decoding of brain signals, pp. 134–146. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2017.05.037.

Elman, J (1990). "Finding structure in time". In: *Cognitive Science* 14.2, pp. 179–211. ISSN: 03640213. DOI: 10.1016/0364-0213(90)90002-E.

Etard, Octave, Mikolaj Kegler, Chananel Braiman, Antonio Elia Forte, and Tobias Reichenbach (2019). "Decoding of selective attention to continuous speech from the human auditory brainstem response". In: *NeuroImage*. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2019.06.029.

Etard, Octave and Tobias Reichenbach (2019). "Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise." In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 39.29, pp. 5750–5759. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.1828-18.2019.

Federmeier, Kara D. and Marta Kutas (1999). "A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing". English (US). In: *Journal of Memory and Language* 41.4, pp. 469–495. ISSN: 0749-596X. DOI: 10.1006/jmla.1999.2660.

Federmeier, Kara D., Edward W. Wlotko, Esmeralda De Ochoa-Dewald, and Marta Kutas (2007). "Multiple effects of sentential constraint on word processing". In: *Brain Research* 1146, pp. 75–84.

Fedorenko, E., M. K. Behr, and N. Kanwisher (2011). "Functional specificity for high-level linguistic processing in the human brain". In: *Proceedings of the National Academy of Sciences* 108.39, pp. 16428–16433. ISSN: 0027-8424. DOI: 10.1073/pnas.1112937108.

Fedorenko, Evelina, Alfonso Nieto-Castañón, and Nancy Kanwisher (2012). "Syntactic processing in the human brain: What we know, what we don't know, and a suggestion for how to proceed". In: *Brain and Language* 120.2. The Neurobiology of Syntax, pp. 187–207. ISSN: 0093-934X. DOI: https://doi.org/10.1016/j.bandl.2011.01.001.

Fedorenko, Evelina et al. (2016). "Neural correlate of the construction of sentence meaning". In: *Proceedings of the National Academy of Sciences of the United*

States of America 113.41, E6256–E6262. ISSN: 10916490. DOI: 10.1073/pnas.1612132113.

Feldman, Harriet and Karl J. Friston (2010). "Attention, Uncertainty, and Free-Energy". In: *Frontiers in Human Neuroscience* 4, p. 215. ISSN: 1662-5161. DOI: 10.3389/fnhum.2010.00215.

Frank, Stefan L. and Morten H. Christiansen (2018). "Hierarchical and sequential processing of language". In: *Language, Cognition and Neuroscience* 33.9, pp. 1213–1218. ISSN: 2327-3798. DOI: 10.1080/23273798.2018.1424347.

Frank, Stefan L., Leun J. Otten, Giulia Galli, and Gabriella Vigliocco (2015). "The ERP response to the amount of information conveyed by words in sentences". In: *Brain and Language* 140, pp. 1–11. ISSN: 0093-934X. DOI: 10.1016/J.BANDL.2014.10.006.

Frank, Stefan L. and Roel M. Willems (2017). "Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension". In: *Language, Cognition and Neuroscience* 32.9, pp. 1192–1203. ISSN: 2327-3798. DOI: 10.1080/23273798.2017.1323109.

Frank, Stefan L. and Jinbiao Yang (2018). "Lexical representation explains cortical entrainment during speech comprehension". In: *PLOS ONE* 13.5. Ed. by Jonathan R. Brennan, e0197304. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0197304. arXiv: 1706.05656.

Friederici, Angela D. (2002). *Towards a neural basis of auditory sentence processing.* DOI: 10.1016/S1364-6613(00)01839-8.

Friederici, Angela D (2011). "The Brain Basis of Language Processing: From Structure to Function". In: *Physiological Reviews* 91.4, pp. 1357–1392. ISSN: 0031-9333. DOI: 10.1152/physrev.00006.2011.

Friederici, Angela D. and Sonja A. Kotz (2003). "The brain basis of syntactic processes: functional imaging and lesion studies". In: *Neuroimage* 20, S8–S17.

