# A Model of the Network Architecture of the Brain that Supports Natural Language Processing

Sarah Aliko
Student ID: 17105227
Department of Experimental Psychology, University College London
Submitted to University College London for the Degree
of Doctor of Philosophy

December 2021

Supervisors: Dr Jeremy I Skipper
Prof Lewis Griffin

I, Sarah Aliko confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Sarah Aliko
21 December, 2021

# Abstract

For centuries, neuroscience has proposed models of the neurobiology of language processing that are static and localised to few temporal and inferior frontal regions. Although existing models have offered some insight into the processes underlying lower-level language features, they have largely overlooked how language operates in the real world.

Here, we aimed at investigating the network organisation of the brain and how it supports language processing in a naturalistic setting. We hypothesised that the brain is organised in a multiple core-periphery and dynamic modular architecture, with canonical language regions forming high-connectivity hubs. Moreover, we predicted that language processing would be distributed to much of the rest of the brain, allowing it to perform more complex tasks and to share information with other cognitive domains.

To test these hypotheses, we collected the *Naturalistic Neuroimaging Database* of people watching full length movies during functional magnetic resonance imaging. We computed network algorithms to capture the voxel-wise architecture of the brain in individual participants and inspected variations in activity distribution over different stimuli and over more complex language features. Our results confirmed the hypothesis that the brain is organised in a flexible multiple core-periphery architecture with large dynamic communities. Here, language processing was distributed to much of the rest of the brain, together forming multiple communities. Canonical language regions constituted hubs, explaining why they consistently appear in various other neurobiology of language models. Moreover, language processing was supported by other regions such as visual cortex and episodic memory regions, when processing more complex context-specific language features. Overall, our flexible and distributed model of language comprehension and the brain points to additional brain regions and pathways that could be exploited for novel and more individualised therapies for patients suffering from speech impairments.

# Impact statement

The work presented in this thesis describes the first alternative model of the neurobiology of language and the brain in the real world, using a network-based approach. This model is highly flexible and can account for and explain the complexity and variability of language processing in a natural setting, as we detail in the subsequent chapters.

The model we propose here has significant implications for our understanding of the neurobiology of language, arguably the most complex and unique human behaviour. We show that language processing is much more distributed, dynamic and flexible than any existing other model suggests. Regions largely considered as the sole loci of language processing in existing models, here take a specific role in coordinating and directing more distributed language areas. By considering both the roles and dynamics of canonical language regions and distributed regions, our model offers new insights on putative regions and pathways to exploit for the development of novel speech therapies, to help facilitate the recovery of patients suffering from aphasia (or any other speech disability).

Moreover, our network model has important implications for the study of any other complex brain behaviour. Our model is the most detailed brain connectome to date, in terms of spatial and temporal resolution, as well as mathematical description. Indeed, it is based on individual participant networks at the resolution of single voxels, over the scale of a full-length movie. Furthermore, it is described using six different network measures at the level of individual voxels and participants' networks, making it the most comprehensive network model to date. These features make the model ideal for future studies on individual cognitive abilities and cognitive strategies, for identifying biomarkers of mental health in individual participants, and for developing personalised therapeutics and speech therapies.

Finally, as part of this work, we made publicly available the *Naturalistic Neuroimaging Database*, which can be a resource for any neuroscientist investigating brain behaviours in naturalistic settings. The database is currently one of the largest and most varied open-source datasets available, offering limitless possibilities for research into brain behaviour and development of new data-driven approaches.

# Acknowledgments

There are many people that have helped, guided and supported me during my time as a PhD student. First and foremost, my supervisor Dr Jeremy Skipper has been a great mentor, a good friend, and has inspired me to be a better researcher. I am particularly grateful that he has always valued my opinion, including me in important lab decisions and has always treated me as his equal. Thank you, Captain!

I want to extend my gratitude to the BUCNI team (in particular Prof Fred Dick and Dr Tessa Dekker) for their help with this project and for offering their expertise whenever data collection reached a hurdle. I want to thank the UT Dallas HPC director Dr Chris Simmons for kindly granting us special access to the TACC resources and for his patience with my questions. I would like to send special thanks to all of our collaborators for their invaluable advice, support and for helping me grow as a researcher: Prof Steven Small, Dr Chee Ming Ting, Dr Fuad Noman, Dr Chengbin Peng, Dr Chris Baldassano, Dr Eva Wittenber, Dr Richard Golden. I would like to also thank past and present members of the awesome LAB lab team, without whom this journey would have been impossible: you have made this PhD more fun. I look forward to celebrating each and everyone of your successes in future! A particular thank you goes to Annette Glotfelty, for being a great friend and colleague, who has helped me so much in the past year. Another special thanks goes to all the participants who kindly volunteered their time for this work.

Finally and most importantly, I want to dedicate this work to my family. To my loving husband who stood by my side every day of this journey, always with a smile and with immense kindness in his heart. To my caring parents who have made their love and support be felt from afar, and whose life's sacrifices are the reason I am here today. To my sweet brother who constantly reminds me to take myself less seriously and enjoy the little things in life. My family has been my ultimate inspiration and the shining light guiding me through these many academic years. I could not have asked for more generous, patient and loving people in my life. Thank you.

# Table of Contents

# Chapter 1: Introduction

What makes us human? Although there are likely many different answers to this question, it is undeniable that one of the defining features of our species is our ability to communicate with words. As such, a major goal of neuroscience has long been to understand how the brain supports speech production and perception. Although many have investigated the neurobiology of language and proposed various models over the centuries, there is still debate over how language and the brain work, with many questions still left unanswered. In this thesis, we first discuss insights and shortfalls from existing models of the neurobiology of language, then propose a novel network-based model of how language processing functions in the real world.

## Classical and contemporary language models

Interest in the neurobiology of language started early in the field of neuroscience, with the pioneering work by Broca and Wernicke. Broca observed that patients with lesions near the left hemisphere (LH) inferior frontal gyrus (IFG) suffered from speech production impairments, drawing the conclusion that the brain region was associated with speech production (Tremblay & Dick, 2016). Later, Wernicke identified another form of speech impairment whereby patients were unable to comprehend speech due to lesions around the superior temporal gyrus (STG), which was later reduced to an area near the LH posterior sylvian fissure by more contemporary neurologists (Tremblay & Dick, 2016). This small region was deemed the site of language comprehension. The classical model was recapitulated and expanded by Geschwind, who showed that the arcuate fasciculus, a set of fibres connecting temporal and inferior frontal regions, connected Broca's and Wernicke's language areas, thus proposing a more network-like model (Geschwind, 1970) (Fig. 1).

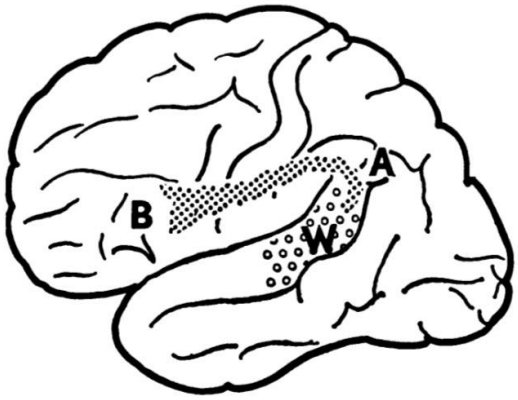



**Figure 1.** Representation of the classical model where B = Broca's area, W = Wernicke's area and A = arcuate fasciculus. Broca's area near the inferior frontal gyrus was considered the site of speech production, while Wernicke's area was considered the site of language comprehension, with the arcuate fasciculus linking the two in the classical model. Image from (Geschwind, 1970).

However, since its conception, the classical model has received criticism from several sources. For one, it was found that lesions to Broca's area, or rather its supposed locus in the brain near the IFG, did not simply cause speech production but also language comprehension deficits; similarly, lesions in Wernicke's area (near the posterior STG) were associated with symptoms of paraphasia, a deficit of speech production, as well as perception deficits, indicating that the behaviours ascribed to each locus in the classical model were inaccurate (Binder, 2015; Dronkers et al., 2017; Hagoort, 2016; Hickok & Poeppel, 2007). Secondly, aphasia disabilities traditionally associated with lesions to Broca's and Wernicke's areas were found to also result from lesions in regions outside these two classical areas, indicating that other areas are also important to language processing (Mesulam et al., 2015; Poeppel et al., 2012). Thirdly, the classical model's anatomical loci could not be precisely located in any individual human brain, due to high intersubject variability and cytoarchitectonic complexity of the loci, raising questions about the structural basis of the model that suggests a universal locus for each of speech perception and production functions across human brains (Amunts et al., 1999). Finally, the model reduced language to overly simplistic 'production' and 'perception' behaviours from lesion observations, but failed to inspect any individual language feature, such as phonemes (Poeppel & Hickok, 2004; Tremblay & Dick, 2016). Therefore, although still widely discussed

and taught (e.g., in psychology and medical schools), the classical model has been largely discredited.

As neuroimaging technologies, such as functional magnetic resonance imaging (fMRI) and electroencephalogram (EEG), emerged, neuroscientists were able to collect more evidence on brain activity as a response to particular stimuli. This allowed for more focus on processes underlying the neurobiology of language and the brain. The most cited model to date is the dual-stream model put forward by Hickok and Poeppel (Hickok & Poeppel, 2007; Poeppel & Hickok, 2004). The origin of the model came from an incongruence in the clinical literature, whereby it was found that some patients presenting damage in frontal and inferior parietal regions were unable to distinguish syllables but could still understand words, and vice versa (Hickok & Poeppel, 2007). Later neuroimaging studies identified another paradox: during various speech perception tasks, regions around both Broca's and Wernicke's areas were activated, whilst damage to either of the two areas resulted most often in speech production deficits rather than perception deficits (Hickok & Poeppel, 2007).

The dual-stream aimed to resolve these paradoxes and provide a unifying and mechanistic model of language comprehension and production. The model proposes that the neurobiology of language is divided into two streams, one ventral and one dorsal, involved in speech perception and production respectively (Poeppel & Hickok, 2004). The dorsal stream relates articulatory and auditory signals and maps to premotor cortex, posterior IFG, Sylvian parietal temporal region (Spt) and insula; the ventral stream maps sound to meaning and activates superior and middle temporal gyri (STG, MTG) in a mostly bilateral fashion (Hickok & Poeppel, 2007). The model allows some convergence between speech production and perception circuits in the superior temporal sulcus (STS) for phonological processing (Hickok & Poeppel, 2007). As can be seen in Fig. 2, the dual-stream model involves few more areas than the classical model, although still primarily in the LH temporal lobe, and these map around what might be postulated as Broca's (e.g., IFG) and Wernicke's (e.g., Spt) regions in the classical model (Nasios et al., 2019).

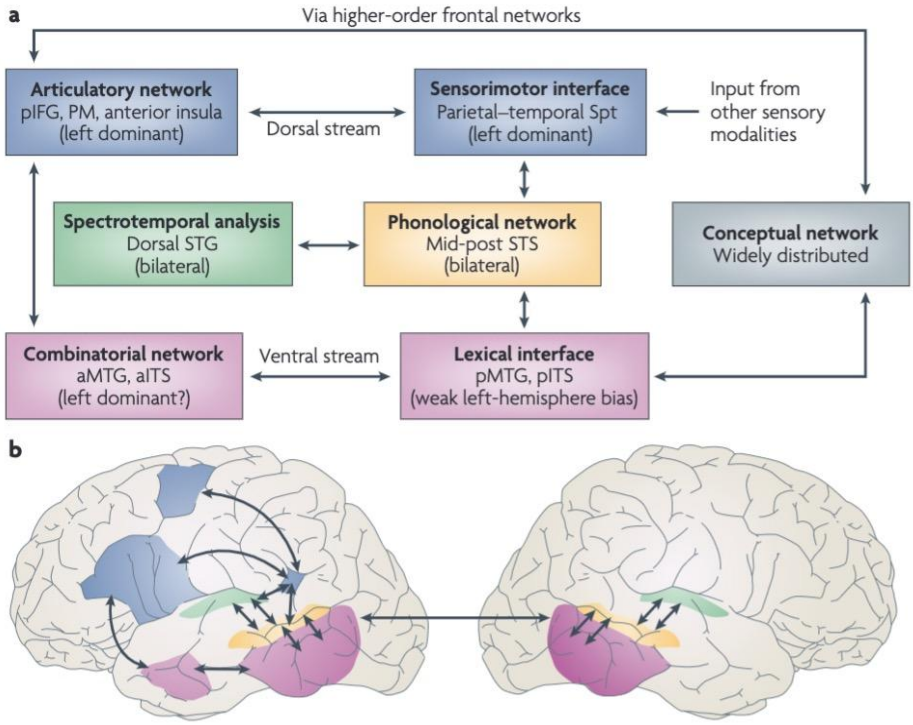



**Figure 2.** Schematic representation of the dual stream model, where areas in purple represent the ventral stream and in blue the dorsal stream. Image from (Hickok & Poeppel, 2007). The ventral stream maps sound to meaning (e.g., language comprehension), while the dorsal stream maps sound to articulatory movements (e.g., speech production).

From a mechanistic standpoint, the network proposed in the dual stream has only been tested by one study, that we are aware of, using diffusion tensor imaging (DTI) and simple speech production and perception tasks (e.g., listen to sentences vs pseudo-sentences) (Saur et al., 2008). The authors first identified 'language' regions by subtraction and maximum peaks, then mapped the underlying white matter tracts to these regions. The results indicated that the arcuate and uncinate fasciculi underlie the dorsal and ventral pathways respectively (Saur et al., 2008). Although this was taken as evidence for the dual-stream model processes, regions, and connectivity, the study had methodological shortcomings that might render the findings questionable, such as (i) using simple tasks as representatives of the complex behaviour of language in the real world; (ii) assuming that task subtraction isolates the specific language component; (iii) assuming that there should be functional specificity for language regions; and (iv) assuming that structural connectivity matches functional connectivity. Perhaps the biggest contribution that the dual-stream model has offered is that it has started a conversation about a

network representation of the neurobiology of language, and has begun to consider language as a slightly more complex human behaviour (Skipper, 2015a).

Although it has not been thoroughly tested as a model, most of the current literature on the neurobiology of language frames results in terms of dual-stream models, assuming that the regions and connections proposed by the model are correct. Indeed, studies on healthy controls and aphasic patients have both endorsed the localised and constrained view of the dual-stream: for instance, lesion studies investigating language processing deficits in stroke survivors have pointed to similar pathways as those proposed by the dual-stream, with deficits that relate to the functional role of those pathways (i.e., lesions to ventral pathway result in comprehension deficits) (Fridriksson et al., 2016), and task-based neuroimaging studies have repeatedly reported those same regions (for an extensive review on this issue see (Price, 2012)). Strikingly, across various language meta-analyses the same regions (roughly IFG, STG, and MTG) plus a few neighbouring areas appear repeatedly, as shown in Fig. 3.

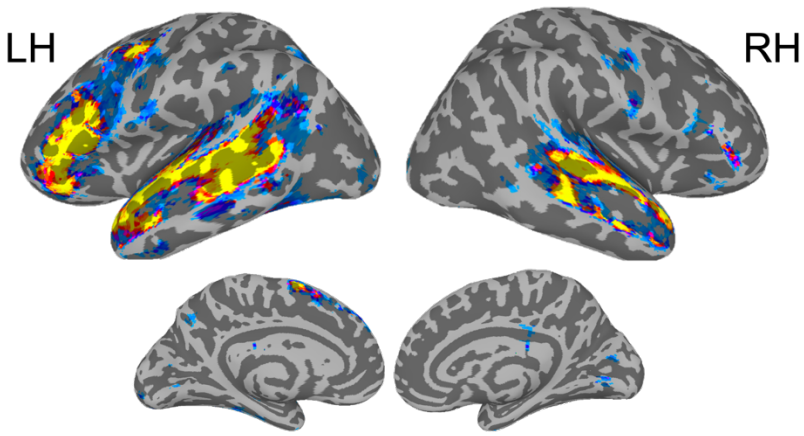



**Figure 3.** Overlap of various language meta-analysis terms from Neurosynth (Yarkoni et al., 2011). These include the meta-analysis terms ‘language comprehension’, ‘comprehension’, ‘sentence comprehension’, ‘speech perception’, ‘language network’, ‘language’. Yellow indicates the highest overlap among all meta-analysis terms, orange/red represents medium overlap, and blue represents unique patterns to a single meta-analysis term. Here, the IFG, STG, MTG and parts of premotor cortex consistently appear as the highest overlapping regions across meta-analyses.

This has given the impression that unique language processes occur in the same brain loci, sharing the same pathways and possibly overlapping each other in discourse comprehension. Although the authors of the dual-stream model have criticised the classical Broca-Wernicke view for its localisational and simplistic explanation of the neurobiology of language, the dual-stream model arguably suffers from similar shortcomings. Here, a variety of complex language processes were simplified and grouped into either 'speech perception' or 'production', rather than having their respective features inspected in detail in contextually meaningful settings.

## Distributed language regions

Are the proposed classical and dual-stream model regions the only frameworks to explore language and the brain? A growing body of literature has begun to inspect more specific and complex processes of language comprehension and individual differences. This work suggests that the neurobiology of language is more distributed and dynamic than what the classical or dual-stream model has presented (Price, 2012; Skipper, 2015a). Here we (non-exhaustively) review some examples of this work.

### *Individual differences*

Studies on individual differences in the anatomy of 'language' regions and the functional organisation of language processing have raised questions about current models of the neurobiology of language processing and the brain. For instance, studies have shown that there is a high amount of intersubject variability across 'language' regions, and that using an averaged 'language' map fails to correctly predict individual variations, especially in the case of aphasic patients' symptoms and recovery outcomes (Ojemann, 1979). For instance, studies have found that the between-subject variability in the cytoarchitecture of Broadmann areas 44/45 (i.e., regions mapping to Broca's area) is significantly higher than the cytoarchitectural within-subject variability (Amunts et al., 1999).

Since individual brains vary anatomically, and not only around 'language' regions, it follows that their functional activity patterns will vary as well, with implications for language processing (Juch et al., 2005). For instance 'language' regions were found to have some of the least functional overlap between subjects when compared to other brain regions (e.g., motor areas) (Frost & Goebel, 2012). Moreover, individual functional variability studies during performance of language tasks have shown that participants' individual frontal and temporal

activity peaks were heterogeneously and widely distributed, and that only the group averaged centre of mass fell within the 'language' regions proposed by neurobiology of language models (Burton et al., 2001). Further supporting this finding, intersubject variability studies on memory retrieval of words have shown that individual participants activate unique and distributed activity patterns, which include among others the supplementary motor area and prefrontal cortex, which relate to the participant's ability to 'visualise' words (Miller et al., 2012) or to individual cognitive strategies during the retrieval process (Heun et al., 2000). These variations are not due to noise, but could result from one of three features: (i) they represent a genetic feature of the individual, (ii) they relate to a difference in cognitive strategy, or (iii) they are a result of changes in contextual information (Seghier & Price, 2018).

*Semantics*

Individual differences represent only one source of variability during language processing. Another source is individual word meaning. Here, neuroimaging studies have also identified regions outside of 'language' areas that are important for processing semantic embeddings of words, revealing that word semantics map to brain regions based on the meaning that each word embodies or evokes. For instance, distinct semantic categories gave rise to very different activity patterns, particularly around sensorimotor areas, matching the perceptual and action meaning that the word embodied (Mitchell et al., 2008; Pulvermüller, 2013).

Supporting this embodied cognition theory, studies have demonstrated that processing of semantic embeddings of words involves activation of not only 'language' regions as 'semantic hubs', but also of sensorimotor regions when listening to words evoking action (e.g., 'kick'), of the olfactory cortex when listening to words related to smell (e.g., 'garlic'), of the visual cortex when listening to words evoking colours (e.g., 'red'), and of the auditory cortex when listening to words evoking sounds (e.g., 'telephone') (González et al., 2006; Kiefer et al., 2008; Martin et al., 1995; Tomasello et al., 2017). Moreover, when relating to time information describing past and present events, which are temporally concrete, these processes map onto visual and parahippocampal cortices usually associated with concrete object processing; whilst when describing future intentions, which are temporally abstract, these activate regions in the medial prefrontal cortex, temporo-parietal junction and posterior cingulate usually associated with the mentalizing network (Gilead et al., 2013). These distributed regions activate within

50-150 milliseconds of the word onset, meaning that they are not likely post-perceptual processes (García et al., 2019; MacGregor et al., 2012; Shtyrov et al., 2014).

Using voxel-wise modelling, Huth et al. showed in higher spatial detail that words belonging to the same semantic category activate unique activity patterns mapping to the perceptual region they relate to, revealing that the overall semantic map extends to most of the rest of the brain (Huth et al., 2016), such as the fusiform, hippocampus, pars orbitalis, cerebellum, superior and middle frontal gyri (Ghosh et al., 2010; Price, 2010).

*Formulaic and overlearned speech*

Another complex feature of language processing that current models of the neurobiology of language fail to address is that of formulaic expressions. Formulaic expressions are multi-word expressions that are overlearned, and these comprise a large portion of our everyday speech. They include, among others, idioms, proverbs, expletives, common speech formulas such as 'I don't know', and overused words or sentences that may vary from individual to individual (Van Lancker Sidtis & Sidtis, 2018). An extensive literature on aphasia has revealed that, even when large portions of the LH temporal lobe are damaged, such as in global aphasia, patients retain the ability to produce formulaic expressions, such as overlearned lists (e.g., Monday, Tuesday, Wednesday, etc.) and swear words (Bridges & Van Lancker Sidtis, 2013). Given that 'language' regions are extensively damaged and therefore cannot be processing formulaic speech in aphasia, three alternative routes have been proposed: (i) RH homologous 'language' regions, (ii) subcortical regions, or (iii) sensorimotor regions are involved in processing formulaic expressions (Sidtis, 2014; Van Lancker Sidtis, 2012).

There is weak support for the first, with evidence coming from studies on damage to RH homologous 'language' regions, showing that this results in significantly less use of formulaic language (Sidtis, 2014). In support of the second, patients suffering from Parkinson's disease (PD), whose subcortical areas are targeted by the disease, progressively lose the ability to produce formulaic expressions (Bridges et al., 2013; Lee & Van Lancker Sidtis, 2020; Van Lancker Sidtis et al., 2015). Moreover, patients suffering from Alzheimer's disease (AD), who have intact subcortical structures, produce significantly more formulaic expressions than healthy controls (Bridges & Van Lancker Sidtis, 2013; Van Lancker Sidtis et al., 2015).

In support of the third, in a recent study we tested the hypothesis that overlearned speech is processed by a uniquely distributed and less fixed network of brain regions. Here, we showed

that overlearned sentences are processed in sensorimotor regions before the typical hemodynamic response rises, and that overlearned sentences are predicted faster and more accurately than previously unheard sentences (Skipper et al., 2021). Moreover, 'language' regions were not involved in processing overlearned speech, with the brain's connectivity profile undergoing significant reconfiguration with increased learning (Skipper et al., 2021). As overlearned sentences are very common in everyday speech, our results suggest that language processing in the real world does not solely involve 'language' regions but is widely distributed, and that specific features of language should be investigated to probe the full extent of the neurobiology of language (Skipper et al., 2021).

In line with some of the above findings, the limited literature on expletives has provided further support for a distributed nature of language processing. Expletives constitute a more complex formulaic language feature, as these are not simply formulaic expressions but are also significantly driven by context. The same swear word, in fact, can have both a negative or positive emotional meaning, depending on the context in which it is uttered (Hansen et al., 2019). Some fMRI studies have revealed a distributed activity in anterior cingulate cortex, insula, and thalamus involved in producing and processing taboo words, with the IFG involved in modulating the emotional meaning and social context of swear words (Hansen et al., 2019; Sulpizio et al., 2019). MRI studies on patients suffering from Tourette's syndrome (TS) showed that the increased swearing in TS patients is likely a result of reduced IFG activity, basal ganglia dysfunction, and activity in the insula, thalamus, and cerebellum, pointing to an involvement of subcortical structures in processing swear words (Finkelstein, 2018; Van Lancker & Cummings, 1999). These observations suggest that processing of formulaic expressions happens away from 'language' regions and into sensorimotor and subcortical regions, depending on how much they are overlearned and on the context in which they are presented (Van Lancker Sidtis & Sidtis, 2018; Sidtis et al., 2018).

### *Aphasia*

Current models of the neurobiology of language are largely based on clinical observations of language deficits in aphasic patients, but they have nonetheless failed to consider what happens in the brain during recovery. Immediately after a stroke, spontaneous neuroplasticity starts a rewiring process that aids language recovery (Stemmer, 2015). This process happens heterogeneously in individual aphasic patients across large portions of both brain hemispheres, and leading to various degrees of recovery, suggesting that (i) recovery is

aided by distributed brain regions processing language, and (ii) these vary from person to person (Stemmer, 2015; Wilson et al., 2019). Some of the regions involved in these neuroplasticity processes are thought to involve subcortical and medial regions, such as precuneus and basal ganglia (Schevenels et al., 2020).

Although there is still large debate and little evidence on the exact brain regions that are involved in neuroplasticity after stroke (perhaps due to the heterogeneous nature of the recovery process) (Wilson, 2020), there are arguments in favour of a distributed network of language processing from studies on lesions to other brain regions that also cause language impairments. For instance, thalamic stroke causing lesions to the LH thalamic area also result in aphasia, demonstrating that subcortical regions are involved in some form of language processing (Fritsch et al., 2020). These regions are thought to help in attentional processes underlying the neurobiology of language (Crosson, 2013; Fritsch et al., 2020). Moreover, patients with cerebellar damage have also been reported to experience aphasic symptoms, with the cerebellum proposed to help in speech articulation, in temporal sequencing of language, or in predictive processing during language comprehension (De Smet et al., 2013; Mariën et al., 2014; Skipper & Lametti, 2021; van Dun & Mariën, 2016).

## Measures of central tendency

As we have seen, when studies investigate individual variability or more complex language features, they identify distributed language processing regions that paint a complex picture of the neurobiology of language. However, the question remains as to why 'language' regions consistently appear in task-based studies and meta-analyses (see Fig. 3). One obvious answer is that, as the dual-stream model suggests, these are the only true language processing regions, and what the distributed language literature has found are regions that simply share information with language processing ones (e.g., attention or mentalizing network). An alternative, however, is that 'language' regions appear as a result of the combined use of the subtractive method and central tendency measures, which pervade most of the neuroimaging literature.

Indeed, most of the existing neuroimaging literature has not considered (i) language as a complex behaviour, instead relying on simple stimuli/tasks and averaging methods, (ii) individual anatomical and cognitive differences between participants' brains, instead using the aggregate, or (iii) shared processes between language and other cognitive domains, instead using simple stimuli/tasks and subtractive methods.

For example, using simple stimuli/tasks and averaging over them only identifies shared brain regions across all of the conditions. Most likely these include: (i) primary auditory regions related to the listening task and (ii) some domain-general or low-level language processing region that activate with any language stimulus/task (e.g., STG, IFG, MTG).

Adding to this, most studies aggregate over participants to identify some shared patterns of activity. These stable regions are likely to be (i) primary auditory regions as they all hear the same stimuli, (ii) 'language' regions that consistently activate with any language task, and (iii) perhaps some domain-general cognitive strategy regions that are common to all participants. However, individual brains vary both anatomically and functionally (Burton et al., 2001; Frost & Goebel, 2012), indicating that aggregate methods (i) remove individual functional variability (Seghier & Price, 2018), and (ii) mistakenly assume that different individual brains can be accurately mapped onto an aggregate cytoarchitecture (Amunts et al., 1999).

Finally, using simple stimuli/tasks with subtractive methods would map the activity to small regions that are considered to be task-specific, revealing nothing about possible interactions with other cognitive domains that are still important to language processing. Instead, these complexities are likely to be ignored because the subtractive method presumes that the activity from the comparison task (e.g., nonwords) does not overlap with the language feature being investigated (e.g., words). However, the two tasks share some acoustic features and likely both require the involvement of other cognitive domains (e.g., attention, decision-making) for processing, as some studies have indeed shown (Mattheiss et al., 2018). Nevertheless, these overlapping features are considered to be irrelevant to language processing under the assumption that brain functional domains are functionally segregated: this has led to the tautology of using language localisers based on task subtraction to study the neurobiology of language (Blank & Fedorenko, 2020; Fedorenko et al., 2010; Pritchett et al., 2018).

These central tendency and subtractive methods have been used in nearly all the existing literature on language comprehension, along with simple stimuli and tasks. For instance, neuroimaging studies investigating lower-level lexical features, such as phonemes and syllables, typically present participants with stimuli such as 'be' and 'po' sounds and instruct them to press a button when they distinguish a difference (Goranskaya et al., 2016). These in no way represent the lexical features they are attempting to isolate (Skipper, 2015). Other examples of simple task-based methods involve asking participants to listen to single words and pseudo- or nonwords and subtracting the BOLD activity between these groups

(Braun et al., 2015). Again, this method reduces language processing to a segregated function, incorrectly assuming that the activity produced by processing nonwords should be completely separate from that of processing words (Mattheiss et al., 2018). Overall, these have led to a poor representation of language in all its complexity and richness.

## Hubs

Central tendency methods have contributed to the reduction of various language features' dynamic and distributed behaviour into a focus on only 'language' regions, as these represent a consistently shared area, perhaps due to their proximity to the auditory cortex that receives acoustic inputs. Although it may be tempting to conclude that 'language' regions must therefore be the only true language processing areas, a more plausible explanation is that they are a convergence zone where connections from other dynamic language processing brain areas pass through. It is thus likely that 'language' regions constitute high-connectivity hubs that control and direct a wider network of distributed and variable brain regions during language comprehension.

Hubs are defined as regions (or nodes) where a significant number of connections (or edges) go in and out of (Fornito et al., 2016). Hubs can relate to the structural white matter substrate, in which case they represent nodes of high connectivity for neural connections, or to the functional organisation, in which they represent nodes of high functional influence for the rest of the network (van den Heuvel & Sporns, 2013). Most of the research in this area has focused on global brain hubs, identifying these as the cingulate, precuneus, insula, superior frontal and superior temporal cortices in both functional and structural networks (De Domenico et al., 2016; Hagmann et al., 2008; van den Heuvel & Sporns, 2011). These are considered the 'backbone' of the brain network, directing communication and interaction for all other regions (Fornito et al., 2016). These findings seem to indicate that aside from the STG, no other 'language' region acts as a hub, neither in anatomical nor functional networks. However, recent neuroimaging studies have proposed a hierarchical organisation of hubs in the functional brain network, with the above-mentioned regions sitting at the top of the hierarchy, and other regions acting as weaker intermediary hubs, where they exert their influence at a more regional functional level (van den Heuvel & Sporns, 2013). These intermediary hubs have been classed in two categories: (i) either they help in connecting two functional modalities together so that they may share some processing and improve integration (i.e., connector hubs), or (ii) they have a fundamental role in directing information flow and processing within their own

functional domain, increasing coordination (i.e., provincial hubs) (Fornito et al., 2016; Hagmann et al., 2008; Joyce et al., 2010; van den Heuvel & Sporns, 2013).

In this hierarchical organisation of hubs in the brain, 'language' regions were found to act as provincial hubs across participants, in various language and non-language (e.g., motor learning) tasks (Bassett et al., 2013; den Ouden et al., 2012; Li et al., 2020). Studies on the functional connectivity of language processing are few, perhaps because most of the literature to date has assumed that the 'language' regions represent the full extent of the neurobiology of language. The limited existing studies focusing on whole-brain functional connectivity and individual network variability have identified a much more distributed and hierarchical 'language network', within which 'language' regions constitute the top of the regional hierarchy as hubs (Akiki & Abdallah, 2019; Hertrich et al., 2020). Because of their prominent appearance and role in causing aphasia, some of the lesion literature has also now begun to propose that 'language' regions are hubs, indicating that they have a central function in a wider neurobiology of language network (Mesulam et al., 2015; Migliaccio et al., 2016). Indeed, during the early stages of recovery, the brain attempts to rewire around 'language' areas at first and extends to other brain regions in later stages (Schevenels et al., 2020; Stemmer, 2015; Wilson et al., 2019). Overall, this evidence tentatively points to 'language' regions as being part of a hierarchy of hubs with the STG as a global brain hub, and IFG, MTG, etc. as intermediary hubs, with putative other regions around these yet to be properly defined.

## Network organisation of the brain

From this follows that 'language' regions result from averaging and subtracting methods likely because (i) they are a commonly activated feature of all language stimuli/tasks, and (ii) some of the 'language' regions are global hubs (e.g., STG), with the others likely connecting and distributing around these to form intermediary structures. It seems clear, then, that if we want to understand the role of 'language' regions as putative hubs, as well as other regions' contributions to language processing, a better approach to investigating the neurobiology of language may be through the use of graph theory. I will first provide a brief overview of the main network theory measures.

### *Graph theory*

Graph theory aims at describing how elements (e.g., people, brain regions, proteins, etc.) are connected through mathematical models (Rombach et al., 2014). There are three layers

at which a network can be described: global, mesoscale, and local (Rombach et al., 2014). Intuitively, global network measures provide a general overview of the network's properties as a whole: for instance, global efficiency measures the network's ease in communicating across all points (Preti et al., 2017). The mesoscale organisation represents the division into components of the network: for instance, the network may be partitioned into functionally segregated components, namely communities (Newman, 2006). Finally, the local level describes how each node behaves within the network: for instance, one could measure how many connections pass through any given node, namely node centrality (Borgatti & Everett, 2006).

Although the global measures may provide some insights on differences between networks, such as comparing healthy control and patient networks, most of the existing biological research focuses on the mesoscale and local measures, as they provide more detailed representations of a network. Two main algorithms are used to assess the mesoscale organisation, namely community partitioning and core-periphery (Rombach et al., 2014). Community partitioning is an ongoing issue in network theory, given the difficulty in partitioning a network into correct functional communities without any prior knowledge; depending on the algorithm and parameters chosen to partition a network, the resulting community organisation may vary (Fortunato & Barthélemy, 2007). Among the many community partition algorithms, the most widely used and accepted is the Newman-Girvan modularity maximisation algorithm, which separates communities if their intra-connectivity is significantly higher than if they were to be joined to another community (Newman, 2006).

Core-periphery algorithms, instead, separate the network into two components, whereby cores must have high intra- and inter-connectivity, while the periphery should be dynamic and loosely connected to the rest of the network (Borgatti & Everett, 2000). Despite its seemingly simple definition, core-periphery algorithms are scarce, or they inaccurately estimate cores using proxy measures such as node centrality (Borgatti & Everett, 2006; da Silva et al., 2008).

Among centrality measures, the most widely used are: (i) degree, which measures overall connections of a node; (ii) eigenvector, which measures influence of a node with respect to its neighbours; (iii) betweenness, which measures how important a node is for bridging groups of other nodes; and (iv) closeness, which measures how easy to reach a given node is

(Barucca et al., 2015; Telesford et al., 2011). How have these measures been used in neuroimaging?

*Network neuroscience*

Most of the neuroimaging literature focusing on the functional connectivity of the brain is based on group-averaged resting-state networks (RSNs), which are collected in the absence of a task (or rather, the participant is left lying in the scanner) (Sporns, 2013). Research on RSNs has found that the brain network contains highly stable global hubs, as previously described (De Domenico et al., 2016; Hagmann et al., 2008; van den Heuvel & Sporns, 2011). These hub regions have important roles for a network's integration and communication (van den Heuvel & Sporns, 2013). At the mesoscale level, five to eight functional communities were identified in RSNs grossly mapping to regions such as central regions, parieto-frontal regions, medial-occipital areas, fronto-temporal and lateral-occipital cortices, matching partitions of the anatomical network (Chen et al., 2008; Meunier et al., 2010). These functional communities were shown to be relatively immutable over time (Hutchison et al., 2013). The overwhelming consensus from the resting-state network studies seems to be that (i) brain regions have specialised into highly segregated functional modules (i.e., having high modularity), (ii) stable hub regions promote integration between these communities, and (iii) the brain network is highly static.

Is the brain segregated, integrated, or both? And is the brain static? Recent task-based studies on functional connectivity have shown that as we move from rest to more complex tasks, the architecture of the network changes dramatically to become less segregated and more integrated (Shine et al., 2015; Yue et al., 2017). Moreover, these studies demonstrated that the brain architecture is much more flexible than previous RSN studies reported; here, modularity is not a static feature, but is highly related to functional integration (Park & Friston, 2013). This means that although certain regions are more likely to perform a given function, they are not necessarily bound to it; their role depends on how to best minimise energy requirements and increase efficiency for the entire network (Bassett & Bullmore, 2006). These dynamics are tightly controlled and are important for the proper functioning of the brain. They aid during learning and neurodevelopment by undergoing significant rearrangements (Bassett et al., 2011; Gu et al., 2019), but when disrupted may lead to onset of disease (Alexander-Bloch et al., 2012; de Haan et al., 2012).

Control of these dynamics is made possible through a hierarchy of hub regions that act as connectors between two or more communities and are highly flexible (Sporns, 2013). One study on overlapping community organisation found that connector hubs allow communities to temporarily share functions, doing so by increasing integration at the overlap between two or more modules (de Reus et al., 2014).

Although still developing, the task-based fMRI literature is already offering a new perspective on the brain network, indicating that it is more flexible and that it balances both integration and segregation, fluctuating between these two features depending on specific tasks or changes in demand (Cole et al., 2013; Friston & Price, 2011; Shine et al., 2015; van den Heuvel & Sporns, 2013). These findings are further supported by recent research on individual brain networks, which have identified significant restructuring of the connectivity and communities of the brain during changes in cognitive states, during individual decision or cognitive strategies, during neurodevelopment, and during learning (Barnes et al., 2014; Feilong et al., 2018; Gu et al., 2019; Kong et al., 2019; Salehi et al., 2018, 2019; Seghier & Price, 2018; Vindras et al., 2012).

### *Network models of language*

What can these findings about the brain network tell us about language processing? Since the brain is highly dynamic, separated into communities that can evolve over time with tasks and cognitive demands, and these are connected and controlled by a hierarchy of hubs, then it is likely that the neurobiology of language is (i) not static, (ii) not limited to 'language' regions, and (iii) composed of a mixture of both hubs and dynamic regions. However, existing models of the neurobiology of language do not account for this dynamic and complex behaviour, rather perpetuating static and localised views of language processing because of the use of central tendency measures.

The only alternative model to the dual-stream that investigates language processing in some detail through functional connectivity has proposed the existence of a 'language network' that is organised in a core-periphery organisation that maps to the same 'language' regions as the dual-stream model (Fedorenko & Thompson-Schill, 2014). Here, Fedorenko and colleagues proposed that LH 'language' regions acted as the core of the 'language network', with the RH 'language' regions being a periphery (Chai et al., 2016; Fedorenko & Thompson-Schill, 2014). The idea of a 'language network' is gaining support in the literature and has been mentioned

in ~10,000 studies, based on a Google search. Although this work has clearly begun the conversation on possible alternative models of the neurobiology of language using a network framework, it is based on few assumptions: (i) it defines the boundaries of the 'language network' using simple artificial task-based language localisers assuming that these are enough to reflect the complexity of natural language processing, and (ii) it uses subtractive techniques that assume functional specificity. This means that currently we still lack a model of the neurobiology of language that can explain the complexity and variability of language processing in the real world. Here, we review three candidate models.

One possible network organisation supporting language is modularity. Modularity, through its functional segregation properties, would help explain why 'language' regions (e.g., STG, MTG, and IFG) have a clear propensity for language processing functions and appear in all neuroimaging studies. Since modular organisations can still evolve over time, modularity can also identify other regions involved in language processing as a function of task, by mapping regions that join into language modules over time. However, because modularity seeks to find some independence among communities, it falls short on addressing potential shared processes (i.e., overlaps) between communities (Gu et al., 2019). These complex relationships between communities could be important for predicting cognitive demand and context-related changes.

Another alternative is a core-periphery architecture. Core-periphery architectures have significant evolutionary advantages: (i) a highly connected core allows integration of information across the network, (ii) through its redundancy in connections it increases robustness to perturbation; meanwhile (iii) dynamically changing peripheries allow increased variability (Faber et al., 2019; Fornito et al., 2016; Stefaniak et al., 2020). These properties of core-periphery networks could help explain how the brain supports individual differences and neuroplasticity. Moreover, it was shown that, due to their high wiring, large lesions to core nodes result in much more deleterious effects than damage to peripheral nodes (Fornito et al., 2016; Zhao et al., 2011). If 'language' regions are part of the core, this would explain why damage to these regions causes severe language impairments. Although core-periphery addresses more dynamic and distributed behaviours, it cannot explain why brain regions preferentially assume certain functions.

A final organisation may involve modularity and core-periphery simultaneously. Indeed, in various scale-free networks, it was shown that multiple mesoscale organisations tend

to co-exist (Rombach et al., 2014). This may be also true of the brain. Here, a joined modularity and core-periphery organisation would help address the following: (i) individual variability, (ii) neuroplasticity after injury, (iii) distributed language regions, and (iv) the role of 'language' regions. Only one neuroimaging study, to the best of our knowledge, has inspected these two architectures simultaneously. Here, it was demonstrated that a joined mesoscale organisation helps predict neurodevelopment and individual differences much better than a given single measure (Gu et al., 2019).

## Overarching hypotheses

In the present thesis we therefore hypothesise that the brain is organised in a core-periphery and modular architecture that supports language processing in the real world. We propose that language processing is highly dynamic and distributed across the brain, with 'language' regions appearing in the aggregate because they act as cores or hubs.

For each result chapter, the overarching hypotheses are as follows:

1. Result chapter 1: The brain creates unique and distributed sensorimotor representations of individual characters in movies. These unique representations are reactivated during resolution of pronominal references.
2. Result chapter 2: Individual sensorimotor embeddings of words form distributed activity patterns encompassing most of the brain. This variability has been obscured by central tendency measures, which only identify 'language' regions as these are stable hubs.
3. Result chapter 3: The brain is both a core-periphery and modular network where 'language' regions are intermediary cores connecting to a wide and dynamic periphery, together sharing language processing functions.

