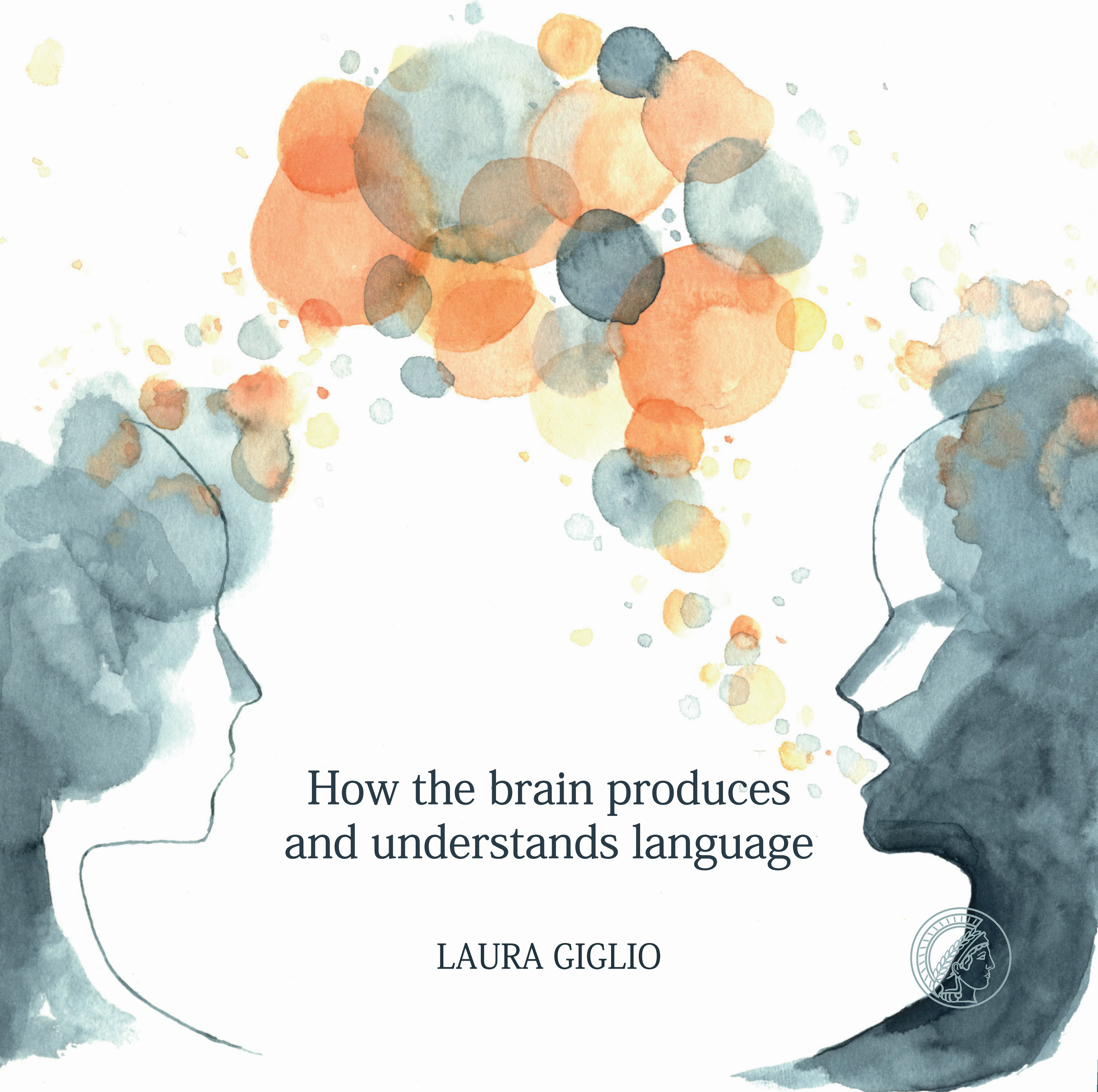
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

Speaking in the Brain



How the brain produces and understands language

LAURA GIGLIO

Speaking in the Brain:

How the brain produces and understands language

Funding body
This research was funded by the Max Planck Society for the Advancement of Science (www.mpg.de/en).

International Max Planck Research School (IMPRS) for Language Sciences
The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs

The MPI series in Psycholinguistics
Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

ISBN: 978-94-92910-48-6
Cover artwork by Michele D’Asaro
Cover design by Ludy Cilissen
Printed and bound by Ipskamp Drukkers, Enschede

The research reported in this thesis was conducted at the Max Planck Institute for Psycholinguistics, in Nijmegen, the Netherlands.

Speaking in the Brain:
How the brain produces and understands language

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 1 juni 2023
om 10.30 uur precies

door
Laura Giglio
geboren op 10 december 1993
te Ivrea, Italië

Promotor:

Prof. dr. Peter Hagoort

Copromotor:

Dr. Markus Ostarek

Manuscriptcommissie:

Prof. dr. Antje Meyer

Prof. dr. Peter Indefrey (Heinrich Heine Universität Düsseldorf, Duitsland)

Dr. Christophe Pallier (Inserm, Frankrijk)

Speaking in the Brain:
How the brain produces and understands language

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Doctorate Board
to be defended in public on

Thursday, June 1, 2023
at 10:30 am

by
Laura Giglio
born on December 10, 1993
in Ivrea, Italy

PhD supervisor:

Prof. dr. Peter Hagoort

PhD co-supervisor:

Dr. Markus Ostarek

Manuscript Committee:

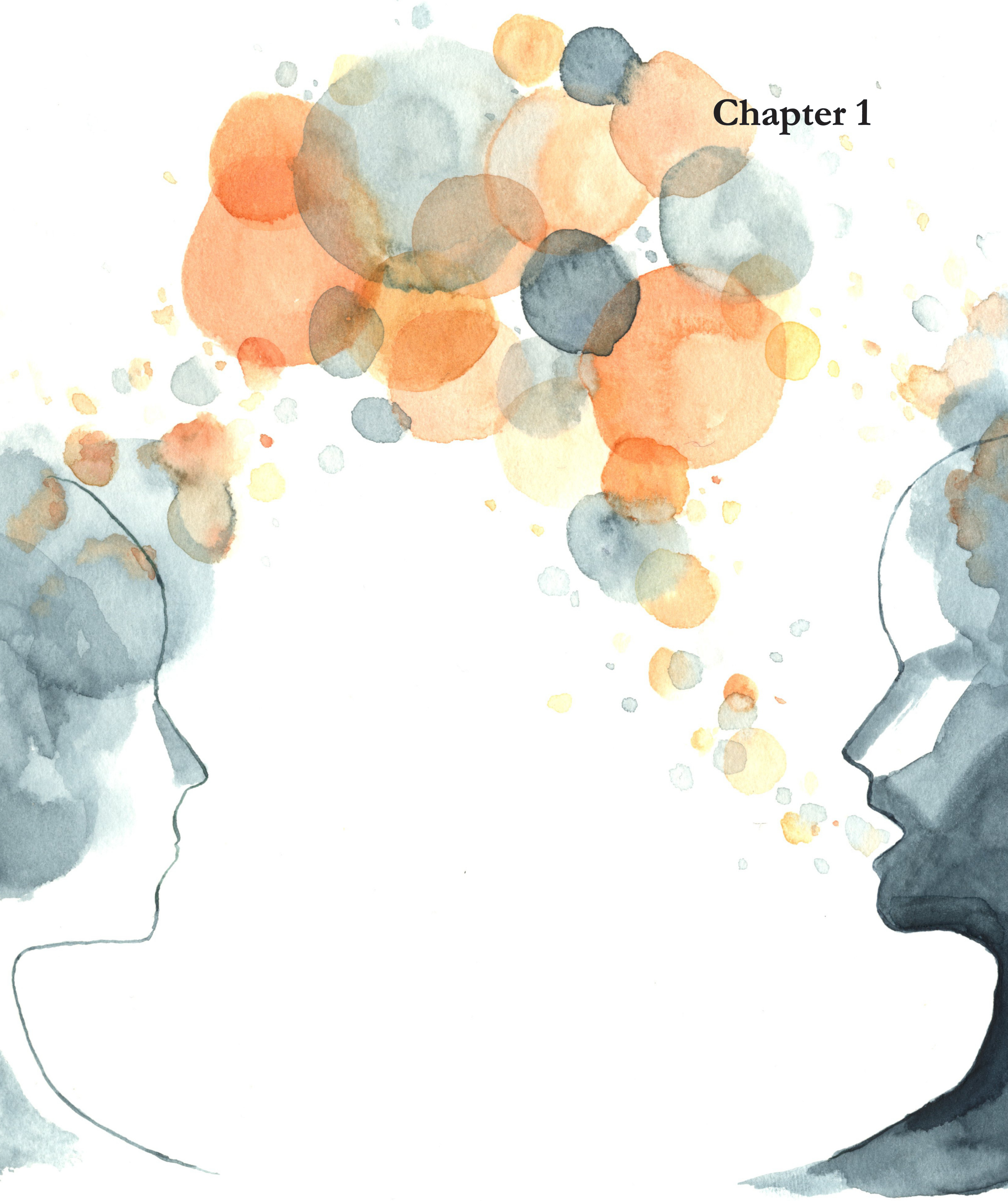
Prof. dr. Antje Meyer

Prof. dr. Peter Indefrey (Heinrich Heine University Düsseldorf, Germany)

Dr. Christophe Pallier (Inserm, France)

Contents

Chapter 1



1 | General introduction

> "All he asked for was a bit of silence, a bit of shush so he could concentrate. He wanted it to be perfectly quiet and still, like the inside of an empty confessional box or the moment in the brain between thought and speech."
> *White Teeth*, Zadie Smith (2001)

When I first read this sentence, I was struck by the fact that it was immediately obvious to me what the writer meant here, while knowing it could only be true by oxymoron – knowing "the moment in the brain between thought and speech" to be a moment where so much happens at the same time that it becomes far-fetched to compare it to the silence of a confessional box. This quote highlights that when we speak we do not notice the computations that are taking place between the intention to speak and the moment of articulation, because they are not accessible to conscious experience (Bock & Levelt, 1994). Perhaps it is precisely because there is so much to compute between thought and speech that this moment is perceived as still: any disruption to the stillness would most likely lead to speech errors. What this thesis sets out to explore is precisely the moment in the brain between thinking and speaking, where a thought is encoded via several processes as a sequence of words that can be articulated. In the following pages, I explore how the brain reflects the computational havoc that must be going on when we communicate.

Speaking is one of the defining acts of the human being, with which we express our ideas, thoughts, emotions, feelings. Speaking is hard: an idea has to be translated into a linear sequence of sounds that has to be understandable for a listener. This requires selecting the correct words, so that it is clear what we refer to, and organizing them in a correct grammatical structure. The other side of the coin is listening, where sounds are translated into messages. In this thesis, I attempt to bridge the gap between our understanding of the "moment in the brain between listening and thinking", and the "moment in the brain between thinking and speaking". Levelt wrote in 1989 that "Language production is the stepchild of psycholinguistics" (Levelt, 1989, p.xiii), because of the relatively fewer studies on production than comprehension in the field of psycholinguistics at the time. While this statement does not reflect the state of psycholinguistics in the 2020s, it is still true if applied to neurobiology of language, which is almost entirely concerned with the study of language comprehension. There are two main reasons for the limited neuroimaging literature on production. First, it is hard to control what a speaker will say (Bock, 1996). What a speaker says and how is one of the defining characteristics of speaking. For the purpose of careful experimentation, however, the speech output needs to be similar enough across participants and trials to reduce confounds. Nevertheless, more recently a few paradigms have been developed that can solve this issue in neuroimaging studies. The second issue in the study of production is the movement elicited by speaking. When articulating, we need to control many muscles in the face and the larynx, eliciting motor artefacts in electrophysiological studies. Speaking also induces movement of the head, causing motion artefacts with functional magnetic resonance imaging (fMRI). The problem of head movement, however, is not impossible to overcome with appropriate preprocessing methods. Therefore, the goal of this thesis is to bring state-of-the-art neuroimaging methods into the realm of language production and thus broaden our understanding of the neural architecture for language. This approach has the potential to identify brain processes that are common to both production and comprehension, as well as to highlight where production and comprehension may diverge.

1.1 The moment in the brain...

The study of the brain infrastructure for language started in the nineteenth century with the association of linguistic impairments with brain lesions following a stroke. Paul Broca first associated a lesion in the left posterior frontal lobe with production deficits. Later, Carl Wernicke found that a lesion in the superior temporal lobe led to an impairment in the comprehension of speech. These findings eventually led to the development of the Wernicke-Lichtheim-Geschwind model, with Broca's area (or left inferior frontal gyrus, LIFG) involved in speaking, and Wernicke's area in listening (Levelt, 2013). Further studies into the linguistic consequences of stroke, tumour and neurodegenerative disorders throughout the twentieth century, together with the extensive study of the healthy brain in the last three decades with the emergence and expansion of neuroimaging methods (Box 1.1), painted a more dynamic and complex picture of the brain structure and function for language (for reviews on current perspectives on Broca's and Wernicke's areas, see Binder, 2017; Hagoort, 2005; Tremblay & Dick, 2016).

Neuroimaging methods like functional magnetic resonance imaging (fMRI) and magneto- and electroencephalography (M/EEG) allowed authors to ground specific components of language in a set of brain areas and the dynamics between them (e.g. Hagoort & Indefrey, 2014; Walenski, Europa, Caplan, & Thompson, 2019; Zaccarella, Schell, & Friederici, 2017).

Current models on the neurobiology of language all agree on the importance of temporal and inferior frontal areas for language processing. There are differences in the specific processes assumed to take place in the different areas and in the dynamics of processing and connectivity. Hagoort (2005, 2013) proposed the temporal lobe to be crucial for the retrieval of phonological, lexico-syntactic and semantic information, which is then unified in the inferior frontal gyrus (LIFG, including BA44, BA45 and BA47) in an interactive network. BA46 in the frontal lobe and the anterior cingulate cortex are proposed to be relevant for control processes needed during language. Friederici (2009; 2012; Friederici & Gierhan, 2013) proposed phonological processing to take place in the superior temporal gyrus, while lexical information and simple phrase-structure building are retrieved and processed in the anterior superior temporal gyrus. The posterior LIFG (BA44), instead, is engaged for complex structure building, such as the computation of long-distance dependency. The anterior LIFG (BA47) is proposed to be involved in semantic processing together with the middle temporal gyrus (MTG) and the anterior temporal lobe (ATL). More recently, the *pars opercularis* of the LIFG was suggested to be involved in syntactic processing in interaction with the posterior temporal lobe (PTL) via the arcuate fasciculus (Zaccarella, Meyer, Makuuchi, & Friederici, 2017; Zaccarella, Schell, & Friederici, 2017). Similarly, Tyler and Marslen-Wilson (2008) stress the importance of connectivity between fronto-temporal regions for combinatorial morphological, syntactic and semantic processing. They additionally highlight the importance of posterior temporal areas for syntactic processing, including inferior parietal areas such as angular gyrus and supramarginal gyrus. Bornkessel-Schlesewsky and Schlesewsky (2013) proposed two processing streams with different functions. A ventral stream along the temporal lobe projecting to the anterior LIFG is involved in the combination of linguistic input in a time-independent way to compose a conceptual meaning. A dorsal stream instead processes the input in an ordered way to compute structured sequences. The LIFG is proposed to integrate ventral and dorsal stream information and provide top-down feedback, thus coordinating linguistic processing with executive control functions. Pylkkänen (2020) recently proposed temporal windows of processing to reach combinatorial meaning, that rely on the ATL for semantic combinatory effects, the PTL for syntactic combinatory effects, the LIFG for processing long-distance dependencies, and the ventromedial prefrontal cortex for a later stage of semantic composition.

Most of these models of language processing in the brain focus on language comprehension, since they are almost exclusively based on evidence from neuroimaging studies of language comprehension. Studies of sentence production have identified a similar fronto-temporal network for syntactic and semantic processing (Blanco-Elorrieta, Kastner, Emmorey, & Pylkkänen, 2018; Collina, Seurinck, & Hartsuiker, 2014; den Ouden, Hoogduin, Stowe, & Bastiaanse, 2008; Golestani et al., 2006; Grande et al., 2012; Hu et al., 2022; Humphreys & Gennari, 2014; Indefrey et al., 2001; Indefrey, Hellwig, Herzog, Seitz, & Hagoort, 2004; Matchin & Wood, 2020; Menenti, Gierhan, Segaert, & Hagoort, 2011; Menenti, Segaert, & Hagoort, 2012; Pylkkänen, Bemis, & Blanco Elorrieta, 2014; Segaert, Kempen, Petersson, & Hagoort, 2013; Segaert, Menenti, Weber, Petersson, & Hagoort, 2012; Takashima, Konopka, Meyer, Hagoort, & Weber, 2020; Thothathiri, 2018). Matchin and Hickok (2020) proposed an integrated architecture for sentence comprehension as well as production, informed mostly by linguistic processing after lesions in aphasia. They argue that the left posterior middle temporal gyrus (LpMTG) is the seat of hierarchical structure building in both production and comprehension, while the LIFG is involved in the linearization of hierarchical structure in production specifically. In comprehension, the LIFG is proposed to be involved in syntactic working memory and top-down prediction of syntactic structure building.

Therefore, while most theories of the neurobiology of language agree on the importance of the temporal lobe and the inferior frontal gyrus for language processing, there is no final agreement on their specific roles. Overall, there is agreement that auditory input is processed in the superior temporal cortex, and organized in a conceptual representation along the temporal lobe (e.g. Bornkessel-Schlesewsky & Schlesewsky, 2013; Hickok & Poeppel, 2007; Lambon Ralph, Jefferies, Patterson, & Rogers, 2017). Lexical-syntactic information is processed in the posterior temporal lobe (in retrieval or structure building, e.g. Hagoort, 2013; Matchin & Hickok, 2020; Pallier, Devauchelle, & Dehaene, 2011; Snijders et al., 2009; Zaccarella, Schell, & Friederici, 2017). The role of the inferior frontal gyrus is more debated, but it is considered to be involved in higher-order processing of sequences according to most accounts, whether it be in unification, integration or control processes (e.g. Bornkessel-Schlesewsky & Schlesewsky, 2013; Hagoort, 2013; Zaccarella, Schell, & Friederici, 2017). In this thesis, I repeatedly probed this network to learn more about the function of these regions in sentence production and comprehension.

Box 1: Functional Magnetic Resonance Imaging (fMRI)

In this thesis, I gathered evidence on brain activity during speaking using fMRI, a neuroimaging method that estimates brain activity based on regional blood flow in the brain. The working assumption is that, when a brain region is involved in a process, it consumes metabolic resources and as a consequence needs more blood inflow. Blood flow fluctuations affect the magnetic signal based on oxygen presence, since oxygen is bound to haemoglobin, which leads to a measure called blood-oxygen-level-dependent (BOLD) signal. The BOLD signal measured with fMRI thus approximates the activity of brain regions during a task. The blood inflow is called the haemodynamic response and usually lags 6 seconds after an event. Therefore, the signal measured with fMRI is unable to distinguish events close in time, lacking temporal resolution, but has relatively good spatial resolution with voxel dimensions of 2-3 mm.

Univariate analysis (Chapters 2 and 4)

Most fMRI evidence is based on univariate contrasts. These are used in studies that measure the change in BOLD signal during a task, defining the effect of a task on a brain region. For example, the BOLD signal is compared during processing of sentences vs. word lists. Voxels that reliably show increased activity during sentences vs. word lists are thought to be involved in sentence processing. Univariate methods focus on activity fluctuations in each voxel separately, but make inferences across regions (or clusters) of voxels that behave in a similar way to account for the multiple comparisons problem (due to over one hundred thousand voxels usually sampled in whole-brain studies). Univariate analyses are thus suitable for studies with task comparisons, where conditions are thought to differ in effort. As a consequence, they rely on assumptions about the processes involved in each task.

Functional Connectivity (Chapter 3)

Functional connectivity analyses can be used to learn about the way a task modulates the connectivity between regions. A common way to estimate the connectivity between two regions is to determine to what extent the correlation between the BOLD activity in a seed region and a target region is modulated by different levels of the task. It should be noted that the functional connectivity analysis used here does not allow for inferences on the directionality of the connectivity.

box continues on next page...

Multi-voxel pattern analysis (MVPA) (Chapter 5)
MVPA methods have been used more recently to characterize brain activity in terms of its patterns of activity across voxels. The activity across voxels is thus expected to differ based on conditions, but not necessarily in a uniform way (increase vs. decrease). This allows for contrasting conditions that are not expected to differ in effort. For example, the response to the visual presentation of objects vs. faces can be compared by analysing the pattern of responses in relevant areas (Haxby et al., 2001). This type of analysis is suitable for ROI analyses, but also whole-brain analyses that take advantage of "searchlights" that focus on a few voxels at a time iteratively (Kriegeskorte, Goebel, & Bandettini, 2006).

Voxel-wise encoding models (Chapter 5)
Encoding models enlarge the hypothesis space in a different way. While MVPA focuses on multiple voxels at a time, but compares conditions or items along a single dimension, encoding models focus on each voxel separately, but characterize brain activity simultaneously for different dimensions. For example, a sentence can be characterized by several predictors on different aspects of its structure and meaning: length, word meaning, syntactic structure, predictability, etc. All these predictors are used to optimally estimate each voxel's activity. These models differ from univariate analyses in aiming to generalize to new data. Learnt parameters for each predictor are used to predict activity for held out materials, and predicted and observed activity are then compared. Similarly to univariate analyses, inferences can be drawn over regions or clusters of voxels with similar predictive behavior.

1.2 ...Between thinking and speaking

Many processes are involved in translating an intention to speak into a motor plan, which results in speech sounds that can be comprehended by a listener. The most complete account of the processes taking place during speaking goes back to Levelt's seminal book "Speaking: from intention to articulation" (1989). Since then, the extensive psycholinguistic research into these processes has been able to confirm and extend some of the original hypotheses. There is general agreement that speaking starts with an intention to convey a message to an addressee. Therefore, the first step is to ground the intention to communicate about an experience (event, situation) into a preverbal message that can be later converted into a linguistic signal. In Slobin's words, "'Thinking for speaking' involves picking those characteristics that (a) fit some conceptualization of the event, and (b) are readily encodable in the language" (1987, p.435). The message can be thought to fit within the boundaries of a conceptual structure that identifies the entities and their relations within an event (Jackendoff, 1992). A debate dating back to the nineteenth century highlights general disagreement on whether the speaker first identifies a whole idea for a sentence ("the sentence as a simultaneous cognition into a sequential structure", Wundt, 1900/1970) or a linking of individual concepts ("the sentence as the expression of the combination of several ideas", Paul, 1886/1970). In more recent terms, a hierarchical form of message generation (where the relations between concepts need to be encoded) is opposed to linear message generation (where concepts can be identified sequentially) (Bock & Ferreira, 2014).

The evidence on message generation is relatively limited, due to the difficulty to experimentally access prelinguistic conceptual representations. This is because conceptual representations are eventually encoded in a linguistic sequence, leaving open whether task manipulations affect the conceptual or the linguistic level (Konopka & Brown-Schmidt, 2014; Papafragou & Grigoroglou, 2019). Eye-tracking studies have been able to show that during picture description experiments there is an event apprehension phase lasting 300-400 ms, before the attention is moved to the first item that is mentioned in the sentence (Z. M. Griffin & Bock, 2000). Similarly, attention is moved to the entity in the scene that is most informative for action encoding before sequential sentence planning starts (Konopka, 2019). Therefore, the gist of the sentence is encoded before sentence production, providing evidence for a hierarchical model of sentence planning. In addition, entities in an event are bound to their role in the event (i.e. a thematic role) very quickly during scene processing (Hafri, Papafragou, & Trueswell, 2013). This rapid binding suggests that, even if concepts are selected before the event is specified, their role in the proposition can already be determined (e.g. agent, patient).

Once the message (or part of a message) is generated, grammatical encoding takes place. Theories of sentence production identify two steps to grammatical encoding based on patterns of speech errors (Bock & Levelt, 1994; Garrett, 1980; Levelt, 1989). First, during functional processing, the lexical items corresponding to the concepts identified in message generation are selected together with their syntactic information (part-of-speech). Lexical items are then assigned syntactic roles (e.g. subject, object). Until now, the elements are not ordered in linear structure. During positional processing, the elements are bound in the syntactic structure and fixed to a specific order (constituent assembly), to be then inflected (e.g. *-ed* addition for past tense). Next, the inflected linear sequences are phonologically encoded to derive a motor plan for a speech sequence that obeys stress and prosodic constraints (Levelt, 1989).

Each of the processes just described, i.e. message generation, functional and positional processing, are thought to happen sequentially. However, each process does not need to be completed for the whole sentence before the next process can start, due to the incrementality of sentence production. Therefore, once a chunk of the message is generated, it can already be grammatically encoded while the rest of the message is being generated (Guhe, 2007; Levelt, 1989). The extent of the incrementality during production is not fixed. Sentence production can proceed with different unfolding dynamics based on context and time pressure (for a review see Bock & Ferreira, 2014). During grammatical encoding, word-driven incrementality (also called linear incrementality) builds the structure of the sentence from individual words. It is opposed to structure-driven incrementality (or hierarchical incrementality) that instead focuses on the relations between items and builds a structural scaffold first. Both strategies are thought to be used during production depending on different constraints (e.g. codability of the event, word accessibility, Kuchinsky, Bock, & Irwin, 2011; van de Velde, Meyer, & Konopka, 2014). A larger planning scope, in line with hierarchically incremental planning, takes place when the internal argument of a verb (e.g. a patient) is mentioned before the verb (as in passive sentences) (Momma & Ferreira, 2019; Momma, Slevc, & Phillips, 2016). Therefore, sentence production appears to be a highly dynamic and flexible process that proceeds highly incrementally.

Novel neuroimaging evidence on the dynamics of sentence production will be presented in Chapters 2 and 4, while the generation of higher-level message representations will be considered in Chapter 5.

1.3 Linking production and comprehension

1.3.1 The nature of the debate

One fundamental issue that has not been resolved in the psychology of language refers to the extent to which the linguistic representations and processes in each modality[1] are shared. Language production and comprehension have been traditionally studied in isolation, and as a consequence, their processes are often dealt

---

[1] In this thesis, I refer to production and comprehension as language “modality”. The term “modality” should not be confused with other contexts in which it has been used before, such as in reference to input modality (auditory, visual) or to output modality (speech, sign).

with separately (Gambi & Pickering, 2017; Meyer, Huettig, & Levelt, 2016). Production and comprehension representations and processes used to be considered separate based on evidence for separate neural infrastructures and different developmental trajectories (e.g. Clark & Hecht, 1983; Grodzinsky, 2000; Tremblay & Dick, 2016). However, most models and discussions now converge on considering their representations (i.e. the linguistic material, the what of language) to be shared (Gambi & Pickering, 2017; Momma & Phillips, 2018; Pickering & Garrod, 2004, for a review). Models that actively attempted to link production and comprehension stress that production and comprehension processes are interwoven but still separate. For example, Pickering and Garrod (2013) call production processes the ones that map from higher to lower levels (e.g. semantics to syntax to phonology), and comprehension processes the ones mapping from lower to higher levels. They then propose that 'production' and 'comprehension' processes take place during both speaking and listening. During production, 'comprehension processes' are a form of self-monitoring at multiple linguistic levels. During comprehension, 'production processes' allow for active prediction. As a result, production and comprehension processes are used during both production and comprehension, but serve different purposes. Dell and Chang (2014) similarly propose that production abilities arise from learning to predict, meaning that prediction is production, but still separate from comprehension.