Friederici, Angela D. and Claudia Männel (2014). "Neural correlates of the development of speech perception and comprehension". In: *The Oxford Handbook of Cognitive Neuroscience*, pp. 171–192.

Friederici, Angela D and Jürgen Weissenborn (2007). "Mapping sentence form onto meaning: The syntax–semantic interface". In: *Brain Research* 1146, pp. 50–58. ISSN: 0006-8993. DOI: https://doi.org/10.1016/j.brainres.2006.08.038.

Fries, Pascal (2005). "A mechanism for cognitive dynamics: neuronal communication through neuronal coherence". In: *Trends in Cognitive Sciences* 9.10, pp. 474–480. ISSN: 13646613. DOI: 10.1016/j.tics.2005.08.011.

— (2015). "Rhythms for Cognition: Communication through Coherence". In: *Neuron*. ISSN: 10974199. DOI: 10.1016/j.neuron.2015.09.034. arXiv: 15334406.

Frisch, Ragnar and Frederick V. Waugh (1933). "Partial Time Regressions as Compared with Individual Trends". In: *Econometrica* 1.4, pp. 387–401. ISSN: 00129682, 14680262.

Friston, Karl (2005). "A theory of cortical responses." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360.April, pp. 815–836. ISSN: 0962-8436. DOI: 10.1098/rstb.2005.1622.

— (2010). "The free-energy principle: a unified brain theory?" In: *Nature Reviews Neuroscience* 11.2, pp. 127–138. ISSN: 1471003X. DOI: 10.1038/nrn2787.

Friston, Karl (2012). “Prediction, perception and agency”. In: *International Journal of Psychophysiology* 83.2, pp. 248–252. ISSN: 0167-8760. DOI: 10.1016/J.IJPSYCHO.2011.11.014.

— (2018). “Does predictive coding have a future?” In: *Nature Neuroscience* 21.8, pp. 1019–1021. ISSN: 15461726. DOI: 10.1038/s41593-018-0200-7.

Friston, Karl and Stefan Kiebel (2009). “Predictive coding under the free-energy principle”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1521, pp. 1211–1221. ISSN: 0962-8436. DOI: 10.1098/rstb.2008.0300.

Gagnepain, Pierre, Richard N. Henson, and Matthew H. Davis (2012). “Temporal predictive codes for spoken words in auditory cortex.” In: *Current Biology* 22.7, pp. 615–621. ISSN: 09609822. DOI: 10.1016/j.cub.2012.02.015.

Geschwind, N (1970). “The organization of language and the brain.” In: *Science (New York, N.Y.)* 170, pp. 940–944. ISSN: 0036-8075. DOI: 10.1126/science.170.3961.940.

Ghitza, Oded (2011). “Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm.” In: *Frontiers in psychology* 2.June, p. 130. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2011.00130.

— (2017). “Acoustic-driven delta rhythms as prosodic markers”. In: *Language, Cognition and Neuroscience* 32.5, pp. 545–561. ISSN: 2327-3798. DOI: 10.1080/23273798.2016.1232419.

Giraud, Anne-Lise and Luc H Arnal (2018). “Hierarchical Predictive Information Is Channeled by Asymmetric Oscillatory Activity.” In: *Neuron* 100.5, pp. 1022–1024. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2018.11.020.

Giraud, Anne-Lise and David Poeppel (2012). “Cortical oscillations and speech processing: emerging computational principles and operations”. In: *Nature Neuroscience* 15.4, pp. 511–517. ISSN: 1546-1726. DOI: 10.1038/nn.3063.

Gorman, Kyle, Jonathan Howell, and Michael Wagner (2011). “Prosodylab-aligner: A tool for forced alignment of laboratory speech”. In: *Canadian Acoustics* 39.3, pp. 192–193.