## Thesis organisation

The present thesis is organised in three subsequent chapters, of which one will describe three original research studies testing our hypotheses. Below are brief overviews of each chapter's contents:

- **Chapter 2**: Here we will introduce a new open-source dataset, namely the Naturalistic Neuroimaging Database (NNDb), that was collected to help address language as a complex behaviour, which is currently missing in existing neurobiology of language

models. We will describe the data collection and preprocessing methods involved in creating the NNDb, which is now one of the largest and most varied open-source naturalistic fMRI datasets available.

- **Chapter 3.1**: Here we aimed at investigating the distribution of the neurobiology of language during a specific language feature, namely pronoun resolution. We show that the brain builds unique perceptual representations of a character through sensorimotor (mainly auditory and visual cortices) regions. We further demonstrate that these activity fingerprints are later reactivated during retrieval and resolution of pronoun references to the specific character.
- **Chapter 3.2**: Here we further investigate the extent of distribution of the neurobiology of language during processing of sensorimotor embeddings of words. We show that language processing recruits most of the rest of the brain, forming unique activity patterns for each sensorimotor embedding. We further show that aggregate methods result in 'language' regions, because these act as connectivity hubs. We finally demonstrate that 'language' regions connect directly to the distributed sensorimotor embedding regions.
- **Chapter 3.3**: Here we investigate a network architecture that best supports language processing in the real world. We propose a novel joined mesoscale architecture of the brain, with a simultaneous multi-core-periphery and modularity organisation, whereby known language regions form intermediary cores and link to a wide and dynamic periphery, together forming one or more communities in individual subjects at different times. We then reproduce findings from the RSN literature using group-averaged networks, and show that these are devoid of any individual variability and do not represent any single participant's network.
- **Chapter 4:** Here we discuss in more depth how our model of the neurobiology of language processing supports language in the real world, taking into account individual variability and language as a complex behaviour.

# Chapter 2: Methods


## Abstract

Neuroimaging has advanced our understanding of the neurobiology of language using simple and artificial stimuli that do not account for the complexity and richness of the real world. To address these methodological limitations, we collected and made publicly available the '*Naturalistic Neuroimaging Database*' (NNDb) to allow for a more complete understanding of the neurobiology of language and the brain, as well as other cognitive domains, under more ecological settings. Eighty-six participants underwent behavioural testing and watched one of 10 critically acclaimed full-length movies during functional magnetic resonance imaging. The timeseries were preprocessed using standard neuroimaging techniques and the resulting data is shown to be of high quality. The NNDb can be used to answer questions previously unaddressed with standard neuroimaging approaches, progressing our knowledge of how language and the brain operate in the real world.


## Introduction

For centuries neuroscientists have attempted to investigate how the brain supports language processing, arguably the most complex human behaviour. Progress towards understanding the neurobiology of language and the brain has been made using task-based functional magnetic resonance imaging (fMRI), and more recently using resting-state networks coupled with task-based meta-analyses. Here, research has identified a set of regions mostly mapping to inferior frontal and temporal lobe cortices, that are considered the loci for speech perception and production in the brain (Hickok & Poeppel, 2007). While these studies have led to a number of important discoveries and have started the conversation on mechanisms underlying language processing, we review evidence indicating that more naturalistic stimuli and tasks are required to understand the complexities and contextual dependencies of language in the real world.

### *Task-fMRI*

For task-fMRI, general behavioural processes are decomposed into discrete component processes that can theoretically be associated with specific activity patterns from linguistic features. To ensure experimental control and because of reliance on the subtractive method (Friston et al., 1996), these putative components are studied with stimuli that often do not

resemble things participants might naturally encounter and language tasks they might actually perform in the real-world (a topic long debated) (Brunswik, 1943, 1955; Neisser, 1976). For example, language comprehension has been broken down into component processes like phonology and semantics. These individual subprocesses are largely localised in the brain using isolated auditory-only 'speech' sounds (like 'ba') in the case of phonology and single written words in the case of semantics (Skipper, 2015b). Participants usually make a meta-linguistic judgement about these stimuli, with a corresponding button response (e.g., a two-alternative forced choice indicating whether a sound is 'ba' or 'pa').

The result of relying on these 'laboratory style' stimuli and tasks is that our neurobiological understanding of language derived from task-fMRI may not be representative of how the brain processes linguistic information in the real world. This is perhaps one reason why fMRI test-retest reliability is so low, with an average intraclass correlation (ICC) of 0.1-0.5 across various studies and fMRI setups (Bennett & Miller, 2010; Gorgolewski et al., 2013, Elliot et al., 2020). Indeed, more ecologically valid stimuli like movies have higher reliability than resting- or task-fMRI, with studies showing significantly higher intersubject correlations and lower head motion in movie paradigms (ICC > 0.7) (Vanderwal et al., 2015; Wang et al., 2017). This is not only because they decrease head movement and improve participant compliance (Greene et al., 2018; Madan, 2018; Vanderwal et al., 2015). Rather, naturalistic stimuli have higher test-retest reliability mostly because they are more representative of operations the brain normally performs and provide more constraints on processing (Burton et al., 2001; Chen & Small, 2007; Miller et al., 2002, 2009; Vanderwal et al., 2017; Wang et al., 2017).

*Resting-fMRI*

There has arguably been a significant increase in our understanding of the network organization of the human brain, and how this may support complex functions such as language processing, because of the public availability of large resting-fMRI datasets, analysed with dynamic functional connectivity measures coupled with task-based meta-analyses (Bullmore & Sporns, 2009; Preti et al., 2017). These include the INDI '1000 Functional Connectomes Project' (Biswal et al., 2010), 'Human Connectome Project' (HCP) (Van Essen et al., 2013) and UK Biobank (Miller et al., 2016). Collectively, these datasets have more than 6,500 participants laying in a scanner 'resting'. Resulting resting-state networks are said to represent the 'intrinsic' network architecture of the brain, i.e., networks that are present even in the

absence of explicit tasks. These networks are often claimed to be modular, meaning they support segregated functions (Gonzalez-Castillo & Bandettini, 2018).

As participants are left lying in the scanner, they are likely switching between staying awake, mentalizing, trying not to think and engaging in inner speech (Gonzalez-Castillo & Bandettini, 2018; Hurlburt et al., 2015). Thus, resting-fMRI cannot be truly considered at 'rest'. Though some of these behaviours are 'natural' (e.g., mind-wandering), unlike task-fMRI, there is no verifiable way to label resulting regional or network activity patterns (Sonkusare et al., 2019; Vanderwal et al., 2019). At best, reverse inference through meta-analyses is used to obtain gross labels, such as 'auditory' and 'attention' networks (Skipper & Hasson, 2017; Smith et al., 2009; Tahedl & Schwarzbach, 2020). Despite claims that these 'intrinsic' networks constrain the connectivity of task-fMRI networks, it is increasingly suggested that this is not necessarily so (Gonzalez-Castillo & Bandettini, 2018). The brain becomes less modular with task (Di et al., 2013), particularly with increased task demands (Braun et al., 2015; Kitzbichler et al., 2011; Vatansever et al., 2015). Indeed, up to 76% of the connections between task- and resting-fMRI differ (Kaufmann et al., 2017). Furthermore, more ecological stimuli result in new sets of networks that are less modular and only partly constrained by resting networks (Kim et al., 2018; Simony et al., 2016). For instance, during a natural vision fMRI study, functional networks behaved significantly more dynamically, through splitting and merging the networks observed during resting state (e.g., the dorsal attention resting state network split into two clusters during natural vision), forming new functional groupings that varied with the changing cognitive demands (Kim et al., 2018).

### *Naturalistic-fMRI*

Based on considerations like these, there is a growing consensus that taking a more ecological approach to neuroscience might increase our understanding of language and the brain, as well as other brain behaviours (Eickhoff et al., 2020; Hasson et al., 2010; Hasson & Honey, 2012; Krakauer et al., 2017; Maguire, 2012; Matusz et al., 2019; Olshausen & Field, 2006; Skipper, 2015b; Spiers & Maguire, 2007; Vanderwal et al., 2021; Varoquaux & Poldrack, 2018). This includes conducting more neuroimaging studies with 'naturalistic' stimuli. Similar to prior definitions (Bottenhorn et al., 2019; Sonkusare et al., 2019), 'naturalistic' might be defined on a continuum from short, static, simple, decontextualised, repeating, unisensory stimuli with low ecological validity (as described above) to long, dynamically changing, complex, contextualised, continuous, typically multisensory stimuli

with high ecological validity. We identified at least 16 (DuPre et al., 2019) (with more being added (di Oleggio Castello et al., 2020)) publicly available fMRI datasets using 'naturalistic' stimuli more on the latter end of this continuum. However, there are no datasets with a large number of participants, long naturalistic stimuli and stimulus variability. Although most studies collect data from ~20 participants, it was shown that small sample sizes have low statistical power, meaning inflated rates of false negatives (Lohmann et al., 2017). Moreover, fMRI studies on effect sizes of samples ranging from 20-80 subjects replicating typical group-level analyses (e.g., general linear models), showed that a minimum of 40 subjects are necessary for detecting high effect sizes and a minimum of 80 participants are required for detecting medium effect sizes and producing replicable task-fMRI results (Geuter et al., 2018; Turner et al., 2018). Naturalistic-fMRI datasets with larger participant numbers tend to use short (~10 minute) audio or audiovisual clips. However, studies on functional connectivity during naturalistic viewing indicate that for single subject studies a minimum of 25 min scan time are preferred, with continued significant improvement in test-retest reliability over scans up to 4 hr long (Anderson et al., 2011). Moreover, longer scanning sessions, from 1.5 hr to multiple daily sessions, showed significantly higher intraclass correlation values in resting state data as well (Gordon et al., 2017; Laumann et al., 2015; Xu et al., 2016).

Longer duration fMRI datasets using more naturalistic stimuli have a small number of participants and one stimulus (though see (Nastase et al., 2019)). These include 11 people watching 'Raiders of the Lost Ark' (Haxby et al., 2011) and 20 listening to an audio description of 'Forrest Gump' during fMRI (Hanke et al., 2014, 2016). These datasets have currently mostly been used to develop and test novel analytical models for neuroimaging data, such as hyperalignment for individual subject functional network analyses (Haxby et al., 2011). However, with only one movie, generalisability is limited. More movies would not only increase generalisability to specific behaviours, but they would also increase the variety of stimuli and contexts to better inspect individual variability of language features. These could then be used to label finer grained patterns of activity, e.g., making machine learning/decoding approaches more feasible (Combrisson & Jerbi, 2015; Khosla et al., 2019; Varoquaux, 2018).

Indeed, there is no a priori reason participants need to watch the same movie (or listen to the same audio). Existing long datasets might use one stimulus because intersubject correlation is a commonly used method for analysing fMRI data from more naturalistic stimuli that are difficult to model (Hasson et al., 2004). Though this is a powerful 'model-free' approach (for an overview, see (Nummenmaa et al., 2018)), it requires participants to watch

the same movie. However, ‘model-free’ (more data-driven) methods like independent component analysis (Bartels & Zeki, 2004), regional homogeneity (Zang et al., 2004), hidden Markov model (Baldassano et al., 2017) and dynamic mode decomposition (Casorso et al., 2019) and more model-based analysis involving convolution/deconvolution, can be also done at the individual participant level with *different* movies. This would increase generalisability and the possibility of more detailed analyses through more varied stimulus annotations.

*NNDb*

To fill these gaps in publicly available data, we collected and made publicly available a '*Naturalistic Neuroimaging Database*' (NNDb) from 86 people who each completed a battery of behavioural tests and watched a full-length movie during movie naturalistic-fMRI. We sought to reach a balance that promotes generalisability, allows a large variety of linguistic features and events to be investigated and supports studies on individual variability as well as intersubject correlations. To achieve this, our participants watched 10 different movies from 10 different genres they had not previously seen. This is because studies have shown that repeated movie viewings might change the functional network architecture of the brain (Andric et al., 2016). We validated that the data was of high quality, and that this increased further with preprocessing, by calculating voxel-wise temporal signal-to-noise ratio (tSNR) and inter-subject correlation (ISC) across and within movies. tSNR on fully preprocessed data ranged between 13.37-98.03 ($M = 63.82$, $SD = 20.79$). Moreover, similar to prior work, the maximum ISC was $r = 0.28$; when the entire dataset was split in half, the results were largely spatially indistinguishable from each other ($r = 0.96$) (for specific details and figures, see original publication (Aliko et al., 2020)). Fig. 4 provides an overview of the study and preprocessing steps.

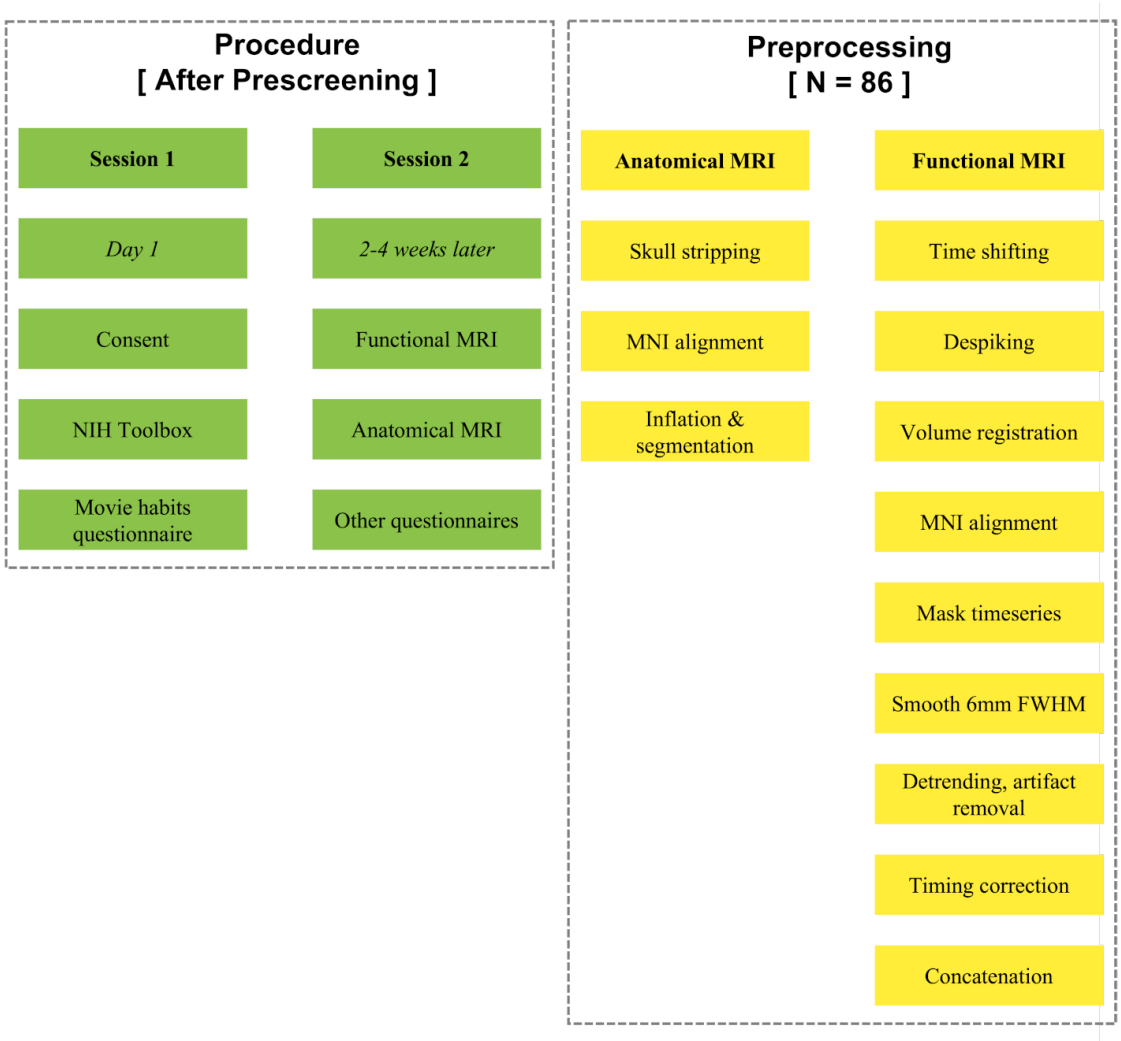



**Figure 4.** Schematic overview of the naturalistic neuroimaging database procedures, preprocessing and data validation. Procedures (green) occurred over two sessions separated by about three weeks on average. Session one consisted primarily of a battery of behavioural tests to quantify individual differences. In session two, functional magnetic resonance imaging (MRI) was acquired while participants watched one of 10 full length movies followed by anatomical MRI.

Data discovery is nearly unlimited with the NNDb as there are a vast number of annotations that can be made from the movies and approaches to analysis, thus supporting studies investigating various brain behaviours in the real world. This includes more than replicating prior findings with more ecologically valid stimuli. That is, there are a number of broad open questions that the NNDb can be used to address for the first time, like the systematic study of how context is used by the brain (Skipper, 2015b). Given the lack of robust neuroimaging biomarkers for mental illness (Boeke et al., 2019; Kapur et al., 2012), the NNDb might also help increase the pace of clinically relevant discovery, e.g., by uncovering labelled network patterns that predict individual differences (Eickhoff et al., 2020).

## Methods

*Participants*

The initial goal for the NNDb was to collect fMRI data of 84 participants watching 10 full-length movies from 10 different genres. Specifically, we set out to collect data on 18 subjects for 2 of the movies, and on 6 subjects for the remaining 8 movies (2x18 + 6x8 = 84 subjects), based on sample size considerations reviewed above (Anderson et al., 2011; Geuter et al., 2018; Gordon et al., 2017; Laumann et al., 2015; Lohmann et al., 2017; Turner et al., 2018; Xu et al., 2016) and to have stimulus variability. Overall, the rationale was to have power for across movies analyses (thus 84 participants), but at the same time have two larger datasets to test hypotheses on a more typical sample of participants.

To reach 84 individuals, we recruited 120 possible participants using the pool management software (http://www.sona-systems.com/). The following inclusion criteria were applied: no permanent metal implants, right-handedness, native English speaker, no claustrophobia, no history of psychiatric or neurological disorders, not taking medication, without hearing impairment and unimpaired or corrected vision. Furthermore, we ensured that each participant had not seen at least two of the 10 movies. From this initial screening, out of the 120 recruits, 91 met the inclusion criteria. Two of the 91 recruits were excluded from the main NNDb dataset as they were determined to be left-handed after all, but were later added in the event that other researchers had less stringent inclusion criteria for their studies; two participants were excluded because they requested to stop the scan, and one had low data quality due to excessive motion.

This resulted in a final dataset of 86 participants (42 females, range of age 18–58 years, M = 26.81, SD = 10.09 years). These were pseudo-randomly assigned to a movie they had not previously seen, (usually) from a genre they reported to be less familiar with in a pre-scan questionnaire. Table 1 provides a summary of participant demographics by movie. At the conclusion of the study, participants were paid £7.5 per hour for behavioural testing and £10 per hour for scanning to compensate for their time (receiving ~£40 in total). The study was approved by the ethics committee of University College London (Project ID FMRI/2013/002) and participants provided written informed consent to take part in the study and share their anonymised data.

| N | Movie | Age | % | | | ≥ Bachelor's (%) | |
|---|---|---|---|---|---|---|---|
| | | | Female | BAME | Monolingual | Participant | Mother |
| 20 | 500 Days of Summer | 27.70 | 50.00 | 30.00 | 85.00 | 75.00 | 60.00 |
| 18 | Citizenfour | 27.00 | 50.00 | 41.18 | 77.78 | 61.11 | 66.67 |
| 6 | 12 Years a Slave | 27.17 | 50.00 | 66.67 | 50.00 | 50.00 | 50.00 |
| 6 | Back to the Future | 22.17 | 50.00 | 40.00 | 66.67 | 66.67 | 83.33 |
| 6 | Little Miss Sunshine | 35.67 | 33.33 | 66.67 | 66.67 | 50.00 | 16.67 |
| 6 | The Prestige | 34.17 | 50.00 | 0.00 | 100.00 | 83.33 | 33.33 |
| 6 | Pulp Fiction | 22.67 | 50.00 | 83.33 | 66.67 | 33.33 | 0 |
| 6 | The Shawshank Redemption | 22.17 | 50.00 | 100.00 | 50.00 | 50.00 | 83.33 |
| 6 | Split | 22.67 | 50.00 | 50.00 | 83.33 | 66.67 | 33.33 |
| 6 | The Usual Suspects | 23.17 | 50.00 | 66.67 | 83.33 | 100.00 | 33.33 |

**Table 1.** Description of participants in the *NNDb*. N is the sample size for each of the 10 movies. In total, 86 participants were included in the final dataset. Age is expressed as the average in each movie. Gender is expressed as percent (%) female. Ethnic diversity is expressed as percent Black, Asian and Minority Ethnic (BAME). Most participants were monolingual English native speakers. Educational attainment of both the participant and the participant's mother is expressed as percent with a Bachelor's degree or higher.

*Procedure*

The 86 participants attended two sessions on separate days. During session one, participants completed the cognitive and emotional batteries and a hearing test from the sensory battery of the National Institute of Health (NIH) Toolbox (Gershon et al., 2013). These represent standardised tools and tests for assessing individual neurological and behavioural functions, such as working memory and language (Gershon et al., 2013). Some tests, such as motor and other sensory tests, were excluded from the Toolbox as they required more complex setups and physical implementations (e.g., walking). Moreover, we collected demographic data, such as age, ethnicity, educational attainment etc. (see Table 1). Participants completed the NIH tests in a sound shielded testing room using headphones and an iPad. At the end of the NIH tests, participants completed a questionnaire on movie habits that was used to determine which of the 10 movies they would watch in the scanning session.

After 2-4 weeks (M = 20.36 days; SD = 23.20), participants were invited for the second session, consisting of (i) functional and (ii) anatomical MRI scans and (iii) a final questionnaire related to aspects of the movie they watched. Except for one case, this was the order in which the session was conducted. Prior to entering the scanning suite, participants reporting corrected vision were provided with MRI-safe glasses, matching their eye prescription to the nearest 0.5 value. Once in the scanning suite, participants chose a comfortable earbud size to be fitted on the noise-attenuating headphones. Next, participants were placed in the head-coil with pillows under the head, covering the ears and under the knees for both comfort and to reduce movement during the scans. Participants were then asked to select a comfortable and clear audio-stimulus volume. Participants were given a bulb in their right hand and instructed to squeeze if something was wrong, or they needed a break during the movie. They were instructed to stay as still as they could throughout scanning as movement would render the scans unusable.

fMRI movie scans were acquired with one to three breaks on average, depending on the length of the movie: longer movies had more breaks. During breaks, participants were told that they could relax but not move. To ensure that this was the case, and that participants were awake and comfortable we monitored participants via a camera over their left eye. If they appeared drowsy or seemed to move too much during the movie, we gave them a warning over the intercom by producing a beep or speaking to them. In rare cases we stopped the scan to speak with the participant. After the movie, an anatomical scan was collected, and once out of the scanner participants filled out a questionnaire. Finally, participants were paid and sent home.

### *Movie Stimuli*

The movies were selected from 10 different genres, in order to have varied stimuli in the final dataset. Criteria for selection included: having an average critical acclaim score of >70% on publicly available metrics of success (e.g., IMDb, Rotten Tomatoes) and having received nominations for cinematic awards (e.g., Academy Award). Table 2 provides an overview of the 10 movies participants watched during fMRI.

| Movie | Genre | Year | Length (sec) | Scores (%) | | |
|---|---|---|---|---|---|---|
| | | | | IMDb | Meta | RT |
| 500 Days of Summer | Romance | 2009 | 5470 | 77 | 76 | 85 |
| Citizenfour | Documentary | 2014 | 6804 | 81 | 88 | 96 |
| 12 Years a Slave | Historical | 2013 | 7715 | 81 | 96 | 96 |
| Back to the Future | Sci-fi | 1985 | 6674 | 85 | 86 | 96 |
| Little Miss Sunshine | Comedy | 2006 | 5900 | 78 | 80 | 91 |
| The Prestige | Thriller | 2006 | 7515 | 85 | 66 | 76 |
| Pulp Fiction | Action | 1994 | 8882 | 89 | 94 | 94 |
| The Shawshank Redemption | Drama | 1994 | 8181 | 93 | 80 | 91 |
| Split | Horror | 2016 | 6739 | 73 | 62 | 76 |
| The Usual Suspects | Crime | 1995 | 6102 | 86 | 84 | 95 |

**Table 2.** Description of the movies used in the *naturalistic neuroimaging database*. Ten full length movies were chosen from 10 genres. These were required to have been successful, defined as an average *Internet Movie Database* (IMDb, https://www.imdb.com/), *Metacritic* (Meta, https://www.metacritic.com/) and *Rotten Tomatoes* (RT, https://www.rottentomatoes.com/) score greater than 70%. IMDb scores were converted to percentages for this calculation. Movie lengths are given in seconds (s), also equivalent to the number of whole brain volumes collected when participants watched these movies during functional magnetic resonance imaging.

All movies were purchased and stored as '.iso' files. Relevant sections of the DVD (i.e., excluding menus and extra features) were directly concatenated to an 'mpg' container using the command:

*ffmpeg -i concat:VTS_01_1.VOB\| ... VTS_01_8.VOB -c copy -f dvd movie_name.mpg*

Where '-c' copies the codec and '-f' specifies the DVD format. The DVD video size and quality are as follows:

- Video (codec): MPEG-PS
- Audio (codec, sampling rate, bits per sample, channels): AC-3, 48.0 kHz, 16, 6
- Resolution (pixels): 720 x 576 (except Citizenfour which was 720 x 480)

- Aspect Ratio: 16:09 (except The Usual Suspects and Pulp Fiction which were 2.40:1 4:3, respectively)
- Frame rate (fps): 25 (except Citizenfour which was 23.976)

The audiovisual files were screened using full-screen mode through a mirror reversing LCD projector to a rear-projection screen measuring 22.5 cm x 42 cm with a field of view angle of 19.0°. The screen was placed at the back of the MRI bore and was reflected through a mirror above the participants' eyes. High quality audio was presented in stereo via a Yamaha amplifier through Sensimetrics S14 scanner noise-attenuating insert earphones (https://www.sens.com/products/model-s14/).

**Movie Pausing**

Movies were played with as few breaks as possible to allow for a natural viewing experience, avoid misalignments between one scan and the other due to participants moving, and reduce discontinuity in the hemodynamic response. Additionally, continuous viewing reduces chances of either hardware or human error in matching the movie on the computer to the stimulus presented to the participant, and therefore allows to match movie features to brain responses. Since the scanning sequence we used required breaks at least every 1 hr (see 'Acquisition' below), we played each movie in ~45 min segments, identifying points in the plot where no important action nor dialogue was happening. In rare cases, the participants signalled to stop themselves, in which case we would later still stop at the predetermined breaks to maintain all datasets as similar as possible. To maintain continuity and allow for these stopping times, we created a script using an Arduino device to allow us to stop the scanner and pause the movie at any time and resume where the movie left off when the scanner was restarted.

Specifically, a Linux BASH script opened movies using '*MPlayer*' (http://www.mplayerhq.hu/). The script then went into a state of waiting for a TTL (transistor-transistor logic) pulse from the scanner, which would indicate that scanning had begun. Pulses were received through a USB port connected to an Arduino Nano programmed to read and pass TTL pulses from the scanner to the script. When the scan sent the first TTL pulse, eight seconds were allowed to elapse before the movie began to play, to let the scanner reach a state of equilibrium. When the scanner was paused, the movie pausing BASH script stopped the movie within 100 ms: this was because the script monitors for TTL pulses every 50 ms, but if an initial pulse was not registered, the script required that the next pulse also did not arrive thus reaching

100 ms delay. When the scan was restarted, eight seconds were again allowed to pass before the movie was played.

Moreover, the scanner software dropped the last brain volume whenever a movie was paused, leading to up to one second (= 1 TR) to be lost from the fMRI timeseries. To solve this problem, two versions of the script were created as follows:

1. The movie picked up from where it left off affecting N = 29 or 33.72% of participants.
2. The movie was rewound by the amount of time lost when the volume was dropped. To calculate this, the script used three output files that it generated when running: a *MPlayer output* file, *current time* file and *final output* file.

Every 50 ms when the TTL pulse was received, the script would send a command to *MPlayer* to get the time position in the movie, which was provided as a value up to one decimal and stored in the *MPlayer output* file. The script would then read the last line of the *MPlayer output* file and write a new line in the *current time* file. Here, every line consisted of (i) a timestamp formed by the elapsed milliseconds from the end of the previous second in Linux epoch format, and (ii) the newly acquired time position in the movie.

If paused, the movie was then rewound by the amount in (i) by passing a command to *Mplayer* through 'slave' mode. When the scanner was restarted, the movie began within 100 ms of the first TTL pulse (again, because it had to receive at least two pulses). Due to a coding error, version (2) of the script occasionally fast forwarded when it should have been rewound in N = 13 or 15.12% of participants. Because fast forwarding could not be greater than one second and the error affected only 47.44% of the runs for those 13 participants, data timing quality was not compromised more than in version (1) on average. After fixing this error, the movies rewound correctly whenever the scanner was stopped for the remaining participants for the remainder of the study (N = 44 or 51.16% of participants).

This was achieved by using values stored in the *final output* file, that comprised start, pause and calculated rewind times in linux epoch format. For details on how the rewind times were calculated, please refer to the original publication at (Aliko et al., 2020).

**Movie Annotations**

Spoken words from the movie dialogues were annotated for onset/offset in the movies using fully automated approaches. These were used for analyses in the subsequent chapters. To

achieve this, we extracted the audio track in '.wav' format and the subtitle track as a '.txt' file from each movie's '.iso' file. The .wav audio file was transcribed from speech to text using '*Amazon Transcribe*', a machine learning tool from *Amazon Web Services* (AWS; https://aws.amazon.com/transcribe/). The resulting transcripts contained onset and offset timings for individual words, although the algorithm did not transcribe all words or transcribed some incorrectly. In order to obtain a final transcript containing all the correct words with corresponding onset/offset times, we compared the AWS transcripts to the subtitle files, which contained the correct words but lacked individual words' timings. Instead, subtitle files had onset and offset of sentences.

Therefore, to fix errors from the AWS algorithm, a script was written that first uses dynamic time warping (DTW (Giorgino & Others, 2009)) to align word onsets from the speech-to-text transcript to corresponding subtitle words in each individual subtitle page, starting 0.5 seconds before and ending 0.5 seconds after the page to account for possible subtitle inaccuracies. To improve matches between subtitles and transcripts, punctuation was removed, and words were stemmed. Subtitle words that matched or that were similar to the transcriptions during the DTW procedure inherited the timing of the transcriptions and were returned to their original unstemmed form. Non-identical words were assigned the word's transcription timing that had maximum Jaro similarity (given Jaro similarity > .60) with that subtitle word. Here, Jaro similarity measures the distance between strings of letters, and the higher its value the more similar two strings are (Cayhono, 2019). Finally, if multiple words in the subtitles aligned with a single transcript word (e.g., 'is', 'a', 'story' in the subtitles and 'story' in the transcription), we gave the timing of the transcribed word to the matched subtitle or most similar word if the Jaro similarity was > .60.

The remaining unlabelled subtitle words were estimated in one of three ways:

1. 'Continuous' words used the onset and offset times from adjacent words directly, making them the most accurate, e.g., in '*Tom, drive [Jane] home please*' the missing word *Jane* would take as onset the offset of *drive*, and as offset the onset of *home*.
2. 'Partial' estimation meant that more than one word between matched/similar words was missing. In such cases the length of each word was approximated by counting the number of letters in each missing word and dividing the intermediary time proportionally. For e.g., in '*Tom, [drive] [Jane] home please*' the missing words *drive*

and *Jane* have 5 and 4 letters respectively; the time between offset of *Tom* and onset of *home* would be assigned 55% to *drive* and 44.44% to *Jane*.

3. 'Full' estimation occurred when there were no matching/similar words transcribed, and the onsets of missing words were estimated from the onset and offset of the subtitles. As for partial estimation, word onsets were estimated proportionally using word length. However, due to occasional pauses in dialogues, this might result in unreasonably long word durations. For e.g., in '*Tom, drive Jane home please… 10 sec pause… [Be] [careful]'* the words *be* and *careful* would be assigned ~2 sec and ~8 sec length each. In such cases, we truncated estimated words < 10 letters and more than 2.5 standard deviations from the mean word length in conversational speech (i.e.,, > 1000 ms) to the mean (i.e.,, 600 ms, based on (Tucker et al., 2019)). As it is common for words more than 10 letters to be longer than 1 second when spoken, estimated word lengths for words with >10 letters and < 2 sec were kept. Estimations > 2 sec were truncated to 1000 ms.

Finally, words were reorganised based on their onset times, and overlaps in time removed by matching the order of words in the subtitles and re-assigning onset times based on adjacent words to the wrongly labelled word.

*Acquisition*

Functional and anatomical images were acquired on a 1.5T Siemens MAGNETOM Avanto with a 32-channel head coil (Siemens Healthcare, Erlangen, Germany). We used multiband EPI (Feinberg et al., 2010; Feinberg & Setsompop, 2013) (TR = 1 s, TE = 54.8 ms, flip angle of 75°, 40 interleaved slices, resolution = 3.2 mm isotropic), with 4x multiband factor and no in-plane acceleration; to reduce cross-slice aliasing (Todd et al., 2016), the 'leak block' option was enabled (Cauley et al., 2014). Slices were manually obliqued to include as much of the brain as possible, although few aspects of the cerebellum were occasionally lost (see 'Cerebellar Coverage' section later). Since the EPI sequence had a software limitation of 1 hr of consecutive scanning, scans were stopped at around each 1 hr mark. Depending on the length of the movie (see Table 2), between 5,470 and 8,882 volumes were collected per participant. A 10 min high-resolution T1-weighted MPRAGE anatomical MRI scan followed the functional scans (TR = 2.73 s, TE = 3.57 ms, 176 sagittal slices, resolution = 1.0 mm)[3].

*Preprocessing*

MRI data files were converted from IMA to NIfTI format and preprocessed using mainly the AFNI software (Cox, 1996). In the original version we manually programmed each preprocessing step, but in subsequent analyses we used the afni_proc.py standardised approach (Cox, 1996). Below we detail the general step-by-step preprocessing steps from the afni_proc.py script, for anatomical and functional scans separately (see Fig. 4 for overview).

**Anatomical**

The anatomical MRI scan was corrected for image intensity non-uniformity with AFNI's *3dUniformize* command and deskulled using *ROBEX (Iglesias et al., 2011)* in all cases, except for one participant where *3dSkullStrip* performed better. The resulting anatomical image was nonlinearly aligned (using auto_warp.py) to the MNI N27 template brain, an average of 27 anatomical scans from a single participant ('Colin') (Holmes et al., 1998). The anatomical scan was inflated and registered with Freesurfer software using *recon-all* and default parameters (version 6.0, http://www.freesurfer.net) (Destrieux et al., 2010; Fischl, 2012). Resulting automated anatomical parcellations were used to calculate the extent of cerebellar coverage and to create white matter and ventricle (i.e., cerebral spinal fluid containing) regions of interest that could be used as noise regressors (Destrieux et al., 2010). These regions were resampled into functional dimensions and eroded to assure they did not impinge on grey matter voxels. Finally, anatomical images were 'defaced' for anonymity (https://github.com/poldracklab/pydeface).

**Functional**

The fMRI timeseries were corrected for slice-timing differences (*3dTshift*) and despiked (*3dDespike*). Next, volume registration was done by aligning each timepoint to the mean functional image of the centre timeseries (*3dvolreg*). For 23 (or 26.74%) of participants, localiser scans were redone because, for e.g., the participant moved during a break and the top slice of the brain was lost. For these participants, we resampled all functional grids to have the same 'xyz' extent (*3dresample*) and manually nudged runs to be closer together (to aid in volume registration). For all participants, we then aligned the functional data to the anatomical images (align_epi_anat.py). Occasionally, the volume registration and/or this step failed as determined by manual inspection of all data. In those instances, we either performed the same procedure as for the re-localised participants (N = 5 or 5.81%) or reran the align_epi_anat.py script, allowing for greater maximal movement (N = 6 or 6.98%). Finally, the volume-

registered and anatomically aligned functional data were (nonlinearly) aligned to the MNI template brain (*3dNwarpApply*).

We then spatially smoothed all timeseries to achieve a level of 6mm full-width half maximum, regardless of the smoothness it had on input (*3dBlurToFWHM* (Friedman et al., 2006)). These were then normalised. In the older version of the preprocessing, we normalised to have a sum of squares of 1 per run, however this meant that short runs had very large normalised timeseries amplitudes. To fix these issues we performed a different normalisation as suggested by the AFNI team (see https://openneuro.org/datasets/ds002837). Finally, all regressors were detrended (*3dTproject*) in one step. The first included a set of commonly used regressors (Caballero-Gaudes & Reynolds, 2017): 1) Legendre polynomials whose degree varied with run lengths (following a formula of [number of timepoints * TR]/150); 2) Six demeaned motion regressors from the volume registration; 3) A demeaned white matter activity regressor from the averaged timeseries in white matter regions; and 4) A demeaned cerebrospinal fluid regressor from the averaged timeseries activity in ventricular regions. The second, involved ICA artefacts that were manually selected (see next section 'ICA artefact removal').

**Timing Correction**

To match to the stimuli, timing correction was done to account for delays caused by the movie pausing script to assure that fMRI timeseries and movies are well aligned. Specifically, this script introduced a known 100 ms delay that was cumulative for each break in the movie. Furthermore, depending on the versions of the script, there was also a possible additional (cumulative) delay from not rewinding (v1) or occasionally fast-forwarding (v2.1). These delays were calculated as previously described. Furthermore, the script output files allowed us to quantify potentially variable soft and hardware delays and account for these as well. In particular, every voxel of the detrended timeseries was shifted back in time using interpolation to account for all delays, in the same manner as in slice timing correction but over all voxels uniformly (*3dTshift*).

Specifically, in v1 of the script, if the movie stopped at, e.g., 1000.850 and the last full TR was lost, it means that 850 ms of the movie was watched but is missing from the timeseries. To account for the missing information, we added a TR to the timeseries being collected before the scanner was stopped and interpolated the next run backwards in time the amount not covered by this TR. The added TR was created by retrieving the last timepoint of the run in

which the movie was stopped and the first timepoint of the run after the movie was stopped and averaging these. Thus, for the 850 ms of movie watched but dropped, there were 150 ms too much time added to the movie by adding a TR (because our TR = 1 second). Thus, we shifted the next run back this amount so that the timeseries is theoretically continuous again (though this is never really possible). As the details of these calculations are complex and not fundamental to the scope of this thesis, we limit our explanation to the above. For a more comprehensive and detailed explanation of these calculations, please refer to the original publication (Aliko et al., 2020).

**ICA Artefact Removal**

Spatial independent component analysis (ICA) is a powerful tool for detecting and removing artefacts that substantially improves signal-to-noise ratio in movie naturalistic-fMRI data (Liu et al., 2019). Using the first preprocessing version, we concatenated the timeseries after detrending for motion and white matter/cerebrospinal fluid regressors and after timing correction. Here, as in the HCP, we did spatial ICA on this timeseries with 250 dimensions using *melodic* (version 3.14) from FSL (Smith et al., 2013). Next, we manually labelled and removed artefacts from timeseries, following an existing guide (Griffanti et al., 2017). Myself and two other trained authors went through all 250 components and associated timecourses, labelling the components as 'good', 'maybe', or 'artefact'. As described in Griffanti et al. (Griffanti et al., 2017), there are a typical set of 'artefact' components with identifiable topologies that can be categorised as 'motion', 'veins', 'arteries', 'cerebrospinal fluid pulsation', 'fluctuations in subependymal and transmedullary veins' (i.e.,, 'white matter'), 'susceptibility artefacts', 'multi-band acceleration' and 'MRI-related' artefacts. Our strategy was to preserve signal by not removing components classified as 'maybe'. On a subset of 50 datasets (58.14% of the data), another author classified all components to check for consistency. We then discussed discrepancies and modified labels as warranted. It was expected that, similar to prior studies, about 70-90% of the 250 components would be classified as artefacts (Griffanti et al., 2017). Once done, the identified ICA artefact component timecourses were included as additional regressors in the single detrending step in the second preprocessing version, and the timeseries were concatenated (*3dTproject*).

## *Limitations*

First, with respect to data acquisition, the study was conducted at 1.5 T. Had it been conducted at 3 T, signal-to-noise ratios (SNR) would theoretically double. However, in practice

SNR is only about 25% better and susceptibility artefacts are worse at 3 T (Wardlaw et al., 2012). That said, we are soon starting to collect a new version of the database using a 3 T scan.

With regard to stimuli, it should be acknowledged that neither the fMRI setting nor movies themselves are necessarily 'natural' or completely realistic (Carroll, 1985; Carroll & Seeley, 2013). In addition to the somewhat artificial environment of the magnet, there is continual rhythmic noise. Although we did not use noise cancelling headphones, the use of noise attenuating ones and the addition of pillows to cover participant's ears should mitigate this limitation.

There are a few other general issues with using movie stimuli. First, movies are long. Though this does not seem to adversely affect motion, it could be problematic for some (e.g., clinical) populations in future work. Second, for clinical 'biomarker' purposes (Boeke et al., 2019; Kapur et al., 2012), long movies might be too expensive even if patients could sit still for 1.5 hours or more. However, there is no a priori reason that models cannot be trained on (e.g., network-based representations of) NNDb data but tested on shorter excerpts of movies.

Finally, there is a limitation with regard to the participants themselves. 10 participants asked for (unplanned) breaks, and these might thus have a different pattern of activity before breaks. However, if it is assumed that this lasts for 20 seconds, it means that only 0.06% of the data were affected. This is unlikely to have a big impact on the results. Indeed, we censored timepoints during that time in five participants and it made no discernible difference to data quality (see (Aliko et al., 2020)).