Zooming in on structure building, which is the focus on Chapters 2-4, there are arguments why there might be a single system for syntactic processing across modalities. Kempen (2000) argued that a sentence formulator and a syntactic parser have similar processing demands. The same processor can construct syntactic structures from semantic structure (in production) or from word strings recognized in the speech input (in comprehension). Behavioural evidence supported this claim by showing that syntactic structure building in production interferes with parsing in comprehension, which is argued to be possible only if they rely on a common processor (Kempen, Olsthoorn, & Sprenger, 2012). More recently, Momma and Phillips (2018) additionally argued that a single mechanism may be sufficient and suggest that the superficial differences that are noted between production and comprehension may reflect input differences. For example, one difference between parsing and generation relates to their different requirements. In parsing, the listener has to resolve structural ambiguity in the input (hypothesis management, Guhe, 2007). Instead, the speaker has to make structural decisions to encode the message. Therefore, although the context is different, in both parsing and generation the computational goal is to select a structural representation consistent with the input (Momma & Phillips, 2018).

1.3.2 The production-comprehension dichotomy within the brain infrastructure for language

Turning now to how neuroimaging evidence can inform this debate, some of the initial reasons for the separation of production and comprehension processes originated from studies of aphasia where lesions led to distinct linguistic impairments in production and comprehension. In particular, Grodzinsky (2000) argued for a functional separation of production and comprehension mechanisms by noting that a lesion in Broca's area led to distinct syntactic deficits in each modality. An influential series of fMRI studies was then undertaken to precisely address whether there were neural preconditions to posit shared representations for production and comprehension. These studies elegantly took advantage of repetition suppression to ask whether brain responses would adapt to the reuse of linguistic material across modalities (Grill-Spector, Henson, & Martin, 2006). Menenti et al. (2011) found that the repetition of lexical, syntactic and semantic information in each modality led to the adaptation of fronto-temporal brain areas in both production and comprehension (with a conjunction analysis). Instead, the auditory and motor cortex were engaged separately for comprehension and production. Further studies additionally showed that the repetition of syntactic structure led to similar repetition suppression effects within modality and across modalities, providing convincing evidence that the same brain areas are involved in processing syntactic structure in production and comprehension (Schoot, Menenti, Hagoort, & Segaert, 2014; Segaert et al., 2013, 2012). This series of studies showed that neuronal populations adapt to the reuse of structure across production and comprehension. Based on this evidence, sentence production and comprehension engage a similar network peaking in the inferior frontal gyrus and the posterior temporal lobe.

The network for production and comprehension was later questioned in two meta-analyses of both modalities (Indefrey, 2018; Walenski et al., 2019). Both meta-analyses were underpowered in production specifically (17 vs. 133 studies of production and comprehension respectively in Indefrey, 2018; 15 vs. 45 contrasts for production and comprehension in Walenski et al., 2019). Indefrey (2018) found that the *pars opercularis* of the LIFG was the only reliable area active in production, while both *pars opercularis* and *pars triangularis* and the PTL were active in sentence comprehension. Instead, Walenski et al. (2019) found a similar network in comprehension, but identified the PTL and the middle frontal gyrus in production (thus not the LIFG). These meta-analyses therefore indicate inconsistencies in the production findings, and differences in the production and comprehension networks for sentence processing.

In this doctoral work, I repeatedly probed this fronto-temporal functional network for sentence production and comprehension with different tasks and analysis techniques to further elucidate how it is regulated by production and comprehension. While I was not able to directly address whether the same neuronal populations were engaged during production and comprehension, as previously done with repetition suppression, I could uncover similarities and differences between modalities and discuss them in light of their different goals and requirements (Momma & Phillips, 2018). The aims of this work are two-fold. On the one hand, I aimed to better characterize the neural infrastructure for language by using both comprehension and production, with the goal to elucidate the network for language processing independent of modality. On the other hand, I addressed the question of the separability of production and comprehension resources, representations and processes by discussing the consequences of potential differences or similarities in their neural implementation. It should be noted that this approach only works by exclusion. Finding no evidence for modality differences does not prove their shared neural implementation, but it allows for an interpretation of shared representations. On the contrary, finding differences between production and comprehension elicits a discussion of how their respective requirements affect neural activity.

1.4 Outline of the thesis

In this thesis, I ran four neuroimaging studies of language processing in production and comprehension. I aimed to (i) further characterize the response of left fronto-temporal areas to language processing[2]; (ii) learn more about differences and similarities of the brain response to production and comprehension; (iii) characterize the dynamics of structure building; (iv) focus on both syntactic and conceptual representations.

In Chapter 2, I addressed the inconsistencies in the response to sentence production and comprehension highlighted by the meta-analyses (Indefrey, 2018;

---

[2] With the four empirical chapters I aimed to characterize the engagement of fronto-temporal areas (the ‘language-network’) in production and comprehension, but I did not mean to imply that these are the only brain regions involved in language processing.

Walenski et al., 2019), relative to the adaptation studies (Segaert et al., 2012). Visual probes were used to elicit the production of word sequences organized in different constituents. I then focused on the response of the whole-brain to the production and comprehension of these stimuli of increasing constituent size. I found a robust network that responds to sentence processing in both production and comprehension. Additionally, I characterized differences in the response of this network to production and comprehension, which were further explored in Chapters 3 and 4. In Chapter 3, I investigated how the connectivity between the LIFG and the temporal lobe was affected by constituent size and modality in the same experiment. I found increased fronto-temporal connectivity in relation to constituent size that was not modulated by modality.

To reduce task effects present when studying production experimentally, in Chapter 4 I investigated the brain response during spontaneous production and naturalistic comprehension. Syntactic processing was measured with incremental parsers that make different predictions about the timing of syntactic structure building. I additionally explored whether production-specific parsers are informative for the incrementality of structure building in production, and complemented these findings with an analysis of speech fluency during production.

In Chapter 5, I moved away from syntactic processing and focused on the representation of semantic structures in the brain. I investigated whether and how sentence features encoded during message generation are represented in the brain. The neural infrastructure that supports compositional processing in production was characterized using encoding models that predict brain activity during the production of individual sentences based on their thematic role structure. The results were then discussed in comparison to a corresponding study of relational structure in comprehension (Frankland & Greene, 2020b).

In Chapter 6, I summarised and integrated the results of Chapters 2-5 to review what we learnt about the separability of production and comprehension from studies on the brain infrastructure for language. I additionally summarised what these findings mean for the interpretability of the function of fronto-temporal brain regions for language processing. Finally, avenues for further research and outstanding questions are discussed.

Chapter 2

2 | Commonalities and Asymmetries in the Neurobiological Infrastructure for Language Production and Comprehension[1]

Abstract

The neurobiology of sentence production has been largely understudied compared to the neurobiology of sentence comprehension, due to difficulties with experimental control and motion-related artifacts in neuroimaging. We studied the neural response to constituents of increasing size and specifically focused on the similarities and differences in the production and comprehension of the same stimuli. Participants had to either produce or listen to stimuli in a gradient of constituent size based on a visual prompt. Larger constituent sizes engaged the left inferior frontal gyrus (LIFG) and middle temporal gyrus (LMTG) extending to inferior parietal areas in both production and comprehension, confirming that the neural resources for syntactic encoding and decoding are largely overlapping. An ROI analysis in LIFG and LMTG also showed that production elicited larger responses to constituent size than comprehension and that the LMTG was more engaged in comprehension than production, while the LIFG was more engaged in production than comprehension. Finally, increasing constituent size was characterized by later BOLD peaks in comprehension but earlier peaks in production. These results show that syntactic encoding and parsing engage overlapping areas, but there are asymmetries in the engagement of the language network due to the specific requirements of production and comprehension.

---

[1] Adapted from Giglio, L., Ostarek, M., Weber, K., & Hagoort, P. (2022). Commonalities and Asymmetries in the Neurobiological Infrastructure for Language Production and Comprehension. *Cerebral Cortex, 32*(7), 1405–1418. `https://doi.org/10.1093/cercor/bhab287`

2.1 Introduction

Since the association of lesions in the left inferior frontal gyrus (LIFG) and aphasia in the nineteenth century, scientists have tried to understand the relationship between the language faculty and the brain. Early reports called the LIFG or Broca's area a "speech movement centre" and the left superior temporal gyrus (LSTG) or Wernicke's area a "sensory speech centre" (from Wernicke, 1892, as described in Levelt (2013)). Since then, the field moved forward from a production-comprehension dissociation to the understanding that both areas are critical for language more generally, and that they do not subserve strictly segregated receptive or productive linguistic functions (Tremblay & Dick, 2016). A wealth of neuroimaging studies and lesion-symptom mapping studies advanced the characterization of brain function greatly, which resulted in a general understanding of the contributions of core regions in the language network (e.g. Friederici & Gierhan, 2013; Hagoort & Indefrey, 2014; Price, 2012; Wilson, 2017).

The LIFG (i.e., Broca's area and adjacent cortex) has been implicated in sentence-level processes in several neuroimaging studies. These include sentence vs. word list comprehension (Fedorenko, Nieto-Castañon, & Kanwisher, 2012; Matchin, Hammerly, & Lau, 2017; Snijders et al., 2009; Zaccarella, Schell, & Friederici, 2017), phrase structure building (Chang, Dehaene, Wu, Kuo, & Pallier, 2020; Pallier et al., 2011; Schell, Zaccarella, & Friederici, 2017; Zaccarella, Meyer, et al., 2017), compositional processes in naturalistic language comprehension (Bhattasali et al., 2019; Henderson, Choi, Lowder, & Ferreira, 2016), and processing of noncanonical sentence structure (Bornkessel-Schlesewsky, Schlesewsky, & Cramon, 2009; Europa, Gitelman, Kiran, & Thompson, 2019; Hirotani, Makuuchi, Rüschemeyer, & Friederici, 2011; Mack, Meltzer-Asscher, Barbieri, & Thompson, 2013; Santi & Grodzinsky, 2010). In different neurobiological models of language processing the LIFG was thus proposed to have a role in combinatorial (Unification) processes in multiple domains of language and cognition (Hagoort, 2005, 2013, 2019); in processing complex syntax (Friederici, 2012); or in sentence processing due to its role in working memory (Matchin, 2018; Rogalsky, Matchin, & Hickok, 2008).

Within the temporal lobe, posterior regions have been implicated in several aspects of comprehension, from auditory to phonological and morphological processing along the superior temporal gyrus and sulcus (Friederici, 2012; Hickok & Poeppel, 2007; Lee et al., 2018). In addition, the posterior middle temporal gyrus (LpMTG) has been associated with the retrieval of lexical-syntactic frames

('Memory' processes, Hagoort, 2005, 2013) and syntactic processes (Flick et al., 2018; Matchin & Hickok, 2020). The anterior temporal lobe (ATL) has been associated with conceptual operations (e.g. Bemis & Pylkkänen, 2013; Boylan, Trueswell, & Thompson-Schill, 2017), also based on findings of ATL atrophy leading to semantic dementia (Lambon Ralph et al., 2017; Mesulam et al., 2014; Wilson et al., 2013).

All studies mentioned above, however, are based on linguistic processes in *comprehension*. The involvement of the main nodes of the language network in sentence *production* is less clear. This is mainly for two reasons: (i) the challenge of achieving good experimental control in sentence production studies, which also limited psycholinguistic studies of production processes (Bock, 1996), and (ii) the obstacle of motion artefacts in neuroimaging as a consequence of movement during speech, which is however not impossible to overcome with state-of-the-art neuroimaging techniques (Willems & Gerven, 2018). These methodological difficulties have led to far fewer studies on the characterization of brain involvement in production than comprehension. As a consequence, the few meta-analyses that attempted to characterize the language network in the two modalities were severely underpowered in production (Indefrey, 2018; Walenski et al., 2019).

The network obtained by meta-analyses of sentence production studies does not fully or consistently overlap with the sentence comprehension network discussed above. Sentence production studies found activity within a left-lateralised fronto-temporal network but not consistently across studies (Collina et al., 2014; den Ouden et al., 2008; Golestani et al., 2006; Grande et al., 2012; Haller, Radue, Erb, Grodd, & Kircher, 2005; Humphreys & Gennari, 2014; Indefrey et al., 2001, 2004; Kircher, Oh, Brammer, & McGuire, 2005; Matchin & Hickok, 2016; Pylkkänen et al., 2014; Takashima et al., 2020; Thothathiri, 2018; Thothathiri & Rattinger, 2015). A recent meta-analysis on some of these studies found left middle frontal gyrus, LpMTG and lateral occipital cortex to be reliably involved in sentence production, but did not find evidence for LIFG involvement (Walenski et al., 2019). Another meta-analysis, instead, found the LIFG to be the only area reliably active in sentence production and for syntactic contrasts across some of those studies, thus lacking temporal lobe involvement (Indefrey, 2018). There is thus disagreement on whether inferior frontal areas or temporal areas are reliably engaged during sentence production, while they are both reliably found in sentence comprehension. The contradictory results of these meta-analyses show that more work is needed to robustly determine the neural correlates of sentence production. Interestingly, these results suggest there are some discrepancies in the networks engaged by linguistic processes in production and comprehension that raise the question whether the same neural resources are used in production and comprehension.

An important line of work addressed the question of a shared or distinct neural infrastructure between sentence production and comprehension. This question is relevant in the context of a long-standing debate in psycholinguistics. There are different views on if, and to what degree, production and comprehension share phonological, lexical, syntactic and semantic representations (Gambi & Pickering, 2017; Meyer et al., 2016; Momma & Phillips, 2018; Phillips, 2013). Support for distinct representations comes from the production/comprehension asymmetries in language acquisition. Comprehension is seen to precede production in many linguistic domains, with some exceptions (Clark & Hecht, 1983; Hendriks & Koster, 2010). This dissociation in acquisition is more easily accounted for by models that keep production and comprehension representations separate. Accounts of syntactic deficits in agrammatic patients also suggest that different processes are compromised in comprehension and production (i.e. tree pruning in production vs. trace deletion in comprehension) (Grodzinsky, 2000). However, there are also views supporting a single processing mechanism that argue that the differences between modalities may be superficial and may instead reflect input differences (Momma & Phillips, 2018). Behavioural evidence has shown that syntactic representations are shared between production and comprehension (Kempen et al., 2012). Also, repetition suppression (Grill-Spector et al., 2006) was used in fMRI to understand which areas adapt to the repetition of linguistic material and whether the adaptation occurs only within one language modality or also across modalities (i.e. from comprehension to production and vice versa). The LIFG, precentral gyrus, LMTG and inferior parietal lobule were found to adapt to syntactic and lexical repetition across sentence production and comprehension (Menenti et al., 2011, 2012; Segaert et al., 2013, 2012), suggesting that production and comprehension share neural resources. This evidence for shared resources in production and comprehension is, however, challenged by the inconsistent and partly contradictory neuroimaging results in production.

In the current study, therefore, we examined the sentence production network in a high-powered study with the aim to further clarify the brain organization of sentence production. To address this issue, our study investigated language production in analogy to a seminal study on constituent structure building in comprehension (Pallier et al., 2011). Constituents are the syntactic building blocks of sentences. By using a constituent size manipulation, we could focus on the processes that allow for encoding of increasingly larger structures, while keeping lexico-semantic, phonological and articulatory processes constant between conditions. Following Pallier et al. (2011), we expected neural activity to gradually increase with the addition of each new node to the constituent structure of the stimuli. We used visual prompts to elicit the production of utterances that had three levels of constituent structure which differed in complexity. The simplest one consisted of one- and two-word sequences; the intermediate condition consisted of intransitive sentences; the version with the most complex structure had participants produce a sentence with a complementizer phrase embedded in the main clause. In their comprehension study, Pallier et al. (2011) showed that LIFG and the left anterior and posterior temporal lobe were responsive to constituent size. Based on previous comprehension evidence we therefore expected to find a gradual involvement of at least LIFG and LMTG with increasing constituent size.

Additionally, we directly compared the sentence production and comprehension networks with the aim to further clarify to what extent they overlap. Few studies so far used both production and comprehension in the same experiment, including a direct comparison between modalities. In particular, it is still unclear whether sentence production and comprehension rely on core regions of the language network to the same or to a different extent. Humphreys and Gennari (2014) found that frontal and subcortical regions were more engaged in production, while the LpMTG was more engaged in comprehension. Indefrey et al. (2004) found the LIFG to be responsive to syntactic processing in production but not in comprehension. Matchin and Wood (2020) instead found similar activity in LIFG for syntactic production and comprehension, and larger activity in LMTG for syntactic comprehension than production. We therefore selected LIFG and LMTG as regions of interest to better characterize their involvement in sentence processing across modalities. In short, there is no clear answer to the question whether production and comprehension recruit frontal and temporal regions similarly or differently. In this study we attempted to answer this question.

2.2 Materials and Methods

Participants

Forty-six right-handed native Dutch participants (28 females, mean = 23.8 years, range 19-35 years) participated in the experiment in return for monetary compensation after giving written informed consent. The study was approved by the ethical committee for Region Arnhem-Nijmegen. Participants had no history of neurological or language-related disorders, and reported having normal or corrected-to-normal vision and hearing. Six participants were excluded for the following reasons: technical problems during preprocessing of the MRI data (n = 1); failing to complete the experiment (n = 2); too many motion artefacts (n = 3). Forty participants were included in the analyses. This number was based on an a priori power calculation for the detection of an effect for the production of passive vs. active sentences in the LIFG and LMTG in a previous study (Segaert et al., 2012), using fMRIpower (Mumford & Nichols, 2008). Even though the specific manipulation was different, it allowed us to estimate the number of participants needed for the detection of a syntactic effect in production in the two regions of interest.

Materials

In our study, we had three levels of constituent structure (see Table 1). The condition with the smallest constituent size (C1) consisted of two verbs and two noun phrases leading to four constituents with one (content) word (C1: “klappen, slapen, de jongen, het meisje”, “to clap, to sleep, the boy, the girl”). The intermediate condition (C2) involved the combination of a verb and a noun phrase leading to two constituents with two content words forming intransitive sentences (C2: “de jongen slaapt, het meisje praat”, “the boy sleeps, the girl talks[2]”). The most complex sentence condition (C4) consisted of the combination of the four content words into a complementizer phrase embedded in the main clause (C4: “de jongen hoort dat het meisje klapt”, “the boy hears that the girl claps”). Critically, the conditions were almost identical in the total number of words to be produced, but they differed in constituent structure. The embedded sentence condition included the additional word “dat” (*that*), which in Dutch

---

[2] The one provided is a literal translation of the Dutch sentence. A more natural translation would be “the boy is sleeping, the girl is talking”, since the present tense in Dutch is also used for continuous events.

is obligatory in complementizer sentences[3]. We did not expect function words to affect sentence planning but they might involve articulation-related processes (Ferreira, 1991). An additional filler condition was added to avoid too many verb repetitions. Filler sentences consisted of a sentence with one transitive verb ("de man helpt de vrouw", "the man helps the woman").

Table 2.1: Example sentences used for each condition.

Condition	Stimuli (in Dutch)	English translation
C1	klappen, slapen, de jongen, het meisje	clap, sleep, the boy, the girl
C2	de jongen slaapt, het meisje praat	the boy sleeps, the girl talks
C4	de jongen hoort dat het meisje klapt	the boy hears that the girl claps
Filler	de man helpt de vrouw	the man helps the woman

To induce the production of the sentences in the different conditions, participants were shown pictures with written verbs (see Fig. 2.1). Crucially, the conditions differed in the configuration of boxes around the verbs and the pictures of human figures. The boxes instructed the participants about the production output that was expected. In condition C1, there were four boxes, one around each item, signalling that the production of four separate items was expected. In this condition the actors and verbs should not be combined to form a sentence. In condition C2, there were two boxes, each around a verb and an actor, indicating that two separate sentences had to be produced. In condition C4, there was one box around all items on the screen, indicating that one single sentence was expected with the first verb heading an embedded clause formed by the second noun and verb. For filler sentences, there was only one box around all items on the screen. In this case there was only one verb, indicating that a transitive sentence was expected. Participants had no problems understanding the task and producing the correct output. By eliciting sentence production in this way, we could minimize the visual differences between conditions: pictures or videos would lead to very large differences in the visualization of word sequences vs. complementizer phrases. This type of speech paradigm elicitation is not unusual in the neuroimaging sentence production literature (e.g. Matchin & Hickok, 2016; Takashima et al., 2020).

The verbs were always presented in their root form, so that the production of the syntactically correct inflections was required in all conditions. In the C1 condition, the verb had to be produced in its infinitival form (generally, by addition

---

[3] The example sentence provided "the boy hears that the girl claps" would also work with a bare infinitival complement construction ("the boy hears the girl clap", "de jongen hoort het meisje klappen"), but many other verbs used in condition C4 do not support this construction (e.g. "klagen", to complain, "dromen", to dream). The word "dat" was thus required for all sentences in condition C4.

of “en”: *denk* to *denken*); in the other conditions, the verb had to be inflected in the third person singular of the present tense (generally, by addition of “t”: *denk* to *denkt*).

Since verbs allowing for a complementizer phrase (CP-verbs) and intransitive verbs (INT-verbs) are inherently different in their use and meaning, we selected a few verbs of each type that were repeated 8 times across the experiment. The verbs were matched in frequency (mean ± std: INT-verbs = 1.38 ± 0.88, CP-verbs = 1.46 ± 0.77, $t = 0.59$, $p = 0.56$) based on SUBTLEX-NL values (Keuleers, Brysbaert, & New, 2010), and concreteness (mean ± std: INT-verbs = 3.26 ± 0.67, CP-verbs = 3.21 ± 0.47, $t = 0.27$, $p = 0.79$) (Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014). Each condition consisted of 80 trials. In C4, we used 20 CP-verbs, repeated 4 times. The CP-verbs were always in first position to allow for the embedded sentence production; each of the 40 INT-verbs in C4 was repeated twice. In C2, we used the same 40 INT-verbs, each presented twice in first position and twice in second position. In C1, there was always one CP-verb and one INT-verb, with alternating first and second positions. Each CP-verb was repeated 4 times in this condition, and each INT-verb was repeated twice. The filler verbs consisted of 80 transitive verbs, each shown only once. We created 4 lists of stimuli that consisted of the same verb combinations for each condition, but for each list the verb was paired with a different picture. Across lists each verb combination was paired with each actor. The actors could be “the boy”, “the girl”, “the man”, “the woman”, with each presented 160 times in total.

In addition to the production condition, we included a comprehension condition that included half of the materials used for production but from a different list (hence with different actor-verb pairings). In the comprehension condition, each verb was repeated only 4 times in total, with 40 trials per condition. Instead of producing the sentences, participants had to listen to recorded stimuli, which started 1 sec after picture onset and lasted a maximum of 4 seconds (mean duration (in seconds): C1 = 3.14, C2 = 2.46, C4 = 2.46, Fillers = 1.79). The absence of an explicit task during the comprehension runs kept the production and comprehension runs as similar as possible without the introduction of effects unrelated to constituent size.

Experimental Procedure

The experiment started with a behavioural practice session to familiarise participants with the task. They read instructions for each condition and had to practise producing the sentences. The experimenter gave feedback to make sure

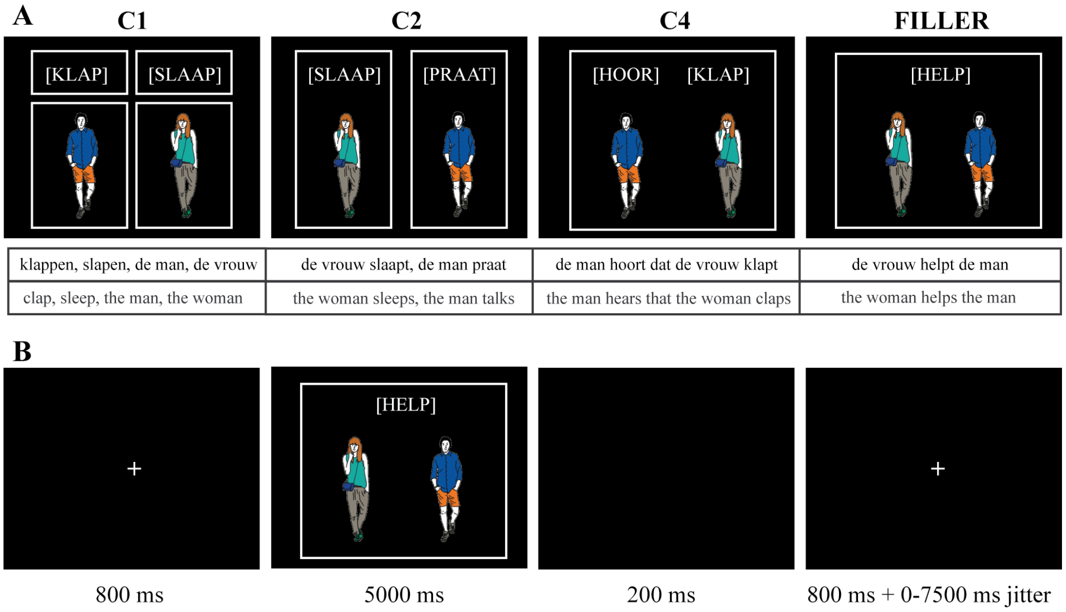


Figure 2.1: Stimulus presentation. **A**: Example of the screen that participants would see for each condition (identical in production and comprehension) with the corresponding expected output. The boxes clarified the type of output that was required. **B**: Screen sequence for each trial. The length of the fixation cross presentation was based on jittering optimized for contrast detection. In comprehension, during picture and verb presentation, a sound recording of the sentence started after 1000 ms.

that the participant understood the task correctly. After the practice session was concluded, the fMRI experiment started. The production lists were divided into 8 runs of 40 trials, each including 10 trials per condition, with as few verb repetitions as possible (per block 5-6 verbs were repeated once out of the 60 verbs presented (excluding fillers)). The comprehension lists were divided into 4 runs of 40 trials each. Production and comprehension runs alternated with two production runs always followed by a comprehension run. There were 12 acquisition runs in total. Each run lasted about 5 minutes. A fixation cross was presented for at least 800 ms before the picture screen was presented (see Fig. 2.1). Participants had 5 seconds to produce the answer. This was followed by a blank screen for 200 ms. We jittered the onset of trials by 0 – 7500 ms (mean 1500 ms), by varying the length of presentation of the fixation cross. The order of conditions and length of jitter was based on design optimization for contrast detection, made with optseq2 (Dale, 1999).

fMRI acquisition

MR data were acquired in a 3T MAGNETOM PrismaFit MR scanner (Siemens AG, Healthcare Sector, Erlangen, Germany) using a 32-channel head coil. The MRI protocol included a T1-weighted MRI scan for anatomical reference and several fMRI scans. The T1-weighted scan was acquired in the sagittal orientation using a 3D MPRAGE sequence with the following parameters: repetition time (TR)/inversion time (TI) 2300/1100 ms, echo time (TE) 3 ms, 8° flip angle, field of view (FOV) 256 mm × 216 mm × 176 mm and a 1 mm isotropic resolution. Parallel imaging (iPAT = 2) was used to accelerate the acquisition resulting in an acquisition time of 5 min and 21 sec. Whole-brain functional images were acquired using a multi-band (accelerator factor of 3) multi-echo T2*-weighted sequence with the following parameters: TR 1500 ms, TEs 13.4/34.8/56.2, flip angle 75°, FOV 84 mm x 84 mm x 64 mm, voxel size 2.5 mm isotropic. Fieldmap images were also acquired to correct for distortions. We acquired 12 fMRI runs per participant.