Gramfort, Alexandre et al. (2013). “MEG and EEG data analysis with MNE-Python”. In: *Frontiers in Neuroscience* 7, p. 267. ISSN: 1662453X. DOI: 10.3389/fnins.2013.00267.

Graves, Alex (2013). “Generating Sequences With Recurrent Neural Networks”. In: arXiv: 1308.0850.

Hagoort, Peter (2003). “Interplay between Syntax and Semantics during Sentence Comprehension: ERP Effects of Combining Syntactic and Semantic Violations”. In: *Journal of Cognitive Neuroscience* 15.6, pp. 883–899. DOI: 10.1162/089892903322370807.

— (2016). “MUC (Memory, Unification, Control): A Model on the Neurobiology of Language Beyond Single Word Processing”. In: *Neurobiology of Language*. Ed. by Gregory Hickok and Steven L. Small. Academic Press. Chap. 28, pp. 339–347. ISBN: 9780124078628. DOI: 10.1016/B978-0-12-407794-2.00028-6.

Halgren, Eric et al. (2002). “N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences”. In: *NeuroImage* 17.3, pp. 1101–1116. ISSN: 10538119. DOI: 10.1006/nimg.2002.1268.

Hamilton, Liberty S. and Alexander G. Huth (2018). “The revolution will not be controlled: natural stimuli in speech neuroscience”. In: *Language, Cognition and Neuroscience*, pp. 1–10. ISSN: 2327-3798. DOI: 10.1080/23273798.2018.1499946.

Hasson, Uri, Giovanna Egidi, Marco Marelli, and Roel M. Willems (2018). “Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension”. In: *Cognition* 180, pp. 135–157. ISSN: 0010-0277. DOI: 10.1016/J.COGNITION.2018.06.018.

Hasson, Uri and Pascale Tremblay (2016). “Neurobiology of Statistical Information Processing in the Auditory Domain”. In: *Neurobiology of Language*. Ed. by Gregory Hickok and Steven L Small. San Diego: Elsevier, pp. 527–537. ISBN: 978-0-12-407794-2. DOI: 10.1016/B978-0-12-407794-2.00043-2.

Hauser, Marc D, Noam Chomsky, and W Tecumseh Fitch (2002). “The faculty of language: what is it, who has it, and how did it evolve?” In: *Science (New York, N.Y.)* 298.5598, pp. 1569–79. ISSN: 1095-9203. DOI: 10.1126/science.298.5598.1569.

Heer, Wendy A. de, Alexander G. Huth, Thomas L. Griffiths, Jack L. Gallant, and Frédéric E. Theunissen (2017). “The Hierarchical Cortical Organization of Human Speech Processing”. In: *Journal of Neuroscience* 37.27, pp. 6539–6557. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3267-16.2017. eprint: https://www.jneurosci.org/content/37/27/6539.full.pdf.

Heilbron, Micha (2018). “Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex?” In: *Neuroscience* 389, pp. 54–73. ISSN: 0306-4522. DOI: 10.1016/J.NEUROSCIENCE.2017.07.061.

Helenius, P (1998). “Distinct time courses of word and context comprehension in the left temporal cortex”. In: *Brain* 121.6, pp. 1133–1142. ISSN: 14602156. DOI: 10.1093/brain/121.6.1133.

Helmholtz, H von (1866). “Concerning the perceptions in general”. In: *Treatise on physiological optics,*

Henderson, John M., Wonil Choi, Matthew W. Lowder, and Fernanda Ferreira (2016). “Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading”. In: *NeuroImage* 132, pp. 293–300. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2016.02.050.

Hickok, Gregory and David Poeppel (2007). “The cortical organization of speech processing”. In: *Perspective* 8.May, pp. 393–402.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long Short-Term Memory”. In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://doi.org/10.1162/neco.1997.9.8.1735.

Hudspeth, AJ. (2013). “The Inner Ear”. In: *Principles of Neural Science*. Ed. by Eric R. Kandel, James H. Schwartz, Thomas M. Jessel, Steven A. Siegelbaum, and AJ. Hudspeth. Fifth Edition. The McGraw-Hill Companies. Chap. 30, pp. 654–661.