# Chapter 3: Results

## *3.1 The brain reactivates sensorimotor representations of unique characters during pronoun resolution*


### Abstract

One of the most complex tasks in language comprehension is reference resolution. How does the brain link words such as "she" to a specific person? While one crucial component of reference resolution involves 1) keeping track of ongoing linguistic information in a dialogue, another is that 2) the brain infers the correct referent from probing ongoing situation models, imagistic representations of the events indirectly conveyed by language. Existing linguistic research suggests this is the case, yet there is no direct neurobiological evidence for situation models during language comprehension. Here, we developed a 3D branched deep neural network trained on functional magnetic resonance imaging data collected during movie watching to distinguish between two main characters, achieving a final accuracy of ~93%. The model regions most strongly supporting these predictions included mostly visual and auditory cortices, with subtle differences between characters. The model distinguished which characters are referred to by pronouns using the same sensorimotor regions, as well as the hippocampus and precuneus (involved in episodic memory retrieval) and the medial prefrontal cortex (involved in memory and mentalizing). Overall, our findings indicate that imagistic situation models are reactivated to resolve references during language comprehension. This regular use of situation models in natural language comprehension further suggests that the processes associated with language comprehension are complex and distributed.


### Introduction

Real-world processing involves complex and contextually-rich information that the brain must be able to distinguish, process and retrieve for learning and comprehension over short times. Existing models of the neurobiology of language do not take this contextual complexity of language into account, mostly discussing only general aspects of ‘speech perception’ and ‘speech production’. As such these models are limited to a small set of brain regions but growing evidence suggests that, when more specific elements of language are considered, the neurobiology of language extends to many more brain regions (González et al., 2006; Huth et al., 2016; Price, 2010; Skipper et al., 2021; Xu et al., 2005). As language is a

complex behaviour, we should thus focus on inspecting language features that encompass various aspects of this contextual richness. One circumscribed linguistic characteristic that fits this profile and is well established linguistically, is pronouns.

Pronouns without context provide little information (except for gender, number, case), and are underspecified, but when encountering pronouns in a discourse or narrative the human brain is capable of quickly inferring to whom/what the pronoun refers (Greene et al., 1992). Indeed, anyone can easily identify the referents of the two pronouns in the sentence 'John went to pick up Jane at school. *He* drove *her* home': through the gender information conveyed by the pronouns, we understand that 'he' refers to John and 'her' to Jane. However, in the sentence 'John picked up James to take *him* home. *He* had been at a party', where two male characters are present, we can still easily understand that 'him' refers to James, as likely does the 'he' pronoun, based on the context of the previous sentence. How does the brain know who the pronoun is referring to?

Linguistic models have proposed that pronouns initially trigger a search back in memory, this search is restricted by the gender/number/case of the pronoun (e.g., 'she' can only refer to a female character, and 'it' can only refer to objects), and the interpretation of the context points to a referent (Wittenberg et al., 2021). This interpretation is possibly supported via a process that builds contextual meaning incrementally with each added word in a sentence (Altmann & Steedman, 1988). Indeed, humans do not remember every single word in a discourse, but rather recall the gist or concept of a conversation (Campos & Alonso-Quecuty, 2006). Therefore, it is likely that the brain builds a general contextual representation of a sentence to help the interpretation of a reference later on (Wittenberg et al., 2021).

These representations are so-called situation models, which capture the embodied sensorimotor, emotional, and imagistic concepts of an event, character, location and action (Baldassano et al., 2018; Yarkoni et al., 2008; Zwaan & Radvansky, 1998). Numerous linguistic and fMRI studies have provided evidence for a role of situation models in processing of pronouns and have proposed that the process behind building situation models involves activating sensorimotor, language and emotional representations (Altmann & Ekves, 2019; Bergen et al., 2007; Zwaan et al., 1995; Zwaan & Radvansky, 1998; Zwaan, 2016; Zwaan et al., 2002). For instance, some such studies have shown that only when processing sentences relating to motion in a real space the visual field or motor regions are activated, but not during lexical priming (Bergen et al., 2007; Schuil et al., 2013). The generalised nature of situation

models allows them to track event-based representations in sentences by linking antecedent concepts to the newly encountered ones (Altmann & Kamide, 2009). Since pronouns on their own carry little information about the referent, their resolution requires that the brain link the general representations of the current and preceding context (McMillan et al., 2012), a process that can be supported by an imagistic model such as situation models.

Neuroimaging studies on narrative comprehension have attempted to identify where situation models are retrieved in the brain. Most of the literature has pointed to an important role of the medial prefrontal cortex (mPFC) in activating situation models during retrieval. In particular, in an fMRI study, the changes in activity in this region were a good predictor for classifying specific event schemas (e.g., watching an airport or a restaurant scene), but only if the temporal sequence of an event was intact (Baldassano et al., 2018). This fundamental role of the mPFC in maintaining situation models active in memory seems reasonable, given that this region is part of the mentalizing and Default Mode networks, which are involved in decision-making processes and construction of imagery (Baetens et al., 2014; Euston et al., 2012; Isoda & Noritake, 2013; Xu et al., 2005). However, the mPFC, was also found to be involved in narrative comprehension in general (Fletcher et al., 1995; Hasson et al., 2007), during processing of coherent consecutive sentences (Ferstl & von Cramon, 2002) and understanding of themes in a story (Xu et al., 2005). Since there has been no direct study inspecting a role of the mPFC in activating situation models, it may well be that this region simply processes coherent and consecutive naturalistic events, rather than specifically activate situation models in memory.

Aside from the mPFC, the only other regions considered important for retrieving situation models during linguistic processing include the middle temporal gyrus (MTG), inferior frontal gyrus (IFG), and angular gyrus (AG), all part of the ‘language’ regions in existing neurobiology of language models (Hammer et al., 2007; Hickok & Poeppel, 2007; Li et al., 2018). These ‘language’ regions also consistently appear in various meta-analyses and task-based studies, suggesting a domain-general role in language processing - as we detailed in Chapter 1 - rather than a specific role in reactivating situation models to support pronoun resolution.

Since ‘language’ regions and the mPFC both have domain-general roles (Euston et al., 2012; Hagoort & Indefrey, 2014), it seems reasonable to think that these regions may activate with any linguistic retrieval task, and that there should be other brain regions involved during

referent-specific pronoun resolution. Although there is behavioural and linguistic evidence for the involvement of situation models in resolving context-specific pronoun referents, no neuroimaging study to date, to the best of our knowledge, has found concrete evidence for this process.

However, research on content retrieval has offered some insights into how a stimulus in one modality (e.g., memory of a person) may activate the conceptual representation of that stimulus in the associated modality (e.g., activation of face fusiform area). Some such studies have initially shown that the activity patterns of various visual representations of objects, places and faces are distributed across different regions in the visual cortex, depending on their category, with other studies reconstructing specific human faces from this unique brain activity fingerprints in visual regions, during free recall (Haxby et al., 2001; Norman et al., 2006; VanRullen & Reddy, 2019). These findings indicate two things: (i) that the brain activates distinct perceptual patterns to process different faces, and that (ii) these patterns are reliably reactivated during recall. Although studies on content retrieval have provided evidence for the distinct formation of representations and their retrieval pathways during processing of visual information, there is still lack of evidence linking linguistic retrieval (e.g., pronouns) to imagistic simulations of situation models in the brain.

Here, we investigated the neurobiological mechanisms behind pronoun resolution. We hypothesise that antecedent visual representations of a character activate sensorimotor regions to build unique situation models, and that these regions are later reactivated during pronominal referencing in the absence of a character's visual representation, thus linking the antecedent to the referent. Moreover, we predict that individual character references will elicit mostly overlapping activity patterns, with few distinct voxel distributions that allow the brain to distinguish between character-specific situation models in memory. These context-dependent differences are predicted to be in and around sensorimotor regions rather than in mentalizing regions (e.g., mPFC), as previous research may suggest.

To test these hypotheses, we used fMRI scans of 20 participants watching the full-length movie '500 Days of Summer' from the Naturalistic Neuroimaging Database (NNDb). We labelled faces and pronominal references of the two main characters (*Summer* and *Tom*, a woman and a man respectively) in the movie and used the 3D brain volumes in the relevant visual and pronoun reference timepoints as input to a branched 3D deep neural network model. The model was implemented to first distinguish the two character references in the visual and

pronoun domain separately, and to then find shared activations of the visual and pronoun representations of each character. Finally, we performed guided backpropagation to identify the clusters of voxels that the model used to learn to distinguish each character reference.

## Methods

### *Neuroimaging data*

We used fMRI data from 20 subjects (10 females, range of age 19-53 years, M = 27.7 years, SD = 10.1 years) from our Naturalistic Neuroimaging Dataset (NNDb) (Aliko et al., 2020). All participants watched the movie '500 Days of Summer', selected because it has the largest sample size for a single movie. The data was preprocessed as detailed in Chapter 2 and the original publication (Aliko et al., 2020).

### *Face detection in movies*

To detect character faces, we first selected the five main characters by filtering the five highest grossing actors in the movie from 'The Movie Database' website (themoviedb.org). We created a folder for each actor and manually downloaded images from Google Images, ensuring that the face of each actor was the main focus of the photo and that multiple angles of their face were included (e.g., left, right, up, down). This is important for training face detection models, since in movies characters may be facing cameras at different angles. Moreover, where possible, we downloaded images of actors from the specific movie, because the movie makeup and costumes may significantly change the appearance of an actor. On average, M=27.8 and SD=11.0 images were used for training the model for each actor.

We used an existing face detection algorithm implemented in python from the package *face-recognition* that allows detection of specific actors (github.com/ageitgey/face_recognition). The algorithm works by first encoding (i.e., lowering the dimension) images of faces to 128 dimensions. These are saved to a ".pickle" file and accessed when running the detection algorithm on the movie. Since detecting faces in full-length movies can be computationally intensive and thus slow, we used a multi-threaded batch approach for the encoding step and a multi-process approach for the detection step at each movie frame. To divide the movie frame-by-frame we used the python package *openCV* (github.com/opencv/opencv). Frame counts were then transformed to seconds and results saved to a ".json" dictionary file of this format:

{Frame in sec: [character 1, (character 2...)]}

The face recognition algorithm has some error associated with its predictions, which cannot be calculated unless all character faces are manually labelled frame-by-frame and compared to predictions. To reduce this error without the need for manual labelling, we only considered a character to be on screen if that character was predicted in over 50% of predicted frames in a TR (=1 sec). For instance, if within a TR there were 6/10 frames with predicted faces of *Tom*, then *Tom*'s face would be considered detected within that TR. Secondary characters were ignored here, even if they may have been present with one of the two main characters on screen. We, however, ensured not to select any times where *Tom* and *Summer* were together on screen.

### *Pronominal reference annotation*

Movie audio signals were annotated using the Amazon AWS speech-to-text translator (aws.amazon.com/transcribe/) (see Chapter 2 for details). From the word transcripts we selected the word timings for '500 Days of Summer' and filtered out all possible pronominal references to female and male referents (i.e., 'he/she', 'his/hers', 'him/her'). We matched the pronoun onset to the subtitle start time containing the pronoun and used the integer of the subtitle onset as the TR (=1 sec) of interest. This is because subtitles constituted short sentences spanning ~1-2 TRs. Finally, we manually labelled to which character the pronouns referred to and filtered pronominal instances referring to the two main leads in the movie (i.e., *Tom* and *Summer*). This was done because although 'she' and 'he' pronouns were used to also refer to secondary characters, these instances were not sufficient for training a neural network. Moreover, if a subtitle contained more than one pronoun, and these referred to different characters, these instances were removed from the training data. At the end, the data contained only cases where either *Summer* or *Tom* alone were being referenced through pronouns.

Since we aimed at investigating overlaps between pronoun and visual representation of a character, we deemed it important to maintain some temporal correlation between the selected samples. For this, we matched pronoun and visual samples of *Tom* vs *Summer* if the visual instance of a character happened within 2 min of the upcoming pronoun reference for the same character and no less than 30 sec beforehand: these two time limits were arbitrarily selected. Nevertheless, these were chosen because it seems likely that the situation model retrieved during the upcoming pronoun reference will be most similarly represented by a close (but not

overlapping) antecedent representation of a character. Therefore, we considered the visual reference of a character to be instances when a situation model was built (or updated) through perceptual information. Finally, we ensured that instances of *Tom* and *Summer* were at least 5 sec apart from each other in either past or future time direction to a given sample: this was to reduce potential overlaps between the two main characters either in the visual or pronoun domain, specifically in scenes where they acted together, and the camera may have been switching between the two characters. All other pronoun samples were dropped, and unmatched visual samples were also removed. Although here we chose to focus on *Tom* vs *Summer*, other characters were also occasionally on screen or speaking in a scene along with one of the two main characters.

### *Data preprocessing and feature selection*

In order to account for properties of the hemodynamic response function (HRF), we slid the onset of visual and pronouns representations by 3 sec, after which the HRF begins to peak on average (5-7sec from onset stimulus) (Yeşilyurt et al., 2008). For this reason, we selected from the 4D dataset the 3D brain volumes from 3 to 7 sec and averaged these to produce a less noisy signal centred around the theoretical peak of the HRF.

We identified imbalances between the samples of *Tom* and *Summer*, with *Summer* having more samples than the *Tom* dataset. In order to fix the imbalance, we applied random oversampling with replacement on the *Tom* dataset using the python package *imblearn* (https://github.com/scikit-learn-contrib/imbalanced-learn). This resulted in 21 samples per character for each participant (i.e., total of 420 samples for each character), with the final sample comprising 840 total samples of 3D (64x76x64 voxels) brain volumes for each of the visual and pronoun datasets (i.e., 1,680 brain volumes for the entire model).

For each participant we removed the voxels outside the brain and in white matter/ventricles and then computed a group mask to ensure all brain images had the same number of input features, which is a requirement for the input to a convolutional neural network (Conv) layer (Hashemi, 2019). Then we centred the samples to approximately have mean = 0 and standard deviation = 1, using the formula:

$$(X - \mu_x)/\sigma_x$$

This is a typical preprocessing step for Conv layers that helps the model learn and converge faster (Huang et al., 2020). Finally, we randomly shuffled the 3D volumes and labels, to avoid overfitting. Labels for *Tom* and *Summer* were then converted to one-hot encoding (i.e., vectorised categorical labels) for input into the deep neural network.

*Model selection and training*

The model for the 3D deep neural network (DNN) matched the existing architecture and hyperparameters proposed by (Vu et al., 2020), which was used to classify 4 simple tasks (e.g., motor vs language) in an fMRI experiment. Here, we built upon it by creating two branches, one for visual data and one for pronoun data, that were then merged for output. This was done to (i) have most layers separate to later inspect where in the brain visual vs pronoun referents map onto; and (ii) merge the final layers to identify any shared voxels of visual and pronoun referents.

As per the specifications detailed by Vu and colleagues, each branch consisted of 3 convolutional layers (Conv 1-3). The Conv layers were built with the following hyperparameters: Conv1 had kernel size 7x7x7 and 8 filters, with a stride of 1; Conv2 had kernel size 5x5x5 with 16 filters and stride of 1; Conv3 had kernel size 3x3x3 with 32 filters and stride of 1 (Vu et al., 2020). The first two Conv layers were followed by an average pooling layer with stride 2 to reduce feature dimensions. Then we applied a flattening layer to vectorise convolved features to 1D, and finally added a fully connected layer with 128 nodes. We then added a further dropout layer with 50% retention probability to reduce overfitting. The two branches of visual and pronoun were then concatenated with a further 50% retention probability dropout layer, and finally output into a fully connected layer with 2 nodes (i.e., classes) for prediction. Each Conv layer and the branches' fully connected layers had 'ReLu' activation functions, whilst the output fully connected layer had a 'sigmoid' activation function (Vu et al., 2020). Ridge regression (L2) regularisation was applied to the output layer activity to discourage overfitting.

The loss function for the model was binary cross-entropy, since we only had 2 classes. We applied a stochastic gradient descent (SGD) optimizer with initial learning rate ($Lr_0$) of $10^{-3}$, which was step-decayed using the formula:

$$Lr = Lr_o * rate_{decay}^{(current\_step / decay\_steps)}$$

Where Lr is the new learning rate for a given step, rate of decay was 0.96 and decay steps were set to 100,000. Given that we had a small sample set that may be prone to overfitting in complex models, we tested 3 model complexities:

1. Three Conv layers as in (Vu et al., 2020)
2. Two Conv layers (removed Conv1)
3. One Conv layer (removed Conv1 and Conv2) and added 50% dropout after Conv3 layer

Fig. 5 shows a diagram of the final selected architecture and hyperparameters used.

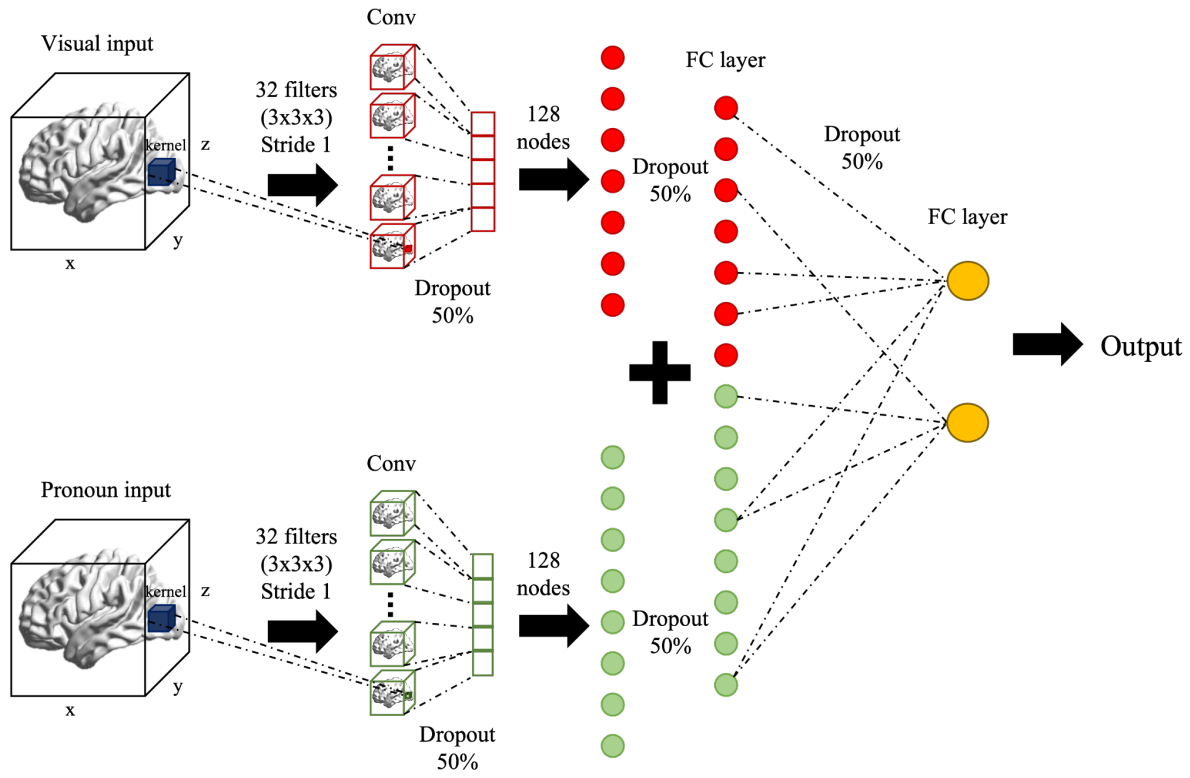



**Figure 5.** Diagram of the final 3D DNN branched model. Separate 3D volume inputs are fed into a visual and pronoun branch. The first layers of each branch consist of a 3x3x3 convolutional layer with 32 filters (stride 1), with a 50% dropout, which are flattened and input into a 128 node fully connected layer. A 50% probability dropout reduces overfitting before merging the visual and pronoun branches. A final 50% dropout and fully connected layer (2 nodes) provide the output predictions.

The model was compared to other 2 deep neural networks for performance, using accuracy and loss as metrics. We compared the model to a 2D ResNet50-based transfer learning model and a 3D DNN with increasing kernel size: the first allows us to test whether using brain volumes instead of slices is better to identify relationships between voxel clusters, the second is useful for comparison of hyperparameters (see Supplementary Materials S1-3 for detailed model information).

Moreover, the reason why simpler models (e.g., support vector machine) could not be tested was because of the two branching inputs, which can only be accommodated by DNN models. The architectures and hyperparameters for the other 2 models that were tested can be found in Supplementary Materials S1-S3. The data was split into 80% training and 20% testing. The performance of the 3 models were tested using 10-fold cross validation over 10 epochs on the training data and for each fold the accuracy and loss were also computed on the hold-out test data. The 3D bi-DNN was then trained on the entire dataset (i.e., 1,680 pairs of samples) over 100 epochs with early stopping to reduce computational load.

*Saliency map visualisation*

In order to visualise which voxels of the 3D images the model was learning from, we computed saliency maps using vanilla guided backpropagation. Given an image *I* belonging to class *c* (either *Summer* or *Tom* here) the class score $S_c(I)$ can be approximated to a linear function using Taylor's expansion rule, such that:

$$S_c(I) = w^T I + b$$

Where *w* is the weight of a voxel and *b* the bias term. Solving the derivative for *w*, we get the following:

$$w = \frac{\delta S_c}{\delta I}\Big|_{I_0}$$

The collection of *w* for each voxel represents the saliency map of an image (Simonyan et al., 2013).

We selected the 3D volumes for *Summer* and *Tom* separately, ran each through vanilla guided backpropagation and averaged the resulting maps for each image to obtain an overall map of each character representation. Due to the branched nature of our model, each character's guided backpropagation resulted in 2 saliency maps for each character: one for pronoun and

one for visual, which shared the higher layers (i.e., concatenation and output layer). We performed paired t-tests of *Summer* vs *Tom* (visual+pronoun) and visual vs pronoun (*Summer*+*Tom*). However, because we could not easily trace back the index of the samples after the shuffling of the data, we could not separate the samples by participant (i.e., separate into 20 clusters), resulting in high degrees of freedom in the t-test. We then thresholded each of the 4 resulting saliency maps to the 95$^{th}$ percentile to ensure that the strongest weights were maintained, thus maintaining the voxels that the model used the most to learn from. 5 voxel clusters in each of the 4 maps were then compared to Neurosynth meta-analysis maps (Yarkoni et al., 2011), in order to identify the highest correlated functional term associated with the cluster of interest.

### *General linear model analysis*

Although the DNN model offers the possibility to detect voxels active during visual vs pronoun instances of a character (and any putative reactivations) even with small sample sizes, it is known to potentially suffer from low interpretability (Sheu, 2020). As such, we applied a general linear model (GLM) using a canonical HRF on the same time points from the DNN model (i.e., start time shifted 3 sec to account for HRF rise), using the AFNI program *3dDeconvolve* (Cox, 1996), in order to further test our hypotheses through more typical analyses. The GLM would output a beta map for each of the following: (i) *Tom* visual instances, (ii) *Tom* pronoun instances, (iii) *Summer* visual instances, (iv) *Summer* pronoun instances. The resulting beta maps for each subject were input into a linear mixed-effects model (*3dLmE*), with subject as a random effect, and age and gender as additional fixed effects. Finally, the group-level maps were corrected for multiple comparisons using a cluster-size approach. First, we estimated the smoothness and autocorrelation function of neighbouring voxels using the *3dFWHMx* command (Cox, 1996). Then we ran *3dClustSim* over 6 uncorrected individual voxel p-values (.05, .02, .01, .005, .002, .001) and an alpha threshold of .01. Using the significant cluster sizes whereby faces or edges need to touch, and voxels are contiguous if they are either positive or negative at each p-threshold, we merged the thresholded maps at each p-threshold to obtain significant voxels ($\alpha$=0.01).

## Results

*Model selection and performance*

We tested three different DNNs for classifying *Tom* and *Summer* using visual/pronominal information. Here, the (i) transfer-learning ResNet50-based model achieved an average validation accuracy of labelling *Tom* vs *Summer* both visually and through pronouns of 47.9% (SD = 10.7%), with M=0.72 (SD=0.05) loss over 10 folds. On the hold-out test data (20% of dataset), the model reached an average accuracy of 50.5% (SD=2.3%) and M=0.72 (SD=0.04) loss. The (ii) 3D increasing kernel size DNN reached an average validation accuracy of labelling *Tom* vs *Summer* both visually and through pronouns of 70.5% (SD = 5.3%), with M=0.64 (SD=0.01) loss over 10 folds. On the hold-out test data, the model reached an average accuracy of 70.1% (SD=3.4%) and M=0.64 (SD=0.01) loss.

Finally, we tested the (iii) 3D decreasing kernel size DNN at various complexities:

1. Three Conv layers: reached an average validation accuracy of labelling *Tom* vs *Summer* both visually and through pronouns of 47.3% (SD = 10.9%) and an average validation loss of 0.69 (SD = 0.003). This last model also achieved an average testing accuracy across folds of 56.8% (SD = 6.4%) and an average loss of 0.69 (SD = 0.0008).
2. Two Conv layers: reached an average validation accuracy of labelling *Tom* vs *Summer* both visually and through pronouns of 71.7% (SD = 11.9%) and an average validation loss of 0.66 (SD = 0.01). This last model also achieved an average testing accuracy across folds of 75.7% (SD = 6.1%) and an average loss of 0.66 (SD = 0.004).
3. One Conv layer: reached an average validation accuracy of labelling *Tom* vs *Summer* both visually and through pronouns of 91.4% (SD = 2.0%) and an average validation loss of 0.34 (SD = 0.04). This last model also achieved an average testing accuracy across folds of 92.6% (SD = 1.1%) and an average loss of 0.34 (SD = 0.02), indicating that it was stable across folds and appropriate for the task (i.e., not overfitting), unlike the former models. We named this model '3D bi-DNN'. The training, validation and testing accuracy per fold for the (i), (ii) and (iii) final models are summarised in Fig. 6.

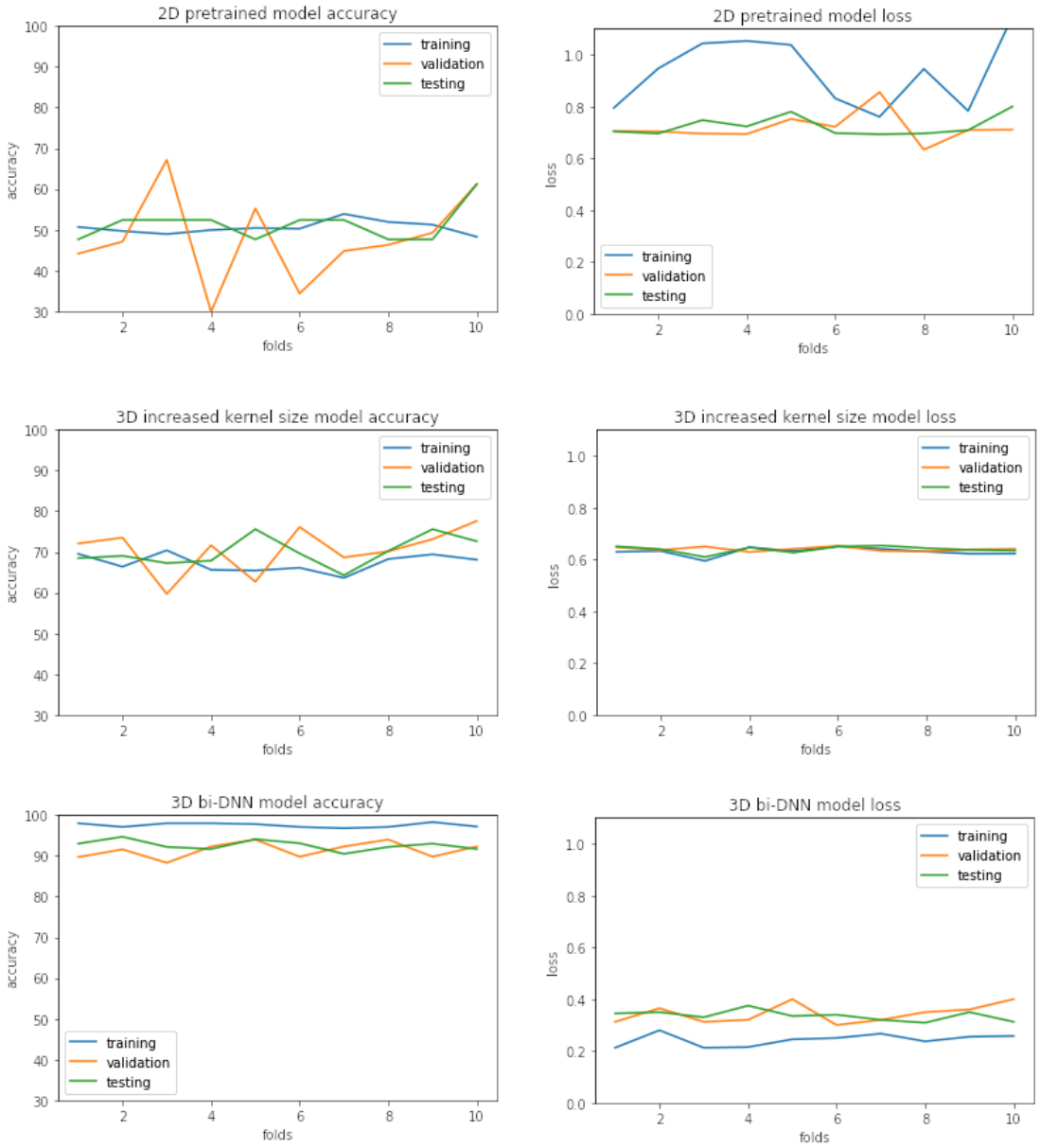



**Figure 6.** Plots of the accuracy (left column) and loss (right column) over 10 folds during cross validation. Over 10 folds: top row represents the RSN pretrained model (47.9% accuracy, 0.72 loss); middle row is the 3D DNN with increasing kernel size (70.5% accuracy, 0.64 loss); bottom row is the final 1-Conv layer 3D bi-DNN model (92.6% accuracy, 0.34 loss). The last model outperforms the others in terms of both increased accuracy and reduced error.

After selecting the 3D bi-DNN with a single Conv layer as the best model, we trained it over 100 epochs on the entire dataset with early stopping to reduce the computational load.

*Saliency maps and GLM maps*

We wanted to then visualise which voxels the model had used to distinguish between *Tom* and *Summer* referents, in order to infer possible visual and pronominal overlaps. Saliency maps resulted in 420 maps for each of the (two) character's faces and 420 for each character's pronoun referent. A paired t-test of (i) *Summer* vs *Tom* and (ii) visual vs pronoun over 840 samples each resulted in all within-brain voxels being significant ($p < 10^{-5}$), due to the high degrees of freedom. The results of the averaged maps thresholded at the 95$^{th}$ percentile are shown in Fig. 7. The most prominent regions the model focused on in the visual branch were parts of the primary visual, secondary visual, STG, superior temporal sulcus (STS), parahippocampus and occipitotemporal cortex (OCT) (Fig. 7A). For the pronoun branch, the regions the model learned from were primary visual, secondary visual, dmPFC, precuneus and hippocampus (Fig. 7B). The two characters overlapped over 38% in the visual and 57% in the pronoun maps respectively and differed primarily in voxels around sensorimotor regions in both. Moreover, within each character's maps, the visual and pronoun branches overlapped for 8% (for both *Tom* and *Summer*), with the overlapping regions being the visual cortex and parahippocampal areas (Fig. 7C). Here the visual references mapped to STG and OCT, while pronoun references mapped to mPFC and precuneus more. From the visual and pronoun maps of each character, we selected the 6 largest clusters in each and ran correlations against functional meta-analysis terms from the database Neurosynth (Yarkoni et al., 2011). Table 3 shows the top 5 functional meta-analysis terms for each of the 6 voxel clusters of choice.

Interestingly, the GLM results did not exactly match the distribution patterns of the saliency maps (Fig. 8). During the visual instances, the GLM revealed significant activation in the occipital lobe, with overlap between the two characters in the parieto-occipital junction, but surprisingly no activation in face fusiform area (FFA) (Fig. 8A). During pronominal instances, the GLM revealed a prominent STG, STS and MTG distribution with large overlap between the two characters, as well as some character-specific distribution in visual areas (e.g., FFA) and overlapping activity in parts of the visual association area (Fig. 8B bottom right).

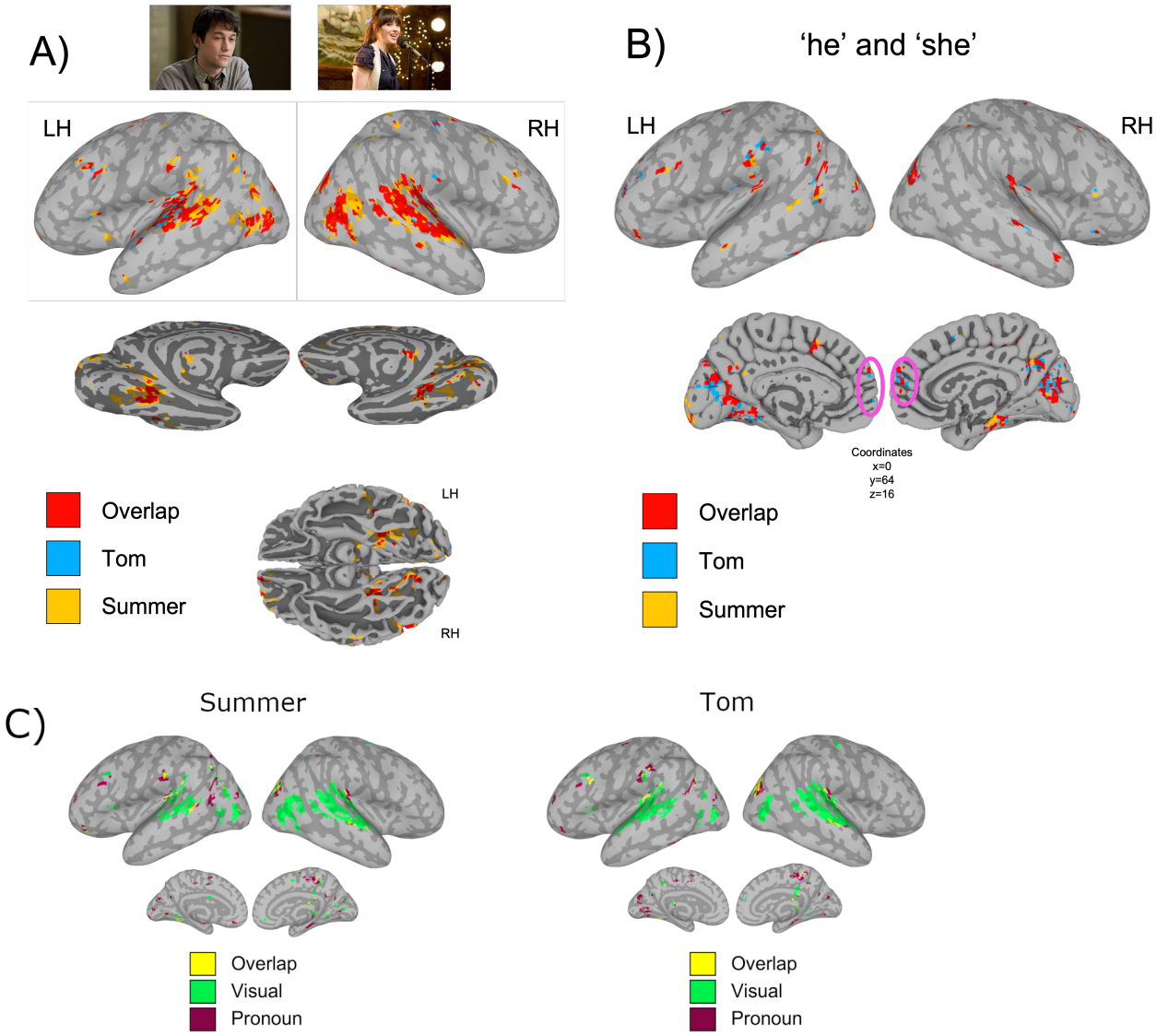



**Figure 7.** Saliency maps from vanilla guided backpropagation. For all figures, threshold = 95th percentile. **A)** Overlap of average visual maps for *Tom* vs *Summer*, where Red = overlap of two characters; Blue = Tom only; Yellow = Summer only. Cluster size = 20. The two characters had high overlap in visual (e.g., FFA in bottom view) and sensorimotor regions during visual references, with small differences in activity around these regions. **B)** Overlap of average pronoun maps for *Tom* vs *Summer*, where Red = overlap of two characters; Blue = Tom only; Yellow = Summer only. Cluster size =20. The two characters overlapped in visual cortices and mPFC. Medial image: medial view showing mPFC (circled in pink) and its coordinates. Here, most voxels overlap between *Summer* and

*Tom* pronouns (i.e., red coloured). **C)** Overlap of average *Summer* and *Tom* maps for visual and pronoun references, where Yellow = overlap of two references; Green = visual only; Purple = pronoun only. Cluster size = 40. Pronoun and visual references overlapped in visual cortex and parahippocampal area, with slightly different patterns of activity.

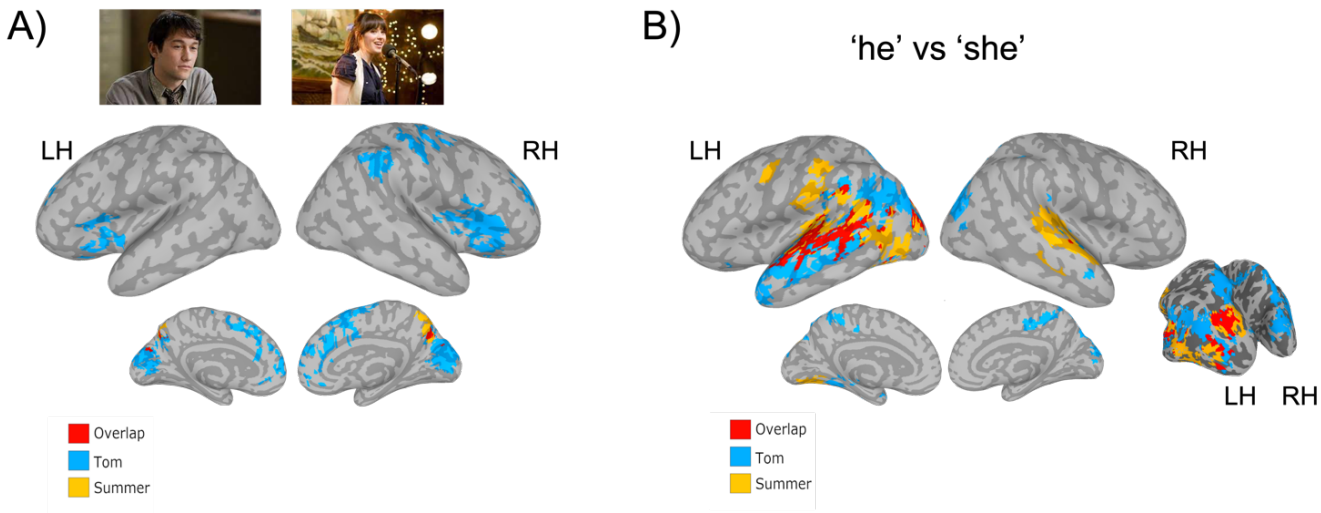



**Figure 8.** GLM of visual and pronominal references. **A)** Overlap of corrected group-level visual beta maps for *Tom* vs *Summer*, where Red = overlap of two characters; Blue = Tom only; Yellow = Summer only. Cluster size =20. The two characters had little overlap in the GLM with Tom showing more distributed patterns. The only overlapping region was at the parieto-occipital boundary. **B)** Overlap of corrected group-level pronoun beta maps for *Tom* vs *Summer*, where Red = overlap of two characters; Blue = Tom only; Yellow = Summer only. Cluster size =20. Here, the distribution was mostly in the left hemisphere, particularly STG/STS, which had the highest level of overlap between the two characters. Some visual association areas (bottom right image), however, also showed overlap between *Tom* and *Summer*. Both *Tom* and *Summer* showed some activity in FFA, although not overlapping.

| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
|---|---|---|---|---|---|---|---|
| Summer Visual | Coordinates | (-44, -74, 4) | (-50, -42, 4) | (50, -68, 6) | (30, -10, 56) | (10, -80, 8) | (-24, -56, -14) |
| | Term 1 | Motion | Sentences | Motion | Execution | Visual | Imagery |
| | Term 2 | Video | Languages | Visual | Movement | Eye Movements | Objects |
| | Term 3 | Action Observation | Construction | Action Observation | Motor Imagery | Video | Finger Tapping |
| | Term 4 | Perception | Word | Videos | Executive Functions | Attention | Navigation |
| | Term 5 | Visual | Syntactic | Gestures | Task | Imagery | Abstract |
| Summer Pronoun | Coordinates | (0, -60, 36) | (0, 16, 56) | (-12, -82, 8) | (-14, -84, 6) | (-26, -48, 8) | (24, -28, -16) |
| | Term 1 | Mentalizing | Task | Eye | Visual | Place | Episodic |
| | Term 2 | Social | Demands | Visual | Lexical Decision | Objects | Autobiographical Memory |
| | Term 3 | Theory Mind | Working | Colour | Abilities | Navigation | Retrieval |
| | Term 4 | Default | Language | Visuo Spatial | Abstract | Face | Experiences |
| | Term 5 | Person | Phonological | Retrieval | Abuse | Encoding | Construction |
| Tom Visual | Coordinates | (-46, -74, 4) | (-60, -20, 4) | (-30, -50, -8) | (-52, -24, 6) | (-52, -76, 2) | (38, -16, 54) |
| | Term 1 | Motion | Sounds | Objects | Auditory | Visual | Movements |
| | Term 2 | Visual | Speech | Face | Sound | Videos | Finger Tapping |
| | Term 3 | Action Observation | Listening | Place | Speech | Action Observation | Hand |
| | Term 4 | Videos | Pitch | Navigation | Listening | Visual Motion | Coordination |
| | Term 5 | Perception | Voice | Shapes | Audiovisual | Video Clips | Motor Task |
| Tom Pronoun | Coordinates | (0, 62, 16) | (-42, 44, 14) | (-30, -80, 32) | (-36, -30, 20) | (-30, -50, -8) | (2, -60, 36) |
| | Term 1 | Referential | Semantic Memory | Visual | Auditory | Objects | Mentalizing |
| | Term 2 | Theory of Mind | Episodic | Object | Sounds | Face | Beliefs |
| | Term 3 | Default | Memory | Episodic Memory | Musical | Place | Default |
| | Term 4 | Personality Traits | Retrieval | Spatial | Listening | Navigation | Theory of Mind |
| | Term 5 | Positive Negative | Episodic Memory | Stream | Hearing | Shapes | Social |

**Table 3.** Table of top 5 Neurosynth functional terms which correlated to the coordinates of centres of six clusters in each character's visual and pronoun reference. The visual references mapped mainly to sensorimotor regions involved in visual, motor and language processing, but also included some elements of imagery/abstraction. These patterns of activity were reactivated during pronoun references. The latter also activated episodic memory and mentalizing regions.