Data Analysis

Behavioural

Speech output in the production fMRI runs was analysed for accuracy and response times. A Dutch native speaker rated the speech for accuracy. Speech was considered correct when the correct actors and determiners were used, the verb was inflected in the correct way, and the correct sentence structure was used. Self-corrections and word repetitions during hesitations were considered errors. Speech onset and offset times were coded using Praat, after scanner noise removal. We analysed onset and durations with linear mixed-effects models (Bates, Mächler, Bolker, & Walker, 2015; Pinheiro & Bates, 2000) and accuracy data using mixed-effects logit models (Jaeger, 2008) with the *lme4* package (version 1.1-21, R version 3.6.2). We used the maximal effect structure that allowed for convergence (Barr, Levy, Scheepers, & Tily, 2013). For accuracy, the model contained the factor Condition (C1, C2, C4) and by-participant and by-item (specifically, verbs) random intercepts. For onset and duration analysis, the model contained the factor Condition and by-participant random slopes for Condition and by-item random intercepts, with log-transformed onset and duration times.

fMRI preprocessing

Preprocessing was performed using fMRIPprep 1.2.6-1 (Esteban, Blair, et al., 2018; Esteban, Markiewicz, et al., 2018). The T1-weighted (T1w) image was corrected for intensity non-uniformity and skull-stripped. Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1 Dale, Fischl, & Sereno, 1999). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov, Evans, McKinstry, Almli, & Collins, 2009) was performed through nonlinear registration using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid, white-matter and gray-matter was performed on the brain-extracted T1w using fast (FSL 5.0.9, Zhang, Brady, & Smith, 2001).

For each of the BOLD runs per subject, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A deformation field to correct for susceptibility distortions was estimated based on a field map that was co-registered to the BOLD reference, using a custom workflow of fMRIPrep. Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson, Bannister, Brady, & Smith, 2002). BOLD runs were slice-time corrected and resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. Multi-echo combination was performed by estimating a T2* map from the preprocessed BOLD by fitting to a monoexponential signal decay model with log-linear regression. For each voxel, the maximal number of echoes with reliable signal in that voxel were used to fit the model. The calculated T2* map was then used to optimally combine preprocessed BOLD across echoes following the method described in (Posse et al., 1999). Estimation of motion artifacts using independent component analysis (ICA-AROMA, Pruim et al., 2015) was performed on the preprocessed BOLD on MNI space time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). The AROMA noise-regressors were later used as confound regressors. The BOLD time-series were resampled to MNI152NLin2009cAsym standard space. Confounding time-series were calculated based on the preprocessed BOLD for framewise displacement (FD) and DVARS (following the definitions by Power et al., 2014). We excluded subjects that had FD values above 2.5 (these were also the subjects that showed highest mean FD and the largest number of volumes with FD values above 1). Additionally, a set of physiological regressors were extracted to allow for anatomical component-based noise correction (aCompCor, Behzadi, Restom, Liau, & Liu, 2007).

Motion-related correction

To prevent excessive motion artefacts due to speaking out loud, participants' heads were secured in a pillow and a tape was attached across their foreheads to provide them with feedback in case of movement, which was shown to reduce motion (Krause et al., 2019). In addition, subjects with FD values above the voxel size were excluded. ICA-AROMA was used to estimate components related to motion that were later added as nuisance regressors together with motion parameters, FD, DVARS and aCompCor in the first-level design matrix.

fMRI analysis

Whole-brain analysis

We used the non-denoised preprocessed BOLD images in MNI152NLin2009cAsym standard space for first-level single-subject analysis. We applied spatial smoothing with an isotropic Gaussian kernel of 4 mm FWHM in SPM12 in Matlab2019a. For the production runs, we computed a general linear model (GLM) in SPM12 with the following condition regressors: correct trials for each of the four conditions, all incorrect trials, temporal derivative, and parametric modulations of speech onset times. For the comprehension runs, the GLM was identical except for the absence of an incorrect trial regressor and parametric modulations. The onset of each trial was set as the picture onset time, and trial duration was set as time until speech offset, hence accounting for differences in duration between individual stimuli and conditions. In addition, we added confound regressors that were computed in fMRIPrep. We included regressors for DVARS, Framewise Displacement, 6 aCompCor parameters and 6 motion parameters. Finally, we added the AROMA noise components computed in fMRIPrep as additional nuisance regressors, to perform non-aggressive denoising. Contrast images for the main effect of constituent size (with weights

[-4 -1 5] based on constituent size of C1, C2 and C4, respectively), main effect of modality (production vs. comprehension) and interaction between constituent size and modality were computed for each participant. For the main effect of constituent size we selected a numerical linear contrast based on Pallier et al. (2011) that reflects activation with a linear increase according to the number of words integrated in a constituent: C1 = one content word per constituent, C2 = 2 content words, C4 = 4 content words. This led to a contrast with weights [-4 -1 5] after mean-centering. By contrasting the three conditions together, the results were less sensitive to other types of differences between individual conditions (e.g., the contrast C4 vs. C2 might be sensitive to verb argument structure differences). The contrast images were tested with a one-sample T-test at the group level following Henson (2015). We thresholded brain responses at the voxel-level at $p = 0.001$ *uncorrected*, and then used $p = 0.05$ Family-Wise Error corrected as the cluster threshold. We also ran a conjunction analysis to specifically look at the overlap between production and comprehension in the response to constituent size. To run the conjunction analysis, we created contrast images for the constituent size effect separately in production and comprehension, and then we entered them into a one-way ANOVA in SPM, with each as a separate cell. By defining separate contrasts for each, we could then run the conjunction analysis for the group-level contrast image of constituent structure in production and comprehension.

ROI analysis

We took functional regions of interest (ROIs) based on the keyword “syntactic” in Neurosynth (https://www.neurosynth.org/, 08/01/2020, Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). This allowed us to select voxels that are reported to be active in multiple studies related to a key search word, here “syntactic”. We downloaded the active voxels with a z-score threshold of 9. This revealed two clusters, one in left IFG and one in left anterior and posterior middle temporal lobe. We extracted mean beta values per participant in each of these ROIs per condition (C1, C2, C4, in production and comprehension) relative to baseline using MarsBar (Brett, Anton, Valabregue, & Poline, 2002) in SPM12. We then compared the beta weights in a mixed-effects model in R (version 3.6) using *lme4* (Bates et al., 2015), with constituent size (C1, C2, C4), modality (Production vs. Comprehension) and ROI (LIFG vs. LMTG) as factors. Deviation coding was used for factors modality and ROI, while a linear contrast with weights [-4 -1 5], as in the whole-brain analysis, was used for constituent size. We added by-participant random slopes for the interaction of ROI and modality and for the main effect of constituent size. We computed the contribution of factors using Type-III Wald tests in car (version 3.0-7, Fox et al., 2020) and pairwise comparisons for significant effects with the package *emmeans* (version 1.4.6, Lenth, Singmann, Love, Buerkner, & Herve, 2020).

Exploratory analysis: BOLD peak latency
As an additional exploratory analysis, we extracted BOLD time courses to determine whether the time to peak was influenced by region, modality and constituent size. To capture a delay in peak times, we used a finite impulse response (FIR) basis set as implemented in Marsbar in SPM12. This allowed us to get estimates of BOLD activity at each TR in the two ROIs for each participant. We then extracted BOLD peak times as the timepoint with highest amplitude between 1.5 and 9 s post stimulus onset for each participant. We ran a linear mixed-effect model with constituent size, modality and ROI as fixed effects, and by-participant random slopes for ROI. We used a linear contrast with weights [-4 -1 5] for constituent size, and deviation coding for modality and ROI. We computed the contribution of factors using *car* (version 3.0-7, Fox et al., 2020) and pairwise comparisons for significant effects with the package *emmeans* (version 1.4.6, Lenth et al., 2020).

2.3 Results

Behavioural results

Accuracy was generally high across participants and conditions (mean percentage correct: C1: 95.4, C2: 96.2, C4: 92.9, Fillers: 95.9; Fig. 2.2). There were slightly more errors in the C4 condition than in the C1 ($\beta = 0.55$, SE = 0.12, Z = 4.7, $p < 0.001$) and in the C2 conditions ($\beta = 0.71$, SE = 0.14, Z = 5.2, $p < 0.001$). Types of errors included using the wrong determiner (in Dutch, *het* is used with *meisje*-girl, and *de* with boy, man and woman; across all sentences for all participants, n = 117), the wrong actor (n = 170), a wrong verb or the correct verb in the wrong inflection/pronunciation (n = 105), the wrong condition (n = 119), not finishing within 5 seconds (n = 77), or other types of errors (n = 62). Unsurprisingly, onset times varied between conditions due to the characteristics of the conditions (mean onset times (in seconds): C1: 1.25, C2: 1.33, C4: 1.39, Fillers: 1.38, Fig. 2.3A). In particular, C1 elicited shorter reaction times than C2 ($\beta = 0.06$, SE = 0.01, $t = 5.7$, $p < 0.001$) and C4 ($\beta = 0.11$, SE = 0.01, $t = 8.2$, $p < 0.001$), as only the first verb had to be planned to initiate speech output. The other conditions, instead, required sentence planning, including subject (determiner and noun) as well as verb planning. The C4 condition elicited longer onset times than C2, too ($\beta = 0.05$, SE = 0.01, $t = 5.1$, $p < 0.001$). Similarly, duration times varied by condition (mean durations (in seconds): C1: 2.46, C2: 1.86, C4: 1.90, Fillers: 1.46; Fig. 2.3B). C1 production was characterized by the separate production of each lexical item, introducing pauses between words, and was thus characterized by the longest durations (vs. C2: $\beta = 0.28$, SE = 0.02, $t = 16.6$, $p < 0.001$; vs. C4: $\beta = 0.27$, SE = 0.02, $t = 15.4$, $p < 0.001$), while C2 and C4 did not differ in duration.

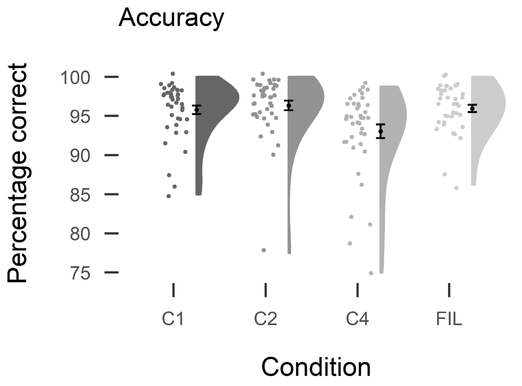


Figure 2.2: Individual and mean accuracy per condition. Black dots indicate mean with standard error of the mean. Grey dots represent individual participants' mean.

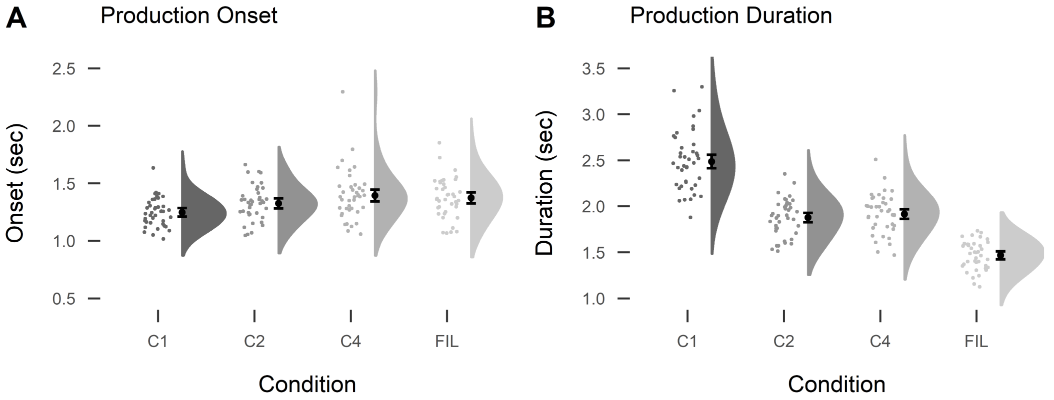


Figure 2.3: Onset (A) and duration (B) times per condition (of correct trials only). Black dots indicate mean with standard error of the mean. Grey dots represent individual participants' mean.

Whole-brain analysis

We focused on the main effects of constituent size in production and comprehension, and on the interaction between modalities (production vs. comprehension) and constituent size. For the main effect of constituent size, a large bilateral network centered around areas of the language network and the corresponding right hemisphere areas, with cerebellar and occipital activity, was found (Fig. 2.4A; Supplementary table S2.1). We found a large left lateralized cluster including peaks in the left IFG, STG, MTG, temporal pole, precentral gyrus, postcentral gyrus, fusiform gyrus and superior parietal lobule. Similar right lateralized activity was found in a cluster in the temporal pole and IFG (*pars orbitalis*), a cluster in postcentral and precentral gyrus and a cluster in superior and middle temporal gyri, a cluster in superior parietal lobule and a cluster in the more posterior parts of the IFG (*pars triangularis* and *pars opercularis*). Additionally, we found clusters in the left and right supplementary motor area, in the left thalamus, left putamen, and right cerebellum.

To evaluate to what extent the activated network was overlapping between comprehension and production, we performed a conjunction analysis of the separate constituent size contrast for production and comprehension. This analysis revealed that in part the constituent size effect was reliably active in both modalities, with clusters in anterior and posterior MTG, LIFG, left precentral gyrus, left fusiform gyrus and right cerebellum (see Fig. 2.4A, Supplementary table S2.1).

We also looked at the main effect of modality to understand if any areas were overall more active in production or comprehension (see Fig. 2.4B, Supplementary table S2.2). We found bilateral frontal areas and parietal areas, as well as subcortical and cerebellar regions, to have larger activity in production than comprehension, partly reflecting articulatory requirements in production. Bilateral superior temporal areas were more engaged in comprehension, which was likely due to auditory processing. Bilateral angular gyrus, precuneus and superior frontal regions were also more engaged in comprehension.

A few areas responded differently to constituent size in production and comprehension (Fig. 2.4C, Supplementary table S2.3). Areas that were more active with larger constituents in production were mainly left lateralized and included the LIFG (*pars triangularis*, *pars orbitalis* and *pars opercularis*), middle frontal gyrus, precentral gyrus, supplementary motor area, inferior and superior parietal lobule, supramarginal gyrus, angular gyrus, and posterior sections of the LMTG. Regions in the right hemisphere included precentral gyrus, postcentral gyrus, superior parietal lobule, supplementary motor area and cerebellum. A complementary network was more active in the comprehension of larger constituents, with peaks in bilateral Heschl's gyrus, STG and temporal pole, and right hemisphere areas, including angular gyrus, precuneus, frontal pole and superior and middle frontal gyri.

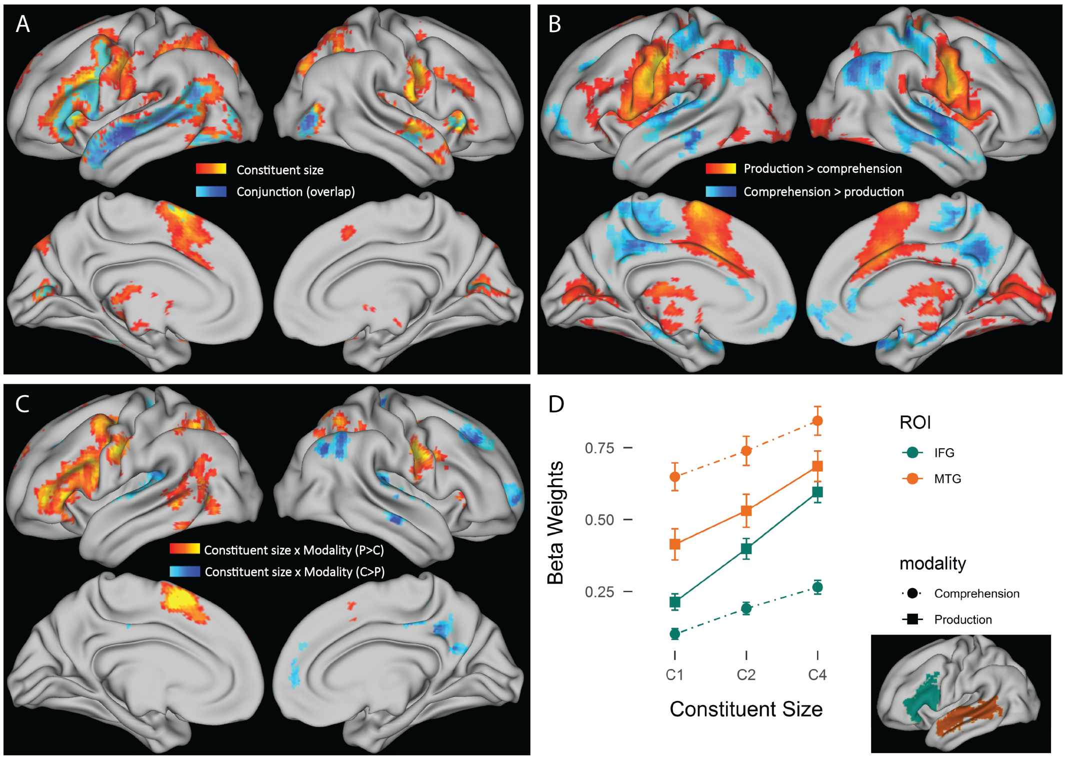


Figure 2.4: Whole-brain and ROI results. **A**: orange: main effect of constituent size with a linear contrast for the three constituent sizes. Blue: conjunction analysis of production and comprehension constituent size effects representing areas active in both production and comprehension following the conjunction of null hypotheses (Friston, Penny, & Glaser, 2005). The blue area is superimposed on the corresponding cluster found as main effect of constituent size. **B**: whole-brain results for the main effect of modality. Orange: areas more active in production than comprehension. Blue: areas more active in comprehension than production. **C**: whole-brain results for the interaction between constituent size and modality. Orange: areas with larger response to constituent size in production than comprehension. Blue: areas with larger response to constituent size in comprehension than production. **D**: mean beta weights extracted from the predefined ROIs (depicted in figure), error bars represent standard error of the mean.

ROI analysis: LIFG and LMTG

We extracted beta weights for the average activity in regions previously associated with syntactic effects to inspect patterns of activation for each condition in production and comprehension (Fig. 2.4D). We ran a linear mixed-effects model on the beta estimates that we extracted per condition per region. We found a main effect of constituent size ($\beta = 0.027$, SE = 0.001, $t = 20.03$, $\chi^2 = 401.04$, $p < 0.0001$), indicating that beta weights increased with larger constituent sizes. Pairwise comparisons indicated that beta weights for C4 were significantly larger than C2 and C1 in all modalities and ROIs (estimates > 0.10, ts > 8.2, ps < 0.0001). We also found a main effect of ROI ($\beta = 0.18$, SE = 0.02, $t = 8.62$, $\chi^2 = 74.4$, $p < 0.0001$), with larger beta estimates in LMTG than LIFG (estimate = 0.35, SE = 0.04, $t = 8.52$, $p < 0.0001$). The effect of ROI interacted with modality ($\beta = 0.11$, SE = 0.009, $t = 11.45$, $\chi^2 = 131.03$, $p < 0.0001$), since there was a larger difference in activity between ROIs in comprehension than in production (MTG – IFG, Production: estimate = 0.14, SE = 0.04, $t = 2.97$, $p < 0.025$; Comprehension: estimate = 0.56, SE = 0.04, $t = 13.16$, $p < 0.0001$). Importantly, there was a three-way interaction between constituent size, modality and ROI ($\beta = 0.004$, SE = 0.0009, $t = 4.24$, $\chi^2 = 18.0$, $p < 0.0001$). Inspection of the slopes for constituent size in each modality and ROI indicated that production elicited the steepest slope in the response to constituent size in the IFG: there was a larger slope difference between modalities in the IFG (Production – Comprehension: estimate = 0.023, SE = 0.003, $t = 9.34$, $p < 0.0001$) than in the MTG (Production – Comprehension: estimate = 0.009, SE = 0.003, $t = 3.37$, $p = 0.005$), and there was a slope difference between ROIs in production (IFG – MTG: estimate = 0.012, SE = 0.003, $t = 4.5$, $p = 0.0001$), but not in comprehension (MTG – IFG: estimate = 0.003, SE = 0.003, $t = 1.47$, $p = 0.46$). These results, therefore, show that: (i) larger constituent structures elicit higher activity in both regions and modalities, (ii) there is a stronger effect of constituent size in production than in comprehension, especially in LIFG, (iii) there is a higher response in LMTG than LIFG overall, and (iv) production elicits stronger activity than comprehension in the LIFG, while the opposite is the case in LMTG: more activity for comprehension than production.

Exploratory analysis: BOLD peak latency

We extracted BOLD times-to-peak for each condition to understand whether the regional and modality-specific effects highlighted by the ROI analysis were also characterized by BOLD time course differences. Pallier et al. (2011) had found that larger constituent sizes were associated with later peak times in the superior temporal sulcus and IFG, in line with the idea that activation is stronger towards the end of a constituent. A model with ROI, modality and constituent size as predictors for time-to-peak showed a main effect of modality ($\beta = 0.33$, SE = 0.06, $t = 5.39$, $\chi^2 = 29.01$, $p < 0.0001$), with comprehension peaking earlier than production (estimate = 0.66, SE = 0.12, $t = 5.34$, $p < 0.0001$). In addition, we found an interaction between modality and constituent size ($\beta = 0.049$, SE = 0.016, $t = 3.002$, $\chi^2 = 9.01$, $p = 0.0027$)[4]. Inspection of the slopes in the response to constituent size showed that comprehension elicited a positive slope, with larger constituent structures peaking later, while production elicited a negative slope, with larger constituent structures peaking earlier (Comprehension – Production: estimate = 0.098, SE = 0.033, $t = 2.98$, $p = 0.0031$) (Fig. 2.5). Therefore, the constituent size effect on peak latency that was found before (Pallier et al., 2011) seems to be dependent on modality, since in production an opposite pattern was found relative to comprehension.

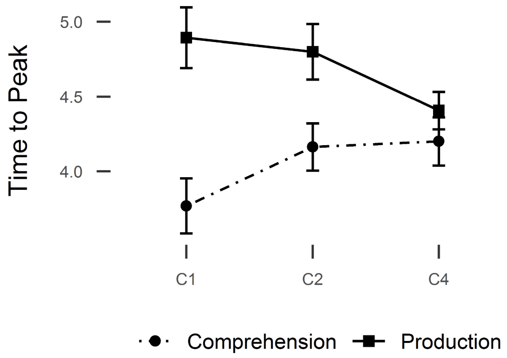


Figure 2.5: BOLD peak latency. BOLD peak times averaged across participants and ROIs for each constituent size in production and comprehension. Error bars represent standard error of the mean.

---

[4] Since peak time extraction provided values in 1.5 sec resolution, we also ran a model with time (as well as constituent size, modality and ROI) as a predictor for percent signal change to ensure that extracted peak times were consistent with the time courses. The model returned a significant interaction between modality and time, and a three-way interaction between modality, constituent size and time, confirming the results of the peak time analysis.

2.4 Discussion

We examined neural responses to the production and comprehension of utterances with increasing constituent size to clarify the neural correlates of sentence production and comprehension. We found that larger constituent sizes engaged areas traditionally part of the language network. These included inferior frontal regions, temporal and inferior parietal regions, mainly in the left hemisphere. Through a conjunction analysis, we confirmed that the LIFG and LMTG responded to constituent size in both comprehension and production. Increased syntactic complexity resulted in stronger activation in these areas. Moreover, we found a modality-specific dissociation, with production recruiting the LIFG more strongly than comprehension, and comprehension recruiting the LMTG more strongly than production. At the same time, the network was found to be differentially responsive to constituent size across modalities. While comprehension elicited similar responses to constituent size in LIFG and LMTG, in production the LIFG was more sensitive to constituent size than the LMTG. Finally, constituent size had opposite effects on BOLD peak latencies in comprehension and production: increasing constituent size elicited later peaks in comprehension but earlier peaks in production.

By demonstrating that the response to constituent size is largely shared between comprehension and production, we extend Pallier et al.'s constituent size effect (2011) to sentence production. Our results are in line with evidence associating sentence-level processes with left inferior frontal and temporal activation (Blank, Balewski, Mahowald, & Fedorenko, 2016; Bornkessel-Schlesewsky et al., 2009; Hagoort & Indefrey, 2014; Henderson et al., 2016; Indefrey, 2018; Segaert et al., 2012; Shetreet, Friedmann, & Hadar, 2009; Walenski et al., 2019). Pallier et al. (2011) found that the LIFG and the posterior superior temporal sulcus were responsive to constituent size also with jabberwocky stimuli, while the ATL and the temporo-parietal junction only responded to stimuli with real words. All of these areas were responsive to constituent size also in the present study. Whether the IFG and the posterior temporal sulcus are sensitive to constituent size also with jabberwocky stimuli in production will have to be determined in future studies specifically designed to address the distinction between syntactic and semantic compositional processes. Finally, the activation delay for larger constituent structures was replicated here, but critically only in comprehension.

Our results, therefore, implicating both LIFG and LpMTG, as well as other areas of the language network, suggest that the inconsistent evidence for sentence production was due to low power in the single studies and in the meta-analyses

(Indefrey, 2018; Walenski et al., 2019). With 40 participants and a large number of trials per condition, we had enough power to detect effects in areas previously linked with sentence processing in comprehension. It is unlikely that the effects we found are reducible to the type of paradigm used to elicit sentence production, since other studies using picture descriptions or sentence reorganization paradigms also found activations in LIFG and/or LpMTG, but, critically, in an inconsistent way (e.g. pictures descriptions, Grande et al., 2012; Indefrey et al., 2001; Menenti et al., 2012; Segaert et al., 2012; sentence generation from words, Collina et al., 2014; Golestani et al., 2006; Haller et al., 2005). Moreover, although our paradigm was partly artificial in eliciting sentence production, it allowed us to cleanly manipulate constituent structure, ensuring consistent behavioural responses across participants. Previous studies used similar types of constrained elicitation paradigms or more constraining ones when more control over the production was required (cf. Matchin & Hickok, 2016; Matchin & Wood, 2020; Takashima et al., 2020).

Crucially, the conjunction analysis showed that production and comprehension engage largely overlapping areas in constituent structure building. An extensive network is engaged in sentence production that does not diverge from the one observed for comprehension in previous studies. The activation pattern, including left anterior and posterior MTG and LIFG is similar to the syntactic adaptation effects found across modalities in fMRI studies with repetition suppression (Menenti et al., 2011; Segaert et al., 2012). Our results do not provide information on whether verb-specific processing is also shared between production and comprehension, since the linear-contrast analysis avoided sensitivity to verb argument structure differences between sentences. Thus, these results confirm shared resources in sentence-level processes across modalities and provide no support for spatial segregation as a basis for distinct processes or representations. These findings, therefore, reconcile the previous inconsistent findings between sentence production and comprehension networks, as shown by meta-analyses (Indefrey, 2018; Walenski et al., 2019), with the adaptation effects across modalities (Menenti et al., 2011; Segaert et al., 2012). Common neural resources provide a neural basis for views of shared linguistic representations and processes, such as retrieval and unification, between production and comprehension (Dell & Chang, 2014; Kempen, 2000; Kempen et al., 2012; Momma & Phillips, 2018).

While the networks overlapped, there were differences in the degree to which each modality recruited core areas. In particular, we found that comprehension engaged the LMTG more than production, and production engaged the LIFG more than comprehension. This finding was consistent with the modality differences in the whole-brain results. Larger activity in the LMTG in comprehension was also found by Humphreys and Gennari (2014), and is likely due to the fact that the auditory input is processed in superior temporal areas with activity spreading in the temporal lobe, whereby the LpMTG might be involved in retrieval and integration of lexical, syntactic and semantic information, given its extensive connectivity patterns (Baggio & Hagoort, 2011; Binder, 2017; Turken & Dronkers, 2011). The clusters showing more activity in production included not only the LIFG, but also more dorsal areas, extending to the precentral gyrus and the supplementary motor area. Together with the cerebellar activation, these latter areas are involved in articulation and motor planning (Basilakos, Smith, Fillmore, Fridriksson, & Fedorenko, 2018; Price, 2012).