Humphries, Colin, Jeffrey R. Binder, David A. Medler, and Einat Liebenthal (2006). “Syntactic and semantic modulation of neural activity during auditory sentence comprehension”. In: *Journal of Cognitive Neuroscience* 18.4, pp. 665–679. ISSN: 0898929X. DOI: 10.1162/jocn.2006.18.4.665.

Hyafil, Alexandre, Lorenzo Fontolan, Claire Kabdebon, Boris Gutkin, and Anne-Lise Giraud (2015). "Speech encoding by coupled cortical theta and gamma oscillations." en. In: *eLife* 4, e06213. ISSN: 2050-084X. DOI: 10.7554/eLife.06213.

Kadir, Shabnam, Chrysoula Kaza, Hugo Weissbart, and Tobias Reichenbach (2019). "Modulation of speech-in-noise comprehension through transcranial current stimulation with the phase-shifted speech envelope". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 1–1. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2019.2939671.

Kanai, Ryota, Yutaka Komura, Stewart Shipp, and Karl Friston (2015). "Cerebral hierarchies: predictive processing, precision and the pulvinar". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1668, p. 20140169. ISSN: 0962-8436. DOI: 10.1098/rstb.2014.0169.

Kandylaki, Katerina D. and Ina Bornkessel-Schlesewsky (2019). "From story comprehension to the neurobiology of language". In: *Language, Cognition and Neuroscience* 34.4, pp. 405–410. ISSN: 2327-3798. DOI: 10.1080/23273798.2019.1584679.

Kandylaki, Katerina D. et al. (2016). "Predicting "When" in Discourse Engages the Human Dorsal Auditory Stream: An fMRI Study Using Naturalistic Stories". In: *Journal of Neuroscience* 36.48, pp. 12180–12191. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.4100-15.2016.

Keitel, Anne, Joachim Gross, and Christoph Kayser (2018). "Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features". In: *PLOS Biology* 16.3. Ed. by Jennifer Bizley, e2004473. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2004473.

Keitel, Anne, Robin A A Ince, Joachim Gross, and Christoph Kayser (2016). "Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks". In: *NeuroImage* 147, pp. 32–42. DOI: 10.1016/j.neuroimage.2016.11.062.

Kielar, Aneta, Jed A. Meltzer, Sylvain Moreno, Claude Alain, and Ellen Bialystok (2014). "Oscillatory Responses to Semantic and Syntactic Violations". In: *Journal of Cognitive Neuroscience* 26.12, pp. 2840–2862. ISSN: 0898-929X. DOI: 10.1162/jocn_a_00670.

Kielar, Aneta, Aya Meltzer-Asscher, and Cynthia K. Thompson (2012). "Electrophysiological responses to argument structure violations in healthy adults and individuals with agrammatic aphasia". In: *Neuropsychologia* 50.14, pp. 3320–3337. ISSN: 00283932. DOI: 10.1016/j.neuropsychologia.2012.09.013.

Kielar, Aneta, Lilia Panamsky, Kira a. Links, and Jed a. Meltzer (2015). "Localization of electrophysiological responses to semantic and syntactic anomalies in language comprehension with MEG". In: *NeuroImage* 105, pp. 507–524. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2014.11.016.

Klein, Dan and Christoper Manning (2003). "In Accurate Unlexicalized Parsing". In: *Proceedings of the 41st Meeting of the Association for Computational Linguistics*. DOI: 10.3115/1075096.1075150.

Knill, David C. and Alexandre Pouget (2004). “The Bayesian brain: the role of uncertainty in neural coding and computation”. In: *Trends in Neurosciences* 27.12, pp. 712–719. ISSN: 01662236. DOI: 10.1016/j.tins.2004.10.007.