## Discussion

Here, we aimed at testing the hypothesis that pronoun resolution reactivates unique sensorimotor character representations through context-dependent situation models, which were built in those same sensorimotor regions during perceptual (i.e., visual) processing of a character.

To test this, we implemented a 3D branched deep neural network that takes as input the 3D brain volumes of matched visual and pronoun references of a character in a movie. Our model achieved ~93% accuracy of distinguishing two main characters in a given movie using both visual and pronoun references. Saliency maps thresholded at the 95th percentile for the visual branch of the 3D bi-DNN model, revealed that the model learned to distinguish between perceptual references of *Tom* and *Summer* through the bilateral involvement of primary and secondary visual regions, STG, STS, parahippocampal area and some primary motor cortex (Fig. 7A). This distribution is in line with our hypothesis that situation models are built via sensorimotor simulations.

Saliency maps thresholded at the 95th percentile for the pronoun branch of the 3D bi-DNN model, showed that the model used voxels in primary visual cortices, FFA, hippocampus and parahippocampal area, precuneus and mPFC to distinguish between *Tom* and *Summer* during pronoun resolution (Fig. 7B). Our findings suggest that pronoun resolution may require a search in memory for the appropriate situation model to reactivate representations of a character, which was built and retrieved through sensorimotor regions.

Results from the GLM analysis indicate a different distribution of activity for visual and pronoun instances compared to the 3D bi-DNN model (Fig. 8). Nonetheless, the pronoun maps also showed activation and overlap between the two characters in visual association areas, as well as separate activity patterns in the FFA, indicating involvement of visual areas for pronoun resolution (Fig. 8B). Surprisingly, the visual map did not have any activity in the FFA (Fig. 8A). Rather, the activity for *Tom* was much more distributed than that of *Summer*, in areas such as posterior parietal, IFG and secondary visual areas, with the two characters only overlapping around the parieto-occipital junction.

### *Model of pronoun resolution*

Prior studies have proposed that narrative-based situation models may be built through sensorimotor regions (Zwaan, 2016), that are activated in the mPFC (Yarkoni et al., 2008) and

in ‘language’ regions during pronoun resolution (Hammer et al., 2007; Li et al., 2018). However, aside from general aspects of the retrieval process, no concrete neuroimaging evidence was found on how pronouns may activate these simulations, nor for the actual existence of these unique simulations in the brain.

Our findings from the 3D bi-DNN model show for the first time that situation models may be simulated through representations in sensorimotor and mentalizing regions when the brain encounters a character perceptually. Based on our findings and the models already proposed by the linguistic literature (Altmann & Kamide, 2009; Wittenberg et al., 2021), we speculate the following model of pronoun resolution based on previous visual information:

1. When a pronoun is uttered, ‘language’ regions engage with episodic memory regions to activate a search in memory for the referent
2. This search involves activating the appropriate situation model in which that referent is represented
3. Once the appropriate situation model is active this will point to a specific character (i.e., referent)
4. This representation reactivates character-specific activity in sensorimotor regions, where the visual-based representation was originally built

Here we dissect some of these proposed processes.

**Pronouns and episodic memory**

Studies on pronoun resolution have identified a set of regions that activate when either (i) the referent is more ambiguous as it can refer to either of two characters; or (ii) the pronoun is incoherent with the antecedent (e.g., ‘Julie was walking home. *He* had been at a party’) (Hammer et al., 2007; Qiu et al., 2012). Such studies have found that these tasks activated mostly the IFG, MTG, and dorsolateral prefrontal cortex (dlPFC) : the IFG is proposed to activate with increasing task demands (e.g., ambiguity), the MTG when resolving incongruent gender of the pronoun and referent, and the dlPFC was proposed to have a general role in decision-making processes to help assign a referent (Hammer et al., 2007, 2011; Hertrich et al., 2021; McMillan et al., 2012; Qiu et al., 2012).

Here, we found no activation of the MTG nor IFG, with very minimal activation of the dlPFC (Fig. 7B). Instead, pronouns for both the male and female character activated predominantly regions in the visual cortex (e.g., FFA) and other circumscribed sensorimotor

regions. Here, the lack of activation in 'language' regions typically engaged during incongruences between pronoun and referent, may be due to the fact that the referents could be more easily resolved through the rich contextual information afforded in the movie.

Instead, we found a network composed of mPFC, hippocampus, parahippocampal area and precuneus likely involved in episodic memory and consolidation of context-dependent representations. Previous studies have suggested a fundamental role of the mPFC in activating situation models (Baldassano et al., 2018; Yarkoni et al., 2008), likely through a role in Theory of Mind and Default Mode networks. Here, the mPFC has a role in decision-making tasks, distinguishing between self and others and creating mental representations (Baetens et al., 2014; Cheetham et al., 2014; Isoda & Noritake, 2013; Moran et al., 2011; Smith et al., 2018; Xu et al., 2005). However, as the mPFC was active along with the hippocampus, parahippocampal area and precuneus in the present study, its role is more likely to be in support of memory processes.

In particular, research has shown that the mPFC is active during memory consolidation (i.e., long-term memory formation), after receiving information from the hippocampus (Euston et al., 2012; Takashima et al., 2006). The latter, instead, is involved in short-term memory reactivation, particularly during context-dependent episodic memory, together with the parahippocampal area and precuneus (Chang et al., 2021; Dickerson & Eichenbaum, 2010; Flegal et al., 2014; Maviel et al., 2004; Michelmann et al., 2021). Once information is fully consolidated, the mPFC inhibits activation of the hippocampus, to avoid building new representations of existing memories (Baldassano et al., 2018; Takashima et al., 2006). We thus propose that the activation of this network in the pronoun branch is as follows: (i) hippocampus, parahippocampal area and precuneus reinstate a situation model from working memory to help resolve the referent, (ii) meanwhile the mPFC updates the situation model with the newly encountered dialogue information and consolidates it in long-term memory. Given that the participants had not previously seen the movie, it is reasonable that the hippocampus would be engaged in the reactivation of context-specific situation models, while the mPFC consolidates these representations. It would be interesting to study possible temporal variations in hippocampus/mPFC over the movie, as well as test this activity in repeated movie viewings.

**Situation models and character representations**

Although much of the literature has discussed situation models for building discourse representations (Zwaan et al., 1995; Zwaan & Radvansky, 1998; Zwaan, 2016), to the best of

our knowledge, there is still no evidence for their existence nor for their involvement in pronoun resolution. Here we were able to detect putative context-dependent situation models for the representation of characters, and their reactivation during pronoun resolution. Situation models containing character representations were generally built in sensorimotor (i.e., primary auditory, visual and some motor) regions (Fig. 7A).

The two characters mainly differed in voxel distributions around the primary visual cortex, OCT region, occipitoparietal area and STG/STS (Fig. 7A). These regions are all involved in visual perception at different levels: at first the ventral pathway forms different distributions in the primary visual cortex that relate to face perception of different characters (Sheth & Young, 2016), then the dorsal pathway engages the occipitoparietal area to process information visually encoded by actions (Freud et al., 2016), while the STG/STS was recently shown to form a third pathway that integrates the previous two to process social interactions afforded by visual information (e.g., gestures, facial expressions etc.) (Manfredi et al., 2017; Pitcher & Ungerleider, 2021).

The character representations were not purely due to visual perception, although these regions were highly active, but also incorporated clusters of voxels involved in imagery, abstraction, attention and construction of representations, as shown by the Neurosynth meta-analysis term correlations (see Table 3). Given that situation models are highly imagistic by nature (Zwaan, 2016), the presence of clusters relating to imagery further suggests that the 3D bi-DNN model has likely isolated processes/regions involved in building situation models of character representations.

**Content retrieval**

Our findings support the hypothesis that pronouns would reactivate sensorimotor fingerprints related to the character representation in different situation models. We found that the primary visual cortex and parts of the OCT area, occipitoparietal cortex, and the STG/STS were reactivated during pronoun resolution (Fig. 7B and C). These suggest that when a pronoun is uttered, a search for the situation model in working memory reinstates the activity distribution specific to the representation of the referent.

Studies on content retrieval have shown that the brain reactivates specific perceptual activity fingerprints when recalling distinct contextual information in the absence of the stimulus. Indeed, there is ample evidence in the episodic memory literature suggesting that the

higher the reactivation of the same activity patterns between antecedent and new event the better the retrieval (Frankland et al., 2019; Oedekoven et al., 2017; Yaffe et al., 2014). For instance, a study using naturalistic stimuli in the form of short videos, showed that increased overlap between antecedent and reinstatement increases vividness of the antecedent video during recall (St-Laurent et al., 2015). The patterns reactivated during free recall are unique to the category they represent (e.g., face, object, place) (Polyn et al., 2005) or the context they refer to (Nyberg et al., 2000), although overlap between activity distributions of categories/contexts proportionally increases with their similarity (Norman et al., 2006). In line with previous research, we found that the sensorimotor activity distributions of *Tom* vs *Summer* character representations mostly overlapped, likely because they both related to the 'face' category or similar contextual situation models (Fig. 7).

Our study builds upon the content retrieval literature, by linking visual information of specific characters to their unique retrieval through pronouns in a naturalistic setting, where the stimuli are complex and continuous and free recall cannot be tested.

*Comparison of GLM and 3D bi-DNN*

Since the computations within DNN models may be difficult to interpret (Sheu, 2020), we additionally conducted a more typical GLM analysis on visual vs pronoun instances to use as a comparison tool for our DNN results. The two models showed differences in the specific voxel activations in both the visual and pronoun instances. Nonetheless, the general trend of distribution of activity was somewhat comparable: for instance, during the visual instances both models identified activity in and around sensorimotor regions. The GLM revealed activations mostly in sensory association and secondary visual areas (Fig. 8A), while the DNN in primary visual, secondary visual and FFA regions (Fig. 7A). These patterns were generally in line with (i) the nature of audio-visual stimuli (i.e., visual cortex activation), and (ii) our hypothesis of building situation models through sensorimotor areas.

During pronoun instances, instead, 'language' regions (e.g., STG, MTG, STS) were significantly active in the GLM model (Fig. 8B). Such activation in 'language' regions would normally be expected for a linguistic task, such as pronoun resolution (Hammer et al., 2007; Li et al., 2018). As previously discussed, however, this activation was not present in the DNN pronoun model. Perhaps, at thresholds lower than the 95$^{th}$ percentile that was applied here, 'language' regions may start to become apparent in the DNN as well. Interestingly, in both GLM and DNN models, pronouns activated parts of the visual cortex, such as FFA, strongly

suggesting that reactivation of visual imagery (or situation models) is necessary to resolve specific pronoun referents. Unlike the DNN pronoun maps, we found no activity in mentalizing regions (e.g., mPFC) and limited activation of parahippocampal areas in the GLM pronoun maps. These results could indicate that the mPFC has a more general role in processing ongoing narrative, rather than specifically activating situation models during retrieval.

### *Limitations*

One limitation of the present study was that, due to the limited number of pronouns in movies, we opted not to remove those pronouns where there was a flashback or on-screen presence of a character referent in the movie. For instance, *Summer*'s face appears in some scenes where other characters are referring to her as 'she', similarly *Tom* may appear when the narrator in the movie refers to him as 'he'. This limitation may bias both GLM and DNN models towards a character's face rather than the pronoun retrieval process. Nevertheless, this happened in 29% of the 84 initially detected pronouns (n.b. some pronouns referring to *Tom* and *Summer* may not be detected by the speech-to-text transcript, therefore would not be included as samples here). Moreover, these instances usually happened as single blocks, and since we limited pronouns to be at least 5 sec apart in either past or future direction, resulting in 21 pronoun samples for *Summer* and 8 for *Tom* prior to balancing, it likely further reduced the co-occurrence of characters on-screen. This issue is mostly relevant for the GLM model, where the addition of penalising functions to discourage the model from detecting visual features is not possible, unlike in DNN models. In future, to improve on this issue, we could ignore instances where the face is on screen during pronoun referencing and collect more participants to increase the sample size.

The small sample size represents a significant limitation, particularly for the GLM model. Low sample and high dimensionality are also known issues of DNNs, but recent studies have suggested that multiple dropout layers successfully offset the risk of overfitting (Liu et al., 2017). Indeed, using multiple stringent dropouts resulted in no overfitting in our DNN model (Fig. 6). Finally, BOLD signals in a fast-event design, such as in movies, present non-linear relationships between variables (Pfeuffer et al., 2003; Vazquez & Noll, 1998), which GLMs cannot model, due to their underlying assumption of linearity between variables. Since DNNs can model nonlinear relationships in the data, the results of our DNN model are likely much more robust that those of the GLM.

*Implications*

In the present study we have shown that pronoun resolution requires the reactivation of unique character representations from perceptually built situation models. This is the first time, to our knowledge, that (i) the existence of situation models for character representations in the brain is demonstrated, and (ii) sensorimotor character representations are shown to be needed for inferring referents.

This has significant implications for our understanding of the neurobiology of language in the real world. Existing models are limited in that they do not account for linguistic complexities and how these may drive multi-modal interactions between language, memory and perception, which are known to be elicited by context (Friston & Price, 2001; Skipper, 2015a). This study demonstrated that when inspecting natural contextual dependencies in specific linguistic features, the distribution of activity includes regions outside classical 'language' areas and requires the interplay between various modalities.

Difficulties in pronoun resolution are a common feature in any type of aphasia, irrespective of language (Arslan et al., 2021), but particularly in agrammatic aphasics (Jarema & Friederici, 1994). Importantly, some aphasic patients have difficulty associating the pronoun to the correct subject referent, highlighting how the process of retrieval may be impaired (Peristeri & Tsimpli, 2013). Our finding that pronoun resolution depends on the reactivation of sensorimotor character representations could offer insights for the development of novel speech therapies, to target more specific language features and associated processes to speed recovery.

*Conclusion*

In this study we demonstrated that pronouns in naturalistic discourse reactivate a set of sensorimotor character representations, that were built as part of situation models when the characters were visually present. These models were built not only perceptually, but also recruited Theory of Mind and Default Mode regions required for forming imagistic simulations. The sensorimotor distribution of character representations mostly overlapped between the two characters, likely because they shared a similar context in the movie. However, character representations also experienced small variations around visual regions, showing that situation models can help point to a unique character representation during pronoun resolution.

Overall, these findings highlight the importance of studying individual language features in a more complex and natural environment and demonstrate that the neurobiology of language is more distributed than existing models suggest. These distributed areas may offer new avenues for novel speech therapies for aphasic patients.

## *3.2 Are 'language regions' an artefact of averaging?*


## Abstract

Neuroimaging studies, meta-analyses, and theoretical models of the neurobiology of language all suggest that several superior and middle temporal and inferior frontal brain regions are responsible for language processing. These observations derive from research that heavily relies on measures of central tendency, such as averaging activity patterns from heterogeneous stimuli, tasks, and participants. We hypothesise that the use of such methods obscures the whole brain distribution of language processing, and that 'language' regions are, rather, network hubs, coordinating other regions whose activity is variable. To test this hypothesis, we used movie functional magnetic resonance imaging data and scored heard words for their sensorimotor properties. Analyses revealed that these properties form unique distributed patterns of activity involving most of the brain. Dynamic functional connectivity analyses identified variable connectivity states, which only resulted in 'language' regions when averaged together. These findings suggest that the natural neurobiology of language forms a whole-brain arrangement composed of hubs and dynamic regions that is made invisible by averaging over very different linguistic categories. Aphasia resulting from damage to hub regions might be better explained by their separation from associated dynamic regions.


## Introduction

Traditional models of the neurobiology of language based on lesion studies proposed that regions in the left posterior sylvian fissure and inferior frontal gyrus (IFG) are the anatomical loci of language comprehension and speech production respectively (Dronkers et al., 2017; Geschwind, 1970; Nasios et al., 2019). Over the last decades, this classical model has been revisited and updated using more modern lesion analyses and, mostly, task-based neuroimaging studies to include bilaterally the superior and middle temporal gyrus (STG and MTG), Sylvian parietal-temporal region (Spt) and premotor cortex (Hickok & Poeppel, 2004, 2007; Rauschecker & Scott, 2009; Rauschecker & Tian, 2000). These regions in the most cited models are said to form dual-streams, involving a dorsal stream for mapping acoustic to articulatory processes and a ventral stream mapping sound to meaning (Hickok & Poeppel, 2007). Regions belonging to the dual-stream model consistently emerge in lesion analyses and studies using various phonological, grammatical, lexical etc. stimuli and tasks.

Why are these regions prominent in language models? Lesion and neuroimaging studies supporting the classical and dual-stream models are usually based on simple stimuli and tasks. In studies of lower-level speech perception, participants might listen to phonemes or syllables and press a button when they detect some instructed difference (Goranskaya et al., 2016). In studies of word processing, participants might read single words vs nonwords (Braun et al., 2015). Further, in studies of sentence processing, participants might listen to normal sentences vs nonword sequences (Fedorenko et al., 2010; Fedorenko et al., 2012). Resulting activation patterns from these simplistic and artificial tasks are then typically subtracted in some manner and averaged within and then across participants. The resulting patterns for all of phonetic, grammatical, lexical and semantic tasks consistently map onto a set of common regions (STG, MTG, and IFG). Further supporting this, neuroimaging meta-analyses also show consistent activation of the IFG, STG, and MTG with some additional variability around these (Fig. 9).

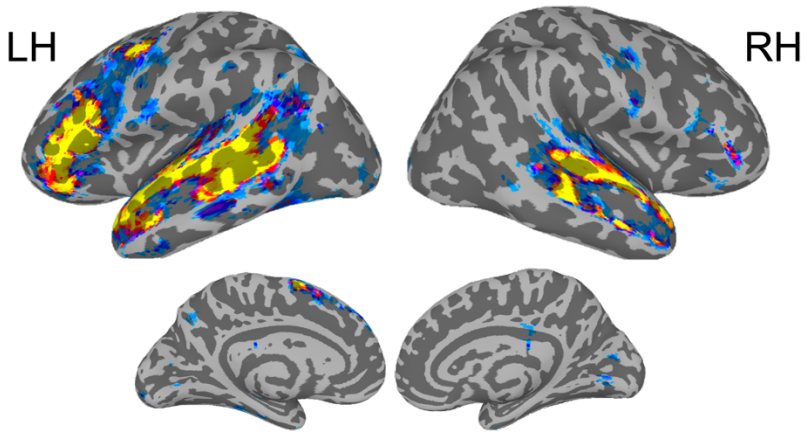



**Figure 9.** Overlap of various language meta-analysis terms from Neurosynth (Yarkoni et al., 2011). These include the meta-analysis terms ‘language comprehension’, ‘comprehension’, ‘sentence comprehension’, ‘speech perception’, ‘language network’, ‘language’. Yellow indicates highest overlap between meta-analyses, Orange/Red represents some overlap, and Blue represents unique patterns. IFG, STG, and MTG consistently appear across meta-analyses.

Why do different task-based studies and meta-analyses result in the same ‘language’ regions? The simplest answer is that the set of regions found in the literature are indeed the sole loci for processing language in the brain, and that this is the true extent of the language

processing network. Alternatively, 'language' regions may result as a product of central tendency measures, such as averaging, and subtracting methods. Indeed, averaging over a multitude of words or sentences that have different meanings and contexts would identify only regions that are common to all the words/sentences: these are likely to be some general language processing regions. Moreover, aggregating over participants would identify activity patterns that they all share: these are likely to be again (i) acoustic processing regions, (ii) domain-general language regions, and (iii) domain-general cognitive strategy regions. Any unique activity pattern related to individual word meanings, individual sentence contexts or individual participants would disappear when applying measures of central tendency, which are instead biased towards finding commonalities across all stimuli, tasks and participants. These central tendency methods have been used in nearly all the existing literature on language comprehension, as they allow to draw conclusions at the population-level.

What happens when we do not use central tendency measures? When investigating individual elements (e.g., individual words, sentences etc.), many studies have found more distributed activity patterns during language comprehension. For instance, studies on word semantics revealed that individual words are represented in brain regions outside known language areas based on the meaning or concept that each word evokes: names of objects related to acoustic features (e.g., 'telephone') activate auditory regions (Kiefer et al., 2008), names of colours (e.g., 'blue') activate colour-processing visual areas (Martin et al., 1995), and words representing object categories activate unique activity patterns in visual cortices, matching the patterns activated when looking at images of the same categories (Shinkareva et al., 2011) even in the absence of the auditory or visual stimulus respectively. In a more detailed study, Huth and colleagues (Huth et al., 2016) mapped the activity of individual semantic categories, showing that these elicit (i) unique activity patterns and (ii) together tile nearly the whole brain. These distributed regions are recruited early during processing of a word, within 150 ms from word onset, meaning they do not represent a post-perceptual process, but are rather an integral component of the processing of the word of interest (García et al., 2019; Kiefer et al., 2008; MacGregor et al., 2012; Shtyrov et al., 2014).

Given that 'language' regions are still present in individual variability maps, it seems unlikely for these regions to be mere artefacts of central tendency measures. An alternative and more plausible explanation is that 'language' regions have a central role in coordinating other highly distributed and dynamic language processing areas, and thus are active during any language task. To explore these putative regional differences, one could apply network

measures. Indeed, network studies have already identified a set of regions that have significant influence on the rest of the brain, and where most of the brain's connections concentrate (Hagmann et al., 2008). These are known as hubs or rich-club and play a fundamental role in network communication and integration (van den Heuvel & Sporns, 2013). Across the whole brain the main identified hubs include the cingulate, precuneus, insula, superior frontal and superior temporal cortices (De Domenico et al., 2016; Hagmann et al., 2008; van den Heuvel & Sporns, 2011). These form the main 'backbone' upon which other regions can communicate and interact (Fornito et al., 2016). Furthermore, studies have shown that a hierarchy of hubs exist in the human brain (van den Heuvel & Sporns, 2013). Intermediary hubs have important roles at more local scales, whereby they may link various subnetworks (i.e., 'connector' hubs for integration), or have a central role within a specific subnetwork (i.e., 'provincial' hubs for coordination) (Fornito et al., 2016; van den Heuvel & Sporns, 2013).

Although 'language' regions do not appear as global hubs (except for the STG), these regions were shown to act as provincial hubs across participants and in various tasks (Bassett et al., 2013; den Ouden et al., 2012; Li et al., 2020). Studies on the network organisation of language and the brain are limited, with most investigating connectivity within 'language' regions through the use of language localisers or pre-selected regions-of-interest (ROIs), leading to the idea that a circumscribed 'language network' exists (Chai et al., 2016; Fedorenko & Thompson-Schill, 2014). However, the few studies focusing on whole brain connectivity and individual variability have identified a much more distributed and hierarchical 'language network', within which 'language' regions constitute the top of the hierarchy (Akiki & Abdallah, 2019; Hertrich et al., 2020). For instance, Akiki & Abdallah (2019) computed nodal consistency of voxels grouped into 22 functional sub-networks; the results showed that 'language' regions had some of the lowest consistency values, which the authors interpreted as these regions acting as connectivity hubs during various tasks. Taken together, these limited findings tentatively suggest that 'language' regions may be hubs of a wider language processing network.

Here, we test the hypothesis that central tendency measures have so far only revealed language hubs. We propose that the neurobiology of language is (i) highly dynamic and distributed in the real world, and that (ii) 'language' regions are intermediary (provincial) hubs. In order to achieve this, we investigate the neurobiology of language processing during a naturalistic movie-watching task, using the Naturalistic Neuroimaging Database (NNDb) data. In the present manuscript we selected 38 participants from the NNDb (20 watched '500 Days

of Summer' and 18 watched 'Citizenfour', average movie length = 6137 sec). In order to test whether the neurobiology of language processing is distributed during sensorimotor representation of words, we scored individual heard words for their sensorimotor embeddings over 11 dimensions (auditory, visual, gustatory, haptic, interoceptive, olfactory, foot-leg, hand-arm, mouth, head, torso) and analysed how these drive activity patterns in the brain. We then computed the average of all heard words and the average of all individual sensorimotor maps, and compared these to the language meta-analysis map from Neurosynth (Yarkoni et al., 2011), to study whether only the aggregate resembles 'language' regions. We then constructed individual voxel-wise networks using a sliding-window approach and measured the strength of connectivity (or centrality) for each voxel. Centrality values were aggregated over time, space and participants to test whether only 'language' regions emerge from the aggregate as provincial hubs.

## Methods

For more details on participants, data acquisition and preprocessing, please refer to Chapter 2 and the original publication (Aliko et al., 2020).

### *Neuroimaging data*

We obtained fMRI data from 38 participants (19 females, range of age 19-58 years, M = 27.4 years, SD = 10.2 years) in the Naturalistic Neuroimaging Dataset (NNDb) (Aliko et al., 2020). We selected only participants who watched either '500 Days of Summer' or 'Citizenfour' from the 10 available in the NNDb; of these, 20 participants watched '500 Days of Summer' and 18 watched 'Citizenfour'. The data was preprocessed as detailed in Chapter 2.

### *Lancaster norm annotations*

Movies were annotated for word onset and duration using available subtitle scripts and the '*Amazon Transcribe*' tool from '*Amazon Web Services'* that performs speech-to-text translation (see Chapter 2). The audio file of the movie was converted into text with timings for on and offset for all words. Because the transcript did not capture all spoken words, some timings were estimated using a script that applied dynamic time warping (Aliko et al., 2020). Finally, words that were contracted (e.g., "he'd" instead of "he would") were modified to their full spelling and the onset and duration of each of the words was estimated by dividing the

original duration by the number of letters in the new words. For instance, if the word “he’s” had originally an onset at 10 seconds and duration of 2 seconds, the new spelling would have duration of:

- “He” has onset at 10 sec and duration of 2 sec/4 letters new spelling = 0.5 sec/letter * 2 letters in word = 1 sec
- “Is” has onset at 10 sec + 1 sec = 11 sec and duration 2 sec - 1 sec = 1 sec

Although in spoken English, the ‘he’ in ‘he’s’ would last longer than the ‘s’, we estimated the duration of the non-contracted version from the full spelling.

In order to investigate the activity distribution of single words, we overlapped the full word annotations to the Lancaster Sensorimotor Norms (LSN) that provide the largest perceptual and action assessment of ~40,000 English words, collected from the average of 3,500 individuals’ scores on a scale 0-5 (Lynott et al., 2020). We are aware of only one similar study to ours that has used LSN in an fMRI setting to study language comprehension. Here, the authors used LSN together with other psycholinguistic scores (e.g., word concreteness and word frequency) and extracted principal components to use as modulators in a naturalistic narrative fMRI study: the findings revealed distributed activation in areas such as DMN, insula, occipito-parietal cortex etc., during language processing (Wu et al., 2022).

On average, 95.8% of the words in the movies had a corresponding entry in the LSN (M = 11,277; SD = 3708.1 words). The resulting scoring from LSN produce 11 regressors for the following sensorimotor entries in order of appearance: auditory (A), gustatory (G), haptic (H), interoceptive (I), olfactory (O), visual (V), foot/leg (Fl), hand/arm (Ha), head (He), mouth (M) and torso (T). Regressors for words in the movie annotations that overlapped the LSN database were separated from words in the movie not classified in the LSN database (M = 4.2%, SD = 0.8% of words in movies), resulting in two text files of the following format respectively:

Regressor 1 (LSN classification present). *Onset*A,G,H,I,O,V,Fl,Ha,He,M,T:duration*

Regressor 2 (LSN classification missing). *Onset:duration*

Two confound regressors for low-level visual and low-level auditory features were also included for each word in the model, to control for effects due to visual stimulation and auditory ones unrelated to words. We selected sound energy as the auditory control regressor. Sound

energy measures the root-mean square acoustic energy of an audio signal, meaning how loud the audio signal is (Shain et al., 2020). Sound energy was calculated every 100 ms using the Python library *librosa* (on average, $M_{value}$= 8x10$^{-3}$ W, $SD_{value}$ = 4.5x10$^{-3}$ W) (McFee et al., 2015). Contrast luminance was selected as the visual regressor: it measures the standard deviation in luma (brightness) values of the pixels in an image (Goodyear & Menon, 1998). Contrast luminance was computed at every frame in the movie using the Python library OpenCV (on average, $M_{value}$ = 53.7 lm, $SD_{value}$ = 17.6 lm) (github.com/opencv/opencv). Both the sound energy and contrast luminance values were averaged over the duration of the words where the LSN classification was present. For instance, if a word had a duration of 200 ms, two values of sound energy (each at 100 ms) would be averaged together and 5000 values of contrast luminance would be averaged together (each at 0.04 ms). In the event that the word duration was smaller than the sampling rate of either control regressor, the value of 1 sampling step was assigned to the word. Thus, for instance, if a word had a duration of 10 ms, the sound energy value assigned to the word would be 100 ms, and the contrast luminance would be the average of 250 values.

Finally, a third control regressor was included in the model, namely word frequency of individual words, in order to remove effects due to the commonality of the word rather than its sensorimotor embedding (Willems et al., 2016). We used the log-transform of word frequency database Subtlex UK (van Heuven et al., 2014), because even though the movies were US productions, our participants lived in the UK at the time of the study (on average, $M_{value}$ = 6.2, $SD_{value}$ = 1.2). One word in 'Citizenfour' did not have an associated word frequency value, because it was missing in the Subtlex database; we therefore assigned a value of '0' frequency to the word. This would not affect the final results of the analysis, since it constituted 1 word out of 13,898 other words within the same movie.

The final file containing words that overlapped the LSN database had the following format:

*Onset*A,G,H,I,O,V,Fl,Ha,He,M,T,luminance,soundpower,frequency:duration*

### *Multiple linear regression and linear mixed effects analysis*

Multiple linear regression using duration and amplitude modulation was performed using the AFNI program *3dDeconvolve* (Cox, 1996) with three regressors: (i) regressor of interest for single words that had LSN scores, sound energy (low-level auditory feature), contrast luminance (low-level visual feature) and word frequency confounds over the duration

of each word (Mante et al., 2005; Moulden et al., 1990; Shain et al., 2020; van Dijk et al., 2020), (ii) regressor for single words without LSN scores, and (iii) regressor for times where no words are present (non-words) (on average, $M_{value}$ = 1707 sec, $SD_{value}$ = 179 sec). The amplitude modulated regression identifies areas of the brain where the BOLD signal varies proportionally with the regressors of interest; while the duration modulated regression identifies areas of the brain where the BOLD signal varies proportionally with the duration of the stimulus (Cox, 1996). The linear regression also outputs the effect of the baseline stimulus (e.g., words) on the BOLD signal, which we called 'words' in this manuscript.

The resulting beta maps from the multiple linear regression with amplitude modulation were input into a linear mixed effects (LME) model using *3dLME*, since the individual words were sampled within-participant (Cox, 1996). In the LME model, we set beta coefficient, age, gender and movie watched for each participant as fixed effects. We set participant as a random effect, whereby the intercept of the slope was allowed to vary by a small random amount compared to the group average for each participant. We computed the baselines for all 11 Lancaster norms and for 'words'.

The results of the LME for each Lancaster norm map and for the 'words' map were corrected for multiple comparisons using a cluster-size correction procedure in AFNI. First, we estimated the smoothness and autocorrelation function of neighbouring voxels using the *3dFWHMx* command (Cox, 1996). Then we ran *3dClustSim* over 6 uncorrected individual voxel p-values (.05, .02, .01, .005, .002, .001) and an alpha threshold of .01. Using the significant cluster sizes whereby faces or edges need to touch, and voxels are contiguous if they are either positive or negative at each p-threshold, we merged the thresholded maps at each p-threshold to obtain significant voxels ($\alpha$=0.01).

### *Centrality analysis*

We constructed time-varying connectivity matrices using a sliding-window approach. First the original fMRI timeseries was resampled to $5mm^3$ to reduce computational complexity of the network analyses. The timeseries was then divided into windows of 1 min length, sliding every 10 sec to allow for a 50 sec overlap between one window and the next. A pairwise Pearson's product moment correlation coefficient was computed on each window using the AFNI program *3dDegreeCentrality* (Cox, 1996). The resulting correlation matrices were

proportionally thresholded to obtain a 10% sparsity in each time window: the top 10% values were considered a connection between two voxels and used to build a connectivity matrix.

Centrality of a node measures how important that node is for the integrity and information-flow of the network. Centrality can be determined using various metrics that provide different information on the role of the node of interest in the network. Four centrality values were measured at each voxel for each window, namely degree, eigenvector, closeness and betweenness. Degree centrality is the sum of inward and outward connections from a node; eigenvector centrality is a measure of influence on a network, meaning that a high-connectivity node linked to nodes of high connectivity will have higher eigenvector centrality (i.e., be more influential) than a high-connectivity node linked to low-connectivity nodes; betweenness centrality measures the shortest paths that pass through a given node; closeness centrality measures the inverse of the distance of shortest paths passing through the node (van den Heuvel & Sporns, 2013). Although these centrality metrics provide different details on a node's importance, they are highly correlated to one-another (Li et al., 2015; Oldham et al., 2019), thus ranking nodes across measures is most informative to create a detailed map of the network nodes' influences (van den Heuvel & Sporns, 2013). We ranked nodes based on each of the four centrality measures, calculated the Spearman's ranking correlation coefficient ($\boldsymbol{\rho}$) between pairwise measures, and clustered nodes using Ward's linkage distance. The clusters were further evaluated using the Davies-Bouldin score, to obtain an optimal clustering of nodes across centrality metrics (Oldham et al., 2019).

Hubs are defined as nodes that are most strongly connected to the rest of the network, therefore having an important structural and possibly functional role (van den Heuvel & Sporns, 2011). If the $\boldsymbol{\rho}$ coefficients for the pairwise centrality measures resulted to be significant as the literature proposes, we would average the four Z-transformed centrality scores to create a single centrality value per node. There is no consensus measure or method to determine hubs in a network, thus we defined hubs to be the nodes in the 90th percentile of the average centrality score. Although a cut-off of 90th percentile is arbitrary, it allows for strong selectivity of nodes while still maintaining the configuration of clusters of high centrality. The thresholded centrality window maps were input into the Affinity Propagation Clustering (APC) algorithm, in order to determine exemplar configurations that are stable across time (Bodenhofer et al., 2011). The resulting dendrogram that APC outputs was cut in half, such that clusters of interest were the ones surviving the halfway cut-off of the tree. In order to test whether canonical language regions only appear in the group average rather than be very stable

across time in a single participant, we concatenated the most representative windows for each cluster (known as exemplars) at the halfway mark for each participant into one map and Independent Component Analysis (ICA) was computed over 100 dimensions using *melodic* (Woolrich et al., 2009). *Melodic* normalises variance of the timecourses and thresholds maps at $p > .5$, which assumes an equal loss from false positives and negatives (Woolrich et al., 2009). Due to the nature of fMRI data collection and preprocessing, noise cannot be completely removed from the final dataset. Thus, ICA not only outputs stable components, but may also include noise ones that could be due to regular physiological or machine noise. We selected non-noisy stable components manually, identifying them as components that conform to the grey matter forming largely bilateral patterns, and do not fall into regions outside of the brain or in white matter and cerebrospinal fluid areas. Noise components that included areas outside of the brain, were randomly distributed or included white matter and ventricles were discarded. Finally, we measured the spatial correlation coefficient of each exemplar from APC and each ICA component to the language meta-analysis regions.

## Results

### *Distribution of sensorimotor properties of words*

We first tested the hypothesis that word processing results in distributed patterns of activity throughout the entire brain and that this pattern is obscured by the use of measures of central tendency, resulting in 'language' regions (i.e., most commonly, the S/MTG and IFG). To do this, we used a method previously demonstrated to result in distributed patterns of activity when the semantic properties of words are taken into consideration (rather than simply averaged over) (Huth et al., 2016). Differently from the previous study, we used the LSN database to score words in movies based on 11 sensorimotor embeddings and used these as modulators of the BOLD fMRI signal. Below is an example of how two words are scored in the LSN on a scale of 0-5 (red = highest score category for the word).

| Word | A | G | H | I | O | V | Fl | Ha | He | M | T |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *LOVE* | 2.056 | 0.722 | 3 | **4.389** | 1.056 | 2.667 | 1.5 | 2.444 | 3.667 | 3.611 | 3.333 |
| *TABLE* | 0.684 | 0.053 | 3.263 | 0.158 | 0.158 | **4.737** | 1.65 | 2.45 | 1.75 | 0.55 | 1.35 |

This analysis produced 12 beta maps, one for 'words' and 11 for each of the sensorimotor modulators of those words. For the group level analysis, 'words' and sensorimotor beta coefficient maps were input into a linear mixed-effects and corrected for multiple comparisons (Chen et al., 2013). The corrected maps for words and the individual effects of Lancaster norms are shown in Fig. 10. Each LSN map shows a unique distribution, although in patterns not related to their perceptual reference (e.g., Olfactory map did not activate olfactory cortex).

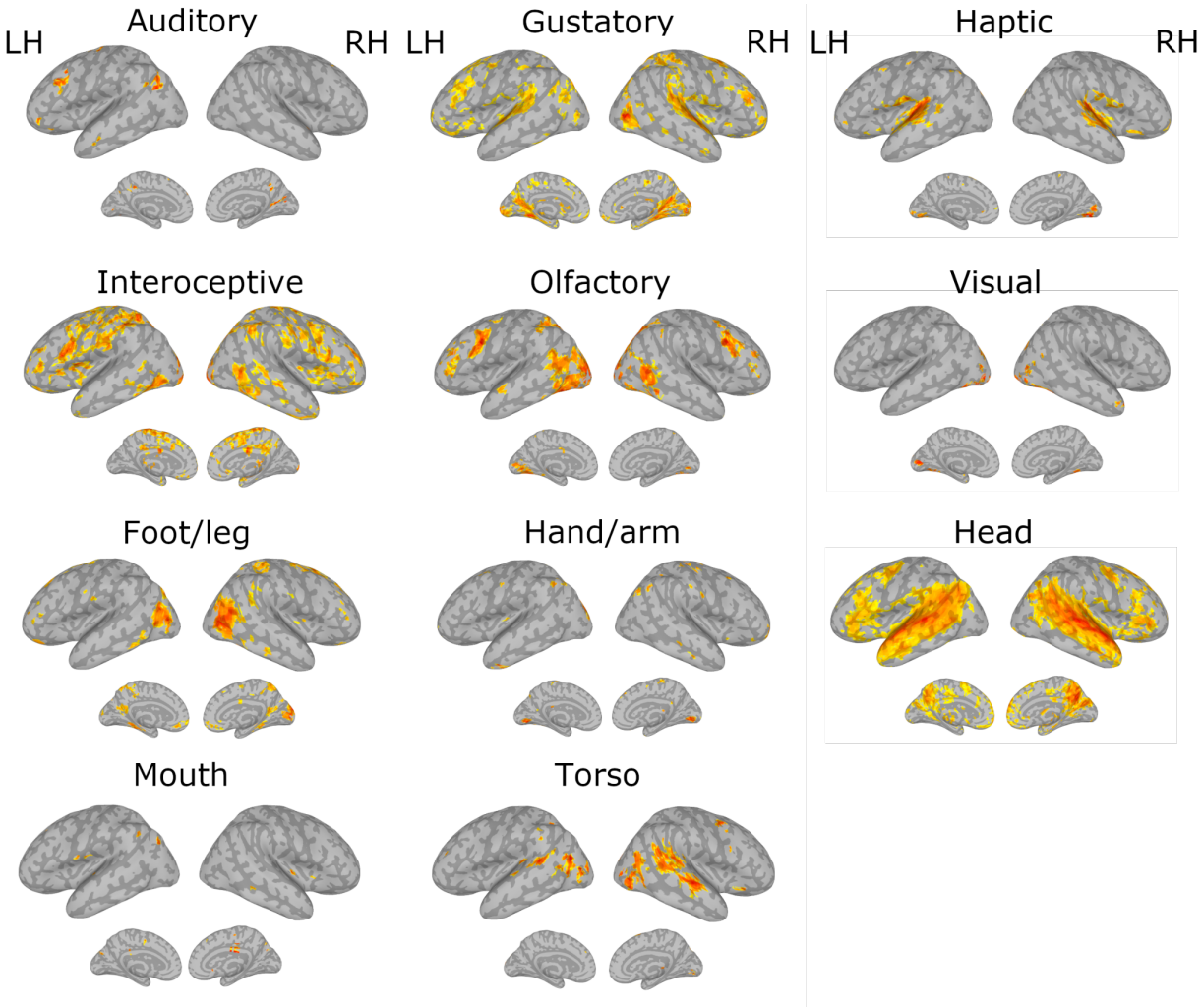



**Figure 10.** Maps for each of the 11 sensorimotor embeddings. Maps were corrected for multiple comparisons with a cluster-size correction and multiple thresholds approach ($\alpha$ = 0.01). Each map formed unique and distributed activity patterns, mainly around (i) primary auditory, motor, visual and premotor areas, (ii) subcortical regions, (iii) some frontal regions. For all maps, cluster size = 20.

In order to determine whether the 'words' map closely resembled 'language' meta-analysis regions and whether sensorimotor maps were more distributed, we computed the spatial correlation of each positive map to the 'language' meta-analysis map from Neurosynth (Yarkoni et al., 2011). The 'words' map had r = 0.42 spatial correlation with the 'language' meta-analysis map (Fig. 11). The primary difference was the relative lack of IFG in the 'words' map. In contrast, the individual sensorimotor maps had poor spatial correlation (on average $M_r$ = 0.09, SD = 0.11) with the 'language' meta-analysis map.