The greater involvement of inferior frontal regions in production than comprehension is likely attributable to stronger sentence planning requirements, also reflected in the stronger effect of constituent size in the LIFG and to a smaller extent in the LpMTG in production than comprehension. In production, the syntactic structure of sentences needs to be fully and correctly computed in order to produce a well-formed utterance (Garrett, 1980, 1982; Indefrey, 2018). In comprehension, instead, inferring sentence meaning can often be done by retrieving word meanings and world knowledge, bypassing the need for a full syntactic analysis of the input (cf. good-enough processing, Ferreira, Bailey, & Ferraro, 2002). For instance, it has been shown that passive or object-relative sentences are sometimes interpreted in line with world knowledge but not necessarily in agreement with the syntactic structure (Ferreira, 2003; Flinker et al., 2015). Therefore, reduced sensitivity to constituent structure in comprehension may signal reduced syntactic processing in reaching the conceptual interpretation for these sentences. This fundamental difference between production and comprehension on the importance of 'getting it right' may also explain the larger engagement of the default mode network in comprehension (in particular, right angular gyrus, right precuneus, right superior frontal gyrus and right frontal pole). We speculate that production disengaged the default mode network more than comprehension in responding to constituent size, due to the stronger requirements for accurate sentence planning (Raichle et al., 2001; Raichle & Snyder, 2007).

The interaction effects between constituent size and modality cannot be reduced to task differences between modalities and in particular to the absence of an explicit task in comprehension. On the one hand, the constituent size effect in comprehension and the finding of larger comprehension activity in the LMTG confirm that participants processed the input even in the absence of a task (see Figure 2.4D). On the other hand, the task requirements in production were very similar across levels of constituent size: what varied was the linguistic complexity of the output. Differences between modalities may instead show task effects, including cognitive control differences. However, as mentioned above, production is inherently a 'task' as opposed to comprehension being more passive also in naturalistic situations. Task effects thus need not reflect spurious task differences due to the current design, but could be related to inherent differences in cognitive control between production and comprehension. Studies of spontaneous production may be able to address to what extent cognitive control is needed during naturalistic production as opposed to comprehension.

An additional dissociation in the response pattern for production and comprehension was found in the BOLD time courses. Production and comprehension elicited opposite profiles of response latencies in relation with constituent size. Larger structures were characterized by later peaks in comprehension, confirming previous evidence suggesting that larger structures take longer to be computed (Pallier et al., 2011). In contrast, larger structures elicited earlier peaks than smaller structures in production. This was likely due to planning differences between conditions. Reaction time analyses showed that onset times increased with constituent size, with C2 taking longer than C1, and C4 taking longer than C2. Since high-level processing can be initiated for the whole clause before speaking (M. Smith & Wheeldon, 1999), it is likely that more extensive planning at the message or structural level took place in early stages for the more complex structures, inducing early peaks in BOLD activity. In contrast, in the conditions with smaller constituent size the structures to be computed were smaller and planning may have been in a word-by-word fashion interleaved with articulation, hence inducing sustained activity with later peaks. Since this was an exploratory analysis for which the stimuli and the design were not optimized a priori, future studies will need to clarify whether BOLD peak latencies in production are indeed influenced by planning scope and if the inverse relationship between onset times and production peak latencies holds with different stimuli and paradigms.

Overall, the current results are striking in showing how production and comprehension share resources but modulate them differently. Spatially, frontal and temporal regions are engaged in both modalities, but to different extents. Temporally, constituent size affects BOLD peak latencies in both modalities but in opposite directions. Rather than providing support for a distinction of core processes and representations between modalities (Meyer et al., 2016), this unbalanced sharing of resources reveals a ‘computational asymmetry’ (Matchin & Hickok, 2020) or ‘directional’ differences (Gambi & Pickering, 2017; Pickering & Garrod, 2013). In production, linguistic processes map from higher to lower linguistic levels, i.e. meaning to phonology, and in comprehension from lower to higher linguistic levels, i.e. phonology to meaning (Pickering & Garrod, 2013). This directional difference implies that the inputs and outputs of each modality are opposite in production and comprehension, which results in differences in recruitment patterns within the shared language network (Indefrey, 2018; Momma & Phillips, 2018), reflected not only in different regional levels of activity, but also in timing patterns.

In conclusion, the current results extend the constituent structure effect found in comprehension (Pallier et al., 2011) to production, and robustly show the involvement of both LpMTG and LIFG in constituent structure building in production, helping to clarify the inconsistencies in the previous studies on the neurobiology of language production. Additionally, the results confirm that the neural resources for sentence production and comprehension are largely overlapping, supporting accounts of shared representations between modalities. Finally, our results also highlight modality-specific differences in regional and time course patterns that underline inevitable differences in the requirements of speaking and listening.

S2 Supplementary Information

Table S2.1: fMRI whole-brain summary of cluster peak coordinates and statistics for the main effect of constituent size and conjunction analysis for production and comprehension constituent size effects. P values are FWE-corrected.

Contrast	**Cluster**		**Peak Voxel (MNI)**				**Peak Anatomical Location**
	p	**Size**	***Z***	**x**	**y**	**z**	
Constituent size	0	7656	Inf	-48	18	24	LIFG (pars triangularis)
			7.66	-58	-2	-13	Superior Temporal Gyrus
			7.42	-28	28	0	Left Anterior Insula
			7.28	-51	20	-6	LIFG (pars orbitalis)
			7.28	-51	-14	-8	LMTG
			6.79	-51	16	-23	Temporal Pole
			6.52	-46	8	37	Left Precentral Gyrus
			6.46	-46	-40	-20	Left Inferior Temporal Gyrus
			6.41	-26	-77	32	Left Middle Occipital Gyrus
			6.32	-44	-57	-16	Left Fusiform Gyrus
			6.27	-54	-10	40	Left Postcentral Gyrus
			6.21	-61	-57	14	LSTG
			6.15	-38	-50	50	Left Inferior Parietal Lobule
			6.11	-54	33	14	LIFG (pars triangularis)
			5.07	-11	-77	57	Left Superior Parietal Lobule
			4.79	-28	0	67	Left Middle Frontal Gyrus
	0	1953	Inf	14	-70	-28	Right Cerebellum
			6.59	-16	-67	-18	Left Cerebellum
	0	642	6.81	-4	8	62	Left Supplementary Motor Area
	0	440	6.69	32	30	0	RIFG (pars orbitalis)
			5.52	49	18	-13	Right Temporal Pole
	0	204	6.12	49	-74	2	Right MTG
	0	194	5.72	64	-4	-3	Right STG
			4.89	49	-14	-13	Right MTG
	0	388	5.69	54	-10	40	Right Precentral Gyrus
			5.39	66	-7	14	Right Postcentral Gyrus
	0	422	5.64	-8	-82	10	Left Calcarine Sulcus
			5.31	9	-72	12	Right Calcarine Sulcus
	0	61	5.55	-34	-27	10	Left Heschl's Gyrus
	0	455	5.3	-11	-30	-13	Brainstem
			5.08	-14	-2	20	Left Caudate
			4.99	-11	-17	10	Thalamus
			4.94	6	6	0	Right Caudate
	0	707	5.22	29	-80	40	Right Superior Occipital Gyrus
			5.07	34	-57	47	Right Angular Gyrus
			5.02	19	-60	60	Right Superior Parietal Lobule
	0.032	28	5.18	-46	-57	-26	Left Cerebellum
	0	98	5.03	36	-22	10	Right Insula
	0.001	52	4.69	32	-10	67	Right Superior Frontal Gyrus
	0	76	4.68	-31	-12	-6	Left Putamen
	0	264	4.53	56	18	32	Right IFG (pars opercularis)
			4.41	42	23	22	Right IFG (pars triangularis)
	0	67	4.53	-8	58	37	Left Superior Frontal Gyrus

continued ...

. . . continued

Contrast	Cluster		Peak Voxel (MNI)				Peak Anatomical Location
	p	Size	*Z*	x	y	z	
Conjunction	0	2757	7.73	-54	-12	-8	LMTG
			6.74	-58	16	17	LIFG (pars opercularis)
			6.27	-48	18	-18	Left Temporal Pole
			6.07	-46	28	-6	LIFG (pars orbitalis)
			5.91	-46	-57	20	Left Angular Gyrus
			5.58	-56	33	7	LIFG (pars triangularis)
			5.38	-41	-42	-23	Left Fusiform Gyrus
			4.66	-44	8	30	Left Precentral Gyrus
			4.54	-46	-84	4	Left Middle Occipital Gyrus
	0	344	7.01	19	-80	-40	Right Cerebellum
	0	117	5.47	49	-74	0	Right MTG
	0	69	4.7	-6	-80	12	Left Calcarine Sulcus
	0	59	4.22	-28	-74	32	Left Middle Occipital Gyrus
			3.6	-24	-82	44	Left Superior Parietal Lobule
	0.01	37	4.11	-4	8	70	Left Supplementary Motor Area
	0.001	58	4.11	12	-74	10	Right Calcarine Sulcus
	0.001	57	4.07	36	23	-3	Right Insula
	0.001	56	4.05	54	-4	-13	Right STG

Table S2.2: fMRI whole-brain summary of cluster peak coordinates and statistics for the main effect of modality. P values are FWE-corrected.

Contrast	Cluster		Peak Voxel (MNI)				Peak Anatomical Location
	p	Size	*Z*	x	y	z	
Production > Comprehension	0	14691	Inf	-51	-12	40	Left Postcentral Gyrus
			Inf	49	-7	30	Right Postcentral Gyrus
			Inf	22	-64	-23	Right Cerebellum
			Inf	-14	-17	4	Left Thalamus
			Inf	-16	-67	-18	Left Cerebellum
			Inf	-31	-17	0	Left Putamen
			Inf	32	-14	-3	Right Putamen
			Inf	14	-17	7	Right Thalamus
			7.57	39	6	4	Right Insula
			7.41	-34	-12	20	Left Insula
			6.54	22	0	17	Right Caudate
			6.45	-54	6	44	Left Precentral Gyrus
			5.83	14	-62	7	Right Calcarine Sulcus
			5.83	-11	-70	7	Left Calcarine Sulcus
			5.59	-41	23	0	LIFG (pars triangularis)
			4.87	32	-92	-10	Right Inferior Occipital Gyrus
			4.79	-54	-54	-16	Left Inferior Temporal Gyrus
			4.73	46	-10	57	Right Precentral Gyrus
			4.7	34	-57	-16	Right Fusiform Gyrus
			4.3	39	8	27	Right IFG (pars opercularis)
			4.22	-41	-64	-6	Left Inferior Occipital Gyrus
			4.03	-38	40	24	Left Middle Frontal Gyrus
			3.91	-14	18	-13	Left Medial Orbital Gyrus
			3.8	-31	8	27	LIFG (pars opercularis)
			3.78	19	-44	-6	Right Parahippocampal gyrus
			3.78	-28	-72	-10	Left Fusiform Gyrus
			3.19	-64	-30	34	Left Supramarginal Gyrus

continued . . .

. . . continued

Contrast	**Cluster**		**Peak Voxel (MNI)**				**Peak Anatomical Location**
	p	**Size**	***Z***	**x**	**y**	**z**	
	0	2087	Inf	4	16	37	Right Middle Cingulate Cortex
			Inf	4	0	67	Right Supplementary Motor Area
			7.06	-4	10	57	Left Supplementary Motor Area
			5.27	-4	-4	40	Left Middle Cingulate Cortex
	0	128	7.39	19	-30	64	Right Precentral Gyrus
	0	128	6.95	-21	-32	62	Left Postcentral Gyrus
	0	109	5.39	-21	-97	-10	Left Inferior Occipital Gyrus
	0	517	4.91	-26	-72	44	Left Inferior Parietal Lobule
	0.002	53	4.76	62	-34	22	Right Supramarginal Gyrus
	0	63	4.02	-58	-44	27	Left Supramarginal Gyrus
	0.01	39	3.94	22	-67	50	Right Superior Parietal Lobule
	0.029	32	3.87	-41	-87	2	Left Middle Occipital Gyrus
Comprehension >Production	0	1907	Inf	46	-14	2	Right Transverse Temporal Gyrus
			6.43	64	-10	-3	Right STG
			6.09	49	-10	-16	Right MTG
			5.87	39	-22	14	Right Heschl's Gyrus
			5.8	39	18	-26	Right Superior Temporal Pole
			5.38	26	-20	-16	Right Hippocampus
			3.8	16	-32	-13	Right Parahippocampal gyrus
	0	2110	Inf	-46	-14	2	Left Transverse Temporal Gyrus
			6.83	-58	-27	12	Left STG
			6.39	-54	-62	42	Left Angular Gyrus
			6.25	-61	-34	-10	Left MTG
			5.07	-41	16	-26	Left Superior Temporal Pole
			4.58	-48	-24	-18	Left Inferior Temporal Gyrus
	0	615	6.98	-18	-84	-30	Left Cerebellum
	0	198	6.83	-18	-17	30	Left Postcentral Gyrus
	0	1849	6.68	-1	-44	40	Left Precuneus
			5.87	-4	-24	50	Left Paracentral Lobule
			5.76	14	-52	37	Right Precuneus
	0	525	6.31	42	-32	60	Right Postcentral Gyrus
	0	627	6.25	-1	53	-6	Left Medial Frontal Cortex
	0	911	6.2	56	-54	40	Right Angular Gyrus
	0	119	6.18	9	-47	-46	Right Cerebellum
	0	408	6.16	14	-82	-28	Right Cerebellum
	0	351	6.02	-36	63	2	Left Lateral Orbital Cortex
	0	131	5.55	-26	36	50	Left Middle Frontal Gyrus
	0	264	5.44	29	66	10	Right Superior Frontal Cortex
	0	250	5.36	42	23	50	Right Middle Frontal Gyrus
	0	311	5.31	-36	-40	64	Left Postcentral Gyrus
	0	76	5.3	2	38	2	Right Anterior Cingulate Cortex
			4.87	-1	28	14	Left Anterior Cingulate Cortex
	0	86	4.98	-21	-27	-18	Left Parahippocampal Gyrus
	0.029	32	4.1	36	38	-10	Right Lateral Orbital Gyrus

Table S2.3: fMRI whole-brain summary of cluster peak coordinates and statistics for the main effect of modality. P values are FWE-corrected.

Contrast	**Cluster**		**Peak Voxel (MNI)**				**Peak Anatomical Location**
	p	**Size**	***Z***	**x**	**y**	**z**	
Interaction Production > Comprehension	0	2307	6.41	-51	20	-6	LIFG (pars orbitalis)
			6.35	-51	10	32	Left Precentral Gyrus
			5.83	-44	43	2	LIFG (pars triangularis)
			5.67	-28	-7	50	Left Middle Frontal Gyrus
			5.29	-58	13	12	LIFG (pars opercularis)
	0	1108	5.53	-26	-54	42	Left Inferior Parietal Lobule
			5.51	-34	-64	57	Left Superior Parietal Lobule
			5.49	-26	-74	37	Left Angular Gyrus
			4.97	-51	-50	54	Left Supramarginal Gyrus
	0	1032	7.74	12	-82	-40	Right Cerebellum
	0	388	7.29	-4	16	57	Left Supplementary Motor area
			3.98	12	10	50	Right Supplementary Motor Area
	0	271	5.52	52	-4	32	Right Postcentral Gyrus
	0	783	5.47	-61	-52	10	LMTG
			5.37	-56	-67	20	Left Angular Gyrus
			4.14	-51	-74	-10	Left Inferior Occipital Gyrus
	0	277	5.28	29	-54	42	Right Inferior Parietal Sulcus
			4.62	32	-62	62	Right Superior Parietal Lobule
			4.16	42	-34	44	Right Supramarginal Gyrus
	0	93	4.77	44	8	30	Right Precentral Gyrus
			3.72	56	18	32	Right IFG (pars opercularis)
	0.001	50	4.74	36	26	2	Right Anterior Insula
	0.003	44	4.72	2	-62	-38	Right Cerebellum
	0.002	46	4.5	-54	-50	-16	Left Inferior Temporal
	0	67	4.36	-11	56	34	Left Superior Frontal Gyrus
			3.62	-8	46	47	Left Superior Medial Frontal Gyrus
	0	94	4.33	34	-70	34	Right Middle Occipital Gyrus
			3.47	26	-72	54	Right Superior Parietal Lobule
	0.02	31	4.04	-16	-67	-20	Left Cerebellum
Interaction Comprehension >Production	0	534	5.57	-38	-24	7	Left Heschl's Gyrus
			4.42	-36	-32	22	Left Posterior Insula
			4.35	-58	-17	7	Left Superior Temporal Gyrus
			4.23	-48	8	-10	Left Superior Temporal Pole
	0	350	5.57	44	-14	0	Right STG
			4.61	36	-24	14	Right Heschl's Gyrus
	0	374	5.51	24	36	42	Right Superior Frontal Gyrus
			4.12	42	30	42	Right Middle Frontal Gyrus
	0	371	5.27	59	-54	40	Right Inferior Parietal Lobule
			4.9	44	-64	32	Right Angular Gyrus
	0	168	5.14	66	-17	-6	Right MTG
	0.039	27	4.6	14	-27	64	Right Precentral Gyrus
	0.001	53	4.59	-31	38	40	Left Middle Frontal Gyrus
	0	179	4.56	26	66	4	Right Frontal Pole
			3.64	2	63	4	Right Superior Medial Frontal Gyrus
	0	100	4.52	9	-44	37	Right Precuneus
	0.001	51	4.5	6	-57	24	Right Precuneus
	0	70	4.48	9	-24	42	Right Middle Cingulate Gyrus
			3.66	9	-37	52	Right Precuneus
	0	95	4.06	14	53	0	Right Superior Medial Frontal Gyrus

Chapter 3

3 | Connectivity of the Fronto-temporal Network in Syntactic Structure Building during Speaking and Listening

Abstract

The neural infrastructure for sentence production and comprehension has been found to be mostly shared. The same regions are engaged during speaking and listening, with some differences in their loading depending on modality. In this study, we investigated whether modality affects the connectivity between regions previously found to be involved in syntactic processing across modalities. We asked how constituent size and modality affected the connectivity of the *pars triangularis* of the left inferior frontal gyrus (LIFG) and of the posterior temporal lobe (LPTL) with the *pars opercularis* of the LIFG, the anterior temporal lobe (LATL) and the rest of the brain. We found that constituent size reliably increased the connectivity among and within these frontal and temporal ROIs. The connectivity between the two LIFG regions and the LPTL was enhanced for all constituent sizes in both modalities, and it was upregulated in production possibly because of linearization and motor planning coordination in the frontal cortex. The connectivity of both seed ROIs with the LATL was lower and only enhanced for larger constituent sizes, suggesting a contributing role of the LATL in sentence processing in both modalities. These results thus show that the connectivity among fronto-temporal regions is upregulated for syntactic structure building in both sentence production and comprehension, providing further evidence to accounts of shared neural resources for sentence-level processing across modalities.

3.1 Introduction

In moving away from the classical Broca's-Wernicke understanding of the neurobiology of language, it has become clear that the division of labour between fronto-temporal regions in processing language fundamentally rests on the connectivity between these and other brain regions (Dick, Bernal, & Tremblay, 2014; Hagoort & Beckmann, 2019; Hagoort & Indefrey, 2014). Much of the brain's volume is constituted by white matter tracts that connect more and less distant regions to one another. The advance of diffusion tensor imaging (DTI) has enlarged the understanding of the orientation of white matter fibers and their terminations, adding to tracing and dissection studies in macaques and post-mortem brains (Rilling et al., 2008). In parallel, resting-state fMRI has advanced the understanding of large-scale networks independently of monosynaptic connections (Buckner, Krienen, & Yeo, 2013). However, the effective modulation of networks by specific linguistic processing has only been studied sporadically, especially in language production, even though it has the potential to enable a deeper and broader understanding of how the coupling between brain regions is related to task-specific language production and comprehension (Friston, 2011; Hagoort & Beckmann, 2019).

The current understanding of the core white matter pathways that connect fronto-temporal regions centers on a dorsal and a ventral stream with separate functions (Dick et al., 2014; Friederici, 2012; Saur et al., 2008; Shekari & Nozari, 2022). The dorsal stream consists of the superior longitudinal fasciculus and the arcuate fasciculus, which have terminations in the posterior temporal lobe, inferior parietal cortex and posterior frontal cortex. These connections, especially the temporal projections, have been found to be enlarged in the human brain compared to the chimpanzee and macaque brains (Rilling et al., 2008; Sierpowska et al., 2022). These tracts have been suggested to have a role in phonological (Hickok & Poeppel, 2007) and syntactic processing (Friederici, 2012; Papoutsi, Stamatakis, Griffiths, Marslen-Wilson, & Tyler, 2011; Zaccarella, Schell, & Friederici, 2017). The ventral stream consists of multiple fasciculi (uncinate fasciculus, extreme capsule, middle longitudinal fasciculus, inferior longitudinal fasciculus and inferior fronto-occipital fasciculus), whose terminations and functions are still a matter of debate, but there is tentative consensus that they are involved in semantic processing (Dick et al., 2014; Friederici, 2012; Shekari & Nozari, 2022).

The functional connectivity between core language regions, which has been claimed to rely on the white-matter pathways introduced above (e.g. Friederici, 2011; Glasser & Rilling, 2008; Hagoort, 2014, 2017), has been investigated with functional neuroimaging studies that identify patterns of correlations between regions of interest (ROIs) and the rest of the brain (Biswal, Zerrin Yetkin, Haughton, & Hyde, 1995; Friston, Frith, Frackowiak, & Turner, 1995). Resting-state fMRI provides an overview of large-scale functional networks that are co-activated at rest, leading to a measure that is in-between anatomical connectivity and dynamic (transient) coupling (Buckner et al., 2013). At rest, the inferior frontal gyrus is functionally connected to parietal and temporal regions in a topographical organization that was suggested to indicate phonological, syntactic and semantic sub-networks across hemispheres, with the *pars opercularis* and the *pars triangularis* of the left inferior frontal gyrus (LIFGoper and LIFGtri) engaging the largest functional networks (Bulut, 2022; Przeździk, Haak, Beckman, & Bartsch, 2019; Xiang, Fonteijn, Norris, & Hagoort, 2010). Several regions in the temporal lobe also have wide co-activation patterns in resting-state fMRI, with the posterior MTG having the broadest connections to frontal and inferior parietal regions (Turken & Dronkers, 2011), resulting in a network that is highly similar to the regions usually found to be activated during language processing (e.g. Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010; Giglio, Ostarek, Weber, & Hagoort, 2022; Hagoort & Indefrey, 2014; Pallier et al., 2011; Segaert et al., 2012). The functional coupling of networks in response to a task is instead studied with task-dependent or effective connectivity in task fMRI using, for example, psychophysiological interactions (PPI, Friston et al., 1997). A PPI investigation of language processes could model how the connectivity between frontal and temporal regions is modulated by the effect of a linguistic stimulus. For example, the connectivity between LIFGoper and the left posterior temporal lobe (LPTL) was found to increase when resolving ambiguity in sentence comprehension (Snijders, Petersson, & Hagoort, 2010). Semantic composition has also been shown to drive increased coupling of LIFGtri, presupplementary motor area and posterior angular gyrus in comprehension (Graessner, Zaccarella, & Hartwigsen, 2021).

Connectivity between language critical brain regions thus can be enhanced by specific linguistic processing during comprehension, but there is a lack of understanding of how the connectivity between fronto-temporal regions that is found in resting-state studies and comprehension task connectivity studies is regulated by language modality (i.e. production and comprehension). One study found that the LIFGoper was connected to a larger brain network in production than comprehension in a competition task involving object-relative clauses

(Humphreys & Gennari, 2014), suggesting that functionally different networks may be engaged during production and comprehension. Here, we asked whether speaking and listening affect the connectivity between fronto-temporal areas in a task that loaded on syntactic structure building (Chapter 2, Giglio et al., 2022). In a previous study we found that the production and comprehension of the same stimuli engaged a similar network, but production elicited a stronger response to syntactic complexity, and frontal and temporal areas were engaged to a different extent in each modality. In particular, in production frontal areas had a stronger response than in comprehension, while comprehension elicited stronger responses in the temporal lobe (Giglio et al., 2022).

These regional differences across modalities could be due to production and comprehension relying on different functional sub-networks of the language system. Binder (2017) proposed a coherent temporal network engaged in sentence comprehension, centered in the posterior MTG for integration, that is supported by the extensive connectivity of the pMTG to other language-relevant regions relative to other temporal regions (Turken & Dronkers, 2011). This network could be functionally less engaged in production, as the final goal is to produce a motor output, after linearizing and articulating, which rely on frontal regions (Hickok & Poeppel, 2007; Kearney & Guenther, 2019; Matchin & Hickok, 2020). While the involvement of inferior frontal regions in production is generally agreed upon, with some debate on their specific function (Matchin & Hickok, 2020), their role in comprehension is not settled. Proposals on the function of the LIFG in linguistic processing range from unification (independent of language modality, Hagoort, 2005, 2013) and phrase-structure building (unspecified for modality, Zaccarella, Meyer, et al., 2017; Zaccarella, Schell, & Friederici, 2017), to linearization (only in production, Matchin & Hickok, 2020) and working memory or top-down prediction in comprehension (Matchin et al., 2017; Matchin & Hickok, 2020). Coordination with inferior frontal regions for articulation and motor planning is nevertheless needed in production (Kearney & Guenther, 2019; Price, 2010), indicating a potentially stronger coupling between temporal and frontal regions during speaking than listening. It is unclear whether fronto-temporal coupling would be modulated differently by syntactic processing in each modality.

In the current study, we directly addressed the question whether the connectivity of temporal and frontal regions previously found to be involved in linguistic processing (Chapter 2) differs between production and comprehension. To this end, we conducted a task-dependent connectivity analysis using generalized PPI (gPPI, McLaren, Ries, Xu, & Johnson, 2012) on a dataset that was previously collected to identify the linguistic network responsive to syntactic complexity in production and comprehension (Giglio et al., 2022). We focused on the connectivity of two seed regions of interest, LIFGtri and LPTL, that were previously found to be involved in syntactic processing in both modalities (Chapter 4, Giglio et al., 2022; Matchin & Hickok, 2020). We chose LIFGtri due to its greater involvement relative to LIFGoper in phrase-structure building in a recent study of spontaneous production (Chapter 4). In addition, LIFGoper was found to connect to other areas of the brain more broadly, which could suggest a lack of specificity for linguistic processing in its connectivity pattern and involvement in more general control and executive processes (Bornkessel-Schlesewsky & Schlesewsky, 2013; Fedorenko, Duncan, & Kanwisher, 2012; Humphreys & Gennari, 2014). We investigated the connectivity of seed regions LIFGtri and LPTL with two target ROIs, LIFGoper and LATL. To reduce individual variability, we focused on voxels that were responsive to constituent size in the previous study, and investigated how connectivity patterns from LIFGtri and LPTL to LATL and LIFGoper were affected by modality and syntactic complexity. We additionally tested for connectivity changes at the whole-brain level as a function of modality and complexity.