Koelsch, Stefan, Peter Vuust, and Karl Friston (2019). “Predictive Processes and the Peculiar Case of Music”. In: *Trends in Cognitive Sciences* 23.1. ISSN: 1879307X. DOI: 10.1016/j.tics.2018.10.006.

Kumar, T Krishna (1975). “Multicollinearity in Regression Analysis”. In: *The Review of Economics and Statistics* 57.3, pp. 365–366.

Kuperberg, Gina R. (2007). “Neural mechanisms of language comprehension: Challenges to syntax”. In: *Brain Research* 1146.1, pp. 23–49. ISSN: 00068993. DOI: 10.1016/j.brainres.2006.12.063.

Kuperberg, Gina R., Tatiana Sitnikova, David Caplan, and Phillip J. Holcomb (2003). “Electrophysiological distinctions in processing conceptual relationships within simple sentences”. In: *Cognitive Brain Research* 17.1, pp. 117–129. ISSN: 09266410. DOI: 10.1016/S0926-6410(03)00086-7.

Kutas, Marta and Kara D. Federmeier (2011). “Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP)”. In: *Annual Review of Psychology* 62.1, pp. 621–647. ISSN: 0066-4308. DOI: 10.1146/annurev.psych.093008.131123.

Kutas, Marta and Steven A. Hillyard (1980). “Event-related brain potentials to semantically inappropriate and surprisingly large words”. In: *Biological Psychology* 11.2, pp. 99–116. ISSN: 03010511. DOI: 10.1016/0301-0511(80)90046-0.

Lakatos, Peter, Chi-Ming Chen, Monica N. O’Connell, Aimee Mills, and Charles E. Schroeder (2007). “Neuronal Oscillations and Multisensory Interaction in Primary Auditory Cortex”. In: *Neuron* 53.2, pp. 279–292. ISSN: 0896-6273. DOI: 10.1016/J.NEURON.2006.12.011.

Lakatos, Peter, George Karmos, Ashesh D Mehta, Istvan Ulbert, and Charles E Schroeder (2008). “Entrainment of neuronal oscillations as a mechanism of attentional selection.” In: *Science (New York, N.Y.)* 320.5872, pp. 110–3. ISSN: 1095-9203. DOI: 10.1126/science.1154735.

Levy, Roger (2008). “Expectation-based syntactic comprehension”. In: *Cognition* 106.3, pp. 1126–1177. ISSN: 00100277. DOI: 10.1016/j.cognition.2007.05.006.

Lewis, Ashley G. and Marcel Bastiaansen (2015). “A predictive coding framework for rapid neural dynamics during sentence-level language comprehension”. In: *Cortex* 68, pp. 155–168. ISSN: 0010-9452. DOI: 10.1016/j.cortex.2015.02.014.

Lovell, Michael C. (2011). “A Simple Proof of the FWL (Frisch-Waugh-Lovell) Theorem”. In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.887345.

Maess, Burkhard, Christoph S. Herrmann, Anja Hahne, Akinori Nakamura, and Angela D. Friederici (2006). “Localizing the distributed language network responsible for the N400 measured by MEG during auditory sentence processing”. In: *Brain Research* 1096.1, pp. 163–172. ISSN: 00068993. DOI: 10.1016/j.brainres.2006.04.037.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press. ISBN: 0-262-13360-1.

Marr, D., T. Poggio, and Sydney Brenner (1979). “A computational theory of human stereo vision”. In: *Proceedings of the Royal Society of London. Series B. Biological*

Sciences 204.1156, pp. 301–328. DOI: 10.1098/rspb.1979.0029. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.1979.0029.

Mattys, Sven L., Peter W. Jusczyk, Paul A. Luce, and James L. Morgan (1999). "Phonotactic and Prosodic Effects on Word Segmentation in Infants". In: *Cognitive Psychology* 38.4, pp. 465–494. ISSN: 0010-0285. DOI: https://doi.org/10.1006/cogp.1999.0721.