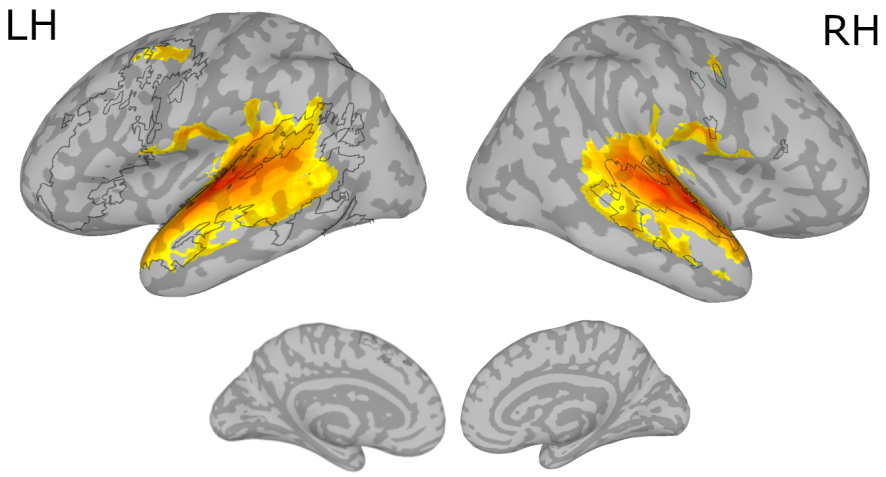



**Figure 11.** 'Words' map corrected for multiple comparisons using a cluster-size correction and multiple thresholds approach (**α** = 0.01). This represents the activity resulting from all words in movies. The 'words' regions are highly correlated to the 'language' meta-analysis map (black outline). However, the IFG is mostly missing in the 'words' map and activity is more equally distributed bilaterally. Cluster size = 20.

We then grouped the sensorimotor maps together to investigate how distributed the overall pattern of activity was. The overall sensorimotor map extended over 63.7% of other brain regions outside the 'language' meta-analysis areas (i.e., regions resulting from the difference between all brain voxels and 'language' meta-analysis areas, with the exclusion of white matter and ventricles) (Fig. 12A). Individually, however, sensorimotor norm maps extended outside 'language' meta-analysis regions by M = 8.0% (SD = 8.1%, range = 0.9% - 25.2%). We then thresholded the overall sensorimotor map to the 90$^{th}$ percentile of values in order to test whether multiple applications of central tendency measures and high thresholding would also result in the 'language' regions. Here, the remaining regions map to the STG, IFG

and occipital cortex (r = 0.26 with 'language' meta-analysis map, Fig. 12B). Although the IFG now somewhat appears in the map, the presence of occipital regions and lack of premotor areas likely drive down the spatial correlation value with the 'language' meta-analysis map. Moreover, the activity in the S/MTG is less distributed than in the 'words' map, potentially due to the high threshold (90$^{th}$ percentile) applied here. Indeed, the unthresholded sensorimotor map (Fig. 12A) included these missing areas, as well as the whole IFG.

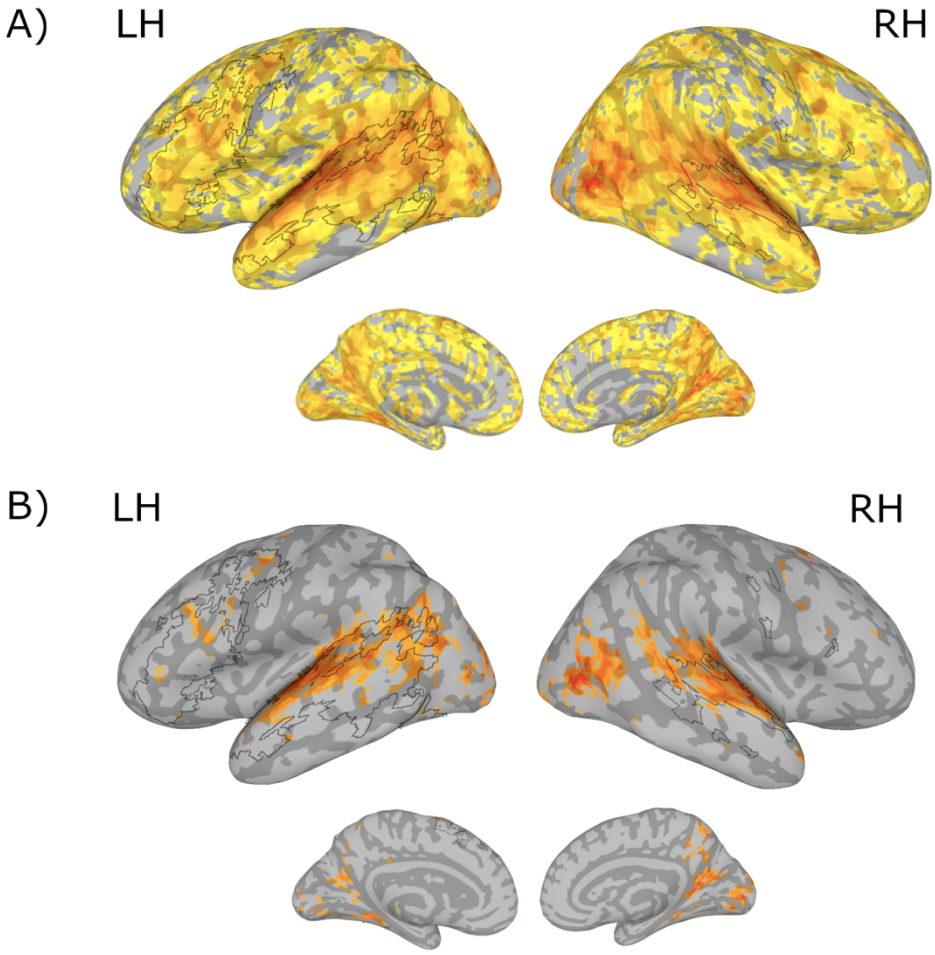



**Figure 12.** Map of distribution of sensorimotor embeddings. **A)** Overall distribution of sensorimotor embeddings after multiple comparisons correction at **α** = 0.01, with Orange/Red = higher average beta values and Yellow = lower average beta values. The distribution encompassed many regions outside 'language' areas (black outline). **B)** Thresholded (90th percentile) values from average distributed sensorimotor map correlate with the Neurosynth 'language' meta-analysis map (black outline). Cluster size = 20.

*Connectivity of canonical language and distributed regions*

The prior results suggest two measures of central tendency would yield primarily STG/MTG and IFG regions. We next tested whether these 'language' regions act as connectivity hubs while distributed regions form a dynamic periphery. To do this, we measured four centrality metrics (degree, eigenvector, betweenness, closeness) for every voxel, or node. Despite differences, the four metrics had a significant ($p < .001$) Spearman's ranking correlation ($M_{rho} = 0.94$, $SD_{rho} = 0.02$) with one another at the group level (i.e., across time windows and across participants). We averaged the Z-transformed centrality metrics and thresholded them to the 90$^{th}$ percentile to obtain the most connected nodes. We then applied APC in order to identify temporal cluster configurations of high-connectivity states (Bodenhofer et al., 2011). This means that if a group of high-centrality nodes recurred over time, it would constitute a stable APC cluster. This method identified on average M = 51.6 (SD = 6.0) temporal hub configurations across participants. Among the ~52 exemplars, we found, on average, low (Mr = 0.11, SDr = 0.05) spatial correlation with the 'language' meta-analysis regions; only 0.2% had a medium correlation comparable to the 'words' map (> .3).

In order to test the hypothesis that 'language' regions appear in the aggregate because they are high-connectivity hubs coordinating distributed regions, we ran independent component analysis (ICA). Although 'language' meta-analysis regions were correlated with few APC exemplars, we hypothesise that they will correlate much more with components of the aggregate. Here, we identified 33 non-noise ICA components. We computed the spatial correlation of each of the 33 states with the 'language' meta-analysis map, to determine whether any of the states matched 'language' meta-analysis regions. Two of the 33 components had r = 0.52 and r = 0.40 correlation value with the 'language' meta-analysis regions respectively (Fig. 13), whilst the other 31 components had $M_r = 0.04$, $SD_r = 0.05$ spatial correlation on average.

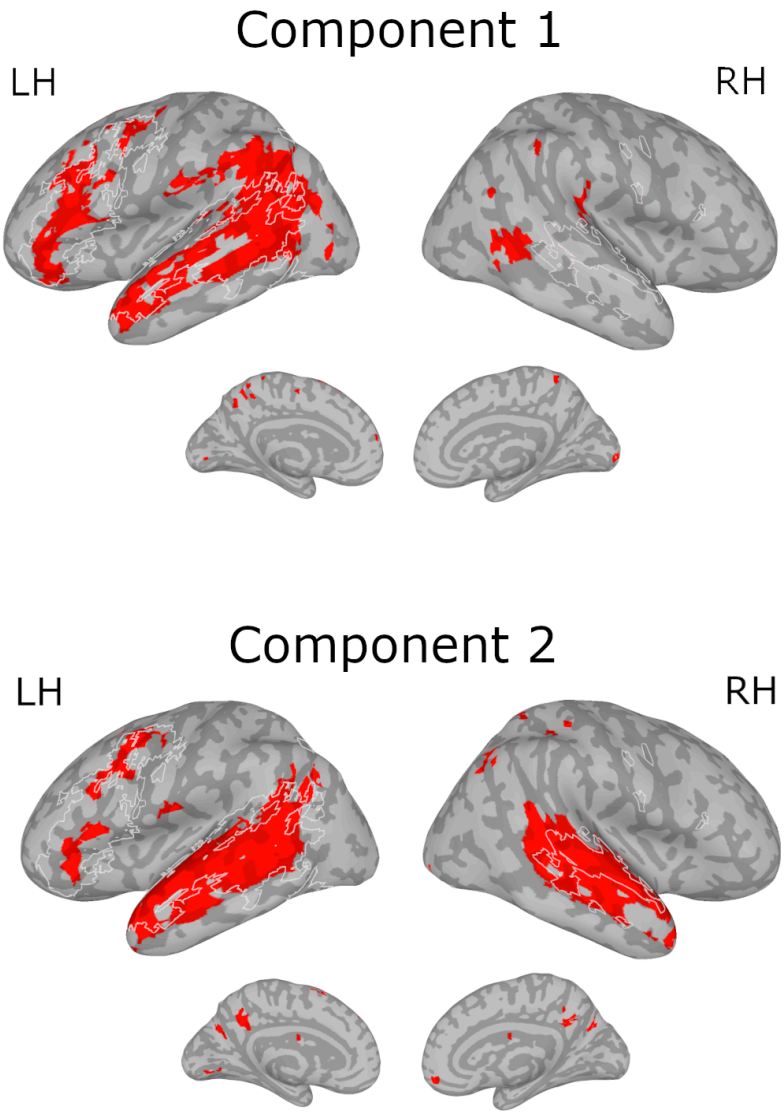



**Figure 13.** Two ICA components from aggregate analyses that had high correlation with the 'language' meta-analysis map from Neurosynth (white outline). Component 1 had r = 0.52 and component 2 had r = 0.40 correlation. Cluster size = 20.

To further investigate potential differences between the connectivity profiles of 'language' regions and sensorimotor map regions, we clustered the original four centrality values using Ward's minimum variance, which consistently divided voxels in each time window into two groups (M = 2.00, SD = 0.02) across all participants: one high-centrality and one low-centrality cluster. In very rare cases, the clustering method detected >2 clusters of centrality, but since the vast majority of windows divided centrality values into 2 groups, we recomputed the few outlier time windows by forcing them to split the data into 2 clusters to

investigate spatial variations in connectivity over time. Fig. 14 shows the average voxel-wise cluster affiliation over time at the group level. Voxels closer to a value of 1 are low-centrality ones across most of the time windows, whilst voxels closer to a value of 2 are high-centrality ones most of the time. Voxels with intermediary values (e.g., 1.5) switch often between a high and low centrality state.

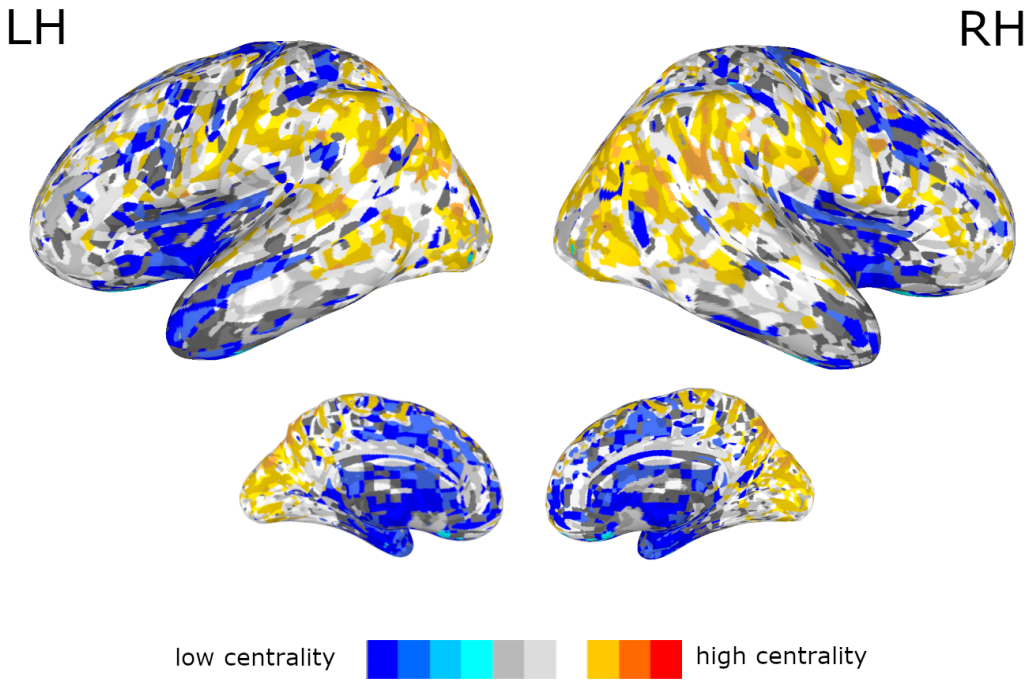



**Figure 14.** Average voxel centrality cluster (high vs low) affiliation over time and participants. Values closer to 2 mean the voxel was a hub most of the time across participants (Red), values closer to 1 mean the voxel was highly dynamic (Blue). Intermediate values (White/Grey) are voxels that change allegiance between low and high centrality (e.g., provincial hubs). MTG and IFG are mostly in intermediary centrality clusters, with STG in high centrality ones. Cluster size = 20.

From the sensorimotor map voxels, we subtracted the 'words' map voxels, in order to maintain only distributed regions not in 'language' areas: this was then called the distributed map. To test the hypothesis that distributed regions are dynamic peripheral nodes most of the time, while 'language' regions are provincial hubs, we computed the mean cluster assignment of the 'words' and distributed map voxels separately for each time window. We considered values >1.7 to be voxels that belonged to high centrality clusters most of the time, since the

maximum cluster value after averaging over windows was 1.863, and therefore >1.7 represents ~90% of the maximum value. We found that the 'words' map voxels had a value >1.7 for M = 24.2%, SD = 10.1% of time windows across participants. In contrast, the distributed map had a value of >1.7 for only M = 3.6%, SD = 2.1% of time windows across participants. Finally, in order to test whether 'words' map regions strongly connect to distributed regions when the former acted as hubs, we inspected the specific connectivity profiles of windows where the mean cluster assignment of 'words' map voxels was >1.7. This analysis showed that when 'words' map voxels were hubs, they shared M = 44.0%, SD = 2.2% of connections with the distributed regions across participants.

## Discussion

Here, we tested the hypothesis that the neurobiology of language processing is highly dynamic and distributed across the brain during natural language comprehension, and that the use of central tendency measures has averaged out activity in the dynamic distributed regions, resulting in only 'language' regions, as these are provincial hubs. Our results showed that each sensorimotor embedding gave rise to unique patterns in the brain, whose activity extended to regions well beyond 'language' regions (Fig. 12). Instead, the overall effect of 'words' and of the thresholded and averaged sensorimotor maps in the brain led to an activity profile that closely resembled 'language' meta-analysis regions (Fig. 11 and 12B), suggesting that only when averaging over stimuli and at the group level we begin seeing patterns resembling current language models. From a network perspective, 'language' regions acted as provincial hubs forming nearly half of all connections with the distributed regions (Fig. 13 and 14). Here, we review each of our findings in more detail in the sections below.

### *Distributed regions in language processing*

We found that individual sensorimotor embeddings of words produce highly distributed patterns of activity that encompass other regions of the brain outside of 'language' areas, adding to similar evidence from the semantic embedding literature (Huth et al., 2016). When grouping all sensorimotor embedding maps, we found that the pattern of activity encompassed large portions (>60%) of the rest of the brain. Although each sensorimotor embedding map overlapped on average only ~8% of the rest of the brain, individual words are represented by multiple sensorimotor embeddings. For instance, the word 'boy' from '500 Days of Summer' scored particularly highly for all of Auditory (2.438), Visual (4) and Head (2.789) domains.

This means that an individual word's embedding will elicit a brain activity pattern that is much more distributed than a single sensorimotor embedding map, and thus will more closely resemble the overall distributed map, with the distribution skewed towards the more significant sensorimotor embedding domain of the word. The distributed regions included parts of the prefrontal cortex, premotor and primary motor regions, insula, posterior cingulate, angular gyrus, precuneus, occipital cortex, some subcortical regions, primary auditory and sensory association areas.

Many other studies investigating individual variability, outside the semantic embedding literature, support our finding on the brain areas forming distributed language regions. For instance, studies on the predictive processes that help in relating previous semantic context to incoming words found an involvement of the hippocampal complex (Maess et al., 2016; Piai et al., 2016), while the medial prefrontal and posterior cingulate were shown to process episodic and semantic memory words (Hertrich et al., 2020). Moreover, the precuneus and temporal lobe bilaterally, were implicated in processing context-specific semantic meanings (Hertrich et al., 2020).

Individual brains vary anatomically, therefore it follows that functional activity patterns will vary as well (Juch et al., 2005). Aside from structural differences, individual differences in cognitive performance, experience of the real-world and cognitive strategy all contribute to increasing functional variability (Van Horn et al., 2008), having implications for language processing. Supporting this idea, intersubject variability studies using memory retrieval of words have shown that individual participants activate different and largely distributed activity patterns (e.g., supplementary motor area, prefrontal cortex, etc.) that relate to their ability to 'visualise' the word (Miller et al., 2012) or to individual cognitive strategies (Heun et al., 2000).

Studies on context of sentences revealed that the brain activates regions related to the meaning of a sentence, either spatially or temporally. For instance, the preceding context to an action verb in a sentence activates primary motor regions in anticipation of the upcoming verb (Schuil et al., 2013). On a temporal scale, sentences describing past and present events map onto occipital and parahippocampal cortices usually associated with concrete object processing, whilst sentences describing future intentions activate regions in the medial prefrontal cortex, temporo-parietal junction and posterior cingulate usually associated with the mentalizing network (Gilead et al., 2013).

Studies on language experience have also identified extended regions of language processing, with different distributions. For instance, some have suggested that formulaic expressions, meaning multi-word expressions that are overused in daily communication, are processed in subcortical regions (Van Lancker Sidtis & Sidtis, 2018; Sidtis et al., 2018). Indeed, usage of formulaic and overlearned language is often maintained during aphasia, even when extensive damage to 'language' regions has occurred (Van Lancker Sidtis & Sidtis, 2018), and this may be because subcortical regions, that may be processing formulaic speech, are preserved in aphasia (Van Lancker Sidtis, 2012). In a recent study, we have specifically demonstrated that overlearned sentences, as opposed to novel sentences, are processed faster and into sensorimotor regions rather than 'language' regions (Skipper et al., 2021). Taken together, this evidence points to a highly distributed network during complex language processing. Our findings offer yet more evidence for an extended neurobiology of language processing that varies with stimuli, with time and across participants.

*Language hubs*

Although variability is clearly important, most neuroimaging studies continue to use central tendency measures to derive stable activity patterns that supposedly represent some feature of language processing. This has led to the notion that 'language' regions are the sole language processing areas across various language tasks.

Here, we showed that central tendency measures applied across stimuli, time and participants inevitably remove all significant variability and reduce language processing to 'language' regions. We found that the 'word' map was highly correlated to current distributions of 'language' regions ($r > .4$). The STG and MTG, but not the IFG, appeared bilaterally in this average 'baseline' map. The pattern in the LH was more similar to the 'language' meta-analysis regions than the RH, with our map showing a more equal distribution in the two hemispheres. This is possibly because words in movies are presented sequentially as part of natural dialogues, while in traditional neuroimaging studies words are presented randomly and in isolation, having no relation to one another. The more bilateral distribution and lack of IFG may be indicative of high semantic context in the movies. This finding reflects previous research showing that homologous language regions in the RH are recruited to process context and narratives (Ferstl et al., 2005; Mitchell & Crow, 2005; Stemmer, 2015), while the IFG is mostly active during higher task demands, such as resolving incongruent references, that would not be required in continuous stimuli, such as movies (Hammer et al., 2007; Martin & Cheng,

2006). Similarly, when averaging and thresholding the grouped sensorimotor maps, the surviving activation fell in and around 'language' regions (r = .26), albeit less than the 'word' map, further suggesting that typical averaging methods mask variability and generally result in 'language' areas.

Our network analyses revealed that the reason for the consistent appearance of these 'language' regions in aggregate analyses, is that they are somewhat stable provincial hubs. Two pieces of evidence support this: (i) we identified two group-level hub components that were highly correlated (r > .4) to the 'language' meta-analysis regions - these hubs involved the STG, MTG (bilaterally in Fig. 13 top and LH in Fig. 13 bottom), left IFG and parts of the premotor cortex; (ii) 'language' regions acted as hubs for ~25% of the time across participants (Fig. 14).

'Language' regions were integrated in a complex connectivity map that included both stable and dynamic regions: the first group involved voxels in STG, occipital, some primary motor and sensory association regions, and the angular gyrus; the second group involved voxels in the insula, cingulate, anterior temporal lobe, supplementary motor area and subcortical regions. Many of the dynamic and intermediary regions were part of the distributed sensorimotor regions, and these areas constituted >40% of all connections to 'language' regions. This result indicates that (i) distributed regions likely share a function with 'language' regions; (ii) distributed regions disappear in the aggregate because they tend to be more dynamic.

Supporting our finding of a hierarchy of stability in the brain, task-based network studies have demonstrated that both spatially and temporally, sensory association and primary regions form strong and stable connections to the rest of the brain, with subcortical areas exhibiting more flexibility (Achard et al., 2006; Bassett et al., 2013; Hwang et al., 2013; Schedlbauer & Ekstrom, 2019). In this organisation, 'language' regions exhibit some variability in connectivity strength, appearing in intermediate layers of the network hierarchy (Bassett et al., 2013; den Ouden et al., 2012; Li et al., 2020).

### *Models*

Current models of the neurobiology of language do not support the distributed and dynamic behaviour of language processing that we have observed here, rather considering language processing as a static and localised network (Fedorenko & Thompson-Schill, 2014).

Instead, a better model should account for all individual variability and treat language as a complex behaviour.

Some recent network studies have identified a novel organisation of the brain network, namely core-periphery, that can unify the hierarchy of connectivity we have observed here (Bassett et al., 2013; Gu et al., 2019). Core-periphery structures combine two network dynamics: the core is a set of highly stable hubs that control a set of flexible regions, or periphery, that vary significantly over time (Csermely et al., 2013; Rombach et al., 2014). Core-periphery networks allow for high complexity, robustness to perturbations and rewiring of connections to maximise energy demands, task requirements and allow recovery after lesion (Cinelli et al., 2017; Csermely et al., 2013).

We take inspiration from these studies, to propose that 'language' regions are part of a global core, tethering a dynamic and flexible periphery of other distributed language processing regions.

*Limitations*

This study inevitably suffered from some limitations, which we will address here. For instance, the distributed sensorimotor embedding regions we have identified may not be performing any language processing, rather activating as (i) a feature of other aspects in the movie or (ii) a post-perceptual language process. Several considerations mitigate against these possibilities. To address the first point, we included sensorimotor embeddings as word modulators to limit the possibility that these were connected to other movie features. Moreover, we added confound and contrast regressors to further control for nuisance from other audio-visual elements of the movies.

To address the second point, previous studies on word semantic processing have shown that distributed regions outside of 'language' areas activate within 50-150ms of the word onset, suggesting their activation is not a post-perceptual process (García et al., 2019; Kiefer et al., 2008; MacGregor et al., 2012; Shtyrov et al., 2014). Moreover, since movies represent a continuous stimulus, there is no opportunity to think or reflect back on the listened words. Finally, we demonstrated that these regions directly and tightly connect to 'language' regions.

A final limitation, with respect to the network analysis, is that we did not inspect specific language features to probe the connectivity profiles, and therefore we may not have

identified language-specific connections or components. In future, we could compare, for instance, high and low word frequency: we expect the former to resemble 'language' features more, and the latter to be more distributed.

*Implications*

In this study we showed that language processing of individual sensorimotor embeddings during real world behaviour forms unique, highly distributed and dynamic patterns of activity. This work adds to a growing body of evidence suggesting that existing neurobiology of language models need to be revisited to incorporate individual variability, contextual variations, etc. (Skipper, 2015). We propose that a better model may be a core-periphery organisation, allowing for (i) high levels of variability through a dynamic periphery, (ii) robustness to perturbations through highly connected cores, (iii) integration of communication for higher task demands through the complex interaction of cores and periphery (Csermely et al., 2013).

This organisation, however, would have ramifications for the way traditional neuroimaging studies are conducted. The consistent use of central tendency measures would obscure the dynamic variability of the periphery, only revealing core structures. Different methods are thus needed to better inspect the network organisation of language and the brain. We suggest methods such as multi-voxel pattern analysis, hyperalignment, deconvolution, cluster-size thresholds, and Bayesian techniques that consider individual variations as well as stable activity and can therefore identify both cores and peripheries (Cohen et al., 2017; Forman et al., 1995; Hasson & Honey, 2012; Haxby et al., 2011).

Finally, our findings have important implications for our understanding of aphasia and its recovery. The high wiring cost of hubs means that damage to these regions would have more deleterious effects on function than damage to dynamic and distributed areas (Fornito et al., 2016; Zhao et al., 2011). As 'language' regions are hubs, this would help explain the symptomatology of aphasia. On the other hand, the presence of the more dynamic distributed regions would explain how the brain mitigates speech deficits via neuroplasticity recovery processes, which are known to happen outside 'language' regions (Hertrich et al., 2020; Kiran & Thompson, 2019). Overall, these findings offer new insights for novel speech therapies into other regions and processes involved in the neurobiology of language.

*Conclusion*

We have demonstrated that when inspecting individual features of words, such as their sensorimotor embeddings, these form unique and distributed patterns of activity encompassing most of the brain. Here, 'language' regions have a role in coordinating these dynamic distributions, acting as provincial hubs. Due to the highly dynamic nature of individual language features, typical central tendency measures have not been able to capture these distributed regions, favouring instead static and localised models of the neurobiology of language.

## *3.3 The brain is a multi core-periphery network with dynamic communities: a flexible model of the neurobiology of language*


## Abstract

Existing models of the neurobiology of language cannot accommodate complex and contextually determined aspects of language processing in the real world. Evidence from studies investigating complex language features points to a distributed and dynamic nature of the neurobiology of language. Thus, a more flexible model of language and the brain is needed to account for this variability and complexity. Three network-based organisations that may support this are (i) a highly segregated organisation, namely modularity, (ii) a highly dynamic organisation, namely core-periphery, or (iii) a combination of both. To account for all complexities of language, we propose that both modularity and core-periphery are needed to support natural language processing, as together they allow for both flexibility and some functional specificity. To test this, we used data from the NNDb and analysed individual time-varying voxel-based networks using core-periphery and modularity algorithms. Results suggest a model whereby 'language' regions are situated in a merged global multi core-periphery and modular network of large, dynamically changing communities. Known 'language' regions constitute one of multiple cores, but only act as such for short time periods. We further demonstrate that distributed brain regions perform language processing, as these form large communities with known 'language' cores, encompassing most of the brain. This organization accounts for the complexity of language processing in the real world and can be informative as to which brain regions and processes have the potential for faster language rehabilitation after lesion.


## Introduction

Language is one of the most complex human behaviours, yet most of the existing neuroscience literature has reduced it to simple task-based studies that in no way represent or account for the natural complexity of language processing. Here, the most cited model of the neurobiology of language, namely the dual-stream model, has perpetuated the notion that language processing is mostly localised to inferior frontal (IFG) and superior and middle temporal gyri (STG, MTG), and that these somehow form two streams for grossly performing 'speech perception' and 'production' (Hickok & Poeppel, 2004). Nevertheless, a growing body of evidence shows that when inspecting more complex features of language processing, such as context or semantics, the neurobiology of language encompasses much of the rest of the

brain, forming distributed and highly variable activity patterns (Huth et al., 2016; Ojemann, 1979; Sidtis et al., 2018; Skipper, 2015; Skipper et al., 2021). This was also demonstrated in previous chapters of this thesis, where we showed that pronoun resolution and sensorimotor word embeddings activate unique and largely distributed activity patterns. In order to account for this high level of complexity and variability, we thus need models of the neurobiology of language that are significantly more flexible. For this, network neuroscience may offer insights on the underlying processes supporting this complex human behaviour. Here, we review network architectures and how they may support the neurobiology of language in the real world.

*Modularity*

Most existing network neuroscience studies describe the network organisation of the brain using resting-state networks (RSNs). These are networks built from BOLD signals of participants in the absence of a task, or rather the participant is left lying in a functional magnetic resonance imaging (fMRI) scanner (Sporns, 2013). RSN studies have indicated that the network architecture best representing the functional organisation of the brain is modularity (Hutchison et al., 2013; Zalesky et al., 2014). Modularity is a property of intermediary (also referred to as mesoscale) network architectures, whereby the network's elements, known as nodes, are grouped into functionally and spatially segregated components, also known as communities (Fornito et al., 2016b) (Fig. 15A). These are defined as clusters with high intra-connectivity density compared to the rest of the network (van den Heuvel & Sporns, 2013).

Research on brain RSNs has shown that the functional connectome is divided into few highly segregated communities that map grossly onto the Default Mode network (DMN) and attention network, as well as to generic behavioural domains (e.g., emotion or perception) (Sporns & Betzel, 2016), and that these are relatively static over time (Hutchison et al., 2013). RSNs have thus portrayed a picture of the brain as a rather static network, with functions clearly separated and localised to specific brain regions. Although RSNs, and modularity, have been shown to well represent the underlying 'baseline' connectivity of the brain at rest (Laird et al., 2011), task-based connectivity studies have identified additional task-evoked dynamics that RSNs could not account for (Cole et al., 2014).

To account for this variability, task-based dynamic functional connectivity studies have proposed a more flexible model of modularity, where modularity is not a static feature, but

rather it describes a quality of functional integration (Park & Friston, 2013). This means that certain regions of the brain have a propensity for a given function, but that they are not necessarily bound to it; their role depends on how best to minimise energy requirements and increase efficiency for the entire network (Bassett & Bullmore, 2006). For instance, the modular organisation of the brain was shown to undergo significant rearrangements during learning (Bassett et al., 2011), during neurodevelopment (Gu et al., 2019), and during disease (Alexander-Bloch et al., 2012; de Haan et al., 2012).

Evidence suggests that these connectivity variations are supported by a hierarchical modular organisation of the brain. Here, larger communities mapping onto general anatomical areas (e.g., occipital, fronto-temporal, and prefrontal) are stable over time, with smaller communities (correlating to multimodal association cortices) experiencing dynamic changes (Meunier et al., 2009, 2010). Further reinforcing the notion of hierarchical modularity, neurological studies have shown that disrupting the organisation of this hierarchy causes weak and random connections to form, leading to deleterious functions (Russo et al., 2014). These findings indicate that hierarchical relationships between larger and smaller communities may support complex brain behaviours, having a fundamental role in maintaining healthy cognitive functions.

How can hierarchical modularity support the neurobiology of language? A possibility is that the neurobiology of language forms a large community in the higher layers of the hierarchy that encompasses smaller dynamic communities at lower hierarchical levels, each of which may support a specific language feature (e.g., semantic or phoneme processing). Although this model would support individual variability of language features, it still does not support complex communication among various language features nor between language processing and other cognitive domains (e.g., attention, emotion, etc.), as communities are by definition segregated. Although inter-community connections do exist, these are usually sparse and mostly serve to integrate information across communities rather than afford functional overlaps (Cherifi et al., 2019; Zalesky et al., 2014).

### *Core-periphery*

Although hierarchical modularity details relationships between communities that may support individual features of language processing, it still holds a rather localisational and semi-static view that does not fully explain the complex relationships between language and

other cognitive domains that task-based studies have identified (Pulvermüller, 2013; Schuil et al., 2013). To address this outstanding issue, a more dynamic and flexible model, such as core-periphery, may be appropriate.

Core-periphery involves two components: a core, whereby a group of nodes is connected to every point of the network, which coordinates a set of dynamic nodes that can only form connections to the core, namely the periphery (Borgatti & Everett, 2000) (Fig. 15B). Although core-periphery structures have been demonstrated in various biological networks, such as protein interactomes, metabolic pathways and cellular signalling pathways (Csermely et al., 2013), they are rarely investigated in the context of functional brain connectivity. This is likely because most neuroimaging studies have mainly sought to isolate stable functional components, such as communities, that also tend to be more consistent across subjects, whilst core-periphery structures relate to dynamic and variable elements of a network (e.g., individual variability) that cannot be detected with central tendency methods (Zalesky et al., 2014).

Core-periphery architectures in such biological networks were shown to have significant evolutionary advantages. First, the core allows for integration of information by controlling a high number of connections; second, a flexible periphery allows for quick environmental adaptations (Faber et al., 2019; Fornito et al., 2016a; Stefaniak et al., 2020). Moreover, due to the high connectivity of the core, the network is afforded significant redundancy and therefore is resilient to perturbation (Cinelli et al., 2017). Indeed, if only few core connections are severed, the redundancy ensures that the function of the core remains largely intact; however, more extensive, and repeated damage to the core causes significantly more disruption to function than damage to peripheral nodes that are loosely connected (Fornito et al., 2016a; Zhao et al., 2011).

How would a core-periphery organisation support language processing? The opposing dynamics of core and periphery explain several complex language behaviours: (i) individual language features may be processed simultaneously in distributed and flexible peripheries, while (ii) 'language' regions could act as a core to help coordinate this distribution; (iii) extensive lesions in 'language' regions (i.e., the putative core) would result in severe aphasic symptoms, but (iv) these functions may be regained through rewiring in peripheral regions. Although this model would support more complex language behaviours, it still assumes that a single core region coordinates a group of peripheral nodes that cannot directly communicate with each other (Borgatti & Everett, 2000).

*Alternative model*

Modularity and core-periphery only partially address the question of how the brain supports language processing in the real world, with each model assuming certain organisational restraints (e.g., non-overlapping communities, or one core and one periphery). An alternative model could include both these organisations with some additional features in each structure.

Core-periphery and modularity have been previously found together in empirical networks, thus the presence of one does not exclude the existence of the other (Rombach et al., 2014; van den Heuvel & Sporns, 2013). For instance, each line in the London Underground network contains some core stations (e.g., Waterloo station) that connect to a wide array of small peripheral stations (e.g., Green Park station), with each metro line representing a single community (e.g., Northern Line) (Rombach et al., 2014). However, the two models exhibit some diverging features, with communities having low inter-connectivity, while the core has high inter-connectivity (Borgatti & Everett, 2000; Newman, 2006). To solve these inconsistencies and allow for both to co-exist, some authors have proposed additional features in each architecture.

For instance, instead of a single core and periphery, networks such as the international airport network were better described by multiple core-periphery pairs, whereby each continent had their own core-periphery group (Kojaku & Masuda, 2017). Unlike classical core-periphery structures, these multiple core-periphery pairs map onto known communities (Kojaku & Masuda, 2017; Yan & Luo, 2019) (Fig. 15D). Conversely, new modularity algorithms allow the existence of community overlaps; here, overlapping areas consist mainly of core regions and better represent complex network relationships (Lancichinetti et al., 2010; Yang & Leskovec, 2014) (Fig. 15C). Although these novel algorithms have identified more detailed and complex features in various networks, only one study, to the best of our knowledge, has investigated the co-existence of these two structures in the resting-state brain to some extent (Gu et al., 2019).

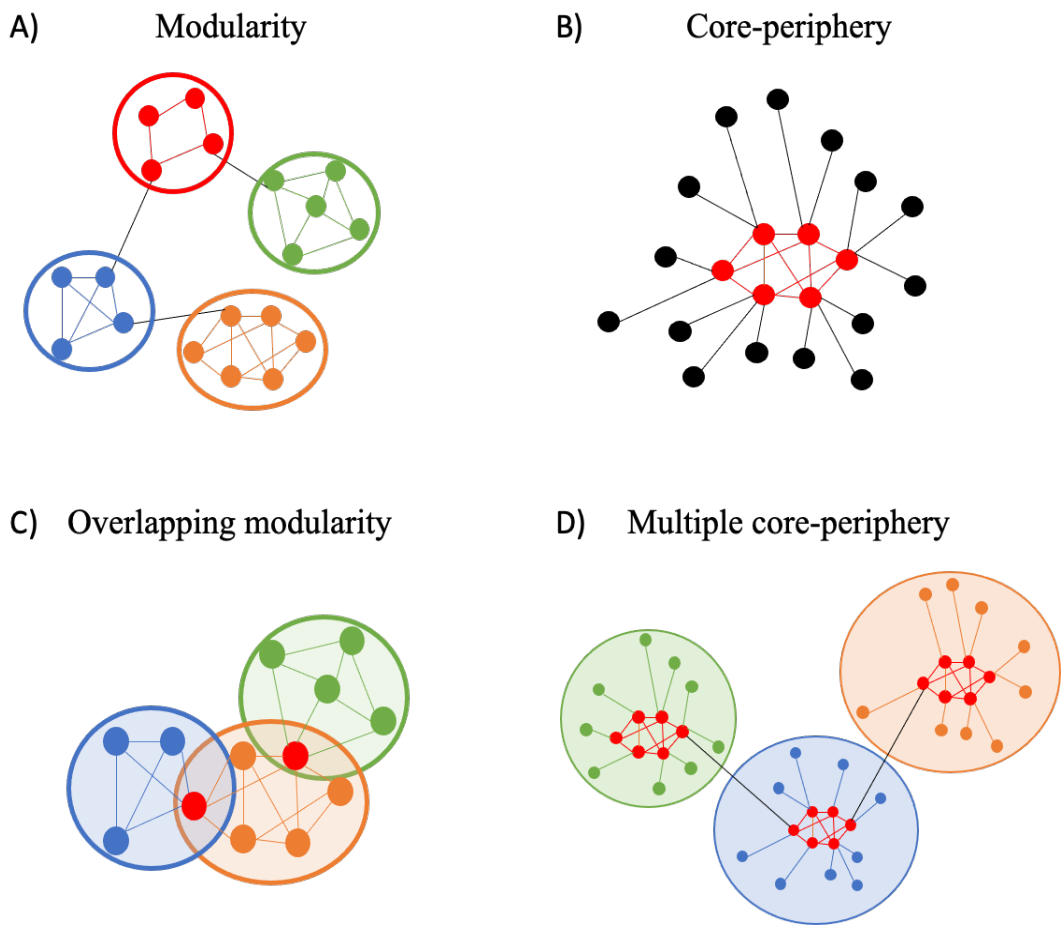



**Figure 15.** Different mesoscale network architectures. **A)** Modularity involves segregation of functions, with communities (red, blue, green, orange) exhibiting high intra-connectivity and low inter-connectivity. **B)** Core-periphery involves a highly connected core (red) and a dynamic and loosely connected periphery (black). **C)** Overlapping modularity allows communities (blue, green, orange) to share some regions (red). **D)** Multiple core-periphery pairs (red and blue/green/orange small circles) map onto separate communities (blue, green, orange large circles).

Here, we propose that the brain network organisation best supporting the complexity and variability of language processing is a combined core-periphery and modular architecture, with multiple dynamic core-periphery pairs and overlapping communities (i.e., Fig. 15D and 15C combined). We propose that these two meso-scale architectures combine such that each community is composed of core-periphery pairs, with cores being more stable components of the network, whilst the periphery drives community evolution. For example, we expect to see

merging and splitting of communities over time, in particular between neighbouring regions. Within this network context, we hypothesise that established 'language' regions act mostly as one of the multiple cores, and that they connect to a large periphery, together encompassing one or more communities. Furthermore, we predict, based on our findings in Chapter 3.2, that primary visual and auditory regions will be the most stable core regions.

In order to test our hypotheses, we used data from 37 participants in the Naturalistic Neuroimaging Database (NNDb), who watched one of two movies: 20 watched '500 Days of Summer' and 17 watched 'Citizenfour' (Aliko et al., 2020). We constructed individual voxel-wise functional connectivity networks using a sliding window approach and analysed these using a novel core-periphery algorithm based on node influence, and a greedy implementation of Newman's modularity algorithm to partition the network into communities (Blondel et al., 2008; Shen et al., 2021). We performed the same analysis on group-averaged networks in order to test the hypothesis that average networks, which are widely used in the literature, have constrained our view of how flexible brain networks are in the real world. Finally, we inspected the specific dynamics of the neurobiology of language. Few studies using voxel-wise networks exist, and these have not analysed individual-level networks (Preti & Van De Ville, 2017; Tagliazucchi et al., 2016; Wink et al., 2012). Ours, to the best of our knowledge, is the first study to analyse voxel-wise individual networks for different brain mesoscale organisations.

## Methods

### *Network construction*

We obtained fully preprocessed fMRI data of 37 participants (right-handed, range of age 19-58 years, $M_{age}$ = 27.5 years, $SD_{age}$ = 10.2 years, 19 females) watching one of 2 movies ('500 Days of Summer' or 'Citizenfour') from the NNDb (Aliko et al., 2020). Originally, the dataset comprised 38 participants, but one participant from 'Citizenfour' was removed post-hoc due to issues with their network construction (further explanation below).