3.2 Methods

Participants

The analysis run in the current study was based on the data collected for a previous study (Chapter 2, Giglio et al., 2022). Forty-six right-handed native Dutch speakers participated in the study after giving written informed consent (28 females, mean = 23.8 years, range 19-35 years). The study was approved by the ethical committee for human research Region Arnhem-Nijmegen (CMO2014/288). Participants reported having no language-related disorders and normal or corrected-to-normal vision and hearing. Six participants were excluded due to technical problems during preprocessing (n = 1); failing to complete the experiment leading to too little data (n = 2); and motion artefacts (n = 3). Forty participants remained for the analysis.

Materials

The materials consisted of word sequences organized in three levels of constituent structure (C1, C2, C4), with 80 trials per conditions. In C1, two verbs had to be produced in infinitival form, followed by two content words preceded by their determiner, leading to four constituents of one content word (C1: "klappen, slapen, de jongen, het meisje", meaning "to clap, to sleep, the boy, the girl"). C2 consisted of two intransitive sentences composed of a subject and its verb, forming two constituents with two content words (C2: "de jongen slaapt, het meisje praat", "the boy sleeps, the girl talks"). In C4, the second clause was embedded in the main clause with a complementizer phrase, leading to one sentence formed by four content words (C4: "de jongen hoort dat het meisje klapt", "the boy hears that the girl claps"). The three conditions thus consisted of the same number of content words but differed in their constituent structure. The verbs were presented in root form (e.g. "klap", "slaap") and had to be inflected in all conditions. The verbs partly differed between conditions, since verbs allowing for a complementizer phrase (CP-verbs) are often not used in intransitive form. We selected 20 CP-verbs for the C4 condition, and repeated each 4 times in the C4 condition and 4 times in the C1 condition. We also selected 40 intransitive verbs, that were repeated 4 times in the C2 condition, twice in C2 and twice in C1. The CP-verbs and the intransitive verbs were matched in frequency (mean ± std: INT-verbs $= 1.38 \pm 0.88$, CP-verbs $= 1.46 \pm 0.77$, $t = 0.59$, $p = 0.56$) (Keuleers et al., 2010) with SUBTLEX-NL, and concreteness (mean ± std: INT-verbs $= 3.26 \pm 0.67$, CP-verbs $= 3.21 \pm 0.47$, $t = 0.27$, $p = 0.79$) (Brysbaert et al., 2014). There was also a filler condition to avoid too many verb repetitions, that used 80 transitive verbs used only once ("de man helpt de vrouw", "the man helps the woman"). The nouns used in the sentences were always "the boy", "the girl", "the man", "the woman" ("de jongen", "het meisje", "de man", "de vrouw").

The sentences were elicited by showing pictures for the nouns and written verbs (Fig. 2.1 in Chapter 2). The configuration of boxes around the verbs and pictures indicated the structure that had to be used by showing which elements should be combined in a sentence. In C1, four boxes indicated the production of four separate items (i.e. word sequences). In C2, two boxes surrounded one verb and one noun each, to form two sentences including two items. In C4 and for fillers, one box surrounded all items to form one sentence that included all the items. Participants had no problems producing the correct output (mean accuracy was above 90% for all conditions, see Chapter 2 for behavioural analysis and results, Giglio et al., 2022).

Experimental Procedure

First, there was a short behavioural practice session where participants read instructions for the task and practiced producing the sentences. The experimenter gave feedback to ensure correct understanding of the task. After the practice session, the fMRI experiment started. The experiment consisted of 8 production runs interleaved by 4 comprehension runs, with a comprehension run always following two production runs. Each run included 40 trials and lasted approximately 5 minutes. Each trial began with a fixation cross for 800 ms, followed by a 5 second presentation of the picture screen, during which participants were instructed to produce or listen to the same materials. The trial ended with a blank screen for 200 ms. Trial onset was jittered by 0-7500 ms (mean 1500 ms) based on design optimization for contrast effect detection (optseq2, Dale, 1999). In the comprehension runs, participants listened to recorded stimuli, which were presented 1 second after picture onset and lasted a maximum of 4 seconds (mean duration (in seconds): C1 = 3.14, C2 = 2.46, C4 = 2.46, Fillers = 1.79).

fMRI acquisition and preprocessing

MR data were acquired in a 3T MAGNETOM PrismaFit MR scanner (Siemens AG, Healthcare Sector, Erlangen, Germany) using a 32-channel head coil. We collected a T1-weighted MRI scan for anatomical reference and several fMRI runs. The T1-weighted scan was acquired in the sagittal orientation using a 3D MPRAGE sequence with the following parameters: repetition time (TR)/inversion time (TI) 2300/1100 ms, echo time (TE) 3 ms, 8° flip angle, field of view (FOV) 256 mm × 216 mm × 176 mm and a 1 mm isotropic resolution. Parallel imaging (iPAT = 2) was used to accelerate the acquisition resulting in an acquisition time of 5 minutes and 21 seconds. Whole-brain functional images were acquired using a multi-band (accelerator factor of 3) multi-echo T2*-weighted sequence with the following parameters: TR 1500 ms, TEs 13.4/34.8/56.2, flip angle 75°, FOV 84 mm x 84 mm x 64 mm, voxel size 2.5 mm isotropic. Fieldmap images were also acquired to correct for distortions along the phase-encoding axis. We acquired 12 fMRI runs per participant.

Preprocessing was performed using fMRIPrep 1.2.6-1 (for explicit details on fMRIPrep pipelines, see Chapter 2; Esteban, Blair, et al., 2018; Esteban, Markiewicz, et al., 2018; Giglio et al., 2022). Each brain was spatially normalised to the ICBM 152 Nonlinear Asymmetrical template version 2009c. Susceptibility distortion correction was performed on all BOLD runs, which were slice-time corrected. Motion artifacts were estimated with ICA-AROMA (Pruim et al., 2015), whose noise regressors were later added to the first-level design matrix, together with head-motion parameters, framewise displacement, DVARS (Power et al., 2014), and anatomical component-based noise correction (aCompCor, Behzadi et al., 2007).

fMRI connectivity analysis

Seed ROI selection

Two seed regions were selected for the connectivity analysis. These were the *pars triangularis* of the LIFG and a region within the left posterior temporal lobe (Fig. 3.1). For the LIFGtri ROI, we used an anatomical mask of the LIFGtri based on the Harvard Oxford cortical atlas (Desikan et al., 2006). For the LPTL ROI, we used a mask based of the posterior temporal part of the group conjunction effect of constituent size in production and comprehension in Chapter 2, since that region overlapped with the posterior MTG, the posterior STS and the posterior STG. We selected posterior voxels by selecting those that were posterior to Heschl's gyrus, according to the Harvard Oxford cortical atlas. For both ROI masks, we then extracted voxels with t values > 2 at the subject level for the main effect of constituent size based on Giglio et al. (2022), thus focusing on voxels that were found to respond to constituent size independent of modality in each subject.

Whole-brain gPPI analysis

A generalized psycholophysiological interaction analysis (McLaren et al., 2012) was conducted in AFNI (version 22.1.09; Cox, 1996) to assess connectivity between regions as a function of constituent size and modality. The procedure was as follows. First, we extracted the detrended timeseries for each seed ROI with AFNI programs 3dmaskave and 3dDetrend. The seed time series was then upsampled by a factor of 6 (0.25 TR) and deconvolved with 3dTfitter to calculate the neuronal response. The interaction timeseries were then obtained by masking the neuronal response time-course with a condition mask from each experimental condition. Condition masks had the value of 1 when a condition was present and 0 when a condition was absent. Interaction terms thus calculated for each condition were then convolved with the haemodynamic response function originally used in the first-level model, and then downsampled back to the original TR.

The interaction timeseries for each run, condition and modality and the seed timeseries were then included as additional regressors in the first-level design matrix (which was otherwise identical to the design matrix in Chapter 2). Two parallel analyses were run for the interaction timeseries extracted from each seed ROI (LIFGtri and LPTL). The preprocessed BOLD images in MNI standard space were smoothed with 4 mm FWHM Gaussian kernel in SPM12 (version 7771, Penny, Friston, Ashburner, Kiebel, & Nichols, 2011) in Matlab2021a. The first-level design matrix included for each run as condition regressors the interaction timeseries of each condition and the seed timeseries of one ROI, correct trials for each of the four conditions, incorrect trials, the temporal derivative of each condition and parametric modulations of speech onset times (incorrect trials and parametric modulations were absent in the comprehension runs). Trial onset was picture onset time, and trial duration was the time from picture onset until speech offset. As nuisance regressors we included DVARS, framewise displacement, 6 aCompCor parameters, 6 motion parameters and the AROMA noise components. Following parameter estimation, we extracted contrast images for each participant for the functional connectivity from each seed ROI independent of conditions, as well as for the main effect of constituent size of the interaction regressors (with weights [-4 -1 5] based on the constituent size of the conditions), main effect of modality (production vs. comprehension) and the interaction between constituent size and modality. The contrast images were then tested at the group level with one-sample T-tests following Henson (2015), with voxel-level threshold at $p = 0.001$ uncorrected, and a $p = 0.05$ family-wise error correction as a cluster threshold.

ROI gPPI analysis

In the ROI gPPI analysis, we assessed connectivity seeding from the LIFGtri and the LPTL and targeting the *pars opercularis* of the LIFG and the anterior temporal lobe (ATL), which were found to respond to constituent size in Chapter 2 (Giglio et al., 2022). We also tested the connectivity between the seed ROIs. A similar procedure was used to identify the target ROIs as was used for the identification of the seed ROIs. First, we selected masks for the LIFGoper and the LATL. For LIFGoper, we used a mask of the *pars opercularis* based on the Harvard Oxford cortical atlas. For the LATL, we used the voxels anterior to Heschl's gyrus in the conjunction effect of constituent size cluster in the temporal lobe at the group level based on Chapter 2 (Giglio et al., 2022). The ROIs were then restricted to include only those voxels that had a t value > 2 for the main effect of constituent size in each participant in each of these masks.

We then extracted mean beta values per participant for the interaction regressors for each condition and modality using MarsBar (Brett et al., 2002) in SPM12. We extracted the mean beta values in LIFGoper, LATL and LPTL for the LIFGtri interaction regressors; and the mean beta values in LIFGoper, LIFGtri and LATL for the LPTL interaction regressors. The beta weights were then compared with two separate mixed-effects models in R (version 4.0.3) for each seed ROI using *lme4* (version 1.1-26, Bates et al., 2015). We used a linear contrast with weights [-4 -1 5] for constituent size. For modality, we used deviation coding. We used Helmert coding for the ROI factor, contrasting the ROIs by lobes and within lobes (i.e. for the LIFGtri model, we contrasted LIFGoper to the LATL and LPTL, and the two temporal regions to each other; for the LPTL model, we contrasted the LATL to the LIFG ROIs, and LIFGoper and LIFGtri to each other). We added by-participant random slopes for the interaction for ROI and modality and for the main effect of constituent size. We computed the contribution of factors using *car* (version 3.1-0, Fox et al., 2021), and pairwise comparisons with *emmeans* (version 1.7.5, Lenth et al., 2022).

3.3 Results

ROI results

First, we focused on the connectivity between the two seed ROIs, LIFGtri and LPTL, and two target ROIs, LIFGoper and LATL (Fig. 3.1). Note that because the analysis included two seed ROIs, p values below 0.025 are significant (Bonferroni corrected), but we report uncorrected p values for completeness. The model of the seed LIFGtri region connectivity with LATL, LIFGoper and LPTL indicated a significant main effect of ROI ($\beta = 0.0018$, SE = 0.0002, $t = 7.37$, $\chi^2 = 56.4$, $p < 0.0001$), of constituent size ($\beta = 0.0002$, SE = 0.00005, $t = 3.69$, $\chi^2 = 13.6$, $p < 0.0003$) and a trend for an interaction between modality, ROI and constituent size ($\beta = 0.00006$, SE = 0.00003, $t = 1.8$, $p = 0.072$, p corrected = 0.146). Pairwise comparisons for the main effect of ROI indicated that the connectivity between LIFGtri and LATL was significantly lower than the connectivity of the LIFGtri with LIFGoper and LPTL (LIFGoper - LATL, LPTL - LATL:

estimates > 0.0037, SE < 0.0007, $t > 5.6$, $p < 0.0001$). The constituent size effect was present overall, but it was reduced with LIFGoper in comprehension, as indicated by the interaction. The slope for constituent size was marginally more positive in production than comprehension in LIFGoper (production – comprehension: estimate $= 0.00034$, SE $= 0.00016$, $t = 2.08$, $p = 0.038$, p corrected $= 0.076$), but did not differ in the other ROIs. Therefore, constituent size was related to an increase in the connectivity of the LIFGtri with all ROIs in both modalities. The connectivity between LIFGtri, LIFGoper and LPTL was higher than between LIFGtri and LATL.

The model for the seed LPTL region connectivity with LATL, LIFGoper and LIFGtri showed a similar pattern (Fig. 3.1B). There was a significant main effect of constituent size ($\beta = 0.00014$, SE $= 0.00005$, $t = 2.6$, $\chi^2 = 6.9$, $p = 0.008$), of modality ($\beta = 0.0011$, SE $= 0.0003$, $t = 3.29$, $\chi^2 = 6.8$, $p = 0.009$) and ROI ($\beta = 0.0023$, SE $= 0.00025$, $t = 9.3$, $\chi^2 = 86.9$, $p < 0.0001$). Again, the connectivity of the LPTL with the LATL was significantly lower than with LIFGoper and LIFGtri (LIFGoper – LATL, LIFGtri – LATL: estimates > 0.0063, SE < 0.00083, $t > 8.3$, $p < 0.0001$). The connectivity with LIFGoper was also marginally larger than with LIFGtri (LIFGoper – LIFGtri: estimate > 0.0012, SE < 0.00048, $t > 2.4$, $p < 0.055$, p corrected $= 0.11$). There was also a marginally significant interaction between modality and ROI ($\beta = 0.00039$, SE $= 0.00019$, $t = 2.02$, $p = 0.048$, p corrected $= 0.096$). Pairwise comparisons indicated that connectivity was significantly higher in production than comprehension with LIFGoper and LIFGtri (production – comprehension: estimates > 0.0028, SE $= 0.001$, $t > 2.73$, $p < 0.0095$, p corrected < 0.019), but not with the LATL (production – comprehension: estimate $= 0.0006$, SE $= 0.0007$, $t > 0.98$, $p = 0.33$). Therefore, the connectivity of the LPTL with all ROIs increased with constituent size, and was higher with the LIFG ROIs than with the LATL. In addition, the connectivity between LPTL and LIFG ROIs was higher in production relative to comprehension.

The ROI results thus indicate that the connectivity between and within frontal and temporal regions increased as a function of sentence processing. However, especially in the LPTL connectivity, the constituent size effect did not always seem linear upon inspection. Therefore, we explored whether a quadratic effect would better model the data. We ran the same models for the connectivity of the LPTL and the LIFGtri but now the contrast for constituent size included both a linear and a quadratic term. The connectivity of the LIFGtri in response to constituent size was significantly linear ($\beta = 0.0014$, SE $= 0.0003$, $t = 4.1$, $p =$

0.0001) but not quadratic ($\beta = 0.0005$, SE = 0.0004, $t = -1.5$, $p = 0.13$). The connectivity of the LPTL in response to constituent size instead was significantly linear ($\beta = 0.0011$, SE = 0.0004, $t = 2.98$, $p = 0.0048$) as well as quadratic ($\beta = 0.0013$, SE = 0.00005, $t = -2.6$, $p = 0.012$), but neither contrast interacted with modality or ROI. Therefore, although the effect is not as consistently linear as it was in Chapter 2, it is also not reliably quadratic in the connectivity with any ROI.

The results additionally showed that the LIFG and the LPTL form a stronger network that seems to connect with the LATL only for the sentence conditions (C2 and C4). Finally, the connectivity within the temporal lobe was not affected by modality, but the connectivity between LPTL and frontal regions, and less robustly within frontal regions, was increased during production. It should be noted that the increased connectivity between LPTL and LIFG in production was only found with LPTL as seed region, but not when LIFGtri was seed region as noted above. As gPPI does not allow for directed inferences, it is presently unclear what significance this result has for the description of this network. For example, the results might be explained by a network configuration where LIFGoper functions as an intermediate region between LPTL and LIFGtri, but potential explanations for this asymmetry cannot be adjudicated by the current experiment or analysis.

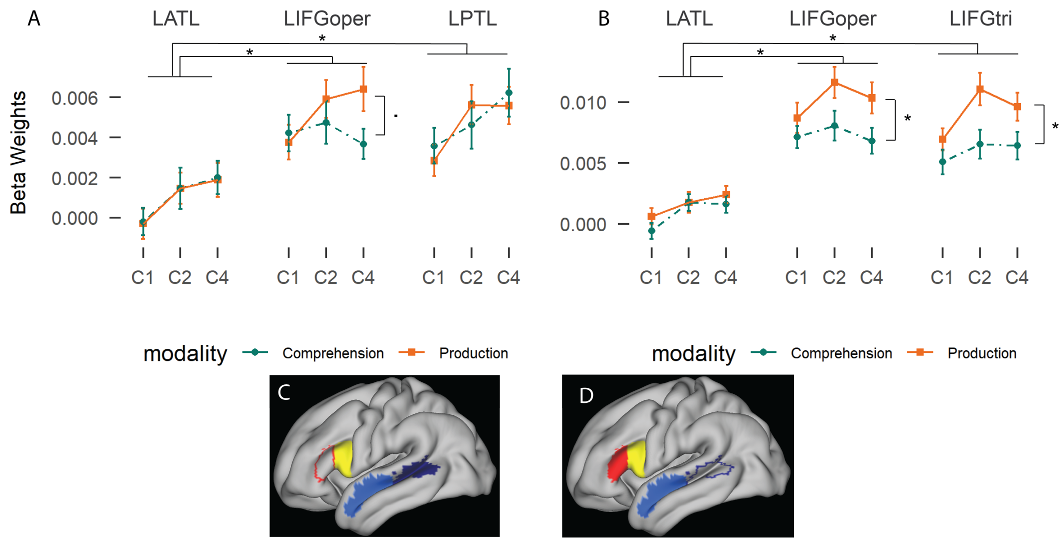


Figure 3.1: ROI gPPI results. Average beta estimates per condition, modality and ROI from seed ROI LIFGtri in **A** and LPTL in **B**. **C** and **D** show the regions used for ROI analysis, with the seed ROIs in outline. Red is LIFGtri, yellow is LIFGoper, dark blue is LPTL, light blue is LATL. Error bars represent standard error of the mean. * indicates $p < 0.05$, . indicates $p < 0.1$.

Whole-brain results

In addition to investigating the connectivity between regions in the frontal and temporal lobes, we investigated the connectivity of the LIFGtri and the LPTL with other areas of the brain. First, we inspected the functional connectivity of the LIFGtri independently of task effects (Fig. 3.2A, Table 3.1). We found the LIFGtri timeseries to be correlated with activity in the superior left and less strongly right IFGoper, with the posterior MTG bilaterally and with a cluster in the left inferior temporal gyrus, with the left inferior and superior parietal lobules and with the left supplementary motor cortex. These were all regions co-activated with the LIFG in the main effect of constituent size in Giglio et al. (2022, Chapter 2).

The left PTL was found to be correlated in activity with a large cluster in LIFGtri, LIFGoper and precentral gyrus, that also extended bilaterally, with the supplementary motor cortex, with the inferior and superior parietal lobules bilaterally (Fig. 3.2B). It also connected with a corresponding right posterior MTG cluster and with smaller clusters in the left posterior temporal lobe. There were also a few smaller clusters in left precuneus, mid cingulate cortex, left caudate and bilateral thalamus. Again, this network was partly overlapping with the constituent size network.

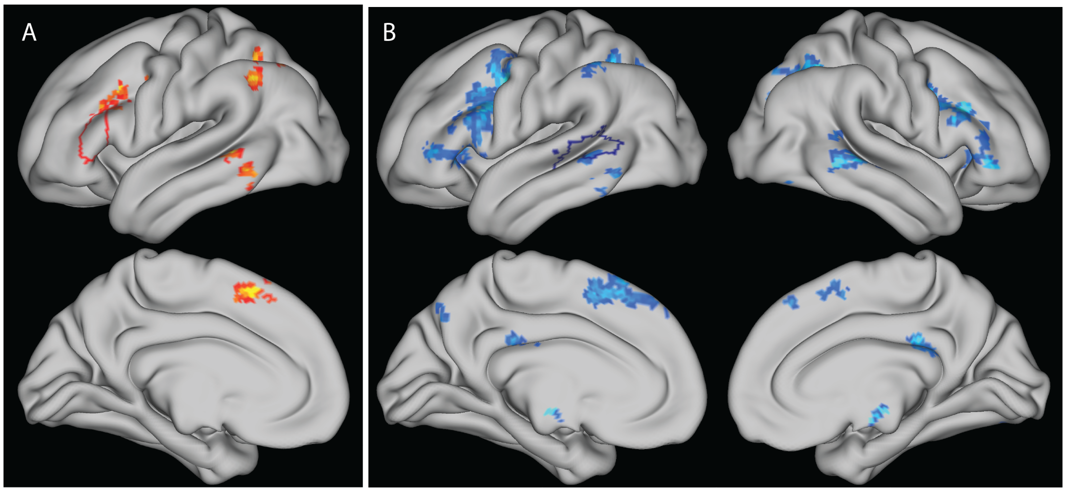


Figure 3.2: Whole brain task independent functional connectivity. Connectivity from LIFGtri seed in **A** in warm colors, and LPTL seed in **B** in cold colors. Seed ROIs are in outlined in red (LIFGtri) and blue (LPTL). Clusters are thresholded at $p < 0.05$ FWE-corrected.

We then assessed connectivity as a function of modality and constituent size. We found a main effect of modality for the LIFGtri seed, showing increased connectivity in comprehension over production to the left and right precentral to postcentral gyri (Fig. 3.3A). The LPTL seed instead was more strongly connected to the posterior occipital cortex in comprehension than in production (Fig. 3.3C). These comprehension effects seem to suggest task-related effects, possibly related to visual processing of the pictures.

The LPTL also had increased connectivity to a small cluster in the LIFGtri in production compared to comprehension, indicating that the temporal-frontal connectivity was upregulated during production, as already indicated by the ROI results (Fig. 3.3B). There was also increased connectivity to a right-lateralised cluster in the superior temporal cortex, possibly indicating monitoring needs in production (Kearney & Guenther, 2019). There were no significant clusters for the main effect of constituent size or for the interaction between constituent size and modality.

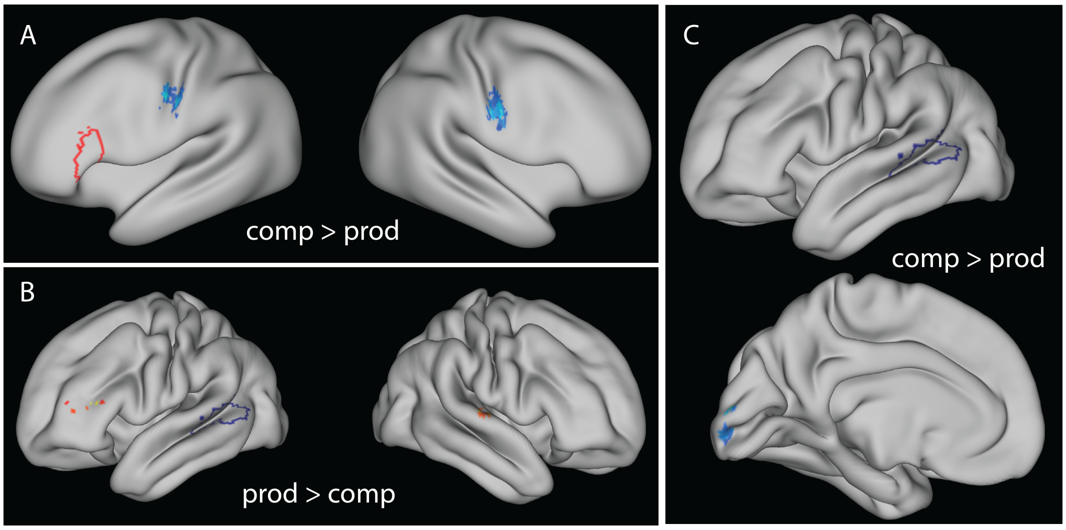


Figure 3.3: Whole brain results for the main effect of modality. From seed LIFGtri in **A** (on inflated brain to show extent of clusters) and seed LPTL in **B** and **C**. Seed regions are shown in outline in red and blue, while significant clusters (thresholded at $p < 0.05$ FWE-corrected) are colored in warm colors for production > comprehension effects, and in cold colors for comprehension > production effects.

3.4 Discussion

We analysed task-based functional connectivity among regions of the fronto-temporal language network to obtain further insight into the coupling of regions involved in linguistic processing as a function of modality. We investigated the connectivity of seed region LIFGtri with target regions LIFGoper and two temporal clusters, and to the whole-brain. We also investigated the connectivity of the left posterior temporal cluster with the anterior temporal cluster and the two

Table 3.1: fMRI whole-brain summary of cluster peak coordinates and statistics. P values below 0.025 are significant after correcting for multiple tests (from two seed ROIs), but all peaks are shown for completeness. P values are FWE-corrected.

Seed ROI	**Contrast**	**Cluster**		**Peak Voxel (MNI)**				**Peak Anatomical Location**
		p	**Size**	***t***	**x**	**y**	**z**	
LIFGtri	Functional Connectivity	0	85	6.07	-56	-57	44	L Inferior Parietal Lobule
		0	84	5.79	-4	18	54	L Supplementary Motor Area
		0	207	5.5	-41	8	22	LIFGoper
		0	55	5.37	59	-40	0	RMTG
		0	161	5.25	-34	-64	44	L Superior Parietal Lobule
		0	150	5.05	-66	-50	-8	LMTG
		0.041	25	4.66	32	-67	44	R Angular Gyrus
		0	55	4.47	54	20	27	RIFGtri
		0	52	4.42	-28	-74	30	L Middle Occipital Gyrus
LPTL	Functional Connectivity	0	953	7.41	-38	6	24	LIFGoper
		0	121	6.58	66	-42	-3	RMTG
		0	293	6.51	29	-54	44	R Angular Gyrus
		0	248	6.23	-4	23	62	L Supplementary Motor Area
		0	372	6.19	-31	-47	37	L Inferior Parietal
		0.037	26	5.86	2	-12	-8	Thalamus
		0	97	5.85	-58	-57	-10	L Inferior Temporal Gyrus
		0	456	5.76	44	23	24	RIFGtri
		0.003	41	5.61	14	-14	10	R Thalamus
		0	61	5.05	2	-32	30	R Mid Cingulate Cortex
		0.005	37	4.96	-14	-2	17	L Caudate
		0.005	38	4.96	-11	-10	10	L Thalamus
		0.001	46	4.85	-66	-47	4	LMTG
		0	71	4.67	-51	-44	47	L Inferior Parietal
		0.031	27	4.67	32	26	-3	R Insula
		0.045	25	4.34	-54	30	20	LIFGtri
		0.018	30	4.23	-4	-72	42	L Precuneus
		0.045	25	4.19	39	-72	-13	R Inferior Occipital Cortex
		0.026	28	3.9	54	-44	10	RMTG
LIFGtri	Comprehension - Production	0	83	4.67	54	-7	30	R Postcentral Gyrus
		0	64	4.62	-48	-10	37	L Precentral Gyrus
LPTL	Comprehension >Production	0	55	5.07	-6	-100	12	L Occipital Pole
LPTL	Production > Comprehension	0.006	37	5.49	-44	50	12	L Anterior Middle Frontal Gyrus
		0.006	37	4.61	62	-12	7	RSTG
		0.035	27	4.58	-28	8	62	L Posterior Middle Frontal Gyrus

LIFG regions. All of these regions were previously found to be modulated by constituent size during production as well as comprehension (Giglio et al., 2022). We found that the functional connectivity among these regions was enhanced for larger syntactic constituents. In addition, the connectivity between the posterior temporal lobe and the inferior frontal gyrus was enhanced in production relative to comprehension. The anterior temporal cluster, while showing enhanced connectivity to LPTL and LIFGtri for the sentence conditions, was relatively less coupled to the LPTL and LIFGtri than they were connected to each other and to the LIFGoper.