McClelland, James L. and Jeffrey L. Elman (1986). "The TRACE model of speech perception". In: *Cognitive Psychology* 18.1, pp. 1–86. ISSN: 00100285. DOI: 10.1016/0010-0285(86)90015-0.

Mesgarani, Nima and Edward F. Chang (2012). "Selective cortical representation of attended speaker in multi-talker speech perception." In: *Nature* 485.7397, pp. 233–6. ISSN: 1476-4687. DOI: 10.1038/nature11020.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781.

Mikolov, Tomas, Martin Karafiát, Lukás Burget, Jan Černocký, and Sanjeev Khudanpur (2010). "Recurrent neural network based language model". In: *Interspeech*.

Mikolov, Tomas, Stefan Kombrink, Lukás Burget, Jan Černocký, and Sanjeev Khudanpur (2011). "Extensions of recurrent neural network language model". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5528–5531.

Miller, George A, George A Heise, William Lighten, William Lichten, and William Lighten (1951). "The Intelligibility of Speech as a Function of the Context of the Test Material". In: *Journal of experimental psychology*.

Miller, George A. and Stephen Isard (1963). "Some perceptual consequences of linguistic rules". In: *Journal of Verbal Learning and Verbal Behavior* 2.3, pp. 217–228. ISSN: 0022-5371. DOI: 10.1016/S0022-5371(63)80087-0.

Molinaro, Nicola, Horacio A. Barber, and Manuel Carreiras (2011a). "Grammatical agreement processing in reading: ERP findings and future directions". In: *Cortex* 47.8, pp. 908–930. ISSN: 0010-9452. DOI: https://doi.org/10.1016/j.cortex.2011.02.019.

— (2011b). "Grammatical agreement processing in reading: ERP findings and future directions". In: *Cortex* 47.8, pp. 908–930. ISSN: 0010-9452. DOI: 10.1016/J.CORTEX.2011.02.019.

Molinaro, Nicola, Paulo Barraza, and Manuel Carreiras (2013). "Long-range neural synchronization supports fast and efficient reading: EEG correlates of processing expected words in sentences". In: *NeuroImage* 72, pp. 120–132. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2013.01.031.

Nieuwland, Mante S. et al. (2020). "Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1791, p. 20180522. ISSN: 0962-8436. DOI: 10.1098/rstb.2018.0522.

Obleser, Jonas and Christoph Kayser (2019). "Neural Entrainment and Attentional Selection in the Listening Brain". In: *Trends in Cognitive Sciences* 23.11, pp. 913–926. ISSN: 1364-6613. DOI: 10.1016/J.TICS.2019.08.004.

Oertel, Donata and Allison J. Doupe (2013). “The Auditory Central Nervous System”. In: *Principles of Neural Science*. Ed. by Eric R. Kandel, James H. Schwartz, Thomas M. Jessel, Steven A. Siegelbaum, and AJ. Hudspeth. Fifth Edition. The McGraw-Hill Companies. Chap. 31, pp. 682–711.

Oostenveld, Robert, Pascal Fries, Eric Maris, and Jan Mathijs Schoffelen (2011). “FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data”. In: *Computational Intelligence and Neuroscience* 2011. ISSN: 16875273. DOI: 10.1155/2011/156869.

Osterhout, Lee, Phillip J. Holcomb, and David A. Swinney (1994). “Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20.4, pp. 786–803. ISSN: 1939-1285. DOI: 10.1037/0278-7393.20.4.786.

Osterhout, Lee, Judith McLaughlin, and Michael Bersick (1997). “Event-related brain potentials and human language”. In: *Trends in Cognitive Sciences* 1.6, pp. 203–209. ISSN: 13646613. DOI: 10.1016/S1364-6613(97)01073-5.

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (2012). “On the difficulty of training recurrent neural networks”. In: *ICML*.

Patten, William (1910). *International short stories (Vol.2)*. P.F. Collier & Son.