To reduce computational load in network analyses, which are highly computationally costly, the voxel resolution was downsampled from $3mm^3$ to $5mm^3$, resulting in 66,424 total voxels for each participant, of which M = 15,889.7, SD = 471.6 were in-brain voxels after masking. In order to investigate the dynamic functional connectivity of the brain during movie-watching, we divided the fMRI timeseries into 1 min windows with a 10 sec step size (in a

typical sliding-window approach), resulting in a total of 'movie length - 59/10' windows for each participant. Specifically, the movie '500 Days of Summer' resulted in 5470-59/10 = 542 windows, while the movie 'Citizenfour' in 6804-59/10 = 675 windows. There is no agreement on the correct window length and time step to use, thus we tested lengths from 30-60 sec and time steps from 1-10 sec and selected 60 sec length and 10 sec step as the most appropriate for our data and computational resources: in particular, <60 sec windows resulted in inclusion of too much noise, while <10 sec step size resulted in exponentially slower algorithm performance. The choice of window length and step size is also in agreement with the literature, which has mostly used window lengths between 30 sec and 1 min, and window/step ratios <50 (Preti et al., 2017; Zalesky et al., 2014).

The adjacency matrix was constructed for each window using the AFNI program *3dDegreeCentrality*, which computes the pairwise Pearson's correlation coefficient for every voxel (Cox, 1996). We applied a proportional threshold to each matrix, in order to maintain the same edge density across participants and make comparisons between participants watching different movies more robust (Garrison et al., 2015). Since there is no consensus on the thresholding value to use, we tested a range of threshold values (5-30%): at 5% the matrix was too sparse and few connections survived the threshold, while at >15% the matrix was too dense with no discernible patterns and requiring large computational resources. We therefore applied a more appropriate threshold of 10%, meaning that the top 10% of correlation coefficients would constitute a connection (corresponding to 1 in the binary adjacency matrix), with values in the bottom 90% set to zero.

One participant in 'Citizenfour' was removed from the dataset due to issues in constructing their network. Each window should have the same number of nodes at the end of thresholding, but we encountered ~20 windows with less nodes than expected in this participant, possibly a defect introduced during the anatomical alignment preprocessing step prior to time-correction. Adding these voxels back to the adjacency matrix as disconnected nodes would result in network algorithms identifying disjointed elements; alternatively, adding them as a connection would create false relationships in the network. We therefore opted for discarding the participant's data as an outlier (further investigation is ongoing), which resulted in the final 37 participants being included in the present manuscript.

*Individual network analyses*

A network can be described globally, at the meso-scale and at the individual node level. Here, we sought to investigate the meso-scale features of the architecture of the network and the functional relationships between groups of voxels. We therefore computed various graph theoretic measures on every window for each participant. From now on we will refer to single voxels as nodes, for simplicity.

The meso-scale architecture of a network provides information on how groups of nodes are functionally related or clustered. Various algorithms can be used to identify different meso-scale structures. Here, we chose to compute core-periphery and community partitioning algorithms. A core-periphery structure implies that a network is divided into a cluster of nodes with high inter- and intra-connectivity (core) and a group of loosely connected and dynamic nodes (periphery) (Borgatti & Everett, 2006; Verma et al., 2016). We applied the core-periphery algorithm that we developed in a recent publication (Shen et al., 2021), which is able to detect core-periphery structures at higher accuracy and at higher efficiency than other existing algorithms. The algorithm starts by assuming that a node exerts a certain amount of influence on the network, which is calculated using a function derived from a random walk with restart model equation. The resulting node influence vectors are incorporated into a probability matrix of influence scores, with the top 10% of values considered as core nodes (for mathematical proofs see (Shen et al., 2021)).

Community partitioning is an ongoing issue in the field of graph theory, due to its computational complexity (Newman, 2006). The fundamental concept of community partitioning is to identify clusters or modules of nodes in a network that share a common function (Blondel et al., 2008; Lancichinetti & Fortunato, 2012). Many algorithms exist for partitioning the network, but the most widely used is based on the optimization of Newman's modularity function, which clusters nodes into modules if their in-module connectivity is higher than the connectivity between clusters, and compares the value against a null network model (Newman, 2006; Sporns, 2013). The modularity score $Q$ is computed as follows:

$$Q = \frac{1}{2m}\sum_{ij}^{N}[A_{ij} - \frac{k_i\, k_j}{2m}]\, \delta(c_i, c_j)$$

Where $N$ is the number of nodes, $m$ is the sum of edges in the network, $A_{ij}$ is the edge between nodes $i$ and $j$, $k_i$ and $k_j$ are the sum of edges of nodes $i$ and $j$ respectively, and $\delta(c_i, c_j)$ is Kronecker's function for clusters of nodes $i$ and $j$ respectively. Here the null model is the Newman-Girvan matrix ($k_i k_j/2m$) (Bassett et al., 2013; Lancichinetti & Fortunato, 2009). Since optimizing modularity is computationally intensive, we used the greedy modularity algorithm developed by Blondel et al. (Blondel et al., 2008), which we will refer to as the *Louvain* algorithm from now on. The algorithm works in two phases:

1. Initially each node is assigned to a single community. Neighbouring nodes $i$ and $j$ are joined, and $Q$ is calculated. If joining $i$ and $j$ increases the value of $Q$ compared to keeping them separate, then nodes $i$ and $j$ are assigned to the same community. The algorithm stops when changes in assignment can no longer improve $Q$.

2. The clusters are then considered as single nodes, with edges within modules represented as self-loops. Phase 1 is applied to this new network.

The two phases are iterated until a maximum $Q$ score is reached. Since *Louvain* is non-deterministic, it can produce slightly different partitions every time it is applied to the network (Bassett et al., 2013). We therefore performed 100 iterations of *Louvain* for each window and built a consensus matrix $D_{ij}$, where each entry $ij$ is the probability of finding nodes $i$ and $j$ in the same module across iterations. *Louvain* was then run a further 50 iterations on each thresholded $D_{ij}$ matrix. Here, $ij$ pairs that have a probability of being in the same community lower than a thresholding parameter are removed from the $D_{ij}$ matrix prior to re-applying *Louvain*. We tested a range of values for the thresholding parameter $\tau$, specifically values of .1, .2, .3 and .4, the latter being the maximum recommended $\tau$ value for *Louvain* in the literature (Fornito et al., 2016b; Lancichinetti & Fortunato, 2012).

Moreover, modularity is known to suffer from a resolution limit, in that it cannot detect smaller modules because they do not maximise the modularity score (Fortunato & Barthélemy, 2007; Lancichinetti & Fortunato, 2011). One solution to this problem is the addition of a parameter $\gamma$ before the null model term, that allows to resolve smaller clusters (Fornito et al., 2016b). Here we tested a range of values of $\gamma$, namely 1 (default), 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9 and 2, and selected the parameter value that generated the highest similarity score across iterations. To measure partition similarity across iterations, we computed the normalised

mutual information score (NMI), which outputs a value in the range [0,1] with 1 being identical partitions and 0 being different partitions (Taya et al., 2016).

Overall, then, at each $\gamma$ we tested four $\tau$ values; we selected the $\gamma$ value providing on average the highest NMI score; within the chosen $\gamma$ we selected the $\tau$ value producing on average the highest $Q$. These parameters were tested on one randomly selected matrix from each participant (i.e., total of 37 matrices), and the optimal values selected by averaging across all 37 matrices. From these tests, we identified a single $\gamma$ and $\tau$ as optimal, and used them to run *Louvain* on all matrices and participants. Although ideally, we would have run the parameter tests on all windows for all participants, this was computationally unfeasible. From this testing we determined that a $\gamma = 1$ and $\tau = .4$ produced the most consistent partitioning with the highest modularity scores ($M_{NMI} = .74$, $SD_{NMI} = .10$; $M_Q = .65$, $SD_Q = .06$).

Since modularity optimization may produce a high $Q$ score even for random networks, such as Erdos-Renyi random networks (Guimerà et al., 2004), we measured the significance of the partitions found by the *Louvain* algorithm using a non-parametric permutation test. We randomly shuffled the community assignments for all nodes in a window, maintaining the number of clusters and their size the same, and calculated the new $Q$ score (Betzel & Bassett, 2017). We repeated this process for 100 iterations, and measured the p-value as follows:

$$p = \Sigma\,(Q_{permuted} > Q_{real})/\,iterations$$

We considered a significant partitioning as one with p-value < .001. Ideally, we would have run this test 1,000 - 10,000 times, but due to its high computational requirements it was unfeasible.

### *Identification of stable core states*

In order to determine relationships between core configurations over time, we performed Affinity Propagation Clustering (APC) on the coreness values across time windows for each participant (Bodenhofer et al., 2011). APC is a data-driven clustering technique that does not require setting a priori parameters of cluster size, therefore valuable when a ground truth of the data is missing. APC clusters data by their relationship, meaning that the algorithm infers a hierarchy from the data (Bodenhofer et al., 2011). This can be represented as a dendrogram tree with relationships as branches and individual states as leaves. The lowest hierarchical layer represents the most divergent states, with branches in higher layers being

states that are more similar to one another. Here, we selected the middle branches of the hierarchy, meaning the dendrogram tree was cut in half and the top hierarchies were maintained, with their leaves considered as states of interest. This was done because the lowest branches resulted in clusters with very few time points included, while the higher layers contained most of the data and thus did not identify enough time configuration clusters. These states thus represent different temporal configurations of the network's core.

In order to compare core-periphery configurations at the individual and aggregate level, we performed group spatial Independent Component Analysis (ICA) over 100 dimensions on the concatenated APC states using *melodic* (Smith et al., 2013), to determine stable core configurations that were shared across participants (Yeo et al., 2014). We chose 100 components as opposed to the maximum of 500, because we were interested in broader spatial clusters; at the same time, we did not select < 100 components, to avoid including noise or individual variability in the components. We therefore manually selected components of interest based on whether they conformed to the grey matter in a mostly bilateral way, to remove any possible noise or individual variability. Altogether, these two methods help us identify the most robust core states for each participant (APC) and across participants (ICA) respectively. In a traditional core-periphery network we would expect only one core state, whereas in a multi-core-periphery network we would expect multiple configurations of the core (Verma et al., 2016; Yan & Luo, 2019).

To further investigate how cores vary over time, we analysed the change in core assignment across time windows at the participant level. We joined all time windows for one participant in a single matrix of the form $NxT$, where $N$ are the number of nodes and $T$ the number of time windows. We then calculated $cores_i/T$, meaning the probability of a given node $i$ being assigned a 1 (core) value across all time windows. The values were then averaged at the group level and thresholded at 90th, 80th, 70th and 50th percentiles to identify stable configurations.

### *Community evolution*

Putative changes happening in communities were investigated using an algorithm for greedy Jaccard similarity over time windows (Thompson et al., 2017). Since the *Louvain* algorithm is non-deterministic and since it was applied on static time windows, the community label assignments may vary from one time step to the next. The greedy Jaccard algorithm re-

assigns labels from one window to the next based on how similar the largest community at time (t+1) was to the largest community at time *t*, determining whether a community has significantly changed by either: (i) splitting into smaller communities; (ii) joining another community. After re-assigning community labels, we collated all time windows for a participant to form a *NxT* matrix, where *N* is the number of nodes and *T* the number of windows. We then computed the probability of nodes *i* and *j* appearing in the same community over time windows, saving the results in a new *NxN* matrix. We thresholded the matrix again at .4 to maintain only the higher probability values and ran 50 iterations of *Louvain* to identify temporal communities.

### *Group-level community partitioning*

We sought to investigate possible differences between the individual and group-level community partitions at each time window. For this purpose, we computed an anatomical mask containing only shared voxels from the adjacency matrices of all participants in a movie. This resulted in an anatomical mask with 13,217 voxels for 'Citizenfour' and one with 13,568 for '500 Days of Summer'. For each participant we ran the AFNI program *3dDegreeCentrality* as before with a 10% edge density proportional threshold (Cox, 1996). The resulting correlation matrices were averaged across participants to create a single *NxN* matrix (*N* = number of nodes). The group matrix was further thresholded at a low correlation value of r = .1 to remove any possibly remaining weak connections: this was done because proportional thresholding leads to some participants possibly having weaker connections than others, that then drive down some correlations in the average, leading to disjointed components (Garrison et al., 2015). The resulting matrix was transformed into a binary adjacency matrix of [0,1] values representing connectivity (1 = connection, 0 = not connected). The *Louvain* and core-periphery algorithms were computed on the group-level matrix using the same parameters as above (i.e., $\gamma = 1$ for 100 iterations, $\tau = .4$ thresholding and 50 further iterations for *Louvain*).

In order to compare individual-level and group-level communities, we first calculated the total number of communities detected in a single time window in the two networks. To examine in more depth the differences in community partition, we re-assigned community labels using the previously described greedy Jaccard similarity algorithm. Then, we calculated how many temporal communities were identified in the two networks (i.e., how often a community was reassigned label). Finally, normalised mutual information (NMI) was calculated for each individual-level partitioning against the group-level one, to measure how

similar the individual communities were to the group average. Since NMI only works on same-size vectors, we matched group and individual-level number of nodes by first finding shared voxels and then removing nodes in each participant that were not in the group network.

*Community organisation of language processing*

We aimed at understanding more in-depth the community and core-periphery partitioning of the neurobiology of language processing. For this, we computed a multiple linear regression using a canonical hemodynamic response function over each word in the two movies for every participant. The word regressor for '500 Days of Summer' included 8,985 words, while the one for 'Citizenfour' included 14,606 words in total. Each word regressor consisted of the start time of the individual word onsets in milliseconds. We included a contrast regressor for non-word timings, meaning times when no word was spoken in the movies, in our analysis. These non-word regressors were composed of 1,834 timepoints for '500 Days of Summer' and 1,581 timepoints for 'Citizenfour'. The regression analysis was performed using AFNI's *3dDeconvolve* function (Cox, 1996). Subsequently, we performed a mixed effects model analysis using AFNI's command *3dMEMA*, and thresholded the resulting t-statistic map at $\boldsymbol{\alpha}$ = .001. The resulting map corresponds to the activation produced by all words on average across two movies, which we called 'words' map.

We extracted the voxel 'xyz' coordinates of the 'words' regions and used these to count how many unique community labels overlapped this region, with the caveat that the overlap had to include at least 10% of the 'words' map voxels. This means that if a community overlapped < 10% of voxels in the 'words' regions, it would not be considered as being involved in word processing. The 10% cut-off was arbitrary, but it eliminated very small clusters. We computed the unique 'words' communities at each window for each participant. We then calculated the percentage of the rest of the brain that were also part of these same 'words' communities: for example, if community 3 (C3) significantly overlapped the 'words' map, we calculated the percentage as follows:

(C3 all_voxels - C3 word_voxels)/ (all_brain_voxels - word_voxels)*100

This was then summed over all communities overlapping the 'word' map and calculated for each window for each participant.

We also investigated how much the 'words' map contributed to the core regions at any given time point, by calculating the total of core 'words' nodes over all possible cores in a window. We then aimed at inspecting connections between core 'words' nodes and its associated periphery. The core-periphery algorithm does output additional information about core-periphery links by pairing these two into groups, much like community partitioning. From this information, we extracted those pairs overlapping core 'words' nodes and calculated the distribution of the associated periphery over the rest of the brain as a percentage.

## Results

### *Core-periphery structure*

In order to investigate whether the brain is organised in a core-periphery architecture, we ran a novel algorithm that identifies core-periphery structures (Shen et al., 2021). Our results show that on average across time windows and participants, there were M = 922 (SD = 107.9) core nodes (corresponding to 5.3% of grey matter voxels at $5mm^3$ resolution), with the remainder being peripheral nodes (or voxels). The number of core nodes varied across time and participants, with the range being 586 - 1257 (~3-7% of grey matter) core nodes.

To determine whether there were different temporal configurations of cores, we ran APC on cores and selected an exemplar for each cluster. This resulted in M = 54.7 (SD = 6.1) core exemplars on average across participants. The core configurations changed every M = 11 time windows (i.e., 160 sec) (SD = 6.5 time windows, or 110 sec) across participants, meaning that the core-periphery distribution in the brain varied every ~2-3 min (+/- 1 SD range = 50 sec - 270 sec) of the movie on average.

In order to determine whether circumscribed sensorimotor and 'language' regions act as the most stable cores, we computed group spatial ICA on core exemplars. Out of the 100 ICA components, we identified 16 stable non-noise components, which we then correlated with meta-analysis maps from the Neurosynth database (Table 4, Fig. 16). The results show that at the individual level (i.e., APC analysis) there were more core configurations than in the aggregate (i.e., ICA analysis).

| Component ICA number | Functional term 1 | Functional term 2 | Functional term 3 |
|---|---|---|---|
| 1 | Spatial | Objects | Attentional |
| 2 | Vision | Visual Stimulus | Navigation |
| 3 | Sentences | Comprehension | Linguistic |
| 4 | Memory Retrieval | Episodic | Recognition Memory |
| 5 | Sounds | Listening | Speech |
| 6 | Motion | Perception | Visual Motion |
| 7 | Speech Production | Vocal | Naming |
| 8 | Face | Object | Vision |
| 9 | Default Mode | Autobiographical | Mentalizing |
| 10 | Navigation | Spatial | Theory of Mind |
| 11 | Default Mode | Referential | Self Referential |
| 12 | Retrieval | Memory | Working Memory |
| 13 | Action | Tactile | Action Observation |
| 14 | Working Memory | Task | Calculation |
| 15 | Movement | Motor Imagery | Stimulation |
| 16 | Movement | Motor Task | Tactile |

**Table 4.** Top 3 associated Neurosynth meta-analysis functional terms for each of the 16 core group ICA components. These map mostly to sensorimotor regions (e.g., visual, auditory, motor) and Default Mode regions. The ICA maps were directly submitted to Neurosynth and correlated to existing meta-analysis terms.

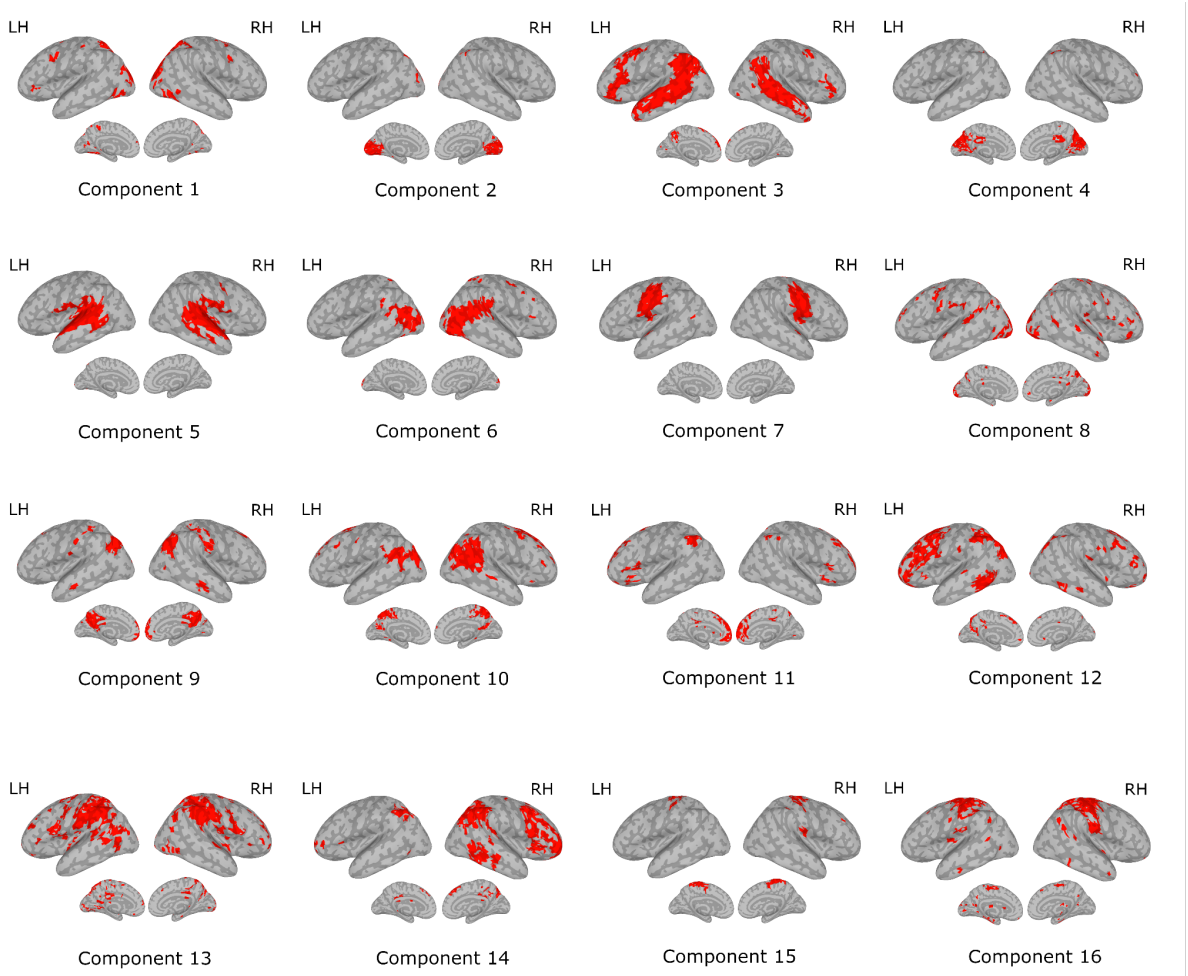



**Figure 16.** Maps of the 16 core ICA components from the aggregate analyses. Multiple spatial configurations of cores existed at various time points. These included 'language' regions, primary visual and primary auditory cortices, sensory association areas, prefrontal areas and precuneus/posterior cingulate. Cluster size = 20.

To further investigate whether auditory/visual sensorimotor regions, in particular, represented the most stable core nodes across subjects we measured how often voxels belonged to a core. To find stable temporal configurations, we considered the 90th, 80th, 70th and 50th percentiles of values across all participants. At the 90th percentile (Fig. 17), voxels in primary auditory and visual, and medial prefrontal areas survived, with a pattern showing r = .24 correlation with the *vision* functional term from Neurosynth, as would be expected in a movie task. Below the 90th percentile, the voxel distributions had poor correlations ($r < 0.1$) with any Neurosynth meta-analysis term, and were not included here.

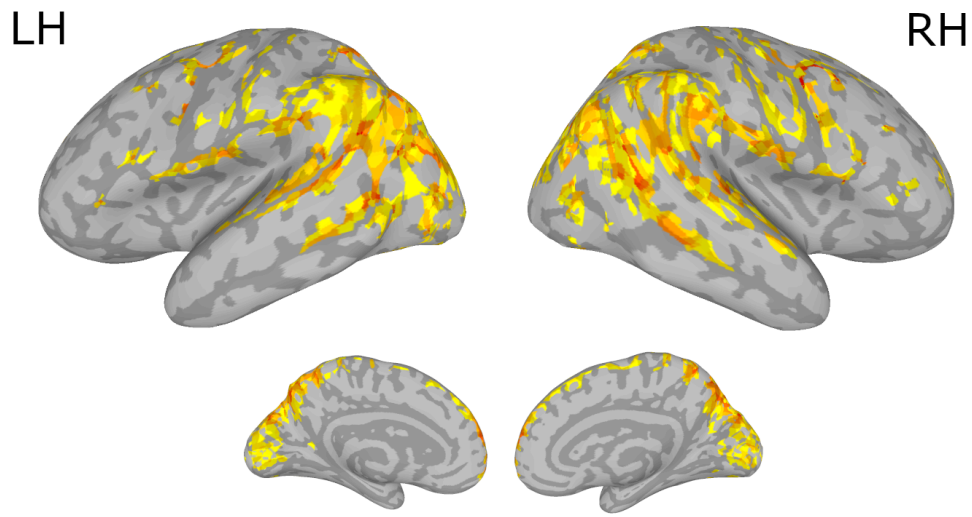



**Figure 17.** Map of most stable temporal cores (thresholded at 90th percentile). The most stable and strong cores appear as sensorimotor regions (e.g., primary visual, auditory, some motor) and some DMN components (e.g., medial prefrontal and angular gyrus). Cluster size = 20.

*Community partitioning*

In order to identify putative functional components of the brain, we ran the modularity optimisation algorithm, *Louvain*. On average, the algorithm found M = 3.7 communities (SD = 0.5) at the final (consensus) step, with average modularity score $M_{QF}$ = 0.66 ($SD_{QF}$ = 0.06) (Fig. 18 shows example communities from consecutive windows of a participant). The final consensus partition modularity score was significantly higher than the initial partitioning score ($M_{QI}$ = 0.23, $SD_{QI}$ = 0.03) across windows and participants (t = 1038.9; $p < 0.001$). Higher *Q* scores are indicative of higher quality of the partitions obtained (Fornito et al., 2016b).

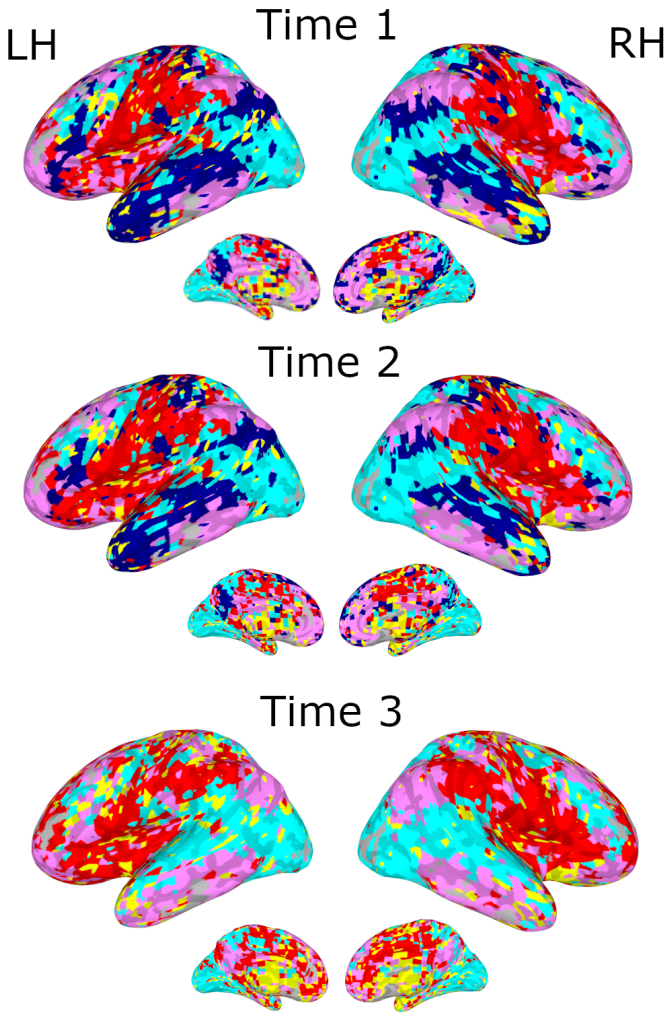



**Figure 18.** Community evolution in consecutive time windows in an example participant brain. Voxels are coloured based on their community allegiance. Time 1 is divided into 5 communities, which are mostly stable until Time 2. At Time 3, the dark blue community disappears, being split between the cyan and pink communities. Cluster size = 20.

To ensure that the communities identified with *Louvain* were non-random, as could be the case with some high *Q* values, we performed a non-parametric permutation test. The results indicated that the original partitions were significant (p-value < 0.001) across all participants and time windows compared to randomly shuffled communities.

We then re-assigned community labels to track their evolution and putative dynamic behaviour over time. Our results show that on average individual-level communities underwent M = 59.4, SD = 8.9 variations over ~609 time windows on average (Supplementary Materials S4 shows an example video of variations).

In order to test whether communities formed stable temporal configurations, we built a matrix D*ij* for each participant where entry *(i,j)* indicates the probability of nodes *i* and *j* being part of the same community over time. In rare cases, the algorithm partitioned a participant's network into large communities and smaller ones of 1-100 voxels; since the latter are most likely outliers due to noise, we only considered communities of size >1000 voxels, which was the minimum size of other participants' communities. This produced M = 5, SD = 0.7 temporal communities on average across participants (Fig. 19 shows an example participant for each movie).

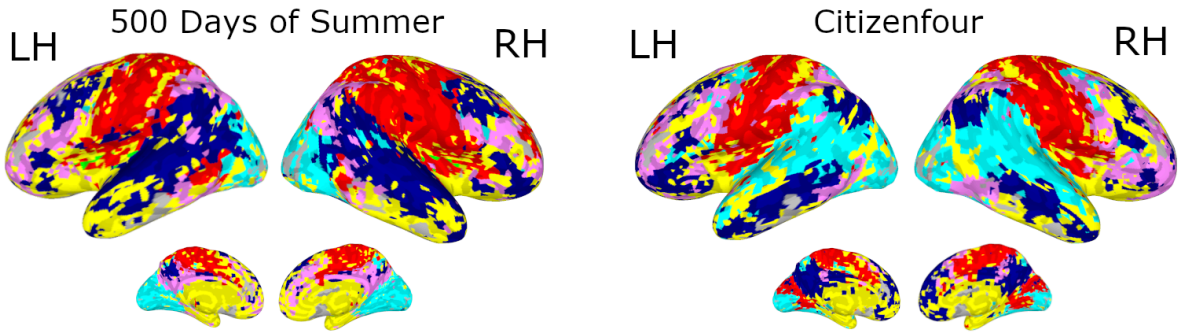



**Figure 19.** Temporal communities from example participants in '500 Days of Summer' and 'Citizenfour'. On average, we identified five communities stable over time. These mostly map to central sulcus (red), temporal (dark blue), occipital (cyan), prefrontal (pink), subcortical (yellow) regions across participants and movies. Cluster size = 20.

*Differences between individual and group-averaged communities*

Most of the network neuroscience literature uses group-averaged adjacency matrices to determine the brain network organisation. We aimed at comparing our community findings at the individual-level to the group average brain network in order to test whether individual networks are much more variable than the group.

For a given time window, the group level had M = 5.5, SD = 0.7 communities on average, compared to the previously found average of M = 3.7 communities (SD = 0.5) at the individual level. Moreover, on average, group-level communities experienced less variations (M = 37.5, SD = 4.5; Supplementary Materials S5 shows a video of variations) than the individual-level ones. These group-level communities did not resemble any individual-level partitions, with a low similarity score ($M_{NMI}$ = 0.05, $SD_{NMI}$ = 0.02).

*Network architecture in language regions*

We next aimed at inspecting the meso-scale network organisation of regions typically associated with language processing. We predicted that 'language' cores would form dynamic communities with distributed peripheral nodes. On average, across windows and participants, we found M = 3 communities (SD = 0.26) overlapping the 'words' map regions, meaning most of the communities in the individual networks overlapped the 'words' regions at any given time. These communities extended, on average, over M = 84.8% (SD = 6.9%) of the rest of the brain, showing largely distributed patterns.

From a core-periphery perspective, M = 5.3% (SD = 1.1%) of core nodes fell within 'words' regions across windows and participants. Moreover, we identified on average M = 37.0% (SD = 7.0%) of core-periphery pairs across windows overlapped 'words' regions, when these were a core. Finally, we found that M = 50.0% (SD = 7.0%) of the rest of the brain (i.e., areas outside 'words' regions) acted as a periphery linked to 'words' core nodes. Note that the reason we find 50% of distributed brain regions whilst in the community analysis we found 84.8% is that in the core-periphery analysis we excluded all other core nodes not in 'words' regions within the same core-periphery pairs, because we were solely interested in peripheral nodes. Therefore, 'words' core nodes were connected to other core nodes elsewhere in the brain, but we did not report the values here.

## Discussion

In the present study we sought to investigate whether a highly flexible and distributed mesoscale architecture of the brain network would better support the neurobiology of language comprehension in the real world. We hypothesised that the model best supporting the complexity and variability of language processing would be a combination of core-periphery and modularity with added features: specifically, we predicted that multiple core-periphery pairs would map onto large and dynamic overlapping communities.

Our results confirmed that in individual brain networks, both multiple core-periphery pairs and dynamic communities co-exist, but due to algorithm limitations we could not inspect overlaps. Here, various components of 'language' regions acted as cores at different times (Fig. 16, components 3, 5, 7). These, in turn, connected to a periphery of other brain regions, together forming multiple large dynamic communities encompassing most of the brain. Overall, the picture is of a highly flexible network architecture that supports complex language features.

*Neurobiology of language comprehension*

Traditional models of the neurobiology of language have proposed a very static and modular organisation of language processing regions, mainly limited to the STG, MTG, some premotor regions and IFG (Hickok & Poeppel, 2007; Poeppel et al., 2012). Within this context, network studies on language processing have assumed an anatomically constrained organisation of language, by using localiser tasks to select 'language' regions a priori for network analysis and building group-averaged networks (Chai et al., 2016; Fedorenko & Thompson-Schill, 2014). Although these studies have introduced some novelty into how we think about language through more process-oriented models, they fall short on methodologies and interpretations by suggesting that constrained and average network models are somehow representative of the richness and complexity of language in the real world (Seghier & Price, 2018).

We have demonstrated that this is indeed not the case when more complex and individual features of language are considered (see Chapter 3.1 and 3.2). This view is supported by a growing body of literature showing that the neurobiology of language is highly distributed, dynamic and variable (Huth et al., 2016; Price, 2010; Pulvermüller, 2018; Skipper et al., 2021). In accordance with this flexibility, our findings on individual participant networks revealed that 'language' regions acted as one of the multi-core structures connecting to a wide periphery of other brain regions. The latter spanned on average 50% of the rest of the brain, suggesting that language processing likely involves largely dynamic and distributed regions.

Our findings further showed that the 'language' core-periphery nodes formed numerous pairs, together clustering into ~3 large communities overlapping >80% of other brain regions at any given time. Moreover, these pairs connected 'language' cores to other core regions elsewhere in the brain, indicating integration and sharing of information with other brain areas. Although the majority of previous studies have only identified at most 2 communities or networks involved in language processing (Fedorenko & Thompson-Schill, 2014; Hickok & Poeppel, 2007), others have demonstrated the existence of a main language network associated with more distributed brain regions still involved in language processing (Hertrich et al., 2020). In Hertrich et al.'s study, the main language network comprised more stable regions that overlapped core 'words' regions in our study, whilst their associated networks overlapped with peripheral nodes in our study (Hertrich et al., 2020). It is clear, then, that more complex and flexible models of the neurobiology of language are needed to account for both these findings.

*Flexible network model*

What model best represents the complexity of language in the real world? Modularity alone tends to support dual-stream or 'language network' models due to its highly segregated and somewhat static nature (Newman, 2006). Although there is increasing evidence suggesting communities vary over time through hierarchical relationships (Bassett et al., 2011; Meunier et al., 2009), these still do not fully account for (i) individual variability and (ii) complex language features (e.g., context).

A core-periphery organisation, instead, can afford high flexibility through a dynamic and loosely connected periphery (Borgatti & Everett, 2000). Possibly due to this high level of variability, this model is understudied in functional brain connectivity, where group-averaged networks are typically used. The only existing evidence for a core-periphery brain architecture has identified sensorimotor regions as highly stable cores across time (Bassett et al., 2013). This finding is replicated in our results, whereby these same regions (i.e., primary visual, auditory and some motor) were the most stable cores across time and participants, suggesting an important role of these regions for the stability of the entire network (Fig. 17).

Although core-periphery networks are more variable and robust, they still assume the existence of a single core connecting to the whole periphery (Borgatti & Everett, 2000). To address these shortcomings, we proposed a combined multiple core-periphery and dynamic modular architecture, as these structures are known to co-exist in various scale-free and empirical networks (Kojaku & Masuda, 2017; Yan & Luo, 2019). Here, we identified 16 different cores that were somewhat stable over time and across participants (Fig. 16). These grossly correlated with components of 'language', Default Mode network, episodic memory and sensorimotor regions from Neurosynth meta-analyses, possibly suggesting that group analyses have so far only identified stable cores rather than the whole distribution of activity arising from a cognitive task (Table 4). Indeed, this stability waned at the individual-level, with the multiple cores showing significant re-configurations every ~2-3 min (50 - 270 sec), possibly matching significant changes in movie stimulus (e.g., scene changes).

Various core-periphery pairs were distributed over 3-4 large communities at any given time. Both these structures varied significantly over time and across participants, indicating a highly dynamic and flexible architecture (Fig. 18). These only became highly segregated when averaging communities over time (Fig. 19). Overall, the flexible network organisation that we

propose here accounts for (i) individual variability, (ii) contextual changes, (iii) shared processes between different cognitive domains.

### *Group-averaged networks*

In order to situate our findings within the existing literature, we also computed group-averaged network analyses, as these are a standard approach in network neuroscience (Gordon et al., 2017; Lehmann et al., 2019). We predicted that group networks would be less dynamic and flexible than any individual network, as the former are stripped of any individual variability

Modularity in group-averaged networks identified the same number (~5-8) of communities as those reported in previous studies on modular structures in both anatomical and functional brain networks, with similar spatial patterns (Fornito et al., 2016b; He & Evans, 2010; Meunier et al., 2010). These were mostly stable across time (Supplementary Materials S5), experiencing less variations than individual networks on average. A direct spatial comparison of time-matched group and individual networks revealed that the two had very little similarity (NMI = 0.05, where 0 = no identity). These results suggest a large divergence between individual and group networks.

The group-averaged networks appear to thus be less flexible and dynamic, not representing any individual brain. This more static behaviour of group networks, and their divergence from individual networks, is well documented in the literature (Gordon & Nelson, 2021). For instance, individual networks were shown to organise into more dynamic and complex structures than group-averaged networks (Braga et al., 2019; Braga & Buckner, 2017; Gordon et al., 2017), whereby these dynamics correctly predicted cognitive abilities and states, whilst group-averaged networks performed poorly (Barnes et al., 2014; Kong et al., 2019).

Overall, these findings suggest that in order to investigate the network organisation that supports the neurobiology of language, individual network variability must be considered, as this is more representative of the flexibility of language in the real world.

### *Limitations*

The present study investigated the functional brain network architecture in a naturalistic setting. Although naturalistic settings capture many complex behaviours, and therefore relate better to the real world (Aliko et al., 2020; Hasson & Honey, 2012), they also pose increased

difficulty of controlling stimuli and creating experimental manipulations (e.g., contrasts between two conditions).

Thus, one limitation in the present study was that we did not test variations in language processing dynamics. For instance, we could have grouped time windows by their average word frequency into low vs. high frequency clusters; then we could have compared the dynamics of high and low frequency words to inspect language processing networks in more depth. As such, our findings on the specific dynamics of the neurobiology of language remain somewhat speculative, and we have thus planned to investigate these further in future work.

Similarly, we have not shown definitively that core-periphery pairs overlapping 'words' regions are actually processing language. This may be achieved through similar methods as the one detailed above, where we contrast a linguistic and a non-linguistic stimulus while controlling for the effects of other movie features. Nevertheless, there is strong evidence and support for proposing that the peripheral regions associated with language core regions are processing language, as (i) these have correlated nodes by definition, and (ii) they together form communities that by definition represent functional segregation (Sporns, 2013a; Wig, 2017). The only case where these peripheral nodes may not be performing language processing functions is if the language core regions (e.g., STG, IFG, MTG) were not performing language processing themselves. However, this is unlikely to be the case as these regions are consistently active during processing of any language stimulus/task (Chai et al., 2016; Hickok & Poeppel, 2007; Skipper, 2015).

*Implications*

Several aspects of this work make it highly innovative: for one, we inspected the brain network architecture of individual participants at high resolution (voxel-wise), over long time periods (~2 h) and during a naturalistic setting. This resolution and complexity have never been attempted in previous work, to the best of our knowledge. Thus, the flexible and combined core-periphery and modular model that we have proposed provides a much more detailed representation of the brain network in the real world, accounting for individual variability and contextual changes.

Moreover, we presented the first model of whole-brain network organisation that accounts for the natural complexity of the neurobiology of language. The flexible and dynamic organisation can help explain how the neurobiology of language changes during variations in

context; the division of language into large communities containing various core-periphery pairs allows for different language features to be processed simultaneously (periphery) while sharing and integrating information (cores). Furthermore, since 'language' regions were part of the multi-cores, and since the latter have high wiring costs, this model helps explain why lesions in 'language' regions cause language impairments (e.g., in aphasia more extensive damage to core nodes is significantly more disruptive than damage to peripheral nodes) (Fornito et al., 2016a; Zhao et al., 2011).

As 'language' regions in our model connect to a wide periphery performing language processing, the model supports evidence from neuroplasticity studies showing that speech recovery after a stroke is driven by heterogenous rewiring processes that involve large parts of both hemispheres (Crosson et al., 2019; Geranmayeh et al., 2014; Kiran & Thompson, 2019). When cores are severed, peripheries closer to nodes may increase their connections to take on the role of new cores, restoring functions. This possible mechanism of recovery offers new insights into potential regions as targets of novel individualised speech therapies for aphasic patients.

### *Conclusion*

We have presented a flexible model of the brain network architecture that supports language processing in the real world. We show that the brain is organised in a multiple core-periphery network within a modular architecture of large dynamic communities. In this context, 'language' regions act as one of the multi-core structures and connect to a large and dynamic periphery. These two components form multiple dynamic communities together, indicating a shared language function.

# Chapter 4: Discussion and Conclusions

Language processing is a complex brain behaviour that depends and heavily relies on contextual information, an individual's cognitive strategies and social and emotional context (Price, 2010, 2012; Skipper, 2015). As such, a model of the neurobiology of language processing must be able to explain and account for the richness of information and complexity that language encompasses. Existing models of the neurobiology of language have undeniably provided insights into how the brain supports language functions, identifying regions of the brain primarily in superior and middle temporal cortex and inferior frontal cortex that are important for processing of any language feature (Fedorenko & Thompson-Schill, 2014; Friederici, 2002; Hickok & Poeppel, 2007; Poeppel & Hickok, 2004; Rauschecker & Scott, 2009). Although our understanding of language and the brain has advanced with these models, these are still extremely limited in their consideration of more complex language features and individual variability, and therefore do not comprehensively explain language in the real world (Skipper, 2015).

In this work, we aimed at investigating language processing in a more naturalistic environment that better represents the complexity of language in the real world. For this, we collected neuroimaging and behavioural data of participants watching full length movies in a fMRI scanner, as we detailed in Chapter 2. This dataset is now one of the largest naturalistic datasets publicly available (Aliko et al., 2020; Madan, 2021).