The connectivity between the LPTL and the LIFG was upregulated with larger constituent size and in production over comprehension, and it was greater than the connectivity between either region and the LATL. This result highlights a dorsal network that is functionally connected in both modalities as a function of constituent structure building. The connectivity within the dorsal stream thus increases with larger constituents in both production and comprehension. These regions are connected via the dorsal stream through the arcuate fasciculus (Dick et al., 2014; Friederici, 2012), that is found to mature in the developing brain in relation with syntactic abilities (Skeide, Brauer, & Friederici, 2016). In addition, the functional coupling between LIFG and LPTL has been found to be critical for syntactic and compositional processing (den Ouden et al., 2012; Griffiths, Marslen-Wilson, Stamatakis, & Tyler, 2013; Sharoh, Ruijters, Weber, Norris, & Hagoort, submitted). When disrupted in patients, it was seen to lead to syntactic impairments (Papoutsi et al., 2011). These results thus show that the functional network for generating and parsing constituents of increasing complexity is common between production and comprehension.

Our results are agnostic as to the origin and direction of the connectivity between these areas. One study that used dynamic causal modelling to generate predictions about the directionality of connections in comprehension found evidence for a model where the LIFG was the driving input in the connectivity with the LPTL, and syntactic complexity modulated the connectivity in this direction (den Ouden et al., 2012). Another fMRI study instead found that sentence complexity in comprehension drove an increase in the LPTL earlier than in the LIFG (Uddén et al., 2022). In MEG the connectivity between fronto-temporal regions was found to be driven by outflow from middle temporal regions, with frontal regions receiving input and sending output to other frontal regions (Hultén, Schoffelen, Uddén, Lam, & Hagoort, 2019; Schoffelen et al., 2017). Another MEG study found that in verb-object compositionality the LMTG affected responses in LIFG, which in turn connected back to LMTG (Lyu et al., 2019), highlighting bidirectional connectivity. These findings thus are consistent with the hypothesis that the LPTL and LIFG interact in a dynamically reverberating network that supports the integration of memory and unification components of language processing following the MUC model (Baggio & Hagoort, 2011; Hagoort, 2013). Our results, therefore, add to the growing literature on the interconnectivity of frontal and temporal regions by extending the evidence to sentence production. As the current gPPI analysis does not allow for directionality inferences, it is left open whether modality would affect the directionality of interactions. Based on the different input/outputs and goals and requirements of production and comprehension (Gambi & Pickering, 2017; Momma & Phillips, 2018), there could be differences in the interaction dynamics between modalities, but they may nevertheless be too fast to be detectable in fMRI if the network is as dynamically reverberating as suggested (Lyu et al., 2019).

The connectivity between the posterior temporal lobe and the inferior frontal gyrus was increased in production relative to comprehension. This enhanced coupling in production highlights an effect of modality in the connectivity of these core language regions. Interestingly, this modality effect complements the modality asymmetry found in Chapter 2, where the LIFG was more active in production than comprehension, while the LPTL was more active in comprehension. The modality activation imbalance is thus potentially related to the modality-affected coupling between posterior temporal and inferior frontal regions. In fact, the increased connectivity in production may upregulate the engagement of the LIFG. It is important to note that the lack of an interaction between modality and constituent size indicates that the modality effect in LIFG-LPTL connectivity was not related to the complexity of syntactic processing. Rather, it may reflect production-specific computations, such as coordination of sequences for phonetic encoding and articulation (Hickok & Poeppel, 2007). Sentence production ends with a motor output that is coordinated by premotor regions in the frontal cortex and may rely on inferior frontal regions for linearization, while comprehension, especially in the current design, does not culminate in an action (Kearney & Guenther, 2019; Matchin & Hickok, 2020; Ries et al., 2019). This functional production network seems to further rely on the enhanced connectivity between LIFGtri and LIFGoper in the production of larger constituents, possibly suggesting the stronger need to coordinate a sentence-level linked articulatory plan (Basilakos et al., 2018; Flinker et al., 2015; Hickok & Poeppel, 2007).

It was remarkable that, independently of modality, the LATL was less coupled to the LPTL-LIFG network, while being similarly sensitive to constituent size as the other regions. This region is connected to the LIFG via the ventral stream, thought to be involved in semantic processing, distinct from the dorsal stream involved in phonological and syntactic processing (Dick et al., 2014; Friederici, 2012; Hickok & Poeppel, 2007). The reduced connectivity with posterior regions may thus reflect its being part of not only a different anatomical but also functional network. However, its connectivity was enhanced for larger constituents in this region, which suggests that it was nevertheless involved in the large network engaged for language processing in this experiment (that also included parietal and middle frontal regions). The increased response and connectivity of the LATL cluster for larger constituents could be due to its involvement in the access of phrase-structure templates (Friederici, 2012), or in larger semantic composition needs for larger sentences (Bornkessel-Schlesewsky & Schlesewsky, 2013; Hickok & Poeppel, 2007; Matchin & Hickok, 2020; Pylkkänen, 2020).

The constituent size effect, though reliable in the ROI analysis, did not surface at the whole-brain level. Inspection of the slopes for the response to constituent size suggests that the lack of change in connectivity at the whole-brain level could be due to the effect not always being linear. The finding of a quadratic as well as a linear response to constituent size in the connectivity seeding from the LPTL supports this interpretation. In fact, in some ROIs the constituent size effect somewhat decreased after C2, suggesting that rather than being sensitive to constituent size itself, the connectivity may be enhanced with sentences (C2 and C4) relative to word lists (C1), again indicating that functional connectivity increases with the interaction of memory and unification (Hagoort, 2014), but not necessarily as a function of increasing syntactic complexity. Since the shape of the connectivity in response to constituent size was not reliably different between ROIs, it is presently unclear whether the linear and quadratic effects reflect distinct features of the connectivity between fronto-temporal regions.

The whole-brain results for modality were partly surprising, identifying a network for comprehension that was not expected a priori. In fact, the increased connectivity in comprehension of the LIFGtri with a bilateral cluster bordering the central sulcus, and the connectivity of the LPTL with an occipital cluster seemed to be related to task-specific effects rather than higher-level linguistic processing. The greater connectivity between LIFGtri and precentral gyrus would rather be expected during production, where motor plans need to be prepared. However, it is possible that the reduced connectivity in production may be due to motor plans being part of a more complex network, which reduces the direct correlation between anterior parts of the LIFG and precentral cortex, where LIFGoper acts as an intermediate region (Flinker et al., 2015; Hickok & Poeppel, 2007; Kearney & Guenther, 2019). Indeed, articulation is thought to depend on the interaction between more posterior regions of LIFG, such as LIFGoper, and motor cortex (Hickok & Poeppel, 2007). An intracranial EEG study on object naming found negative correlations between LIFGtri and a region of the motor cortex prior to articulation (Conner, Kadipasaoglu, Shouval, Hickok, & Tandon, 2019). It is thus possible that in the current study a positive interaction between LIFGoper and the motor cortex during production co-occured with a negative interaction between motor cortex and LIFGtri. This negative correlation in production may have then manifested as a positive interaction during comprehension.

Similarly, the increased temporo-occipital connectivity in comprehension found in the whole-brain analysis was not anticipated on the basis of the linguistic manipulations used in this experiment. It might be due to there being sustained attention on the visual input during comprehension compared to production, because of lower-level task demands, such as matching auditory input with visual input. This direct connectivity may have been reduced in production again due to a more complex network being involved in producing the sentence. Visual and linguistic information may have been integrated in the ventro-occipitotemporal cortex for reading in production (Sharoh et al., 2019), but not in comprehension where reading was not necessary for the task.

In summary, we found a dynamic posterior temporal – inferior frontal network to be upregulated for syntactic structure building in both production and comprehension, suggesting similar dynamics of sentence processing in both modalities and reiterating the importance of the dorsal stream for syntactic processing. This posterior temporal – inferior frontal connectivity was upregulated in production possibly due to production-specific linearization and articulatory needs. A mid-anterior temporal cluster was also upregulated for constituent structure building but it was functionally less co-activated with the posterior temporal lobe and the inferior frontal gyrus.

Chapter 4

4 Diverging Neural Dynamics for Syntactic Structure Building in Naturalistic Speaking and Listening[1]

Abstract

The neural correlates of sentence production have been mostly studied with constraining task paradigms that introduce artificial task effects. In this study, we aimed to gain a better understanding of syntactic processing in spontaneous production vs. naturalistic comprehension. We extracted word-by-word metrics of phrase-structure building with top-down and bottom-up parsers that make different hypotheses about the timing of structure building. In comprehension, structure building proceeded in an integratory fashion and led to an increase in activity in posterior temporal and inferior frontal areas. In production, structure building was anticipatory and predicted an increase in activity in the inferior frontal gyrus. Newly developed production-specific parsers highlighted the anticipatory and incremental nature of structure building in production, which was confirmed by a converging analysis of the pausing patterns in speech. Overall, the results showed that the unfolding of syntactic processing diverges between speaking and listening.

---

[1] This chapter also appears in Giglio, L., Ostarek, M., Sharoh, D., & Hagoort, P. (2022). *Diverging Neural Dynamics for Syntactic Structure Building in Naturalistic Speaking and Listening*. bioRxiv. `https://doi.org/10.1101/2022.10.04.509899`

4.1 Introduction

Studies on the neurobiology of language typically use highly controlled experimental paradigms far removed from typical language experience in everyday life. However, in the last decade, there has been a shift towards naturalistic studies of language processing. This shift has happened in multiple directions, from the use of virtual environments to increase ecological validity (Huizeling, Peeters, & Hagoort, 2021; Peeters, 2019), to the auditory presentation of audiobooks or narrative reading with neuroimaging (Alday, Schlesewsky, & Bornkessel-Schlesewsky, 2017; Brennan, 2016; Heilbron, Armeni, Schoffelen, Hagoort, & de Lange, 2022; Willems & Gerven, 2018). The increased ecological validity in naturalistic studies opens a window into language processing free of the artificiality of task designs, whose main goal is to isolate specific features of language (Andric & Small, 2015). In traditional settings, experimental control comes at the cost of context, which is heavily reduced to minimize confounds. This contrasts with the highly contextual nature of everyday language use, creating a large gap between the actual object of study and its realization in experiments. Combining naturalistic stimuli and advanced analysis methods, such as audiobooks and probabilistic parsers, has the potential to bring the two much closer together (Brennan, 2016; Hale et al., 2022).

The overwhelming majority of studies on the neurobiology of language is on comprehension, while speaking is largely unexplored. Importantly, while naturalistic studies are becoming more common in the field of language comprehension, studies of naturalistic *production* are still lacking. This is problematic because the gulf between spontaneous production and production in controlled experiments is particularly large. In spontaneous language production, the speaker is by definition in control of what to say. In contrast, experimental paradigms attempt to have as much control over participants' speech as possible. This has usually been achieved with picture description experiments or with the use of visual probes together with written linguistic stimuli (e.g. Giglio et al., 2022; Z. M. Griffin & Bock, 2000; Matchin & Hickok, 2016; Takashima et al., 2020). While these strategies have allowed for controlled investigations of linguistic processing, they may be confounded with the heavy task requirements that make controlled production very removed from everyday speaking.

In this fMRI study, we aimed to study syntactic processing in spontaneous production and comprehension in order to understand whether and how they differ. Previous studies found shared neural resources for production and comprehension (Menenti et al., 2011; Segaert et al., 2012) and a similar network for processing syntactic complexity across modalities (Giglio et al., 2022; Hu et al., 2022). At the same time, production was found to elicit larger responses to syntactic complexity than comprehension, especially in the left inferior frontal gyrus (LIFG) (Giglio et al., 2022; Hu et al., 2022; Indefrey et al., 2004). The differential sensitivity to complexity between modalities may be due to two main factors: 1) Speaking is a form of action, unlike the more passive process of listening. The message needs to be fully and correctly encoded into a linear sequence that results in articulation (Bock, 1982; Garrett, 1980, 1982; but see Goldberg & Ferreira, 2022). In comprehension, instead, one may choose how much attention to pay to the linguistic input, and sometimes may fail to correctly parse ambiguous input (Ferreira, 2003; Ferreira et al., 2002; for discussion, see Ferreira & Lowder, 2016). 2) Alternatively, previous studies may have been affected by unequal task requirements between modalities. The tasks used to elicit controlled production are usually more artificial than the respective tasks in comprehension, where only listening is expected, even if the same stimuli are used. To reduce task differences, here we focused on the neural response of syntactic processing in spontaneous production and compared it to the neural response to comprehension of the same materials. To the best of our knowledge, this is the first fMRI study of syntactic processing with unconstrained speaking.

To model linguistic processing in spontaneous production, we used word-by-word indices that were previously used to successfully predict brain activity in comprehension (e.g. Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016; Lopopolo, van den Bosch, Petersson, & Willems, 2021; Nelson et al., 2017). Widely used continuous indexes of linguistic processing include measures of word surprisal that show the brain's sensitivity to predictability (Henderson et al., 2016; Shain, Blank, van Schijndel, Schuler, & Fedorenko, 2020; Willems, Frank, Nijhof, Hagoort, & van den Bosch, 2016); continuous measures of syntactic tree building (Brennan, Dyer, Kuncoro, & Hale, 2020; Brennan & Hale, 2019; Brennan et al., 2016; Lopopolo et al., 2021; Nelson et al., 2017; Stanojević et al., 2021); and word embeddings modelling semantic features (Wehbe et al., 2014). Focusing on syntactic processing, increasingly sophisticated approaches that characterize continuous structure building have highlighted that left fronto-temporal regions are sensitive to measures of syntactic processing (Bhattasali et al., 2019; Brennan et al., 2020, 2016; Li & Hale, 2019; Nelson et al., 2017; Stanojević et al., 2021). We asked how syntactic processing, modelled with continuous metrics of syntactic tree building, affects brain activity in production and comprehension.

In the current study, we compared two parser models, a top-down and a bottom-up parser strategy. These strategies account for the same structure but make different hypotheses about the timing of syntactic operations. Top-down parsers build nodes at phrase-opening in an anticipatory fashion, whereas bottom-up parsers build nodes at phrase-closing in an integratory fashion. Here, we hypothesized the timing of operations to be the critical difference between production and comprehension, due to the different requirements and inputs of each modality (Momma & Phillips, 2018). In production, the speaker has some knowledge about the upcoming structure, since the structure related to the words that are uttered must have been computed (Bock & Levelt, 1994; Levelt, 1989). In comprehension, instead, after accounting for predictable continuations, listeners need to wait for the input to fully compute the structure. We thus expected the more anticipatory top-down parser to better predict neural activity in production and bottom-up parsing to better predict neural activity in comprehension. In a follow-up exploratory analysis, we explored whether alternative parsing strategies may be more fitting for production, since the parser models discussed so far were developed for language processing in comprehension specifically. In particular, we developed two parsers that assume different levels of incremental processing, by making different predictions about how early phrase-structure building operations occur.

Finally, we investigated which regions responded to syntactic processing load in each modality. In particular, we focused on two parts of the LIFG (*pars opercularis* or BA44, and *pars triangularis* or BA45) and on the left posterior middle temporal gyrus (LpMTG). These regions were all found to be responsive to syntactic manipulations in both modalities (Hagoort & Indefrey, 2014; Lopopolo et al., 2021; Pallier et al., 2011; Snijders et al., 2009; Takashima et al., 2020; Zaccarella, Meyer, et al., 2017; Zaccarella, Schell, & Friederici, 2017), sometimes with differences in their sensitivity to each modality (Giglio et al., 2022; Indefrey, 2018; Matchin & Wood, 2020). In particular, the LIFG was found to be more responsive to syntactic complexity in production than comprehension (Giglio et al., 2022; Indefrey et al., 2004), while the LpMTG was more responsive during comprehension (Giglio et al., 2022; Matchin & Wood, 2020). We thus investigated to what extent these regions are sensitive to continuous metrics of syntactic processing, and whether the modality differences observed in earlier studies (Giglio et al., 2022; Indefrey, 2018; Matchin & Wood, 2020) are replicated in naturalistic settings.

4.2 Results

Incremental metrics of phrase-structure building

To obtain incremental metrics of syntactic processing, we proceeded in two steps. First, we extracted the constituent structure of each sentence with a probabilistic context-free phrase-structure grammar (Stanford parser, D. Klein & Manning, 2003). From the extracted constituent parse, we then computed the parser operations carried out at each word according to different parsing models (Hale, 2014). These parsers incrementally build the syntactic structure of a sentence following different strategies, leading to a hypothesized number of phrase-structure building operations that need to be carried out at each word (Hale, 2014). This results in an incremental complexity metric that corresponds to the number of nodes that are built with each word. A top-down strategy builds the phrase structure from the top of the tree to a given word, such that it predicts increased activity when phrases are opened. In comprehension, it sometimes anticipates nodes before they are unambiguous to the listener, for example in the presence of adjuncts. Bottom-up parsing instead builds the phrase structure only after all the evidence has been seen, that is, after all words attached to each node have been met. It thus predicts increased activity when phrases are closed. Ultimately, both strategies lead to the same node count, but they make different predictions about the timing of syntactic operations and thus of corresponding neural activation (see Methods for more details, Fig. 4.11A-B).

We also quantified the load of processing complexity on working memory with an *open nodes* measure. This measure counts the number of nodes that have been opened (i.e. counted by the top-down strategy) but have not been closed yet (i.e. counted by the bottom-up strategy), tracking the numbers of words that need to be kept in working memory until they can be merged in a constituent (Nelson et al., 2017). In other words, this complexity metric tracks how much of the hypothesized structure needs to be confirmed by upcoming input. We expected this index to predict an increase in activity in comprehension, following Nelson et al. (2017) and Uddén et al. (2022). In production, it would also lead to an activity increase if speakers kept track of the structure that needs to be closed. Finally, to make sure that the syntactic predictors did not simply track word probabilities based on context, we quantified word surprisal from transformer model GPT-2 (Radford et al., 2019).

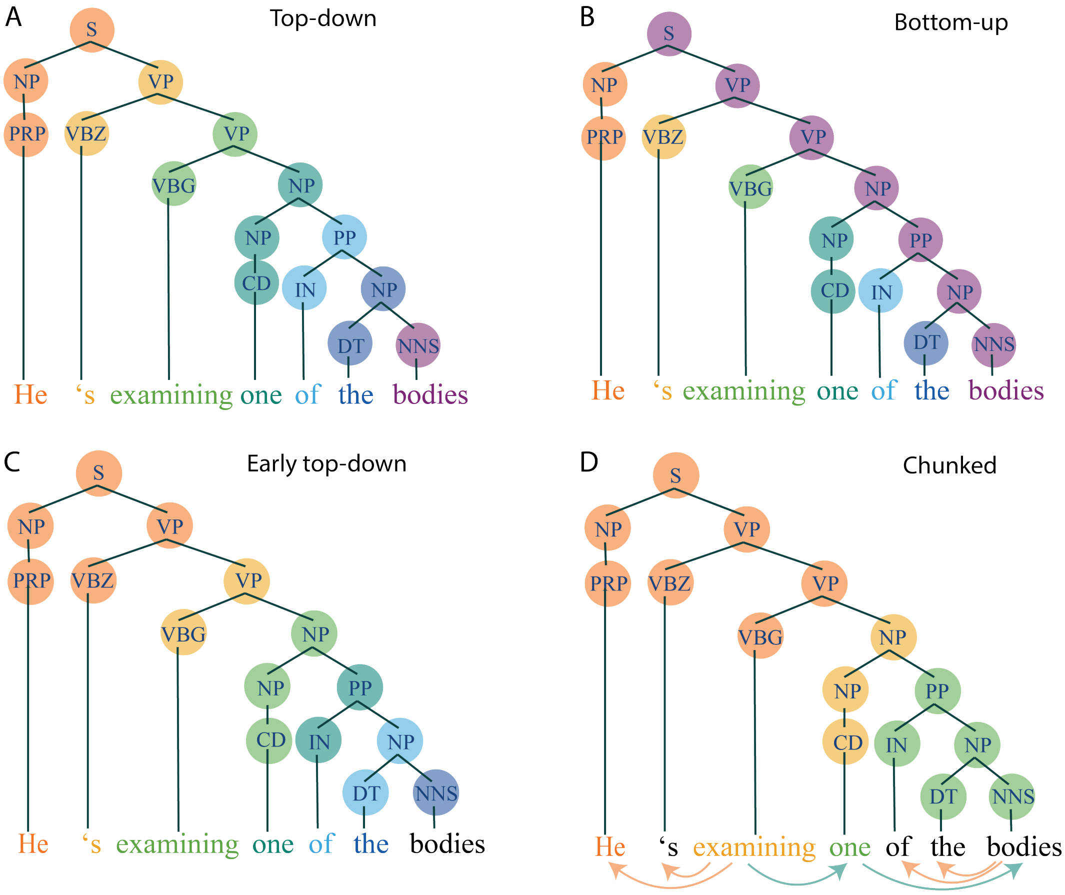


Figure 4.1: Node counting following different parsing strategies. The colored circles refer to the nodes that are built at the time point the word in the same color is uttered or heard. **A**: Colored representation of top-down phrase structure building, with nodes counted from the top of the tree to the word. **B**: Colored representation of bottom-up phrase structure building, with nodes counted from the bottom of the tree (i.e. the terminal nodes) to the top. Only nodes where both daughter nodes have been already met can be counted at each word. **C**: Colored representation of *early top-down* phrase-structure building, assuming operations to take place before word onset (production-specific). **D**: Colored representation of *chunked* phrase-structure building, following a less incremental strategy (production-specific). This node counting strategy is chunked based on the heads of the dependency parse of the same sentence (shown by the arrows below words, also see Supplementary Fig. S4.1). Heads are words from which an arrow originates. The nodes of the same constituent structure used by the other strategies are counted here. The chunked nature of this parser results in phrase-structure building operations to be assigned to some but not all words in a sentence. Black words are words that are not assigned any phrase-structure building operation (e.g. sentence-final words).

Distinct dynamics for phrase-structure building in language production vs. comprehension

To directly compare the word-by-word predictors with BOLD activity with a 1.5 second resolution (thus including several words at each fMRI volume), we convolved the linguistic predictors with the haemodynamic response function and resampled it to the 1.5 repetition time (see Methods for more details, Fig. 4.2). We then regressed average BOLD activity in BA44, BA45 and LpMTG in subject space against the predicted timeseries for each linguistic predictor with linear mixed-effects models.

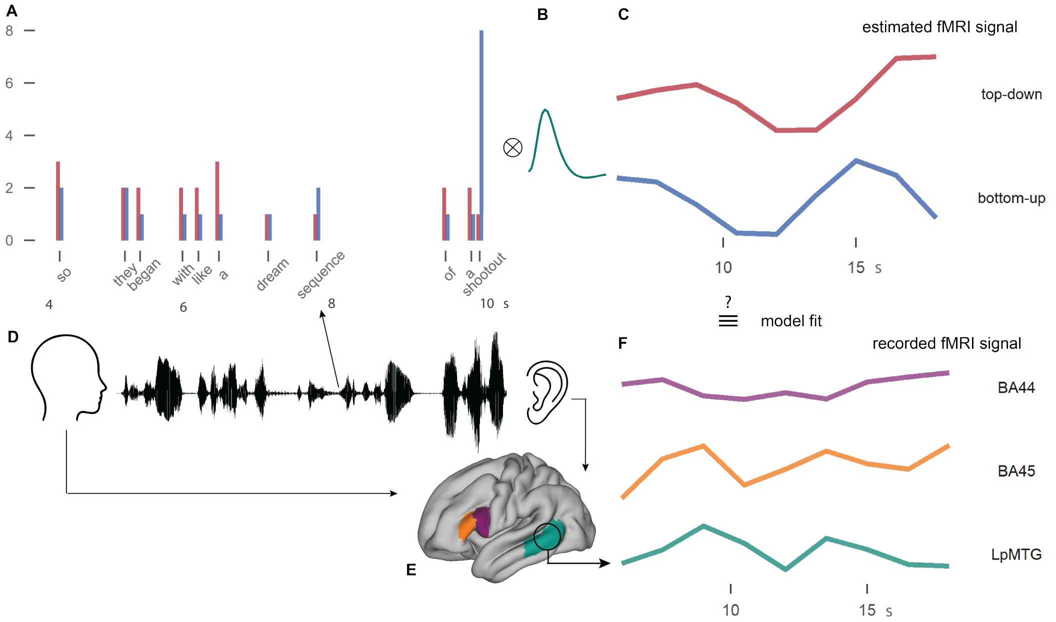


Figure 4.2: Graphical representation of the analysis paradigm to relate word-by-word predictors of linguistic complexity to BOLD activity. **A**: Word-by-word predictors of syntactic complexity were extracted from the constituent structure of the sentence spoken by a participant and listened to by other participants (**D**). The height of the bars in **A** represents the number of phrase-structure building operations expected to take place at each word following top-down and bottom-up parsing strategies (e.g. at “so” 3 nodes are counted for top-down, 2 for bottom-up). The weights of the syntactic predictors were convolved with the haemodynamic response function (**B**) to get predictor timeseries of BOLD activity at 1.5 sec resolution (**C**). These predictors timeseries were then compared to the average BOLD activity (**F**) in the brain of the speaker or the listener (**D**) in the three regions of interest (BA44, BA45 and LpMTG, **E**).

First, we ran a linear model of phrase-structure building operations on neural activity in BA44, BA45 and LpMTG (Fig. 4.3). The model included word rate, syllable rate, word frequency, word surprisal, top-down, bottom-up and open nodes, language modality and ROI (Supplementary Table S4.1). Word rate, word frequency and word surprisal significantly predicted an increase in BOLD activity (βs > 0.11, ts > 2.5, ps < 0.015). Syllable rate also marginally significantly predicted an increase in BOLD activity. More words, longer and less predictable words thus led to an increase in BOLD, while less frequent words, which should be harder to process, led to a decrease in activity. The effect of modality was also marginally significant ($\beta = 0.008$, SE $= 0.004$, $t = 1.8$, $\chi^2 = 3.3$, $p < 0.068$), with production having more positive activity than comprehension. The effect of modality did not interact with the effect of ROI.

Larger word surprisal elicited an increase in BOLD in both modalities (Fig. 4.3, $\beta = 0.16$, SE $= 0.02$, $t = 6.6$, $\chi^2 = 50.7$, $p < 0.0001$). This effect interacted with ROI ($\beta = 0.08$, SE $= 0.02$, $t = 3.6$, $\chi^2 = 15.4$, $p < 0.0005$), since BA44 responded significantly less to surprisal than BA45 and LpMTG (estimates > 0.1, $p < 0.03$) in both modalities. Open nodes also had a significant effect on BOLD activity (Fig. 4.3, $\beta = 0.031$, SE $= 0.015$, $t = 2.1$, $\chi^2 = 8.9$, $p < 0.003$). The effect interacted with modality and ROI ($\chi^2 = 8.9$, $p < 0.015$). It was significant only in comprehension in BA45 and LpMTG (estimates > 0.84, $p < 0.001$), while the estimates approached zero in all ROIs in production. Open nodes track the number of nodes to be kept in working memory until they can be integrated. It thus seems that the amount of structure that needs to be kept in working memory to be confirmed with the input leads to a brain activity increase in comprehension, but not in production.