Peelle, Jonathan E. and Matthew H. Davis (2012). “Neural oscillations carry speech rhythm through to comprehension”. In: *Frontiers in Psychology* 3.SEP, p. 320. ISSN: 16641078. DOI: 10.3389/fpsyg.2012.00320.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). “Glove: Global Vectors for Word Representation”. In: *EMNLP*.

Poeppel, David (2014). “The neuroanatomic and neurophysiological infrastructure for speech and language.” In: *Current opinion in neurobiology* 28C, pp. 142–149. ISSN: 1873-6882. DOI: 10.1016/j.conb.2014.07.005.

Rao, Rajesh P. N. and Dana H. Ballard (1999). “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects”. In: *Nature Neuroscience* 2.1, pp. 79–87. ISSN: 1097-6256. DOI: 10.1038/4580.

Rommers, Joost, Danielle S. Dickson, James J. S. Norton, Edward W. Wlotko, and Kara D. Federmeier (2017). “Alpha and theta band dynamics related to sentential constraint and word expectancy”. In: *Language, Cognition and Neuroscience* 32.5, pp. 576–589. ISSN: 2327-3798. DOI: 10.1080/23273798.2016.1183799.

Rösler, Frank, Thomas Pechmann, Judith Streb, Brigitte Röder, and Erwin Hennighausen (1998). “Parsing of Sentences in a Language with Varying Word Order: Word-by-Word Variations of Processing Demands Are Revealed by Event-Related Brain Potentials”. In: *Journal of Memory and Language* 38.2, pp. 150–176. ISSN: 0749596X. DOI: 10.1006/jmla.1997.2551.

Saffran, Jenny R. (2003). “Statistical Language Learning: Mechanisms and Constraints”. In: *Current Directions in Psychological Science* 12.4, pp. 110–114. DOI: 10.1111/1467-8721.01243. eprint: https://doi.org/10.1111/1467-8721.01243.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport (1996). “Statistical Learning by 8-Month-Old Infants”. In: *Science* 274.5294, pp. 1926–1928. ISSN:

0036-8075. DOI: 10.1126/science.274.5294.1926. eprint: https://science.sciencemag.org/content/274/5294/1926.full.pdf.

Sassenhagen, Jona (2019). "How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression". In: *Language, Cognition and Neuroscience* 34.4, pp. 474–490. ISSN: 2327-3798. DOI: 10.1080/23273798.2018.1502458.

Seabold, Skipper and Josef Perktold (2010). "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*.

Sedley, William et al. (2016). "Neural signatures of perceptual inference". In: *eLife* 5.6, pp. 797–801. ISSN: 2050-084X. DOI: 10.7554/eLife.11476.

Shetreet, Einat, Naama Friedmann, and Uri Hadar (2009). "An fMRI study of syntactic layers: Sentential and lexical aspects of embedding". In: *NeuroImage* 48.4, pp. 707–716. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2009.07.001.

Smith, Nathaniel J. and Roger Levy (2013). "The effect of word predictability on reading time is logarithmic". In: *Cognition* 128.3, pp. 302–319. ISSN: 00100277. DOI: 10.1016/j.cognition.2013.02.013.

Steinhauer, Karsten and John E. Drury (2012). "On the early left-anterior negativity (ELAN) in syntax studies". In: *Brain and Language* 120.2, pp. 135–162. ISSN: 0093934X. DOI: 10.1016/j.bandl.2011.07.001.

Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney (2012). "LSTM Neural Networks for Language Modeling". In: *INTERSPEECH-2012*. (Portland, OR, USA), pp. 194–197.

Tremblay, Pascale and Anthony Steven Dick (2016). "Broca and Wernicke are dead, or moving past the classic model of language neurobiology". In: *Brain and Language* 162, pp. 60–71. ISSN: 0093-934X. DOI: 10.1016/J.BANDL.2016.08.004.

Tse, C.-Y. et al. (2007). "Imaging cortical dynamics of language processing with the event-related optical signal". In: *Proceedings of the National Academy of Sciences* 104.43, pp. 17157–17162. ISSN: 0027-8424. DOI: 10.1073/pnas.0707901104.