We proposed that, when investigating specific language features, language processing would be much more distributed, encompassing many other brain regions as appropriate to the feature, with unique activity patterns depending on context and embodied meaning of words or the goal of the listener. We further proposed that measures of central tendency and subtractive methods have obscured this distribution due to its high variability, instead only resulting in 'language' regions. We hypothesised that the reason that 'language' regions consistently appeared in the aggregate is because these regions act as stable intermediary connectivity hubs forming one of many brain cores in a flexible core-periphery architecture. Here, we predicted that 'language' regions would connect to more distributed areas that form dynamic core-periphery pairs in order to perform context-dependent language processing. Our work showed the following:

- Chapter 3.1: we demonstrated that during pronoun resolution, unique character representations in situation models are reactivated in sensorimotor regions (i.e., mainly primary visual and auditory cortices), through a search in memory in episodic memory

regions (i.e., hippocampus, precuneus), supported by mentalizing areas (e.g., medial prefrontal cortex).

- Chapter 3.2: we showed that individual sensorimotor embeddings of words activate uniquely distributed brain regions. These together encompass most of the rest of the brain. Here, 'language' regions constitute an intermediary hub, which thanks to its high centrality and stability, survives central tendency measures and subtractive methods.
- Chapter 3.3: we identified that the brain network organisation supporting language processing is composed of multiple core-periphery pairs situated in large dynamic communities. Here, 'language' regions act as somewhat stable cores that form multiple communities with largely distributed core-periphery pairs.

Overall, our experiments demonstrated that real-world language processing is better explained by a highly flexible network model composed of a multiple core-periphery architecture with dynamically changing communities, where 'language' regions direct a distributed and variable periphery of other brain regions. This warrants a move away from current dual-stream models of the neurobiology of language, to one that considers language as a complex, dynamic, distributed and flexible behaviour.

In the following sections, we will first review a general description of the proposed model, then discuss in more detail how various components of the model support language as a complex behaviour, and finally suggest implications for our understanding of language and the brain, as well as for speech impairments.

## Network model of language and the brain

In this work, we found that many distributed and variable regions are also involved in language processing, where 'language' regions act as high-connectivity hubs that are somewhat stable over time. As such, we propose that the brain organisation supporting language processing has a multiple core-periphery and dynamically modular architecture.

Core-periphery structures involve two components: a set of highly inter- and intra-connected nodes (core), and a set of loosely connected and dynamic nodes (periphery) (Borgatti & Everett, 2000). The presence of multiple core-periphery pairs affords the network more flexibility than typical core-periphery networks (where a single core and a single periphery exist) because these can form different core-periphery configurations at different times (Yan & Luo, 2019), thus supporting adaptation to a changing environment. The existence of a spatial core-periphery structure in brain networks has mostly been hinted at by previous research,

where a rich-club organisation (i.e., one form of core configuration) was identified (van den Heuvel & Sporns, 2011). The only two research articles, that we could find, inspecting in detail the existence of both core and periphery nodes in the brain identified a spatial and temporal core-periphery organisation where cores primarily involved Default Mode network (DMN) and sensorimotor regions (Bassett et al., 2013; Gu et al., 2019). Our results suggest that these same regions, in particular the anterior medial prefrontal cortex (mPFC) and angular gyrus (AG) of the DMN, and visual association areas, primary auditory and primary motor in sensorimotor regions, were the most stable cores across time.

The majority of the other multiple cores, however, varied frequently over time, likely in response to contextual changes. These less stable intermediary cores involved 'language' regions (e.g., MTG and IFG), somatosensory association cortices, precuneus, posterior cingulate cortex and dorsolateral prefrontal cortex. Here, 'language' regions directly connected to a wide group of periphery nodes, encompassing about half of the rest of the brain and together forming various large communities, thus likely sharing a common function.

The overall network profile of 'language' regions included connections to both peripheral regions and other core regions. Overall, these highly distributed language-related core-periphery pairs included, at different times, large parts of frontal regions, subcortical structures, precuneus, cingulate cortex, medial prefrontal areas, premotor cortex, supplementary motor area, and primary auditory and visual cortices. Together, this comprised > 80% of the brain.

Although we found that communities spanned multiple core-periphery pairs and varied dynamically, due to algorithmic limitations we could not inspect whether communities overlapped nor if they formed hierarchical structures. However, network studies have indicated that core-periphery structures are usually an indicator of overlapping communities, where the overlaps are composed of cores (Yang & Leskovec, 2014). As such, we predict that core regions would sit at the overlap between communities, which was already shown to be the case in edge-based rather than node-based brain communities (de Reus et al., 2014). Similarly, some studies have found that rich-club organisations, which are tightly connected sets of hubs and therefore similar to cores, sit along the overlap between dynamic communities during development (Betzel et al., 2017). Moreover, hierarchically modular structures are a well-established feature of brain networks (Alexander-Bloch et al., 2010; Meunier et al., 2009, 2010). Fig. 20 below shows a diagrammatic view of the complete proposed structure.

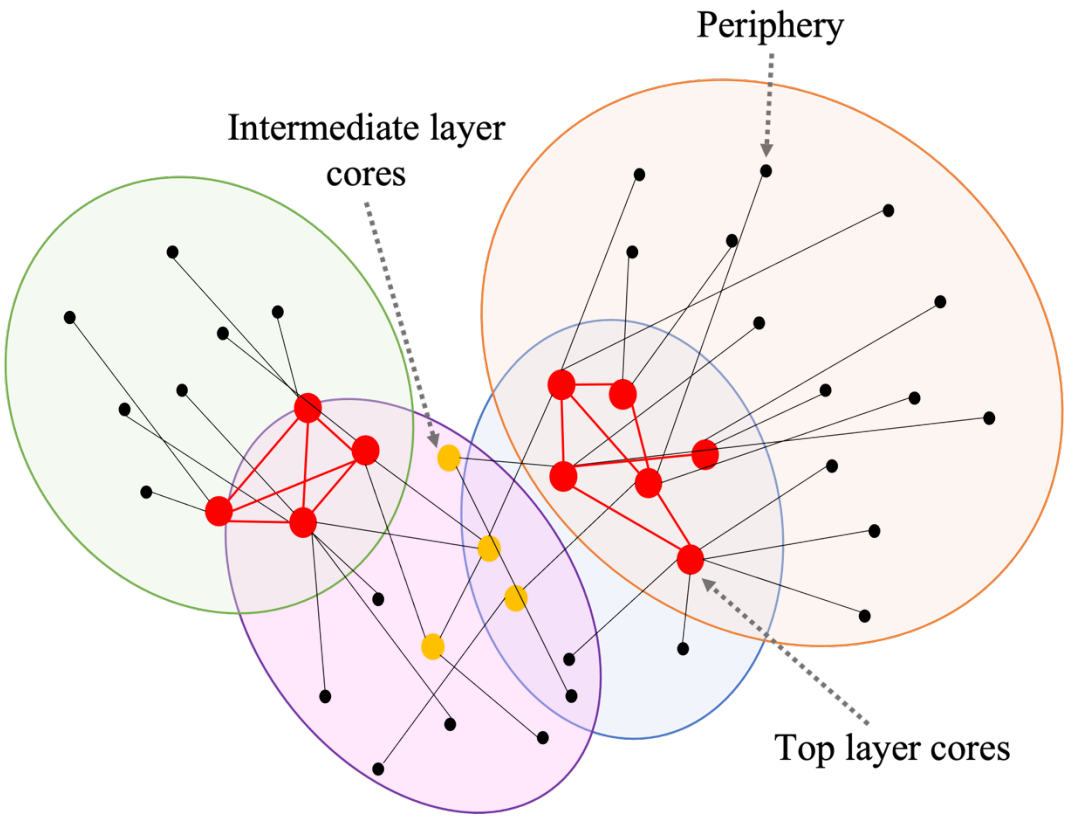



**Figure 20.** Proposed model of the brain network architecture supporting language processing in the real world. Core regions (red circles) sit at the overlap between communities (green, blue, purple and orange circles). The cores connect to various core-periphery pairs in each community (black circles), as well as intermediary hubs (yellow circles) together coordinating the periphery node dynamics. 'Language' regions are likely to behave as the intermediary (yellow) nodes in this diagram, coordinating with top-level cores (red) to direct processing in peripheries (black). Moreover, communities may form hierarchies that support related and progressively more complex language functions: orange and green communities are at the top of the hierarchy with stable cores, purple and blue are smaller communities with intermediary cores.

In what follows we detail how this proposed network architecture with its dynamic and distributed nature may support various complex aspects of language processing.

*Semantics*

To understand the semantic meanings of individual words, the brain must be able to draw information from the embodied or evoked meaning of that word (Pulvermüller, 2013; Pulvermüller et al., 2005). Studies have identified brain regions outside of 'language' areas active when processing the embodied meaning of a word in that modality: for instance, action words activated primary motor regions while words describing colours activated visual cortices

(González et al., 2006; Kiefer et al., 2008; Klepp et al., 2019; Pulvermüller, 2013). Similarly, we found that largely distributed brain regions including primary auditory, primary visual, premotor, middle frontal, somatosensory association cortices and various subcortical structures were activated at different times and to various extents when processing sensorimotor embeddings of words.

A flexible multiple core-periphery and hierarchically and overlapping modular network supports notions of embodied cognition as follows (Fig. 21):

- Different lower-level communities relate to different semantic categories. These communities may contain overlapping regions, most likely in core areas, that connect similar semantic categories to allow them to share information. These shared patterns of activity are known to support semantic category processing (Tomasello et al., 2017).
- The semantic communities tile most regions of the brain forming unique distributions. This extensive and category-specific distribution is supported by various semantic studies (e.g., (Binder et al., 2009; Huth et al., 2016)), as well as our results in this work.
- The modalities to which semantic categories relate to are likely to be composed of a mixture of core and peripheral nodes. Here, the cores would coordinate the general semantic category processing (i.e., semantic hubs), while peripheries would process more fine-grained semantic meanings of individual words. Here, studies have shown that 'language' regions may constitute some of the semantic hubs (Tomasello et al., 2017).

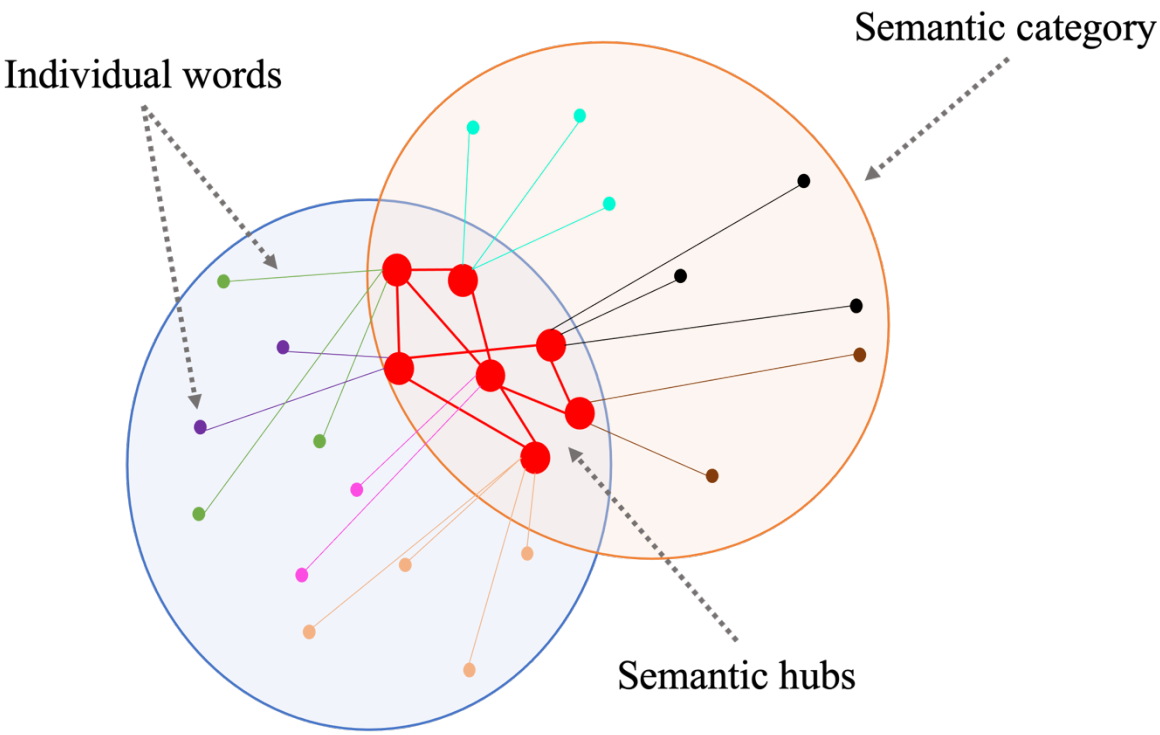



**Figure 21.** Semantic processing model. Cores (red) form semantic hubs for sharing information between semantic categories (orange and blue large circles), each of which constitute a community. Individual core-periphery pairs (multiple colours) within a community of semantically-related items, help process individual words.

*Context*

Context represents a higher-level language feature, building complex relationships between the semantic meaning of single words in a sentence (Xu et al., 2005). Language cannot exist in the absence of context (Skipper, 2015), and as such it should be thoroughly investigated. There are three aspects of context that a comprehensive model of the neurobiology of language must consider: (i) since context is fundamental to language, the brain must constantly process it; (ii) since context can change rapidly, the brain must allow a flexible and dynamic environment; (iii) since the amount of context available varies, the brain must adapt to both higher and lower contextual information.

As such, our model helps explain how these aspects of context are processed as follows (Fig. 22):

- AG and anterior mPFC form highly stable cores that directly connect to 'language' intermediary cores. Indeed, AG and mPFC were shown to be active during processing of coherent narratives, with a particular role in decision-making, understanding concepts and inferring relationships between words and sentences (Ferstl & von

Cramon, 2002; Fletcher et al., 1995; Hasson et al., 2007; Humphries et al., 2006; Newman et al., 2001; Xu et al., 2005). For instance, we found activation of the mPFC during pronoun resolution irrespective of character. These are likely to support any contextual processing function.

- Higher-level larger communities encompass multiple core-periphery pairs. Here, communities can vary dynamically to account for the rapid changes in context. Merging of two communities, for example, could happen when two previously separate contexts are joined in new events.
- Each large community encompasses multiple smaller ones in the hierarchy. The smaller ones were likely involved in semantic category processing, which here are merged at the higher level to support context processing.
- Lack of contextual information can be overcome by allocating more peripheral nodes and integrating various core-periphery pairs to process the ambiguities.
- In the presence of high contextual information, less peripheral nodes would need to be employed to process the current context. As the brain is known to reorganise during predictive processes (Skipper & Zevin, 2017), these peripheries can instead be used to help predict upcoming information.

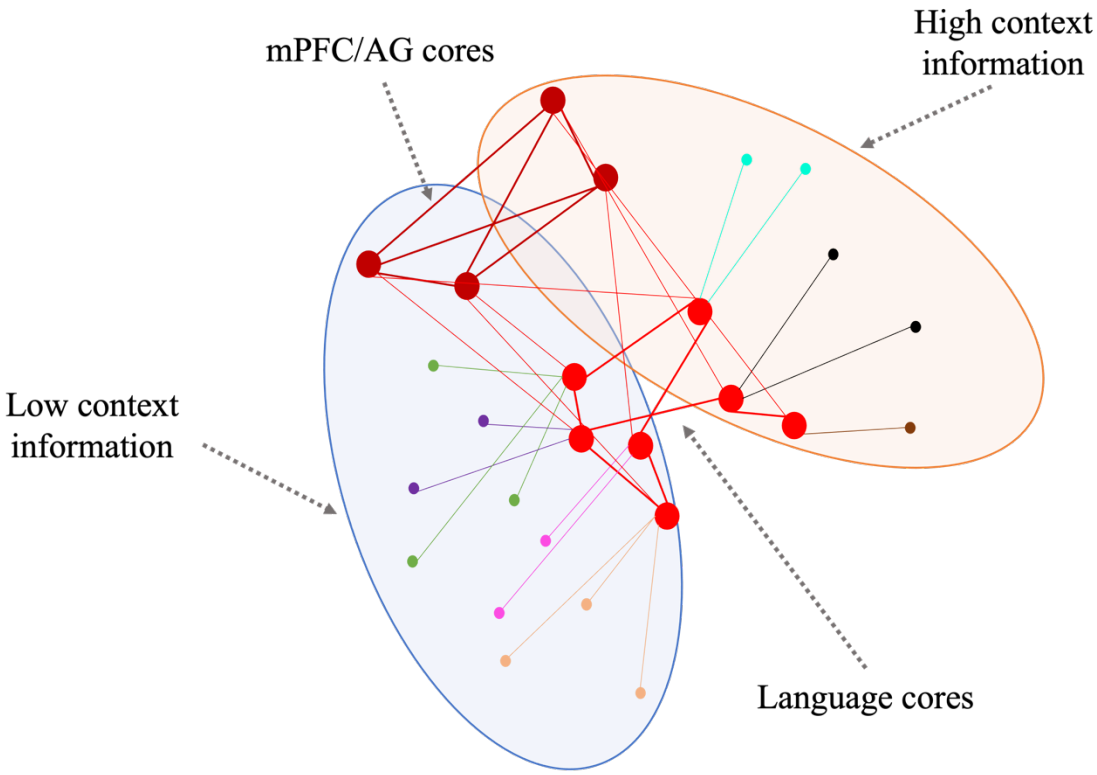



**Figure 22.** Simplified diagram of context processing model. 'Language' cores (bright red) support processing of different contexts (orange and blue large circles), each of which constitute a higher-level community. Many or few core-periphery pairs (multiple colours) are assigned to a community depending on level of contextual information (e.g., low vs high context). Not pictured here are smaller semantic category processing communities, that together form the blue/orange ones here. At the top of the architecture, the mPFC/AG (dark red) oversee these processes by being engaged in any context-processing task and directly connecting to 'language' cores.

*Imagistic representations and memory*

During discourse, the brain builds imagistic situation models of the meaning of the context (Altmann & Ekves, 2019; Zwaan et al., 1995). These are later activated when retrieving information about an event (Baldassano et al., 2018; Wittenberg et al., 2021; Yarkoni et al., 2008). For instance, we showed that situation models active in episodic memory reactivate sensorimotor character representations (e.g., primary visual cortex) when resolving pronouns.

This process requires the brain to build imageries and conceptual representations of dialogues, connect antecedent events to new ones (Piai et al., 2016) and retrieve the correct representations to process the current context (Wittenberg et al., 2021). This particular higher-level language behaviour is understudied (or rather inexistent) in all existing models of the neurobiology of language.

Our model, with its high complexity and flexibility supports the formation and retrieval processes of situation models as follows:

- During the situation model formation, core regions in context-processing and mentalizing regions (e.g., mPFC) connect to visual, auditory and motor cortices, as well as 'language' regions to build various character/event representations (Zwaan, 2016). The sensorimotor regions are composed of both intermediary cores and peripheries: the intermediary cores relate to general category areas (e.g., face vs object), while the peripheries around these cores form variable patterns relating to specific characters, places, events etc. The transfer of character/event specific information is facilitated by 'language' cores.
- After their creation, situation models are kept active in episodic memory regions (e.g., hippocampus, precuneus) (Berkovich-Ohana et al., 2020; Fletcheret al., 1995; Sreekumar et al., 2018; Wang et al., 2010), and continually probed by 'language' cores to which they connect to (Maguire et al., 1999; Oedekoven et al., 2017). These episodic memory regions are likely intermediary cores such that (i) they are continually active to allow language to access situation models, but (ii) are still flexible enough to switch between different situation models where needed. This flexibility is well documented in the literature (Duff & Brown-Schmidt, 2012).
- During retrieval, 'language' cores and episodic memory regions form one community to facilitate transfer of information. This process activates a search in memory for the appropriate situation model. The more ambiguous the referent, the more peripheral nodes between these regions may be recruited to support the search. Once a situation model is identified, the community merges with primary visual/auditory/motor cortices, forming a larger higher-level community, in order to reinstate the pattern of activity for a specific referent through peripheral nodes.

*Individual variability and shared features*

Variability in activity patterns of individual brains for a given task, is an important predictor of individual cognitive abilities and cognitive strategies (Heun et al., 2000; Miller et al., 2012; Seghier & Price, 2018; Szenkovits et al., 2012). Due to the consistent use of central tendency measures in neuroimaging studies, these variations have been mostly ignored in existing models of the neurobiology of language, instead favouring the most stable activity patterns across subjects. This has impoverished the extent of our understanding of how individual brains comprehend various language features and use them to build complex representations and thoughts.

Our model, instead, is primarily based on individual variability and therefore is more fit to explain individual differences that underlie the neurobiology of language. At the same time, we have inspected shared features across participants, thus making our model fit for explaining both individual and shared (i.e., group) dynamics. Thus, we propose that our model supports individual differences and shared language features as follows:

- Individual cognitive strategies are supported by dynamic and flexible peripheries and variable communities at the lower levels of the modular hierarchy. These together form different configurations in different individuals that support the different strategies.
- Individual cognitive abilities are supported by the amount of integration between core-periphery pairs and by levels of overlap between different communities. Supposedly, the higher the overlap of communities, or the higher integration between core-periphery pairs, the higher the cognitive abilities.
- At the group-level, the presence of higher-level large communities and top-layer stable cores, ensure that all human brains have the same underlying structure and perform the same functions. However, the presence of lower layers of the hierarchy for both cores and communities, support the development of cognitive differences in individual brains.

From the above examples, we conclude that our flexible and hierarchical model accounts for various complex language features as well as individual variability. Although we did not exhaustively detail every aspect of language and the brain, we did show that some of the most complex features can be explained by the model. Overall, we conclude that our model robustly addresses and supports both findings from existing neurobiology of language models, as well as evidence from neuroimaging studies of individual variability.

## Implications

Language is arguably the most complex behaviour of the human brain, and as such it requires a highly flexible model to account for its richness. Existing models of the neurobiology of language have not done enough to explain the variability of language processing in the real world, rather perpetuating localisational and static notions of language comprehension.

Here, we proposed the first network-based alternative model of the neurobiology of language that accounts for real-world behaviour: this model is highly flexible, dynamic and distributed, and considers both individual variability and the intricacies of various language features, as we have shown. Our network-based model unifies the notions of previous language models with findings from individual variability studies, identifying a specific role of 'language' regions as coordinative hubs in a wider and more complex language processing network. This has significant ramifications for our understanding of the neurobiology of language, as well as the methods commonly used in neuroimaging: it is clear from our work, that simple stimuli/tasks, central tendency measures and subtractive methods should be at least coupled with more naturalistic experiments, multivariate and inter-subject variability approaches, in order to inspect language features in their natural setting.

As such, our model is much better suited for predicting individual cognitive abilities, for understanding how the brain adapts to task changes, and how neuroplasticity after damage may help re-establish lost functions. In this latter context, our model has significant implications for patients with aphasia, as it can help identify additional brain regions and pathways for rewiring that can be exploited through novel speech therapies. Existing models cannot explain the heterogeneity in symptoms nor recovery of aphasic patients (Geranmayeh et al., 2014), and there has been a push in the aphasia literature for a network-based approach to understanding recovery, as this may help better inspect individual pathways (Kiran & Thompson, 2019).

Therefore, by accounting for individual variability and proposing a network-based approach, our model may help better understand the heterogeneity of aphasic symptoms and recovery pathways (Geranmayeh et al., 2014). Here, our model could significantly boost the development of individualised speech therapies, through the creation of programs aimed at rerouting connectivity to brain regions outside of 'language' areas in aphasic patients.

## Future work

In this work we have proposed a new flexible network-based model of the neurobiology of language. However, although we demonstrated a direct contribution of distributed regions in language processing and proposed pathways for these, we did not inspect the specific connectivity that various language features elicit. As such, we plan to investigate how word frequency and semantic context drive the connectivity profiles of language processing in the future. This would help elucidate the specific contribution in the network model of distributed vs 'language' regions.

Moreover, although we have performed dynamic functional connectivity using a sliding window approach, this method has its limitations as it (i) requires setting arbitrary parameters (e.g., window length), and (ii) relies heavily on estimating connections from Pearson's correlations rather than using more causational techniques. Thus, in order to verify our temporal results further, we have already begun conducting analyses using a novel data-driven approach called temporal delay, that uses a Hilbert transform function to capture the response delay between voxels' timeseries (Saad et al., 2003). The preliminary results show that sensorimotor regions (e.g., primary motor, visual, auditory) experience the longest time delays relative to other voxels, with a temporal profile indicating that these regions take on the sensory input first and also gather the processed information from other brain regions at the end. This is in line with a role as top-layer stable cores. We plan to continue this analysis and include visibility graph analysis (VGA), a technique to estimate the length and dynamics of temporal windows for various brain regions (Sannino et al., 2017), in order to inspect the dynamics of language processing networks in more depth.

Finally, in order to further investigate the relationship between network dynamics and individual cognitive strategies/abilities, we have implemented a long-short-term memory (LSTM) machine learning model that predicts the NIH toolbox cognitive scores from the individual sliding windows' connectivity: this model has already achieved 98% accuracy in predicting various cognitive and emotional scores of individual participants from their connectivity profiles. Interestingly, different graph measures accurately predicted scores at different time intervals, suggesting that the movie stimuli at a particular time engage a given cognitive state through a specific connectivity profile. We plan to investigate this point further, in order to understand how the specific stimulus relates to the network dynamics, and how these translate to behaviour or cognitive state. This particular study would help elucidate how individual participants' network features relate to cognitive strategies and abilities, but also to

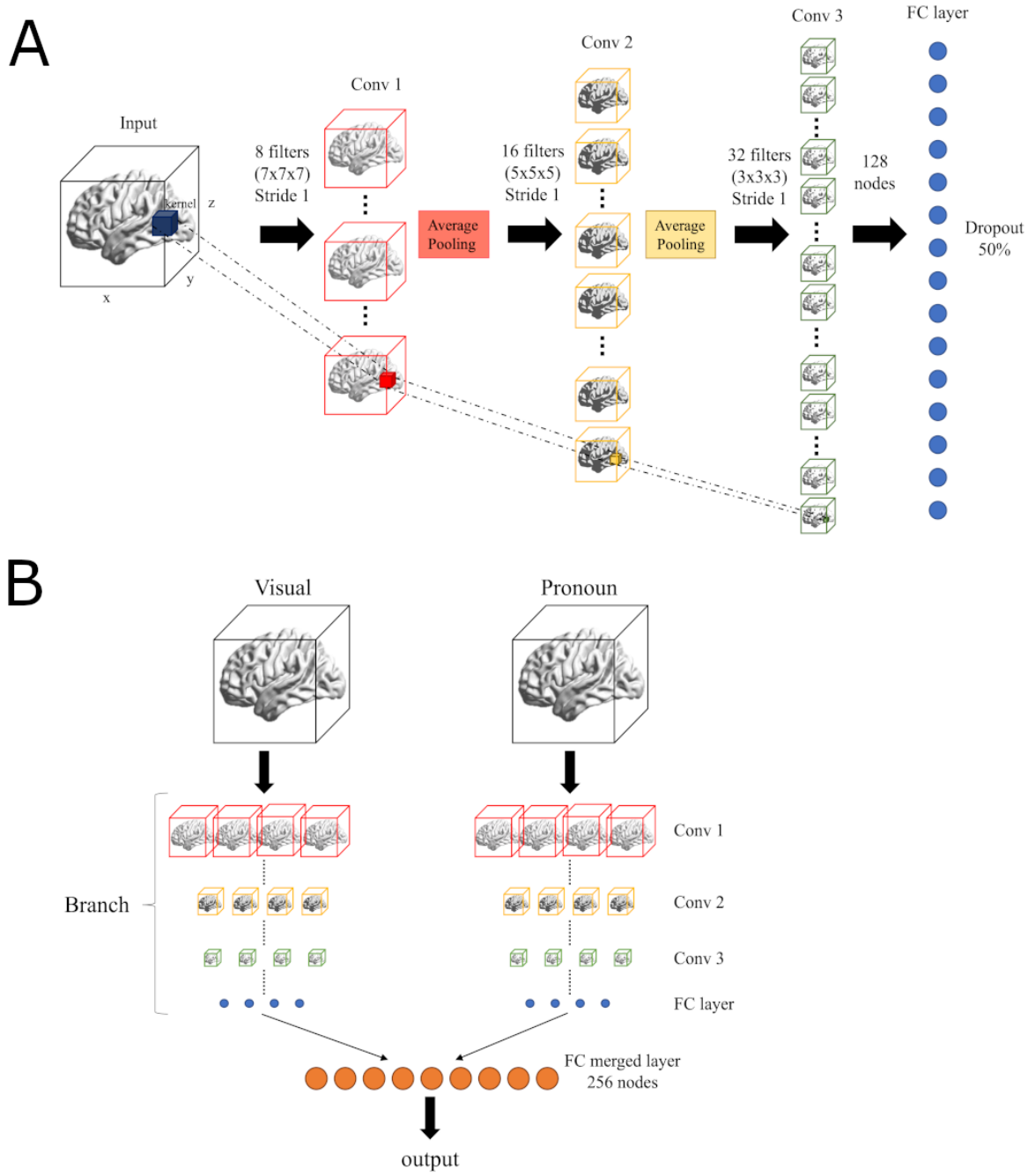
their emotional state. This would allow us to not only understand language and the brain in more depth, but also investigate individual biomarkers of mental health.

## Conclusions

This thesis presented a novel network-based model of natural language processing that supports the complexity and flexibility of language in the real world. We have demonstrated that a multiple core-periphery network architecture coupled with dynamic and largely distributed communities, best encompasses various aspects of language in its natural setting. This model supports both individual participant variability and shared language features and pathways, thus making it ideal for investigating individual differences in cognition, neuroplasticity and disease progression.

# Supplementary Materials

## S1. Diagram of 3D DNN with decreasing kernel with 3 layers




A) Diagram of the 3D DNN with a decreasing kernel size, with hyperparameters and architecture obtained from (Vu et al., 2020). The model had 3 convolutional layers, which were progressively removed starting from the top of the hierarchy (input layer) for testing. The final model included only the last convolutional layer with 3x3x3 kernel size. B) Diagram representing the branched architecture

containing visual and pronoun branches, with 3 convolutional layers; each branch comprised a single convolutional layer in the final model.

## S2. Diagram of 3D DNN with increasing kernel

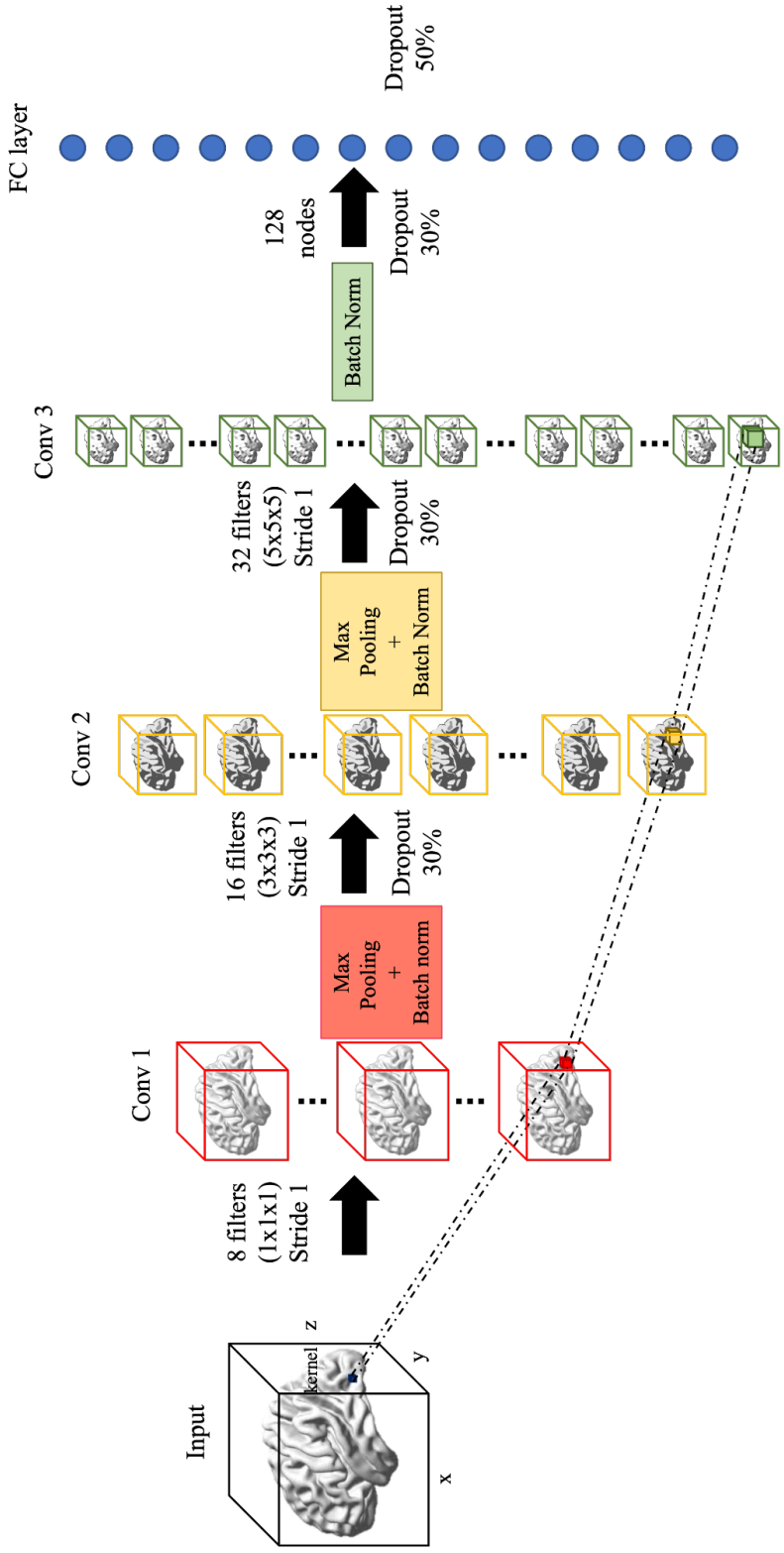


Diagram of the 3D DNN model architecture and hyperparameters with increasing kernel size. The model contained 3 convolutional layers with kernel sizes starting at 1x1x1 and ending at 5x5x5. This architecture constituted one branch of the model, with the final model having the same architecture as S1 above.

## S3. Diagram of 2D RSN pretrained model

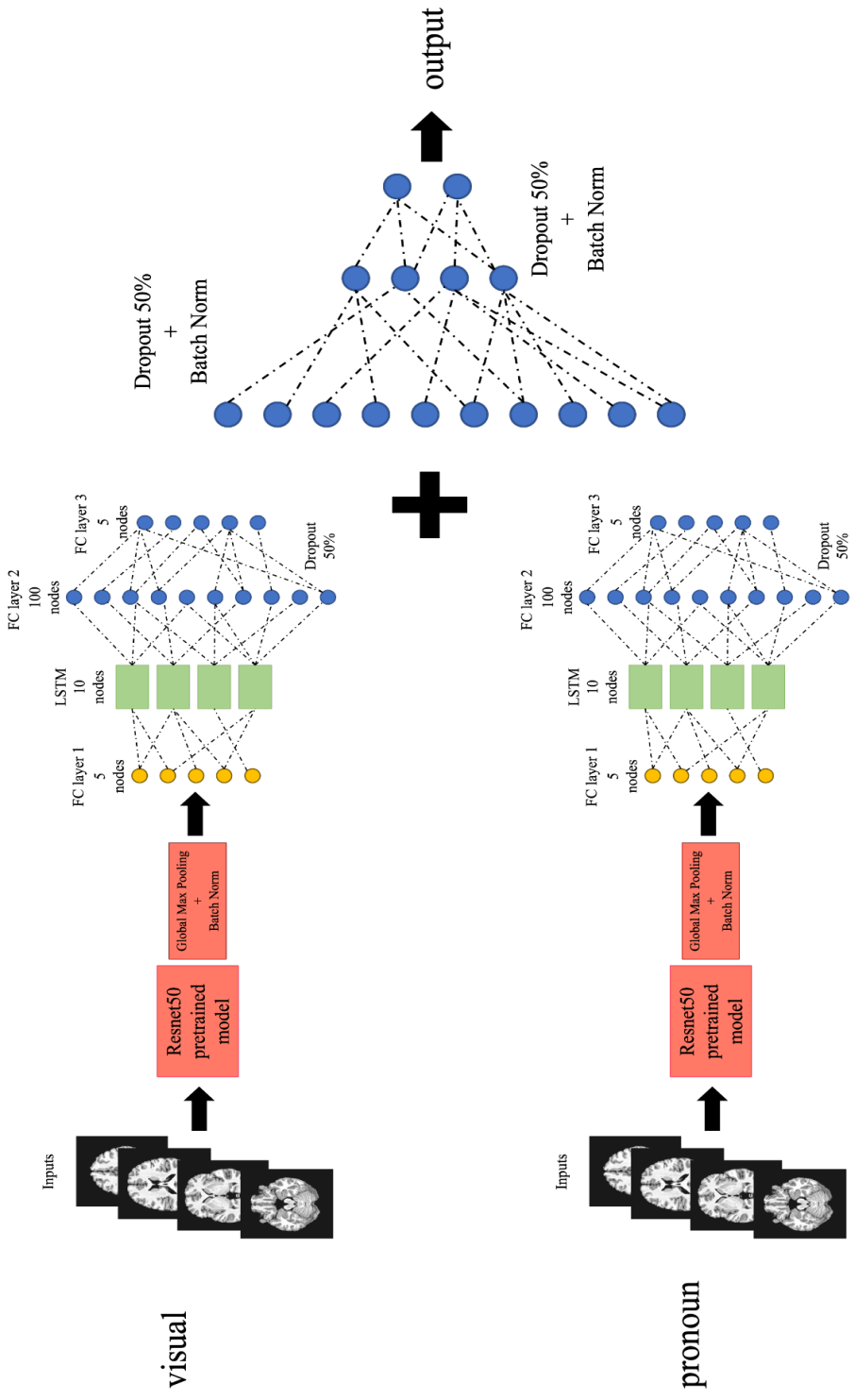


Diagram of the RSN model with its hyperparameters. Here 2D brain images (slices) were input into the visual and pronoun branch, each containing a set of convolutional and fully connected layers, whose weights were pre-trained using the ResNet50 model, and emerging in one long-short-term-memory (LSTM) layer to combine all 2D slices from one sample into one prediction. The two branches were finally merged into deeper fully connected layers.

**S4. Movie of individual level communities**

Link to the movie on an example of individual network communities from '500 Days of Summer', as they change over time. The change in colour from one image to the next indicates that a community has evolved (e.g., has split, was born, was merged or died). Communities varied often, nearing ~60 variations over time.

https://drive.google.com/file/d/1YhlbCgfMCx6Ogx1lOaUoPMqqyeFyEpHD/view?usp=sharing

**S5. Movie of group-averaged network communities**

Link to the movie on group-averaged network communities from the movie '500 Days of Summer', as they change over time. The change in colour from one image to the next indicates that a community has evolved (e.g., has split, was born, was merged or died). Communities did vary, but less often (< 40) compared to individual-level communities.

https://drive.google.com/file/d/1jcM3uI893U2uYum_Uy-v-3xyP0yXM6ZB/view?usp=sharing

**S6. Considerations on network algorithms for future work**

***Network construction***. As is the case for the entire network neuroscience field, there is still no consensus on the appropriate window and step length to use for the sliding window approach. Previous studies suggested window lengths of 30-60 sec and <50 sec step size, but no clear method is available (Preti et al., 2017). It would be interesting to apply various window and step lengths, to test whether we can capture different dynamics of the network. Here, we initially computed pairwise Pearson's correlation values to build adjacency matrices. Various thresholding approaches are available: from absolute thresholding and proportional edge-density ones, to more complex n,K-dependent thresholds (Garrison et al., 2015). We chose a proportional threshold of 10% edge density based on previous literature suggesting that this method is better suited for network comparisons (Alexander-Bloch et al., 2012), and based on qualitative inspection and available computational resources. We propose to try an absolute threshold in future, as it can reveal different features about individual variability, given that the

resulting individual networks may have different densities and thus structures (Garrison et al., 2015). An understudied feature of network construction is the role of negative correlation values, which are removed in all network neuroscience studies. These may still represent important functional connections, which may be important to define feedback connections or disease, but they are always discarded as noise (Kazeminejad & Sotero, 2020; Parente & Colosimo, 2020; Zhan et al., 2017).

Finally, we aimed at obtaining the highest possible resolution of the network which for our data was $3mm^3$. However, due to computational limitations, we chose to resample the network to $5mm^3$ voxel-wise. In future, we aim to increase the resolution of our model further.

***Algorithms***. Our novel core-periphery algorithm represents a significant improvement in detection of this meso-scale structure in various networks (Shen et al., 2021). The original definition of core-periphery was based on the estimation of a 'coreness' value for each node (Borgatti & Everett, 2000), but no algorithm so far is able to estimate this measure, to the best of our knowledge. Instead, studies have used proxy measures, such as k-core decomposition and centrality, to define core nodes (Fornito et al., 2016a). However, while all cores are hubs, not all hubs are cores (Borgatti & Everett, 2006). Our algorithm offers a more accurate estimate of 'coreness' and can detect multiple core-periphery structures that have been described in various networks (Yan & Luo, 2019), making it overall more suitable than existing algorithms to detect complex core-periphery structures (Shen et al., 2021).

Finally, although the *Louvain* algorithm is widely used and has been extensively tested, it cannot detect overlaps between communities (Lancichinetti et al., 2010; Palla et al., 2005). These overlaps can be indicative of a shared cognitive function. Overlaps were recently found in protein-protein interactomes, where they acted as multiple cores of a core-periphery network (Yang & Leskovec, 2014). It would be interesting to then use a different algorithm, such as OSLOM that is designed to find overlapping communities (Lancichinetti et al., 2010), to identify potential overlaps and test whether our multi-cores act as shared nodes between two or more communities. Since OSLOM requires more computational resources, we aim at analysing our networks with this novel algorithm in the future.

# References

Alexander-Bloch, A. F., Gogtay, N., Meunier, D., Birn, R., Clasen, L., Lalonde, F., Lenroot, R., Giedd, J., & Bullmore, E. T. (2010). Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia. Frontiers in Systems Neuroscience, 4, 147.