We next determined whether incremental metrics of phrase-structure building significantly predicted brain-activity in BA44, BA45 and LpMTG. Both top-down and bottom-up parsers added significant contributions to the model, in interaction with modality and ROI (two-way interactions: top-down, $\chi^2 = 10.4$, $p < 0.006$, bottom-up, $\chi^2 = 12.7$ $p < 0.002$). Anticipatory top-down node counts predicted a significant increase in activity in production only in BA45 (estimate $= 0.49$, SE $= 0.2$, $p < 0.02$), while they predicted a decrease in comprehension in BA45 and LpMTG (BA45: estimate $= 0.28$, SE $= 0.15$, $p = 0.063$, LpMTG: estimate $= 0.8$, SE $= 0.15$, $p < 0.0001$). The response to top-down node counts was significantly different among all ROIs in comprehension (difference estimates > 0.58, SE$= 0.23$, ps < 0.03), but not in production (difference estimates < 0.5, SE$= 0.28$, ps > 0.27). Integratory bottom-up node counts predicted an opposite

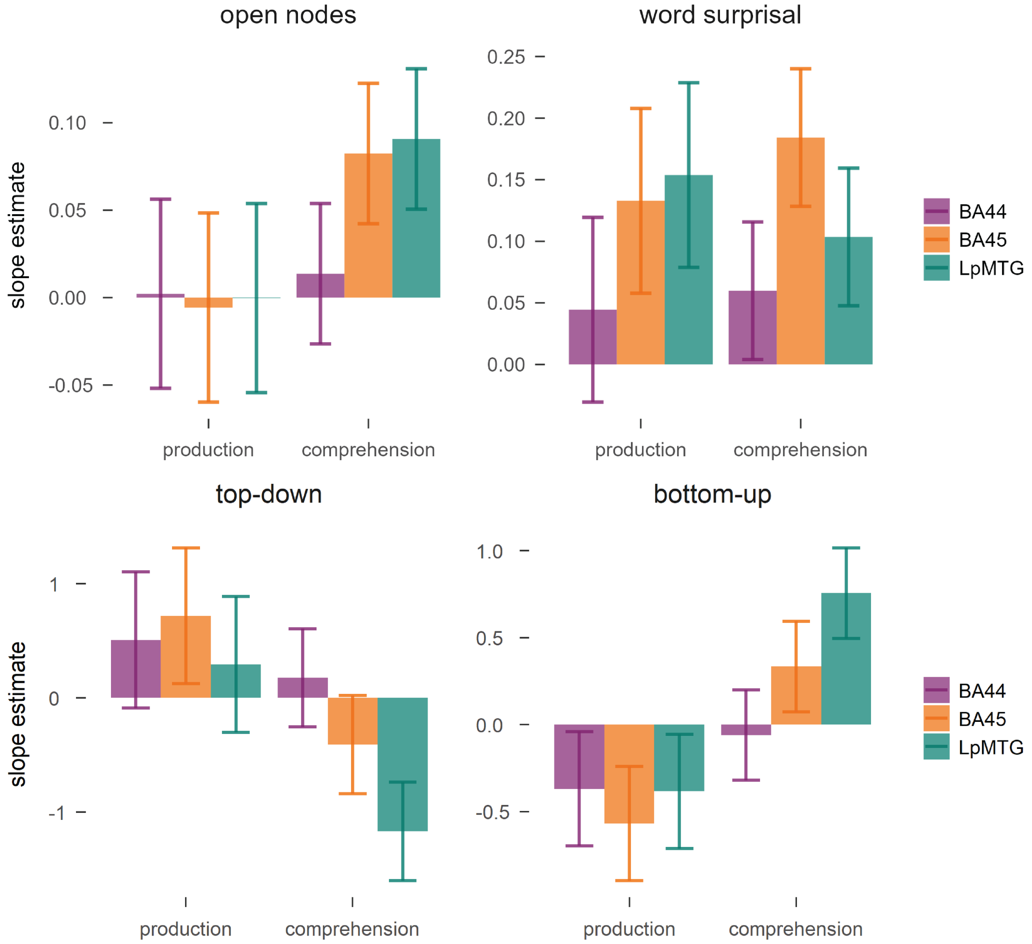


Figure 4.3: Beta estimates for the effect of each predictor of phrase-structure building, open nodes and word surprisal on BOLD activity in the regions of interest. Error bars represent 95% confidence intervals.

pattern of results. Larger bottom-up counts led to a significant decrease in activity in production in all ROIs (estimates > 0.44, SE $= 0.2$, $p < 0.03$), while they predicted an increase in comprehension in BA45 and LpMTG (BA45: estimate $= 0.4$, SE $= 0.16$, $p = 0.012$, LpMTG: estimate $= 0.9$, SE $= 0.15$, $p < 0.0001$). Again, ROIs responded differently to bottom-up counts across modalities. In comprehension, the strongest response was in LpMTG (difference estimates $>$ 0.39, SE $= 0.16$, $ps < 0.035$), while in production the responses were not significantly different among ROIs (estimates < 0.21, SE $= 0.18$, ps > 0.5). Therefore, activity in BA45 increased for different dynamics of phrase-structure building in production and comprehension, while the LpMTG positively responded only to comprehension. Activity in LpMTG and BA44 instead did not respond to anticipatory structure building in production, but decreased for integratory structure building, suggesting that both BA44 and LpMTG were active during sentence production, although not at the latencies predicted by the top-down parser, but their activity decreased at phrase-closing.

The parsers thus revealed marked differences between language production and comprehension. Anticipatory node counts led to an increase during production, but a decrease during comprehension, suggesting that in production syntactic structure building dominates at phrase opening. In comprehension, instead, neural activity may be reduced at phrase opening and may instead increase after phrase opening, when the top-down parser predicts a decrease in activity. This interpretation is confirmed by the activity increase in comprehension with integratory node counts, that predict increased activity at phrase closing. The decrease in activity for bottom-up counts in production suggests that at phrase closing syntactic processing load is reduced. Neural activity thus increases with syntactic structure building in both modalities, although with different dynamics, in BA45 in both production and comprehension, in LpMTG in comprehension only.

Phrase-structure building in production proceeds in a highly incremental fashion

The parsing strategies mentioned so far were developed for comprehension. This is problematic because these parsers assign linguistic operations *at the time* a word is said. This is a reasonable assumption in comprehension, where processing follows the input. However, in production, once the word is articulated, the associated grammatical and lexical encoding has to have taken place already (e.g. Bock & Levelt, 1994; Indefrey et al., 2004). We thus explored two production node building strategies that could better account for the timing of syntactic encoding in production. In both, the syntactic structure related to a word was assumed to be built at the latest while the previous word was articulated, but the two strategies made different predictions about the incrementality of structure building. An *early top-down* model predicted structure building to happen as the previous word is uttered (i.e. at each word we counted the nodes associated with the following word, see Methods for more details, Fig. 4.1C). This strategy leads to an equally incremental node building strategy as the original top-down strategy, but that, critically, happens earlier, more in line with theories of word production (Indefrey et al., 2001, 2004).

While more fitting for production in terms of timing, this view presupposes a highly incremental syntactic encoder. However, studies of sentence planning in production have long debated whether planning is linearly or hierarchically incremental, that is, whether the structure is built from each concept separately or from the relations between concepts (Bock & Ferreira, 2014). Hierarchical models of sentence production consider the verb to be the central node for the syntactic structure (Bock & Levelt, 1994; Levelt, 1989), suggesting that planning proceeds less incrementally. We thus explored whether a less incremental parser would better account for brain activity than a word-by-word incremental parser. We developed a node building strategy that counted all the nodes between words that were identified as heads according to dependency parsing (see Methods for more details, Fig. 4.1D and Supplementary Fig. S4.1). This *chunked* strategy predicts chunks of syntactic processing to happen at focal points, in a less incremental way. We compared the initial top-down parser used in the previous analyses with the early top down model and the *chunked* model by fitting three linear mixed models to the production data, each with of these different predictors of phrase-structure building. The *early top-down* model led to the best model fit (as measured with the Akaike information criterion, lower values indicate better fit: *early top-down*, 170803.9; *top-down*, 170821.3; *chunked*, 170834.5, Supplementary Tables S4.2, S4.3, S4.4).

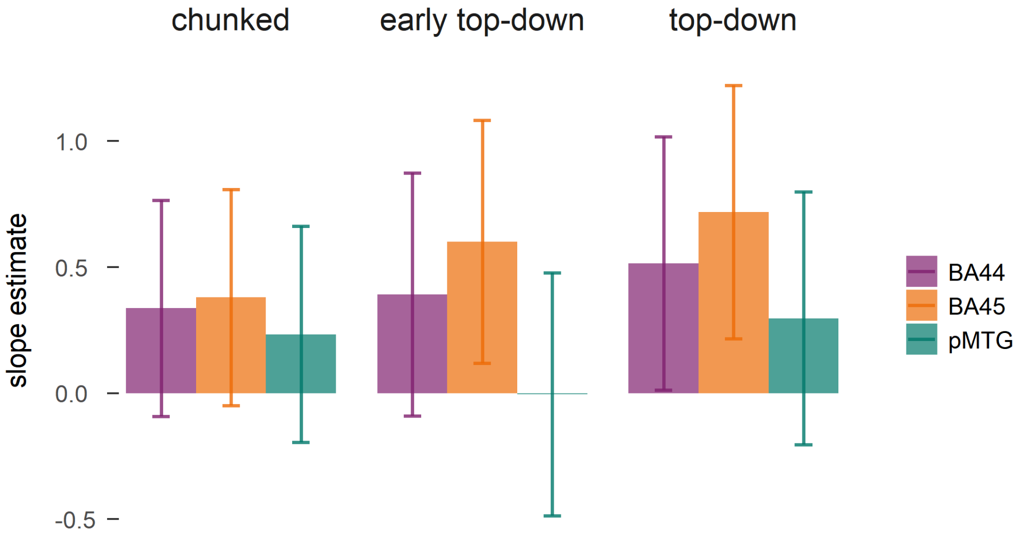


Figure 4.4: Beta estimates of the effect of each predictor of phrase-structure building in production on BOLD activity in the regions of interest. Error bars represent 95% confidence intervals. The *early top-down* model led to the best model fit (Akaike information criterion, lower values indicate better fit): *early top-down*, 170803.9; *top-down*, 170821.3; *chunked*, 170834.5.

Interestingly, top-down and *chunked* models predicted an overall increase in BOLD (*top-down*, $\chi^2 = 7.5$, $p < 0.007$; *chunked*, $\chi^2 = 3.9$, $p = 0.049$), while the *early top-down* main effect was not significant ($\chi^2 = 2.7$, $p = 0.1$), but it interacted with ROI ($\chi^2 = 6.2$, $p < 0.05$) (Fig. 4.4). In particular, *early top-down* counts predicted an increase in BA45 (estimate = 0.44, SE = 0.18, $p = 0.015$), while the effect was absent in LpMTG (estimate = 0.003, SE = 0.18, $p > 0.98$). The estimate for LpMTG thus decreased when phrase-structure building operations were posited to take place earlier, suggesting that the LpMTG responded to node counts later than the LIFG (see Supplementary Information for converging evidence on the latency of the response based on analysis of the temporal derivative, Supplementary Fig. S4.2). This pattern of results, therefore, indicates that in production phrase-structure building operations preferentially took place before word onset in an incremental fashion. An analysis on the pausing patterns throughout the speech additionally revealed that top-down node counts affected pause length before word articulation, providing converging evidence for phrase-structure building to happen before word onset in production (see Supplementary Information and Supplementary Fig. S4.3 for the analysis on pause length and word duration).

4.3 Discussion

In the first study to investigate the neural correlates of syntactic processing during spontaneous production, we modelled incremental phrase-structure building with probabilistic parsers and used them to predict brain activity in BA44, BA45 and LpMTG. We found that phrase-structure operations successfully predicted brain activity during naturalistic speech. A central finding was that the timing of phrase-structure operations differed strongly between production and comprehension. The results suggest that phrase-structure building occurs in an integratory manner in comprehension, mostly in LpMTG and BA45. Phrase-structure building was instead markedly anticipatory in production (occurring predominantly before word onset), as evidenced by anticipatory parser operations predicting pause length before each word during speech, and by a highly anticipatory incremental production parser that best modelled the production data in BA45.

Syntactic processing elicited BOLD activity increases in both production and comprehension, but critically the temporal profiles of brain activity diverged across modalities. This highlights inherent processing differences between language production and comprehension. In production, structure building can proceed by establishing the upcoming structure before words are uttered, which was confirmed by the longer pauses associated with larger numbers of top-down parsing operations. In comprehension, instead, phrase-structure building proceeds in a more integratory manner, waiting for the input to commit to the structure. These results fit with previously obtained evidence on BOLD timing sensitivity to structure complexity and modality, where BOLD peaked earlier with more complex structures in production but later in comprehension, relative to easier structures (Giglio et al., 2022; Pallier et al., 2011; see Pylkkänen, 2020; Pylkkänen et al., 2014 for converging evidence on production and comprehension dynamics of composition in MEG). Thus, the present results converge with previous controlled experiments in showing early syntactic encoding in production relative to later encoding in comprehension. This is likely due to different processing dynamics in production and comprehension, which have opposite inputs, outputs and mappings between linguistic levels (Pickering & Garrod, 2013).

It should be noted that these results only outline coarse processing dynamics, given the low temporal resolution of the BOLD signal, and that they do not aim to faithfully model all processes going on during speaking and listening. For example, these parsers are perfect 'oracles', meaning that they always posit phrase-structure building operations that actually happen, in contrast with potential ambiguities in the input (Hale, 2014). Recent evidence has shown that modelling syntactic ambiguity improves the fit with brain activity (Brennan et al., 2020). In addition, there is substantial evidence that comprehension is sensitive to the predictability of the input, such that some amount of anticipatory syntactic processing is expected in comprehension as well (Heilbron et al., 2022; Henderson et al., 2016; Shain et al., 2020; Willems et al., 2016). Indeed, Brennan et al. (2016) found a positive relationship between top-down operations, syntactic surprisal and BOLD activity in comprehension. Similarly, Coopmans (2023) found that a top-down parser best modelled brain activity during comprehension in MEG. Nelson et al. (2017) instead found bottom-up counts to better model brain activity (measured with electrocorticography) than top-down counts for the comprehension of single sentences. It is possible that different characteristics of the speech input led to this difference between studies. In our case, the input was spontaneous speech that also included disfluencies and corrections, while Brennan et al.'s and Coopmans at al.'s linguistic input were audiobook stories. There is evidence that lexical predictions can be influenced by reading strategies (Brothers, Swaab, & Traxler, 2017). It might have been easier to anticipate the structure in the 'cleaner' audiobook story than in the recall of an unfamiliar story. The reduced contextual information available in Nelson et al. (2017) may also have led to a reduction in anticipatory syntactic processing. Future studies with naturalistic comprehension will need to clarify to what extent the nature of the input determines the strength of anticipatory vs. integratory syntactic structure building.

Previous studies found modality differences in the sensitivity of neural responses to syntactic processing (Giglio et al., 2022; Indefrey et al., 2004). In particular, syntactic processing led to stronger responses in production than comprehension. This difference could have been observed either due to task-related effects or due to modality-inherent differences, such as a stronger need in production to fully compute the syntactic structure to be able to speak correctly, in contrast to good-enough processing in comprehension (Bock, 1982; Ferreira et al., 2002; Garrett, 1980). While we could not directly address this question with modality as a between-subject variable, the results indicate that the different modality load on syntactic processing found in previous studies may in effect be task-related. In this study, syntactic structure building elicited a neural activity increase that was quantitatively similar across ROIs in both modalities, although with different dynamics. This finding provides evidence that, in contexts where production is spontaneous and unconstrained by artificial tasks and comprehension is meaningful and as a consequence more engaging, syntactic parsing and encoding have a similar load on brain activity.

Interestingly, there were some regional differences in the sensitivity to syntactic predictors in each modality. In particular, both LpMTG and BA45 responded to syntactic processing in comprehension. Instead, in production BA45 was the most responsive, with a less direct involvement of LpMTG activity (excluding a potential later activation as suggested by analysis of the temporal derivatives, see Supplementary Information). This finding is somewhat unexpected, in light of previous results showing shared resources and representations across modalities (Giglio et al., 2022; Kempen et al., 2012; Segaert et al., 2012). It is not, however, the first study to show a higher comprehension load in temporal regions and a higher production load in frontal regions (Giglio et al., 2022; Humphreys & Gennari, 2014; Indefrey et al., 2004; Matchin & Wood, 2020). A plausible explanation relates to processing differences between modalities (Momma & Phillips, 2018). Both modalities may rely on the LIFG to coordinate syntactic processing (e.g., unification-type processing, Hagoort, 2013), while engaging the posterior temporal lobe for lexical-syntactic retrieval at different latencies depending on the amount of information available for structure building. In production, as suggested by the top-down parser, structure building may have been initiated before lexical-syntactic retrieval, engaging the LpMTG at later timescales, which is tentatively shown by the later response of the LpMTG in production. In comprehension, as indicated by the bottom-up parser, lexical-syntactic retrieval may have preceded or co-occurred with structure building and have appeared at canonical HRF delays. Interestingly, BA44 responded less strongly than BA45 to most predictors, in many cases with very low activity. This could suggest a reduced involvement of BA44 in syntactic processing in general, in line with Hagoort and Indefrey (2014) proposing a reduced function for BA44 in canonical syntactic processing. Alternatively, the anatomical selection may have been suboptimal for a region that is found to be highly variable between individuals (Fedorenko & Blank, 2020). Previous studies found also the anterior temporal lobe to show sensitivity to phrase-structure building operations (Brennan et al., 2020, 2012, 2016). We chose not to focus on this region due to converging evidence for a role of the anterior temporal lobe in semantic composition over syntactic processing (e.g. Lambon Ralph et al., 2017; Mesulam et al., 2014; Pylkkänen, 2020; Wilson et al., 2013), but we do not exclude that effects of phrase-structure building may have been present in the anterior temporal lobe in this study as well.

Returning to parser-specific modelling of syntactic processing, the parsers discussed so far were developed for syntactic processing specifically in comprehension (Hale, 2014). Since production is thought to proceed with different dynamics, we explored parsers that were more plausible for processing in production (Pickering & Garrod, 2013). In production, syntactic processing is thought to happen before word articulation (Bock & Levelt, 1994; Indefrey & Levelt, 2004; Levelt, 1989). There are different views on whether lexical access guides the structure, or whether the structure encoding the relations between concepts guides the order of lexical access (Bock & Ferreira, 2014). While the evidence provides mixed support for both accounts, suggesting that syntactic encoding is flexible and variable (e.g. Konopka & Meyer, 2014; Kuchinsky et al., 2011; van de Velde & Meyer, 2014), several proposals identify the verb as a central node in sentence planning, suggesting that the syntactic structure until the verb is sometimes computed early on (Levelt, 1989; Momma & Ferreira, 2019; Momma et al., 2016). Cross-linguistic evidence even suggests that in some languages some level of planning happens during the previous sentence (Sarvasy, Morgan, Yu, Ferreira, & Momma, 2022). By taking advantage of brain activity as an index of processing dynamics, we compared more and less incremental models of sentence planning with two parser models that made different predictions on the temporal unfolding of syntactic structure.

An incremental parser that is more anticipatory than the original top-down parser led to the best model fit, leading to the strongest increase in BA45 but not affecting LpMTG. Structure building thus proceeds before word articulation. This was also confirmed by converging results on longer pauses before words associated with more phrase-structure building operations, in line with previous behavioural evidence linking pausing patterns in speech with syntactic complexity (Ferreira, 1991). A less incremental parser that always plans the structures for a few chunks of words at a time provided the worst fit for brain activity. These results suggest that a highly incremental parser may be the more standard planning strategy in production, and that the structure up to the verb does not need to be planned at the start of the sentence. We thus provide the first piece of neuroimaging evidence addressing the long-standing debate on the incrementality of sentence planning. This approach could contribute to the understanding of the dynamics of sentence planning, with the development of models that take into account the variability of each sentence, for example by modelling longer planning scopes only when the verb follows an internal argument (cf. Momma & Ferreira, 2019), or depending on word accessibility (Kuchinsky et al., 2011; van de Velde & Meyer, 2014).

Importantly, with this study we demonstrated the feasibility and benefits of studying production with spontaneous speech. The costs associated with spontaneous production, such as increased variability and disfluencies of the linguistic signal, increased motion artifacts in fMRI and the slow temporal resolution, are balanced by the many advantages. Spontaneous production yields a larger amount of data than controlled tasks. This is the case especially in behavioural analysis but also with fMRI, provided the speech samples are of sufficient length. In addition, with spontaneous speech the artificiality of the task is largely reduced. Although speaking in monologue is not as common as dialogue, it is more ecologically valid than speaking following careful instructions with limited acceptable speech output. In addition, the probability distributions of linguistic inputs and outputs are preserved in spontaneous contexts, in contrast with many experiments (Jaeger, 2010). Finally, neuroimaging studies on spontaneous production allow for potentially new insight into production questions that have been so far mostly addressed with psycholinguistic studies.

In summary, we showed that spontaneous production can be used to study the neural correlates of linguistic processing, providing very rich data that can be directly linked to behavior with the analysis of pause length and word durations. We found that syntactic structure building engages the inferior frontal gyrus in both production and comprehension with diverging dynamics. Phrase-structure building was anticipatory in production but integratory in comprehension. Finally, we provided neural evidence for incremental models of syntactic encoding in production using production-specific parsers.

4.4 Materials and Methods

Data acquisition and preprocessing

Production data

The production data used were collected by Chen et al. (2017; Chen et al., 2018) and made available on OpenNeuro (`https://openneuro.org/datasets/ds001132/versions/1.0.0`). In this experiment, participants watched an episode of the BBC TV series *Sherlock* and then recalled what happened in the episode. Data were collected for 22 right-handed native English participants (10 female, ages 18-26, mean age 20.8). Five participants were excluded due to excessive head motion (2 participants), because recall was shorter than 10 min (2 participants) or for falling asleep during the movie (1 participant). Data for one participant was not shared because of missing data at the end of the movie scan, which left us with 16 participants for the current analysis. Speaking led to an average framewise displacement of 0.32 (average per participant, range = 0.13-0.54), which was higher than the average in the comprehension data (0.22, range = 0.08-0.42) but was corrected for with noise regression (see fMRI data preprocessing for more details).

Participants watched the first 50 minutes of the first episode of the BBC TV series *Sherlock*, after confirming that they had not watched any episode of *Sherlock* before. Participants were told they would be asked to verbally describe what they had seen. After watching the episode, they were immediately instructed "to describe what they recalled of the movie in as much detail as they could, to try to recount events in the original order they were viewed in, and to speak for at least 10 min if possible but that longer was better. They were told that completeness and detail were more important than temporal order, and that if at any point they realized they had missed something, to return to it. Participants were then allowed to speak for as long as they wished, and verbally indicated when they were finished (for example, “I’m done”). During this session they were presented with a static black screen with a central white dot (but were not asked to, and did not, fixate).” (Chen et al., 2017). Their speech was recorded in the scanner with an MR-compatible microphone.

We also used a second production scan for one of these participants, who also recalled an episode of BBC TV series *Merlin*, as part of the data collected and released by Zadbood, Chen, Leong, Norman, and Hasson (2017). This speech sample was used as audio stimulus for the Comprehension data (see below). The procedure and acquisition were the same. Therefore, in total, we used 17 speech samples from 16 participants, since one participant recalled both *Sherlock* and *Merlin*. The 17 recalls were 10-45 minutes (mean = 22 minutes, SD = 8.8 minutes), including on average 2874 words (range = 1666-6230, SD = 1299).

Comprehension data

For the comprehension data, we used the data shared by Zadbood et al. (2017; Zadbood, Chen, Leong, Norman, & Hasson, 2018) on OpenNeuro (`https://openneuro.org/datasets/ds001110/versions/00003`). In this experiment, participants watched an episode of either BBC TV series *Merlin* or *Sherlock* and listened to an audio recording of the story they did not watch. Audio recordings were obtained from a participant that watched and recounted the two movies, here analysed as part of the production data. In this dataset, 52 right-handed native English speakers (age 18-45) were scanned. Fifteen participants were excluded because of head motion (n=4), for falling asleep (n=4), due to poor memory (n=5), for having seen the movie before (n=2). This resulted in 36 analysed participants, 18 that listened to the *Merlin* recall, and 18 that listened to the *Sherlock* recall. The audio recall for *Merlin* was 14.7 minutes long and included 2141 words. The audio recall for *Sherlock* was 17.5 minutes long and included 2468 words.

Data Acquisition

The acquisition parameters were identical in the two datasets. MRI data was collected on a 3T full-body scanner (Siemens Skyra) with a 20-channel head coil. Functional images were acquired using a T2*-weighted echo planar imaging pulse sequence (TR 1500 ms, TE 28 ms, flip angle 64, whole-brain coverage 27 slices of 4 mm thickness, in-plane resolution 3 × 3 mm2, FOV 192 × 192 mm2). Anatomical images were acquired using a T1-weighted MPRAGE pulse sequence (0.89 mm^3 resolution).

fMRI data preprocessing

Preprocessing was performed using *fMRIPrep* 20.2.6 (Esteban, Blair, et al., 2018; Esteban, Markiewicz, et al., 2018), which is based on *Nipype* 1.7.0 (Gorgolewski et al., 2011, 2018).

Anatomical data preprocessing
The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.3.3 (Avants, Epstein, Grossman, & Gee, 2008), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, Zhang et al., 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (A. Klein et al., 2017).

Functional data preprocessing
For each BOLD run, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted, because no fieldmap was acquired. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al., 2002). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD* run in *MNI152NLin2009cAsym* space. First, a reference volume and its skull-stripped

version were generated using a custom methodology of *fMRIPrep*. Automatic removal of motion artifacts using independent component analysis (ICA-AROMA, Pruim et al., 2015) was performed on the *preprocessed BOLD* on MNI space time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). The "aggressive" noise-regressors were collected and placed in the corresponding confounds file. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), the derivative of the relative (frame-to-frame) bulk head motion variance (DVARS) and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al., 2014) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al., 2002). FD and DVARS are calculated for each functional run, both using their implementations in *nipype* (following the definitions by Power et al., 2014). Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for anatomical CompCor (aCompCor). For aCompCor, three probabilistic masks (cerebrospinal fluid (CSF), white matter (WM) and combined CSF+WM) are generated in anatomical space.

Incremental complexity metrics

Phrase-structure parsing

First, we extracted the constituent structure of each sentence with a probabilistic context-free phrase-structure grammar. We used the Stanford parser with CoreNLP in Python 3 via the Natural Language Toolkit (NLTK) package (D. Klein & Manning, 2003; Manning et al., 2014). The transcript provided in the shared dataset was divided in what we considered independent sentences. Since the production was very spontaneous and unconstrained, sentence boundaries were not objective and self-evident as they are in text. In speech, the boundaries can depend on the syntactic structure of the sentence, but also on pausing patterns. For example, coordinated clauses may be considered one single sentence or divided into two separate sentences based on pause lengths. After extensive exposure to the transcripts, it became clear that shorter boundaries better reflect the planning chunks followed by speakers. In particular, some sentences extend over 30 words or more, with many embedded phrases. Participants, however, do not appear to fully keep in working memory the original syntactic structure, which is revealed by their disfluencies and corrections throughout the long sentence. Disfluencies affect how sentence boundaries, which are critical for the performance of automatized parsers, can be placed in the discourse, since they are not as obvious as in text. For example, boundaries could fully track the syntactic structure, also including hesitations and corrections within its boundaries, or they could track speech patterns and 'reset' every time there is a disfluency. In this type of data, with monologues without audience feedback, short disfluency boundaries seemed more appropriate, but it is to be determined if different approaches work better in other contexts. It should be noted that an initial analysis was run on more liberal sentences, which perhaps better tracked the overarching syntactic structure but did not optimally reflect the planning processes of participants. The results were similar with both sentence boundaries approaches, but the conservative approach to sentence length was less noisy.