Turken, And U. and Nina F. Dronkers (2011). "The Neural Architecture of the Language Comprehension Network: Converging Evidence from Lesion and Connectivity Analyses". In: *Frontiers in System Neuroscience* 5.February, p. 1. ISSN: 1662-5137. DOI: 10.3389/fnsys.2011.00001.

Van Den Brink, Daniëlle, Colin M. Brown, and Peter Hagoort (2001). "Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects". In: *Journal of Cognitive Neuroscience* 13.7, pp. 967–985. ISSN: 0898929X. DOI: 10.1162/089892901753165872.

Van Petten, Cyma and Barbara J. Luka (2006). "Neural localization of semantic context effects in electromagnetic and hemodynamic studies". In: *Brain and Language* 97.3, pp. 279–293. ISSN: 0093934X. DOI: 10.1016/j.bandl.2005.11.003.

Wang, Lin, Peter Hagoort, and Ole Jensen (2017). "Language Prediction Is Reflected by Coupling between Frontal Gamma and Posterior Alpha Oscillations". In: *Journal of Cognitive Neuroscience*, pp. 1–16. ISSN: 0898-929X. DOI: 10.1162/jocn_a_01190.

Weiss, Sabine and Horst M. Mueller (2012). ""Too Many betas do not Spoil the Broth": The Role of Beta Brain Oscillations in Language Processing". In: *Frontiers in Psychology* 3, p. 201. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2012.00201.

Weissbart, Hugo, Katerina D. Kandylaki, and Tobias Reichenbach (2019a). "Cortical Tracking of Surprisal during Continuous Speech Comprehension". In: *Journal of Cognitive Neuroscience*, pp. 1–12. ISSN: 0898-929X. DOI: 10.1162/jocn_a_01467.

— (2019b). *EEG recordings and stimuli.* DOI: 10.6084/m9.figshare.9033983.v1.

Wernicke, C (1874). *Der aphasische Symptomencomplex. Eine psychologische Studie auf anatomischer Basis*, pp. 219–283. ISBN: 9781141248926. DOI: 10.1007/978-3-642-65950-8.

Willems, Roel M., Stefan L. Frank, Annabel D. Nijhof, Peter Hagoort, and Antal van den Bosch (2016). "Prediction During Natural Language Comprehension". In: *Cerebral Cortex* 26.6, pp. 2506–2516. ISSN: 1047-3211. DOI: 10.1093/cercor/bhv075. eprint: http://oup.prod.sis.lan/cercor/article-pdf/26/6/2506/17309746/bhv075.pdf.

Yang, Dong-Ping, Hai-Jun Zhou, and Changsong Zhou (2017). "Co-emergence of multi-scale cortical activities of irregular firing, oscillations and avalanches achieves cost-efficient information capacity." In: *PLoS computational biology* 13.2, e1005384. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005384.

Yildiz, Izzet B and Stefan J Kiebel (2011). "A hierarchical neuronal model for generation and online recognition of birdsongs." In: *PLoS computational biology* 7.12, e1002303. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002303.

Yildiz, Izzet B, Katharina von Kriegstein, and Stefan J Kiebel (2013). "From birdsong to human speech recognition: bayesian inference on a hierarchy of nonlinear dynamical systems." In: *PLoS computational biology* 9.9, e1003219. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003219.

Zion Golumbic, Elana M et al. (2013). "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"." In: *Neuron* 77.5, pp. 980–91. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2012.12.037.

Zoefel, Benedikt and Rufin VanRullen (2015a). "EEG oscillations entrain their phase to high-level features of speech sound." In: *NeuroImage* 124.Pt A, pp. 16–23. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2015.08.054.

— (2015b). "Selective Perceptual Phase Entrainment to Speech Rhythm in the Absence of Spectral Energy Fluctuations". In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 35.5, pp. 1954–1964. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.3484-14.2015.