Alexander-Bloch, A., Lambiotte, R., Roberts, B., Giedd, J., Gogtay, N., & Bullmore, E. (2012). The discovery of population differences in network community structure: new methods and applications to brain functional networks in schizophrenia. *NeuroImage*, *59*(4), 3889–3900.

Aliko, S., Huang, J., Gheorghiu, F., Meliss, S., & Skipper, J. I. (2020). A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, *7*(1), 347.

Altmann, G. T. M., & Ekves, Z. (2019). Events as intersecting object histories: A new theory of event representation. Psychological Review, 126(6), 817–840.

Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: eye movements and mental representation. *Cognition*, *111*(1), 55–71.

Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H. B., & Zilles, K. (1999). Broca's region revisited: cytoarchitecture and intersubject variability. *The Journal of Comparative Neurology*, *412*(2), 319–341.

Anderson, J. S., Ferguson, M. A., Lopez-Larson, M., & Yurgelun-Todd, D. (2011). Reproducibility of single-subject functional connectivity measurements. *AJNR. American Journal of Neuroradiology*, *32*(3), 548–555.

Andric, M., Goldin-Meadow, S., Small, S. L., & Hasson, U. (2016). Repeated movie viewings produce similar local activity patterns but different network configurations.

*NeuroImage*, *142*, 613–627.

Arslan, S., Devers, C., & Ferreiro, S. M. (2021). Pronoun processing in post-stroke aphasia: A meta-analytic review of individual data. *Journal of Neurolinguistics*, *59*, 101005.

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2014). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience*, *9*(6), 817–824.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, *95*(3), 709–721.e5.

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of Real-World Event Schemas during Narrative Perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(45), 9689–9699.

Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences*, *11*(8), 327–332.

Bartels, A., & Zeki, S. (2004). The chronoarchitecture of the human brain--natural viewing conditions reveal a time-based anatomy of the brain. *NeuroImage*, *22*(1), 419–433.

Bartels, A., & Zeki, S. (2005). The chronoarchitecture of the cerebral cortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 733–750.

Bassett, D. S., & Bullmore, E. (2006). Small-world brain networks. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *12*(6), 512–523.

Bassett, D. S., Porter, M. A., Wymbs, N. F., Grafton, S. T., Carlson, J. M., & Mucha, P. J. (2013). Robust detection of dynamic community structure in networks. *Chaos* , *23*(1), 013142.

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., & Grafton, S. T.

(2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(18), 7641–7646.

Bassett, D. S., Wymbs, N. F., Rombach, M. P., Porter, M. A., Mucha, P. J., & Grafton, S. T. (2013). Task-based core-periphery organization of human brain dynamics. *PLoS Computational Biology*, *9*(9), e1003171.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*, 133–155.

Berkovich-Ohana, A., Noy, N., Harel, M., Furman-Haran, E., Arieli, A., & Malach, R. (2020). Inter-participant consistency of language-processing networks during abstract thoughts. NeuroImage, 211, 116626.

Betzel, R. F., & Bassett, D. S. (2017). Multi-scale brain networks. *NeuroImage*, *160*, 73–83.

Betzel, R. F., Medaglia, J. D., Papadopoulos, L., Baum, G. L., Gur, R., Gur, R., Roalf, D., Satterthwaite, T. D., & Bassett, D. S. (2017). The modular organization of human anatomical brain networks: Accounting for the cost of wiring. Network Neuroscience (Cambridge, Mass.), 1(1), 42–68.

Binder, J. R. (2015). The Wernicke area: Modern evidence and a reinterpretation. *Neurology*, *85*(24), 2170–2175.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex* , *19*(12), 2767–2796.

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., Dogonowski, A.-M., Ernst, M., Fair, D., Hampson, M., Hoptman, M. J., Hyde, J. S., Kiviniemi, V. J., Kötter, R., Li, S.-J., … Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings*

*of the National Academy of Sciences of the United States of America*, *107*(10), 4734–4739.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. In *arXiv [physics.soc-ph]*. arXiv. http://arxiv.org/abs/0803.0476

Boeke, E. A., Holmes, A. J., & Phelps, E. A. (2019). Toward Robust Anxiety Biomarkers: A Machine Learning Approach in a Large-Scale Sample. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*. https://doi.org/10.1016/j.bpsc.2019.05.018

Borgatti, S. P., & Everett, M. G. (2000). Models of core/periphery structures. *Social Networks*, *21*(4), 375–395.

Borgatti, S. P., & Everett, M. G. (2006). A Graph-theoretic perspective on centrality. *Social Networks*, *28*(4), 466–484.

Bottenhorn, K. L., Flannery, J. S., Boeving, E. R., Riedel, M. C., Eickhoff, S. B., Sutherland, M. T., & Laird, A. R. (2019). Cooperating yet distinct brain networks engaged during naturalistic paradigms: A meta-analysis of functional MRI results. *Network Neuroscience (Cambridge, Mass.)*, *3*(1), 27–48.

Braun, U., Schäfer, A., Walter, H., Erk, S., Romanczuk-Seiferth, N., Haddad, L., Schweiger, J. I., Grimm, O., Heinz, A., Tost, H., Meyer-Lindenberg, A., & Bassett, D. S. (2015). Dynamic reconfiguration of frontal brain networks during executive cognition in humans. Proceedings of the National Academy of Sciences of the United States of America, 112(37), 11678–11683.

Brunswik, E. (1943). Organismic Achievement and Environmental Probability. Psychological Review, 50(3), 255–272.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. Psychological Review, 62(3), 193–217.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews. Neuroscience, 10(3), 186–198.

Burton, M. W., Noll, D. C., & Small, S. L. (2001). The anatomy of auditory word processing: individual variability. Brain and Language, 77(1), 119–131.

Caballero-Gaudes, C., & Reynolds, R. C. (2017). Methods for cleaning the BOLD fMRI signal. *NeuroImage*, *154*, 128–149.

Carroll, N. (1985). The Power of Movies. *Daedalus*, *114*(4), 79–103.

Carroll, N., & Seeley, W. P. (2013). Cognitivism, psychology, and neuroscience: Movies as attentional engines. In A. P. Shimamura (Ed.), *Psychocinematics: Exploring cognition at the movies , (pp* (Vol. 385, pp. 53–75). Oxford University Press, xii.

Casorso, J., Kong, X., Chi, W., Van De Ville, D., Yeo, B. T. T., & Liégeois, R. (2019). Dynamic mode decomposition of resting-state and task fMRI. *NeuroImage*, *194*, 42–54.

Cauley, S. F., Polimeni, J. R., Bhat, H., Wald, L. L., & Setsompop, K. (2014). Interslice leakage artefact reduction technique for simultaneous multislice acquisitions. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *72*(1), 93–102.

Cayhono, S. C. (2019). Comparison of document similarity measurements in scientific writing using Jaro-Winkler Distance method and Paragraph Vector method. *Materials Science and Engineering*.

Chang, C. H. C., Lazaridi, C., Yeshurun, Y., Norman, K. A., & Hasson, U. (2021). Relating the Past with the Present: Information Integration and Segregation during Ongoing Narrative Processing. *Journal of Cognitive Neuroscience*, *33*(6), 1106–1128.

Cheetham, M., Hänggi, J., & Jancke, L. (2014). Identifying with fictive characters: structural brain correlates of the personality trait “fantasy.” *Social Cognitive and Affective Neuroscience*, *9*(11), 1836–1844.

Chen, E. E., & Small, S. L. (2007). Test-retest reliability in fMRI of language: group and task effects. *Brain and Language*, *102*(2), 176–185.

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., & Petersen, S. E. (2014). Intrinsic and task-evoked network architectures of the human brain. *Neuron*, *83*(1), 238–251.

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, *250*, 126–136.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173.

de Haan, W., van der Flier, W. M., Koene, T., Smits, L. L., Scheltens, P., & Stam, C. J. (2012). Disrupted modular brain dynamics reflect cognitive dysfunction in Alzheimer's disease. *NeuroImage*, *59*(4), 3085–3093.

de Reus, M. A., Saenger, V. M., Kahn, R. S., & van den Heuvel, M. P. (2014). An edge-centric perspective on the human connectome: link communities in the brain. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369(1653). https://doi.org/10.1098/rstb.2013.0527

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. NeuroImage, 53(1), 1–15.

Dickerson, B. C., & Eichenbaum, H. (2010). The episodic memory system: neurocircuitry and disorders. Neuropsychopharmacology: Official Publication of the American College

of Neuropsychopharmacology, 35(1), 86–104.

di Oleggio Castello, M. V., Chauhan, V., Jiahui, G., & Ida Gobbini, M. (2020). The Grand Budapest Hotel: an fMRI dataset in response to a socially-rich, naturalistic movie. In bioRxiv (p. 2020.07.14.203257). https://doi.org/10.1101/2020.07.14.203257

Di, X., Gohel, S., Kim, E. H., & Biswal, B. B. (2013). Task vs. rest-different network configurations between the coactivation and the resting-state brain networks. Frontiers in Human Neuroscience, 7, 493.

Dronkers, N. F., Ivanova, M. V., & Baldo, J. V. (2017). What Do Language Disorders Reveal about Brain-Language Relationships? From Classic Models to Network Approaches. Journal of the International Neuropsychological Society: JINS, 23(9-10), 741–754.

Duff, M. C., & Brown-Schmidt, S. (2012). The hippocampus and the flexible use and processing of language. Frontiers in Human Neuroscience, 6, 69.

DuPre, E., Hanke, M., & Poline, J.-B. (2019). Nature abhors a paywall: How open science can realize the potential of naturalistic stimuli. NeuroImage, 116330.

Eickhoff, S. B., Milham, M., & Vanderwal, T. (2020). Towards clinical applications of movie fMRI. *NeuroImage*, 116860.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. Psychological Science, 31(7), 792–806.

Euston, D. R., Gruber, A. J., & McNaughton, B. L. (2012). The role of medial prefrontal cortex in memory and decision making. *Neuron*, *76*(6), 1057–1070.

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*(3), 120–126.

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Gunther, M., Glasser,

M. F., Miller, K. L., Ugurbil, K., & Yacoub, E. (2010). Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. *PloS One*, *5*(12), e15710.

Feinberg, D. A., & Setsompop, K. (2013). Ultra-fast MRI of the human brain with simultaneous multi-slice imaging. *Journal of Magnetic Resonance* , *229*, 90–100.

Ferstl, E. C., & von Cramon, D. Y. (2002). What does the frontomedian cortex contribute to language processing: coherence or theory of mind? NeuroImage, 17(3), 1599–1612.

Fischl, B. (2012). FreeSurfer. NeuroImage, 62(2), 774–781.

Flegal, K. E., Marín-Gutiérrez, A., Ragland, J. D., & Ranganath, C. (2014). Brain mechanisms of successful recognition through retrieval of semantic context. Journal of Cognitive Neuroscience, 26(8), 1694–1704.

Fletcher, P. C., Frith, C. D., Baker, S. C., Shallice, T., Frackowiak, R. S. J., & Dolan, R. J. (1995). The Mind's Eye—Precuneus Activation in Memory-Related Imagery. NeuroImage, 2(3), 195–200.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition, 57(2), 109–128.

Fornito, A., Zalesky, A., & Bullmore, E. T. (Eds.). (2016a). Chapter 6 - Components, Cores, and Clubs. In Fundamentals of Brain Network Analysis (pp. 163–206). Academic Press.

Fornito, A., Zalesky, A., & Bullmore, E. T. (Eds.). (2016b). Chapter 9 - Modularity. In Fundamentals of Brain Network Analysis (pp. 303–354). Academic Press.

Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. Proceedings of the National Academy of Sciences of the United States of America, 104(1), 36–41.

Frankland, P. W., Josselyn, S. A., & Köhler, S. (2019). The neurobiological foundation of memory retrieval. Nature Neuroscience, 22(10), 1576–1585.

Freud, E., Plaut, D. C., & Behrmann, M. (2016). “What” Is Happening in the Dorsal Visual Pathway. Trends in Cognitive Sciences, 20(10), 773–784.

Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, D.-B., & Rorden, C. (2016). Revealing the dual streams of speech processing. Proceedings of the National Academy of Sciences of the United States of America, 113(52), 15108–15113.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. Trends in Cognitive Sciences, 6(2), 78–84.

Friedman, L., Glover, G. H., Krenz, D., Magnotta, V., & FIRST BIRN. (2006). Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. NeuroImage, 32(4), 1656–1668.

Friston, K. J., & Price, C. J. (2001). Dynamic representations and generative models of brain function. Brain Research Bulletin, 54(3), 275–285.

Friston, K. J., Price, C. J., Fletcher, P., Moore, C., Frackowiak, R. S. J., & Dolan, R. J. (1996). The trouble with cognitive subtraction. NeuroImage, 4(2), 97–104.

Frost, M. A., & Goebel, R. (2012). Measuring structural–functional correspondence: Spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage, 59(2), 1369–1381.

Garrison, K. A., Scheinost, D., Finn, E. S., Shen, X., & Constable, R. T. (2015). The (in)stability of functional brain network measures across thresholds. *NeuroImage*, *118*, 651–661.

Geranmayeh, F., Brownsett, S. L. E., & Wise, R. J. S. (2014). Task-induced brain activity in aphasic stroke patients: what is driving recovery? Brain: A Journal of Neurology, 137(Pt 10), 2632–2648.

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH toolbox for assessment of neurological and behavioral function. *Neurology*,

*80*(11 Suppl 3), S2–S6.

Geschwind, N. (1970). The organization of language and the brain. *Science*, *170*(3961), 940–944.

Geuter, S., Qi, G., Welsh, R. C., Wager, T. D., & Lindquist, M. A. (2018). Effect Size and Power in fMRI Group Analysis. In *bioRxiv* (p. 295048). https://doi.org/10.1101/295048

Giorgino, T., & Others. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of Statistical Software*, *31*(7), 1–24.

Gonzalez-Castillo, J., & Bandettini, P. A. (2018). Task-based dynamic functional connectivity: Recent findings and open questions. *NeuroImage*, *180*(Pt B), 526–533.

González, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuán, A., Belloch, V., & Avila, C. (2006). Reading cinnamon activates olfactory brain regions. *NeuroImage*, *32*(2), 906–912.

Goodyear, B. G., & Menon, R. S. (1998). Effect of luminance contrast on BOLD fMRI response in human primary visual areas. Journal of Neurophysiology, 79(4), 2204–2207.

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J. M., Coalson, R. S., Nguyen, A. L., McDermott, K. B., Shimony, J. S., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Nelson, S. M., & Dosenbach, N. U. F. (2017). Precision Functional Mapping of Individual Human Brains. *Neuron*, *95*(4), 791–807.e7.

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., Whittle, I., & Pernet, C. (2013). Single subject fMRI test-retest reliability metrics and confounding factors. *NeuroImage*, *69*, 231–243.

Greene, D. J., Koller, J. M., Hampton, J. M., Wesevich, V., Van, A. N., Nguyen, A. L., Hoyt, C. R., McIntyre, L., Earl, E. A., Klein, R. L., Shimony, J. S., Petersen, S. E., Schlaggar, B. L., Fair, D. A., & Dosenbach, N. U. F. (2018). Behavioral interventions for reducing

head motion during MRI scans in children. *NeuroImage*, *171*, 234–245.

Griffanti, L., Douaud, G., Bijsterbosch, J., Evangelisti, S., Alfaro-Almagro, F., Glasser, M. F., Duff, E. P., Fitzgibbon, S., Westphal, R., Carone, D., Beckmann, C. F., & Smith, S. M. (2017). Hand classification of fMRI ICA noise components. *NeuroImage*, *154*, 188–205.

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *70*(2 Pt 2), 025101.

Gu, S., Xia, C. H., Ciric, R., Moore, T. M., Gur, R. C., Gur, R. E., Satterthwaite, T. D., & Bassett, D. S. (2019). Unifying the Notions of Modularity and Core-Periphery Structure in Functional Brain Networks during Youth. *Cerebral Cortex* . https://doi.org/10.1093/cercor/bhz150

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, *6*(7), e159.

Hagoort, P. (2016). Chapter 28 - MUC (Memory, Unification, Control): A Model on the Neurobiology of Language Beyond Single Word Processing. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 339–347). Academic Press.

Hammer, A., Goebel, R., Schwarzbach, J., Münte, T. F., & Jansma, B. M. (2007). When sex meets syntactic gender on a neural basis during pronoun processing. *Brain Research*, *1146*, 185–198.

Hammer, A., Jansma, B. M., Tempelmann, C., & Münte, T. F. (2011). Neural mechanisms of anaphoric reference revealed by FMRI. *Frontiers in Psychology*, *2*, 32.

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., Nigbur, R., Waite, A. Q., Baumgartner, F., & Stadler, J. (2016). A studyforrest extension,

simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, *3*, 160092.

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*, 140003.

Hashemi, M. (2019). Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data*, *6*(1), 1–13.

Hasson, U., & Honey, C. J. (2012). Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, *62*(2), 1272–1278.

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*(1), 40–48.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640.

Hasson, U., Nusbaum, H. C., & Small, S. L. (2007). Brain networks subserving the extraction of sentence information and its encoding to memory. *Cerebral Cortex* , *17*(12), 2899–2913.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–416.

Hertrich, I., Dietrich, S., Blum, C., & Ackermann, H. (2021). The Role of the Dorsolateral Prefrontal Cortex for Speech and Language Processing. *Frontiers in Human Neuroscience*, *15*, 645209.

Heun, R., Jessen, F., Klose, U., Erb, M., Granath, D., Freymann, N., & Grodd, W. (2000). Interindividual variation of cerebral activation during encoding and retrieval of words.

*European Psychiatry: The Journal of the Association of European Psychiatrists*, *15*(8), 470–479.

He, Y., & Evans, A. (2010). Graph theoretical modeling of brain connectivity. *Current Opinion in Neurology*, *23*(4), 341–350.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402.

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., & Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *Journal of Computer Assisted Tomography*, *22*(2), 324–333.

Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., & Shao, L. (2020). Normalization Techniques in Training DNNs: Methodology, Analysis and Application. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2009.12836

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. Journal of Cognitive Neuroscience, 18(4), 665–679.

Hurlburt, R. T., Alderson-Day, B., Fernyhough, C., & Kühn, S. (2015). What goes on in the resting-state? A qualitative glimpse into resting-state experience in the scanner. *Frontiers in Psychology*, *6*, 1535.

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., Gonzalez-Castillo, J., Handwerker, D. A., Keilholz, S., Kiviniemi, V., Leopold, D. A., de Pasquale, F., Sporns, O., Walter, M., & Chang, C. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage*, *80*, 360–378.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature,

532(7600), 453–458.

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., & Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, *30*(9), 1617–1634.

Isoda, M., & Noritake, A. (2013). What makes the dorsomedial frontal cortex active during reading the mental states of others? *Frontiers in Neuroscience*, *7*, 232.

Jarema, G., & Friederici, A. D. (1994). Processing articles and pronouns in agrammatic aphasia: evidence from French. *Brain and Language*, *46*(4), 683–694.

Juch, H., Zimine, I., Seghier, M. L., Lazeyras, F., & Fasel, J. H. D. (2005). Anatomical variability of the lateral frontal lobe surface: implication for intersubject variability in language neuroimaging. *NeuroImage*, *24*(2), 504–514.

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, *17*(12), 1174–1179.

Kaufmann, T., Alnæs, D., Brandt, C. L., Doan, N. T., Kauppi, K., Bettella, F., Lagerberg, T. V., Berg, A. O., Djurovic, S., Agartz, I., Melle, I. S., Ueland, T., Andreassen, O. A., & Westlye, L. T. (2017). Task modulations and clinical manifestations in the brain functional connectome in 1615 fMRI datasets. *NeuroImage*, *147*, 243–252.

Kazeminejad, A., & Sotero, R. C. (2020). The Importance of Anti-correlations in Graph Theory Based Classification of Autism Spectrum Disorder. *Frontiers in Neuroscience*, *14*, 676.

Khosla, M., Jamison, K., Ngo, G. H., Kuceyeski, A., & Sabuncu, M. R. (2019). Machine learning in resting-state fMRI analysis. *Magnetic Resonance Imaging*, *64*, 101–121.

Kiefer, M., Sim, E.-J., Herrnberger, B., Grothe, J., & Hoenig, K. (2008). The Sound of Concepts: Four Markers for a Link between Auditory and Conceptual Brain Systems.

*Journal of Neuroscience*, *28*(47), 12224–12230.

Kim, D., Kay, K., Shulman, G. L., & Corbetta, M. (2018). A New Modular Brain Organization of the BOLD Signal during Natural Vision. *Cerebral Cortex* , *28*(9), 3065–3081.

Kiran, S., & Thompson, C. K. (2019). Neuroplasticity of Language Networks in Aphasia: Advances, Updates, and Future Challenges. Frontiers in Neurology, 10, 295.

Kitzbichler, M. G., Henson, R. N. A., Smith, M. L., Nathan, P. J., & Bullmore, E. T. (2011). Cognitive effort drives workspace configuration of human brain functional networks. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(22), 8259–8270.

Klepp, A., van Dijk, H., Niccolai, V., Schnitzler, A., & Biermann-Ruben, K. (2019). Action verb processing specifically modulates motor behaviour and sensorimotor neuronal oscillations. *Scientific Reports*, *9*(1), 15985.

Kojaku, S., & Masuda, N. (2017). Finding multiple core-periphery pairs in networks. *Physical Review. E*, *96*(5-1), 052313.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490.

Laird, A. R., Fox, P. M., Eickhoff, S. B., Turner, J. A., Ray, K. L., McKay, D. R., Glahn, D. C., Beckmann, C. F., Smith, S. M., & Fox, P. T. (2011). Behavioral interpretations of intrinsic connectivity networks. *Journal of Cognitive Neuroscience*, *23*(12), 4022–4037.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *80*(5 Pt 2), 056117.

Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *84*(6 Pt

2), 066122.

Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports*, *2*, 336.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., & Fortunato, S. (2010). Finding statistically significant communities in networks. In *arXiv [physics.soc-ph]*. arXiv. http://arxiv.org/abs/1012.2363

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M.-Y., Gilmore, A. W., McDermott, K. B., Nelson, S. M., Dosenbach, N. U. F., Schlaggar, B. L., Mumford, J. A., Poldrack, R. A., & Petersen, S. E. (2015). Functional System and Areal Organization of a Highly Sampled Individual Human Brain. *Neuron*, *87*(3), 657–670.

Lehmann, B. C. L., Henson, R. N., Geerligs, L., Cam-CAN, & White, S. R. (2019). Characterising group-level brain connectivity: a framework using Bayesian exponential random graph models. In *bioRxiv* (p. 665398). https://doi.org/10.1101/665398

Li, J., Fabre, M., Luh, W.-M., & Hale, J. (2018). Modeling Brain Activity Associated with Pronoun Resolution in English and Chinese. *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, 87–96.

Liu, B., Wei, Y., Zhang, Y., & Yang, Q. (2017). Deep neural networks for high dimension, low sample size data. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia. https://doi.org/10.24963/ijcai.2017/318

Liu, X., Zhen, Z., Yang, A., Bai, H., & Liu, J. (2019). A manually denoised audio-visual movie watching fMRI dataset for the studyforrest project. *Scientific Data*, *6*(1), 295.

Lohmann, G., Stelzer, J., Müller, K., Lacosse, E., Buschmann, T., Kumar, V. J., Grodd, W., & Scheffler, K. (2017). Inflated false negative rates undermine reproducibility in task-

based fMRI. In *bioRxiv* (p. 122788). https://doi.org/10.1101/122788

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291.

Madan, C. R. (2018). Age differences in head motion and estimates of cortical morphology. *PeerJ*, *6*, e5176.

Madan, C. R. (2021). Scan Once, Analyse Many: Using Large Open-Access Neuroimaging Datasets to Understand the Brain. *Neuroinformatics*. https://doi.org/10.1007/s12021-021-09519-6

Maguire, E. A. (2012). Studying the freely-behaving brain with fMRI. *NeuroImage*, *62*(2), 1170–1176.

Maguire, E. A., Frith, C. D., & Morris, R. G. (1999). The functional neuroanatomy of comprehension and memory: the importance of prior knowledge. Brain: A Journal of Neurology, 122 ( Pt 10), 1839–1850.

Manfredi, M., Proverbio, A. M., Gonçalves Donate, A. P., Macarini Gonçalves Vieira, S., Comfort, W. E., De Araújo Andreoli, M., & Boggio, P. S. (2017). tDCS application over the STG improves the ability to recognize and appreciate elements involved in humor processing. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *235*(6), 1843–1852.

Matusz, P. J., Dikker, S., Huth, A. G., & Perrodin, C. (2019). Are We Ready for Real-world Neuroscience? *Journal of Cognitive Neuroscience*, *31*(3), 327–338.

Maviel, T., Durkin, T. P., Menzaghi, F., & Bontempi, B. (2004). Sites of neocortical reorganization critical for remote spatial memory. *Science*, *305*(5680), 96–99.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., Matt, M., Battenbergk, E., & Nieto, O. (2015, July 8). *Librosa Audio and Music Signal Analysis in Python*.

https://www.youtube.com/watch?v=MhOdbtPhbLU

McMillan, C. T., Clark, R., Gunawardena, D., Ryant, N., & Grossman, M. (2012). fMRI evidence for strategic decision-making during resolution of pronoun reference. *Neuropsychologia*, *50*(5), 674–687.

Mesulam, M.-M., Thompson, C. K., Weintraub, S., & Rogalski, E. J. (2015). The Wernicke conundrum and the anatomy of language comprehension in primary progressive aphasia. *Brain: A Journal of Neurology*, *138*(Pt 8), 2423–2437.

Meunier, D., Lambiotte, R., & Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience*, *4*, 200.

Meunier, D., Lambiotte, R., Fornito, A., Ersche, K. D., & Bullmore, E. T. (2009). Hierarchical modularity in human brain functional networks. Frontiers in Neuroinformatics, 3, 37.

Michelmann, S., Price, A. R., Aubrey, B., Strauss, C. K., Doyle, W. K., Friedman, D., Dugan, P. C., Devinsky, O., Devore, S., Flinker, A., Hasson, U., & Norman, K. A. (2021). Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nature Communications*, *12*(1), 5394.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536.

Miller, M. B., Donovan, C.-L., Bennett, C. M., Aminoff, E. M., & Mayer, R. E. (2012). Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals. *NeuroImage*, *59*(1), 83–93.

Miller, M. B., Donovan, C.-L., Van Horn, J. D., German, E., Sokol-Hessner, P., & Wolford,

G. L. (2009). Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *NeuroImage*, *48*(3), 625–635.

Miller, M. B., Van Horn, J. D., Wolford, G. L., Handy, T. C., Valsangkar-Smyth, M., Inati, S., Grafton, S., & Gazzaniga, M. S. (2002). Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *Journal of Cognitive Neuroscience*, *14*(8), 1200–1214.

Moran, J. M., Lee, S. M., & Gabrieli, J. D. E. (2011). Dissociable neural systems supporting knowledge about human character and appearance in ourselves and others. *Journal of Cognitive Neuroscience*, *23*(9), 2222–2230.

Nasios, G., Dardiotis, E., & Messinis, L. (2019). From Broca and Wernicke to the Neuromodulation Era: Insights of Brain Language Networks for Neurorehabilitation. *Behavioural Neurology*, *2019*, 9894571.

Nastase, S. A. et al. Narratives: fMRI data for evaluating models of naturalistic language comprehension. *OpenNeuro* https://doi.org/10.18112/openneuro.ds002345.v1.1.1 (2019).

Neisser, U. (1976). *Cognition and Reality: Principles and Implications of Cognitive Psychology*. W. H. Freeman.

Newman, A. J., Pancheva, R., Ozawa, K., Neville, H. J., & Ullman, M. T. (2001). An event-related fMRI study of syntactic and semantic violations. Journal of Psycholinguistic Research, 30(3), 339–364.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology: CB*, *21*(19), 1641–1646.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Nummenmaa, L., Lahnakoski, J. M., & Glerean, E. (2018). Sharing the social world via intersubject neural synchronisation. *Current Opinion in Psychology*, *24*, 7–14.

Nyberg, L., Habib, R., McIntosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(20), 11120–11124.

Oedekoven, C. S. H., Keidel, J. L., Berens, S. C., & Bird, C. M. (2017). Reinstatement of memory representations for lifelike events over the course of a week. *Scientific Reports*, *7*(1), 14305.

Ojemann, G. A. (1979). Individual variability in cortical localization of language. *Journal of Neurosurgery*, *50*(2), 164–169.

Olshausen, B. A., & Field, D. J. (2006). What is the other 85 percent of V1 doing. *L. van Hemmen, & T. Sejnowski (Eds. )*, *23*, 182–211.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814–818.

Parente, F., & Colosimo, A. (2020). Functional connections between and within brain subnetworks under resting-state. *Scientific Reports*, *10*(1), 3438.

Park, H.-J., & Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science*, *342*(6158), 1238411.

Peristeri, E., & Tsimpli, I. M. (2013). Pronoun processing in Broca's aphasia: Discourse–syntax effects in ambiguous anaphora resolution. *Aphasiology*, *27*(11), 1381–1407.

Pfeuffer, J., McCullough, J. C., Van de Moortele, P. F., Ugurbil, K., & Hu, X. (2003). Spatial

dependence of the nonlinear BOLD response at short stimulus duration. *NeuroImage*, *18*(4), 990–1000.

Piai, V., Anderson, K. L., Lin, J. J., Dewar, C., Parvizi, J., Dronkers, N. F., & Knight, R. T. (2016). Direct brain recordings reveal hippocampal rhythm underpinnings of language processing. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(40), 11366–11371.

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, *25*(2), 100–110.

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a new neurobiology of language. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(41), 14125–14131.

Poeppel, D., & Hickok, G. (2004). Towards a new functional anatomy of language. *Cognition*, *92*(1-2), 1–12.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–1966.

Preti, M. G., Bolton, T. A., & Van De Ville, D. (2017). The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage*, *160*, 41–54.

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. Annals of the New York Academy of Sciences, 1191, 62–88.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847.

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and Language*, *127*(1), 86–103.

Pulvermüller, F., Shtyrov, Y., & Ilmoniemi, R. (2005). Brain signatures of meaning access in action word recognition. *Journal of Cognitive Neuroscience*, *17*(6), 884–892.

Qiu, L., Swaab, T. Y., Chen, H.-C., & Wang, S. (2012). The role of gender information in pronoun resolution: evidence from Chinese. *PloS One*, *7*(5), e36156.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience, 12(6), 718–724.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Saad, Z. S., DeYoe, E. A., & Ropella, K. M. (2003). Estimation of FMRI response delays. NeuroImage, 18(2), 494–504.

Sannino, S., Stramaglia, S., Lacasa, L., & Marinazzo, D. (2017). Visibility graphs for fMRI data: Multiplex temporal graphs and their modulations across resting-state networks. Network Neuroscience, 1(3), 208–221.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., & Weiller, C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(46), 18035–18040.

Seghier, M. L., & Price, C. J. (2018). Interpreting and Utilising Intersubject Variability in Brain Function. *Trends in Cognitive Sciences*, *22*(6), 517–530.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. Neuropsychologia, 138, 107307.

Shen, X., Aliko, S., Han, Y., Skipper, J. I., & Peng, C. (2021). Finding core-periphery structures with node influences. *IEEE Transactions on Network Science and*

*Engineering*.

Sheth, B. R., & Young, R. (2016). Two Visual Pathways in Primates Based on Sampling of Space: Exploitation and Exploration of Visual Information. *Frontiers in Integrative Neuroscience*, *10*, 37.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1312.6034

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*, 12141.

Skipper, J. I. (2015a). The NOLB model: A model of the natural organization of language and the brain. In R. M. Willems (Ed.), *Cognitive neuroscience of natural language use , (pp* (Vol. 265, pp. 101–134). Cambridge University Press, xiv.

Skipper, J. I. (2015b). The NOLB model: a model of the natural organization of language and the brain. In R. M. Willems & R. M. Willems (Eds.), *Cognitive Neuroscience of Natural Language Use* (pp. 101–134). Cambridge University Press.

Skipper, J. I., Aliko, S., Brown, S., Jo, Y. J., Lo, S., Molimpakis, E., & Lametti, D. R. (2021). Reorganization of the Neurobiology of Language After Sentence Overlearning. *Cerebral Cortex* . https://doi.org/10.1093/cercor/bhab354

Skipper, J. I., & Hasson, U. (2017). A Core Speech Circuit Between Primary Motor, Somatosensory, And Auditory Cortex: Evidence From Connectivity And Genetic Descriptions. In *bioRxiv* (p. 139550). https://doi.org/10.1101/139550

Skipper, J. I., & Zevin, J. D. (2017). Brain reorganization in anticipation of predictable words. In *bioRxiv* (p. 101113). https://doi.org/10.1101/101113

Smith, R., Lane, R. D., Alkozei, A., Bao, J., Smith, C., Sanova, A., Nettles, M., & Killgore,

W. D. S. (2018). The role of medial prefrontal cortex in the working memory maintenance of one's own emotional responses. *Scientific Reports*, *8*(1), 3460.

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., Kelly, M., Laumann, T., Miller, K. L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A. Z., Vu, A. T., … WU-Minn HCP Consortium. (2013). Resting-state fMRI in the Human Connectome Project. *NeuroImage*, *80*, 144–168.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 13040–13045.

Smith, V., Mitchell, D. J., & Duncan, J. (2018). Role of the Default Mode Network in Cognitive Transitions. *Cerebral Cortex* , *28*(10), 3685–3696.

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714.

Spiers, H. J., & Maguire, E. A. (2007). Decoding human brain activity during real-world experiences. *Trends in Cognitive Sciences*, *11*(8), 356–365.

Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience*, *15*(3), 247–262.

Sporns, O., & Betzel, R. F. (2016). Modular Brain Networks. *Annual Review of Psychology*, *67*, 613–640.

Sreekumar, V., Nielson, D. M., Smith, T. A., Dennis, S. J., & Sederberg, P. B. (2018). The experience of vivid autobiographical reminiscence is supported by subjective content representations in the precuneus. Scientific Reports, 8(1), 14899.

St-Laurent, M., Abdi, H., & Buchsbaum, B. R. (2015). Distributed Patterns of Reactivation

Predict Vividness of Recollection. *Journal of Cognitive Neuroscience*, *27*(10), 2000–2018.

Szenkovits, G., Peelle, J. E., Norris, D., & Davis, M. H. (2012). Individual differences in premotor and motor recruitment during speech perception. *Neuropsychologia*, *50*(7), 1380–1392.

Tahedl, M., & Schwarzbach, J. V. (2020). An updated and extended atlas for corresponding brain activation during task and rest. In *bioRxiv* (p. 2020.04.01.020644). https://doi.org/10.1101/2020.04.01.020644

Takashima, A., Petersson, K. M., Rutters, F., Tendolkar, I., Jensen, O., Zwarts, M. J., McNaughton, B. L., & Fernández, G. (2006). Declarative memory consolidation in humans: a prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(3), 756–761.

Taya, F., de Souza, J., Thakor, N. V., & Bezerianos, A. (2016). Comparison method for community detection on brain networks from neuroimaging data. *Applied Network Science*, *1*(1), 8.

Todd, N., Moeller, S., Auerbach, E. J., Yacoub, E., Flandin, G., & Weiskopf, N. (2016). Evaluation of 2D multiband EPI imaging for high-resolution, whole-brain, task-based fMRI studies at 3T: Sensitivity and slice leakage artefacts. *NeuroImage*, *124*(Pt A), 32–42.

Tomasello, R., Garagnani, M., Wennekers, T., & Pulvermüller, F. (2017). Brain connections of words, perceptions and actions: A neurobiological model of spatio-temporal semantic activation in the human cortex. Neuropsychologia, 98, 111–129.

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language*, *162*, 60–71.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019).

The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, *51*(3), 1187–1204.

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*, 62.

van den Heuvel, M. P., & Sporns, O. (2011). Rich-club organization of the human connectome. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 31(44), 15775–15786.

van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. *Trends in Cognitive Sciences*, *17*(12), 683–696.

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*, 100600.

Vanderwal, T., Eilbott, J., Finn, E. S., Craddock, R. C., Turnbull, A., & Castellanos, F. X. (2017). Individual differences in functional connectivity during naturalistic viewing conditions. *NeuroImage*, *157*, 521–530.

Vanderwal, T., Finn, E., Glerean, E., & Hasson, U. (2021). Naturalistic Imaging: The use of ecologically valid conditions to study brain function. *Neuroimage*. https://www.sciencedirect.com/journal/neuroimage/special-issue/10S14SQ48ND

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., & Castellanos, F. X. (2015). Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage*, *122*, 222–232.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn Human Connectome Project: an overview. *NeuroImage*, *80*, 62–79.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: a

new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* , *67*(6), 1176–1190.

Van Horn, J. D., Grafton, S. T., & Miller, M. B. (2008). Individual Variability in Brain Activity: A Nuisance or an Opportunity? *Brain Imaging and Behavior*, *2*(4), 327–334.

Van Lancker, D., & Cummings, J. L. (1999). Expletives: neurolinguistic and neurobehavioral perspectives on swearing. *Brain Research. Brain Research Reviews*, *31*(1), 83–104.

Van Lancker Sidtis, D., Choi, J., Alken, A., & Sidtis, J. J. (2015). Formulaic Language in Parkinson's Disease and Alzheimer's Disease: Complementary Effects of Subcortical and Cortical Dysfunction. *Journal of Speech, Language, and Hearing Research: JSLHR*, *58*(5), 1493–1507.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*(Pt A), 68–77.

Varoquaux, G., & Poldrack, R. A. (2018). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, *55*, 1–6.

Vatansever, D., Menon, D. K., Manktelow, A. E., Sahakian, B. J., & Stamatakis, E. A. (2015). Default Mode Dynamics for Global Functional Integration. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(46), 15254–15262.

Vazquez, A. L., & Noll, D. C. (1998). Nonlinear aspects of the BOLD response in functional MRI. *NeuroImage*, *7*(2), 108–118.

Verma, T., Russmann, F., Araújo, N. A. M., Nagler, J., & Herrmann, H. J. (2016). Emergence of core-peripheries in networks. *Nature Communications*, *7*, 10441.

Vu, H., Kim, H.-C., Jung, M., & Lee, J.-H. (2020). fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations. *NeuroImage*, *223*, 117328.

Wang, J., Conder, J. A., Blitzer, D. N., & Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. In Human Brain Mapping (Vol. 31, Issue 10, pp. 1459–1468). https://doi.org/10.1002/hbm.20950

Wang, J., Ren, Y., Hu, X., Nguyen, V. T., Guo, L., Han, J., & Guo, C. C. (2017). Test-retest reliability of functional connectivity networks during naturalistic fMRI paradigms. *Human Brain Mapping*, *38*(4), 2226–2241.

Wardlaw, J. M., Brindle, W., Casado, A. M., Shuler, K., Henderson, M., Thomas, B., Macfarlane, J., Muñoz Maniega, S., Lymer, K., Morris, Z., Pernet, C., Nailon, W., Ahearn, T., Mumuni, A. N., Mugruza, C., McLean, J., Chakirova, G., Tao, Y. T., Simpson, J., … SINAPSE Collaborative Group. (2012). A systematic review of the utility of 1.5 versus 3 Tesla magnetic resonance brain imaging in clinical practice and research. *European Radiology*, *22*(11), 2295–2303.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex* , *26*(6), 2506–2516.

Wittenberg, E., Momma, S., & Kaiser, E. (2021). Demonstratives as bundlers of conceptual structure. *Glossa a Journal of General Linguistics*, *6*(1). https://doi.org/10.5334/gjgl.917

Wu, W., Morales, M., Patel, T., Pickering, M. J., & Hoffman, P. (2022). Modulation of brain activity by psycholinguistic information during naturalistic speech comprehension and production. In *bioRxiv* (p. 2022.03.07.483336). https://doi.org/10.1101/2022.03.07.483336

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage*, *25*(3), 1002–1015.

Xu, T., Opitz, A., Craddock, R. C., Wright, M. J., Zuo, X.-N., & Milham, M. P. (2016).

Assessing Variations in Areal Organization for the Intrinsic Brain: From Fingerprints to Reliability. *Cerebral Cortex* , *26*(11), 4192–4211.

Yaffe, R. B., Kerr, M. S. D., Damera, S., Sarma, S. V., Inati, S. K., & Zaghloul, K. A. (2014). Reinstatement of distributed cortical oscillations occurs with precise spatiotemporal dynamics during successful memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(52), 18727–18732.

Yan, B., & Luo, J. (2019). Multicores-periphery structure in networks. *Network Science*, *7*(1), 70–87.

Yang, J., & Leskovec, J. (2014). Overlapping Communities Explain Core–Periphery Organization of Networks. *Proceedings of the IEEE*, *102*(12), 1892–1902.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). *NeuroSynth: a new platform for large-scale automated synthesis of human functional neuroimaging data*. 4th INCF Congress of Neuroinformatics, Boston. https://www.frontiersin.org/10.3389/conf.fninf.2011.08.00058/event_abstract

Yarkoni, T., Speer, N. K., & Zacks, J. M. (2008). Neural substrates of narrative comprehension and memory. *NeuroImage*, *41*(4), 1408–1425.

Yeşilyurt, B., Uğurbil, K., & Uludağ, K. (2008). Dynamics and nonlinearities of the BOLD response at very short stimulus durations. *Magnetic Resonance Imaging*, *26*(7), 853–862.

Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L., & Breakspear, M. (2014). Time-resolved resting-state brain networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(28), 10341–10346.

Zang, Y., Jiang, T., Lu, Y., He, Y., & Tian, L. (2004). Regional homogeneity approach to fMRI data analysis. *NeuroImage*, *22*(1), 394–400.

Zhan, L., Jenkins, L. M., Wolfson, O. E., GadElkarim, J. J., Nocito, K., Thompson, P. M.,

Ajilore, O. A., Chung, M. K., & Leow, A. D. (2017). The significance of negative correlations in brain connectivity. *The Journal of Comparative Neurology*, *525*(15), 3251–3265.

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, *23*(4), 1028–1034.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.

Zwaan, R., Langston, M., & Graesser, A. (09/1995). The Construction of Situation Models in Narrative Comprehension. *Psychological Science*, *6*(5). https://journals.sagepub.com/doi/pdf/10.1111/j.1467-9280.1995.tb00513.x