Following sentence parsing with the Stanford constituent parser, we took a measure of syntactic processing with incremental complexity metrics derived from the number of syntactic nodes that are built with each word. Nodes can be built with different parsing strategies: top-down, bottom-up and left-corner (Hale, 2014). In top-down parsing, nodes are built from the top of the syntactic tree to the terminal node (corresponding to a word). In other words, nodes are counted when phrases are opened. This strategy can lead to the anticipation of nodes that may not always be known to a listener. For example, in the sentence "Mary eats apples daily", a node accounting for the upcoming presence of "daily" is counted already at the word "eats". This anticipation is justifiable in production, where the upcoming structure is presumably known to the speaker in advance, but it might reflect unjustifiable prediction in comprehension. Nevertheless, this implementation of a top-down strategy may be successful in accounting for predictive processes in comprehension.

At the other end of the incremental parsing spectrum is bottom-up parsing, according to which nodes are built from the bottom of the syntactic tree (i.e. from the terminal nodes, corresponding to each word) up to the highest *closed* nodes, i.e. nodes where all daughter nodes have already been met. For example, in Figure 4.1B, the top node in purple (S) cannot be built until its right-branching node VP is built as well, which in this case only happens at the end of the sentence. In other words, bottom-up parsing builds nodes when phrases are closed. This strategy thus predicts increased syntactic processing at the end of clauses and sentences, after all the evidence for the structure is encountered. We expected this parsing strategy to better reflect processing in comprehension than production, because in the latter the structure is presumably already built before the last word is uttered. Neither top-down nor bottom-up parsing strategies fully match human performance (Brennan & Pylkkänen, 2017; Hale, 2014), but they capture aspects of syntactic processing that are expected to differ across modalities. Finally, left-corner parsing needs less evidence than bottom-up parsing to count nodes, but is not as predictive as top-down parsing. After convolving with the haemodynamic response function (HRF), left-corner was highly correlated with the top-down parser (Supplementary Fig. S4.4). Therefore, we decided to only focus on opposite parsing strategies that were most expected to differ between production and comprehension, i.e. top-down and bottom-up.

We also counted the number of nodes that were still open at each word with an *open nodes* measure, similarly to Nelson et al. (2017). Open nodes were the number of nodes that were open at each word: this measure tracked the number of nodes that had been opened up to the word and that had not been closed yet, thus providing an index for the number of nodes that need to be kept in working memory until they can be merged in a constituent (Nelson et al., 2017).

Production-specific parsing operations

To account for the timing that is specific to production, we developed two production-specific parsers. An *early top-down* model counts the nodes that are built for the *next* word. At the first word of the sentence, nodes are counted for the first and second word (even though nodes built for the first word would have been built earlier, we preferred this over making assumptions on when the nodes would be built before the sentence, which could be varying due to different factors). At the second word, nodes are counted for the third word, etc.

For the less-incremental *chunked* parsing, we selected the heads of each sentence following dependency parsing (see Supplementary Information for more information on the analysis on dependency parsing). We considered as heads all words that had a dependent relation attached to them (e.g. the verb is head of subject and object). We then counted all nodes (of the same constituent structure used by the other parsers) encountered from the first word up to and including the next head, then from the head up to and including the next head, and so on. *Chunked* parsing, therefore, builds nodes early on for all the upcoming words that are dependent relations of the next head. For example, at the start of a sentence all the nodes are built for the structure up to and including the verb, usually the first head.

It should be noted that top-down, *early top-down*, and *chunked* measures were highly correlated after convolving with the haemodynamic response function (Supplementary Fig. S4.4). To avoid collinearity, instead of comparing them in the same model, we tested models with only one predictor and determined which model provided the best fit (see Regression analysis for more details).

Word surprisal

We quantified word surprisal from transformer model GPT-2 (Radford et al., 2019). We used GPT-2 XL via the TensorFlow implementation provided by HuggingFace's Transformers package (Wolf et al., 2020). Each word's probability was based on a context of at least 700 words after the first 700 words of each participant's recall. Surprisal was calculated as the negative logarithm of the conditional probability of the word based on context.

Behavioural analysis

To determine if these indexes of processing complexity had an effect on participants' speech patterns, we inspected how they affected word duration and pause lengths in all the production recalls. Recordings were not made available with the Production dataset, but word timestamps for each participant's recall were shared by Janice Chen's lab. Onsets and offsets of each word were obtained with Gentle (`https://lowerquality.com/gentle/`, `https://github.com/lowerquality/gentle`). We ran a linear mixed-effects model with *lme4* (version 1.1-26, Bates et al., 2015) in R (version 4.0.3). We used number of syllables, word frequency, word surprisal, top-down, bottom-up and open nodes as predictors for pause length (before the word characterized by each predictor) and word duration. This analysis allowed us to compare neural effects with behavioural patterns of speech.

fMRI analysis

Predictor timeseries

Each word-by-word predictor was mean-centered (except for the word rate predictor) and convolved with the canonical haemodynamic response function following SPM's double gamma function as computed in *nilearn* (*nilearn/nilearn*, 2022). We thus obtained predictor timeseries temporally resampled to the acquisition TR of 1.5 s, reflecting BOLD increases and decreases following predictor weights time-locked to word onset (Fig. 4.2C).

ROI selection

We selected three ROIs that have been associated with syntactic processing in previous studies: two LIFG ROIs, following the distinction between LIFG *pars opercularis* (BA44) and LIFG *pars triangularis* (BA45), and left posterior middle temporal gyrus (LpMTG). After preprocessing the fMRI data, we selected the ROIs for each participant in their functional space. BA44 and BA45 were extracted following Freesurfer's label creation with the Destrieux Atlas (Destrieux, Fischl, Dale, & Halgren, 2010) and resampled to functional space with bbregister. Freesurfer's MTG ROI is quite long in extension, following the gyrus from very posterior portions to the temporal pole. We therefore extracted this ROI and then masked it with a posterior temporal lobe mask (posterior to Heschl's gyrus) based on the Harvard-Oxford cortical atlas. Examples of these ROIs in MNI brain can be seen in Figure 4.2E.

Timeseries extraction

The BOLD timeseries were extracted with NiftiLabelsMasker from *nilearn* (Abraham et al., 2014; *nilearn/nilearn*, 2022), after confound regression. Framewise displacement, DVARS, motion parameters, aCompCor paramers and ICA-AROMA regressors classified as noise were used for noise regression, to reduce the impact of motion artifacts caused by speaking. The timeseries was extracted from the functional BOLD volumes in functional space as an average of the voxels in each ROI mask.

Regression analysis

To determine to what extent each of these continuous indices of syntactic processing significantly affected brain activity (average BOLD activity in the three ROIs), we used linear mixed-effects models with *lme4* (version 1.1-26, Bates et al., 2015) in R (version 4.0.3). We used a baseline model that included word rate (i.e. a predictor indicating the onset of each word), syllable rate, as an index of articulatory rate, log-transformed word frequency, and word surprisal. All models additionally included modality and ROI as factors. Modality (production vs. comprehension) was contrast-coded with deviation coding. We used Helmert coding for ROI, contrasting LIFG with LpMTG, and the two LIFG *partes* with each other. All other factors were continuous numerical predictors. All models included word surprisal and its interaction with ROI and modality. All models also included by-participant random slopes for syllable rate, frequency, word surprisal and other factors of interest, excluding by-participant random effects and correlations to allow for convergence and avoid singularity issues. In some cases, we had to exclude the random slopes for one of these factors, but never for the factor of interest in that model. We computed the contribution of factors to the models using *car* (version 3.0-10, Fox et al., 2021), and pairwise comparisons with the package *emmeans* (version 1.6.1, Lenth et al., 2022).

The first model determined the contribution of top-down and bottom-up metrics of phrase-structure building to brain activity in the three ROIs and in each modality to a baseline model that included word surprisal and open nodes. The interactions of each metric with ROI and modality were included and the significant contribution of the incremental metric in a region or modality was determined with pairwise comparisons. With this model we also determined to what extent word surprisal and open nodes affected brain activity in each modality.

We then used three models to ask whether metrics of syntactic processing fine-tuned for production would improve model fit. These metrics are not realistic for syntactic processing in comprehension, so the models only included production data. The baseline models all included word surprisal and bottom-up parser operations, and additionally included *top-down*, or *early top-down*, or *chunked* predictors of phrase structure building and their relative by-participant random slopes. Since the three parsers were highly correlated after convolving with the HRF, we separately fitted three linear models. We compared model fit with the Akaike Information Criterion (AIC), where more negative values indicate better model fit (Cavanaugh & Neath, 2019).

S4 Supplementary Information

Dependency parsing

We extracted dependency parsing as an index to perform chunking of each sentence in relational units. We used the dependency parser provided by the Stanford parser via CoreNLP. We identified the heads of the sentence as the words that have a relation attached to them. For example, in the sentence "He's examining one of the bodies", "examining" is the first head, followed by "one" and by "bodies" (Supplementary Fig. S4.1). These three words are the only ones that have a dependency relation attached to them, while the other words are all dependents of one head. The *chunked* parsing strategy counted the nodes intervening between all heads, to model a less incremental strategy to syntactic structure building, following the idea that speakers plan the structure of a few words at a time (e.g. always planning the structure of the verb at the start of the sentence).

Figure S4.1: Dependency parse of the sentence. Left-relations are in orange, right relations in green, heads are in purple. Heads are words on which a dependency relation is attached (i.e. from which an arrow starts).

Temporal derivative

Since the results of production-specific parsers indicated that the LpMTG may have had a later response to top-down operations than BA45, we looked into how the temporal derivatives of the same predictors modelled the data. The temporal derivative of the haemodynamic response function (HRF) is usually used in fMRI analysis to account for small differences in the latency of the BOLD response. An increase in the temporal derivatives means that the BOLD response peaks earlier, while a decrease indicates a later peak. We ran the same linear-mixed effects model used before with the addition of temporal derivatives of all predictors of interest (Supplementary Table S4.5). We found a significant three-way interaction between the top-down derivative, modality and ROI ($\chi^2 = 15.7$, $p < 0.0004$); the surprisal derivative, modality and ROI ($\chi^2 = 18$, $p < 0.0002$); and two-way interactions between the open nodes derivative and modality ($\chi^2 = 4.4$, $p < 0.037$), and the open nodes derivative and ROI ($\chi^2 = 26.8$, $p <$

0.0001). The top-down temporal derivative was marginally significantly positive in BA45 in production (estimate = 0.31, SE = 0.16, $p = 0.056$), indicating an earlier BOLD peak than assumed by the canonical HRF, while it was significantly negative in LpMTG (estimate = 0.32, SE = 0.16, $p = 0.046$), indicating a later peak. It was not significantly different from zero in comprehension, nor did it differ between ROIs. These results suggest that the LpMTG may have been active after BA45 in response to more top-down node counts.

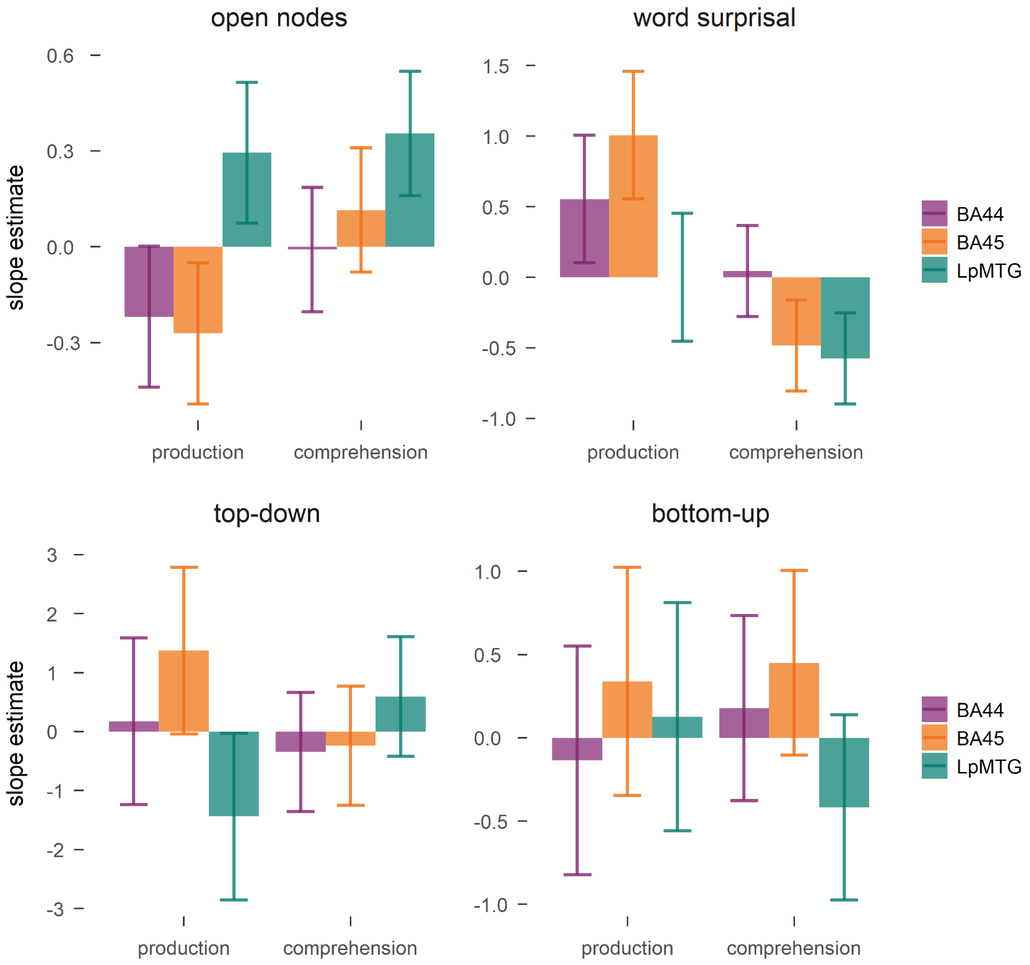


Figure S4.2: Beta estimates for the effect of the temporal derivative of each predictor of on BOLD activity in the regions of interest. Error bars represent confidence intervals. Positive estimates indicate an earlier BOLD response, negative estimates indicate a later BOLD response.

Word surprisal elicited later BOLD peaks in comprehension and earlier BOLD peaks in production, relative to the canonical HRF. In comprehension, BA45 and LpMTG were both significantly related to a delay in activity (BA45: estimate = 0.46, SE = 0.16, $p = 0.003$; LpMTG: estimate = 0.55, SE = 0.16, $p = 0.0005$). In production, both activity in both BA44 and BA45 increased with the temporal derivative for surprisal (BA44: estimate = 0.53, SE = 0.2, $p = 0.16$; BA45: estimate = 0.97, SE = 0.2, $p = 0.0001$). These results suggest that word surprisal elicited earlier activity increases in production than comprehension, which likely relates to the timing of lexical access (before word onset in production, after word onset in comprehension). The open nodes measure showed earlier BOLD responses in the LpMTG (estimate = 1.3, SE = 0.3, $p = 0.0001$), and earlier responses in comprehension than production (estimate = 0.213, SE = 0.1, $p = 0.037$). Again, BA45 and the LpMTG had different BOLD peak latencies, suggesting that BA45 responded earlier to top-down nodes but later to open nodes, while LpMTG responded earlier to open nodes and later to top-down nodes.

Speech fluency: results and discussion

The number of syllables of a word significantly predicted an increase in word duration (unit is seconds, $\beta = 0.08$, SE = 0.004, $t = 22.6$, $\chi^2 = 510.02$, $p = 0.0001$), which was expected, but a decrease in pause length before word articulation ($\beta = -0.02$, SE = 0.004, $t = 4.9$, $\chi^2 = 23.9$, $p = 0.0001$). The shorter pause before articulation of long words is possibly due to the longer time available to plan for later words during uttering of a long word. Higher word frequency instead predicted shorter word duration ($\beta = -0.017$, SE = 0.001, $t = 15.1$, $\chi^2 = 239.8$, $p = 0.0001$), but did not affect pause length ($\beta = 0.0008$). Larger word surprisal increased word duration to a small extent ($\beta = 0.004$, SE = 0.0004, $t = 9.6$, $\chi^2 = 91.9$, $p = 0.0001$), and it had a larger positive effect on pause length ($\beta = 0.02$, SE = 0.002, $t = 10.4$, $\chi^2 = 109.1$, $p = 0.0001$). Less predictable words based on context thus took longer to be initiated and were uttered for a slightly longer time (4 ms), after accounting for their length.

We also determined how predictors of syntactic complexity related to speech fluency. Top-down node counts predicted the largest decrease in word duration ($\beta = -0.045$, SE = 0.002, $t = 20.1$, $\chi^2 = 404.1$, $p = 0.0001$), suggesting that when phrases are opened, information can be conveyed faster, possibly to offload working memory. It also predicted the largest increase in pause length before the word in question is uttered ($\beta = 0.09$, SE = 0.008, $t = 10.9$, $\chi^2 = 119.7$, $p < 0.0001$) suggesting that grammatical encoding related to a word is performed before word articulation, and that nodes are built in an anticipatory way. Bottom-up parser operations predicted an opposite pattern. Larger bottom-up counts increased word duration ($\beta = 0.012$, SE = 0.0009, $t = 12.6$, $\chi^2 = 159.5$, $p < 0.0001$), but decreased pause length ($\beta = 0.021$, SE = 0.002, $t = 11.6$, $\chi^2 = 135.7$, $p < 0.0001$). The shorter pauses suggest that at phrase closing the structure is already computed. Finally, open nodes predicted a significant but very small decrease in word duration (β = -0.002, SE = 0.0004, t = 5.3, $\chi^2 = 28.5$, $p < 0.0001$), and a larger decrease in pause length (β = -0.024, SE = 0.002, t = 14.6, $\chi^2 = 213.9$, $p < 0.0001$), suggesting easier processing the further along in a sentence. In line with the neuroimaging results, this pattern of results suggests that phrase-structure building happens before word articulation and at phrase-opening, with a decrease in pauses the further along in the sentence.

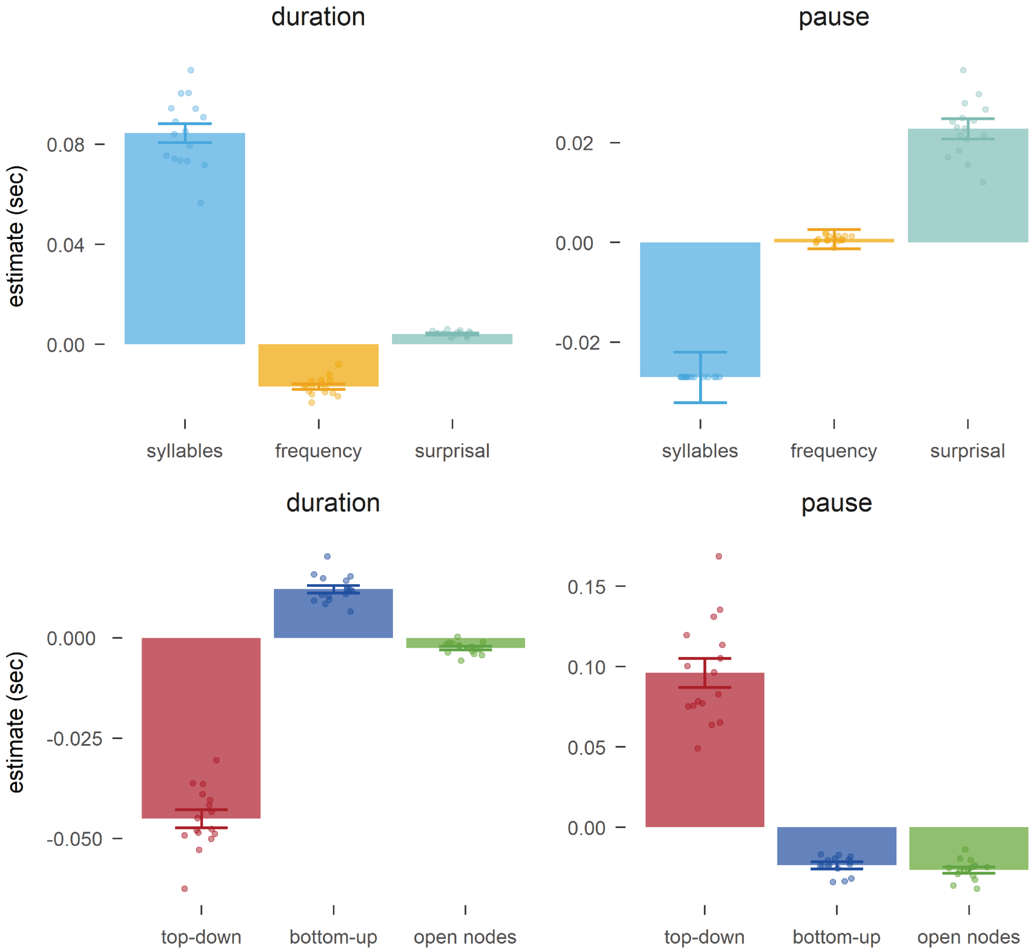
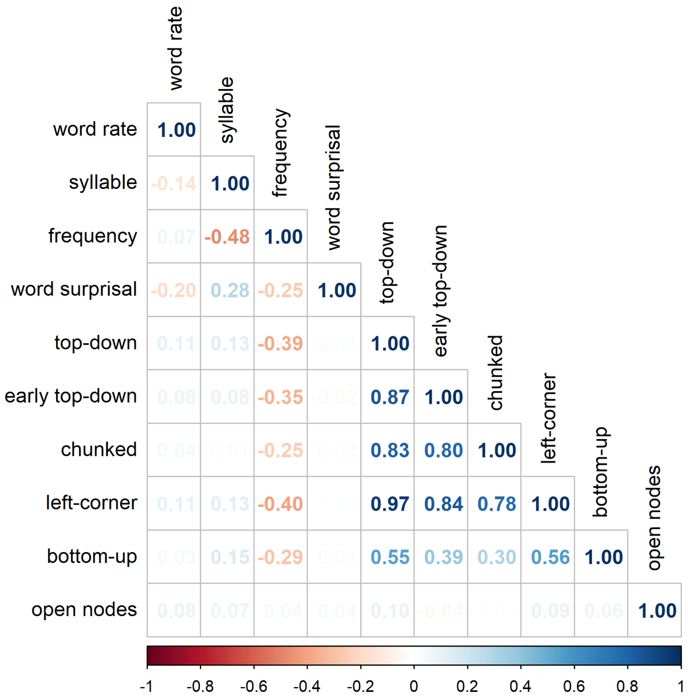


Figure S4.3: Estimates in seconds of the effect of each predictor of word characteristics and phrase-structure building on word durations and pause length before word articulation. Error bars represent standard error of the mean. Individual points represent each participant's estimate as estimated by the random slopes. The model estimated identical random slopes for number of syllables on pause length for each participant.

This was the first study to show an increase in neural activity for words associated with higher surprisal, not only in comprehension but also in production.

Many studies showed sensitivity of brain activity to surprisal in language comprehension computed with several models (Shain et al., 2020; Willems et al., 2016). The neural results are in line with the behavioural results that show an increase in pause length before less probable words and a small increase in their duration, as found previously (Aylett & Turk, 2004). The results thus converge in demonstrating the sensitivity of the production system to the statistical probabilities of the linguistic input and output, both in behavioural and neural patterns. This finding is in line with accounts of efficient language production that propose a uniform distribution of information in discourse (Uniform Information Density, Jaeger, 2010; Jaeger & Levy, 2007; Karimi, 2022; Piantadosi, Tily, & Gibson, 2011). More informative units (in information-theoretic terms, i.e. larger surprisal in the current study) take more time in discourse, while redundant units can be uttered faster or eliminated (e.g. for optional words like complementizer *that*, Jaeger, 2010).

Correlation between model predictors



	word rate	syllable	frequency	word surprisal	top-down	early top-down	chunked	left-corner	bottom-up	open nodes
word rate	1.00									
syllable	-0.14	1.00								
frequency	0.07	-0.48	1.00							
word surprisal	-0.20	0.28	-0.25	1.00						
top-down	0.11	0.13	-0.39	0.01	1.00					
early top-down	0.08	0.08	-0.35	-0.02	0.87	1.00				
chunked	0.04	0.01	-0.25	0.02	0.83	0.80	1.00			
left-corner	0.11	0.13	-0.40	0.00	0.97	0.84	0.78	1.00		
bottom-up	0.03	0.15	-0.29	0.01	0.55	0.39	0.30	0.56	1.00	
open nodes	0.08	0.07	0.04	0.04	0.10	-0.04	0.00	0.09	0.06	1.00

Figure S4.4: Correlation matrix showing Pearson's r correlation among all predictors. Note that not all predictors were used in the same model.

Table S4.1: Summary of model output of BOLD activity in production and comprehension. ROI1 refers to the contrast BA44 vs. BA45, ROI2 refers to the contrast BA44 & BA45 vs. pMTG. Mod stands for modality. AIC stands for Akaike Information Criterion.

			BOLD		
Predictors	**Estimates**	**SE**	**CI**	**Statistic**	***p***
(Intercept)	-0.18	0.01	-0.21 – -0.16	-14.93	0.001
word rate	0.54	0.03	0.48 – 0.61	16.15	0.001
syllable	0.22	0.12	-0.02 – 0.45	1.82	0.069
frequency	0.12	0.05	0.03 – 0.21	2.5	0.012
top-down	0.02	0.16	-0.29 – 0.33	0.13	0.899
ROI1	0	0.01	-0.01 – 0.01	0.26	0.796
ROI2	0	0	-0.01 – 0.01	-0.02	0.98
mod1	-0.01	0	-0.02 – 0.00	-1.81	0.07
bottom-up	-0.05	0.08	-0.21 – 0.11	-0.6	0.547
surprisal	0.16	0.02	0.12 – 0.21	6.67	0.001
open nodes	0.03	0.01	0.00 – 0.06	2.06	0.04
top-down * ROI1	-0.09	0.09	-0.27 – 0.08	-1.04	0.299
top-down * ROI2	-0.23	0.05	-0.33 – -0.13	-4.42	0.001
top-down * mod1	-0.49	0.16	-0.79 – -0.18	-3.13	0.002
ROI1 * mod1	0	0.01	-0.01 – 0.01	0.06	0.956
ROI2 * mod1	0	0	-0.01 – 0.01	-0.11	0.909
ROI1 * bottom-up	0.05	0.06	-0.07 – 0.17	0.81	0.42
ROI2 * bottom-up	0.12	0.03	0.05 – 0.19	3.38	0.001
mod1 * bottom-up	0.39	0.08	0.23 – 0.55	4.84	0.001
mod1 * surprisal	0	0.02	-0.04 – 0.05	0.17	0.868
ROI1 * surprisal	0.08	0.02	0.04 – 0.12	3.63	0.001
ROI2 * surprisal	0.01	0.01	-0.01 – 0.04	0.92	0.359
ROI1 * open nodes	0.02	0.01	0.00 – 0.03	2.02	0.044
ROI2 * open nodes	0.01	0	-0.00 – 0.02	1.69	0.091
mod1 * open nodes	0.03	0.01	0.00 – 0.06	2.15	0.032
top-down * ROI1 * mod1	-0.2	0.09	-0.37 – -0.02	-2.21	0.027
top-down * ROI2 * mod1	-0.12	0.05	-0.22 – -0.02	-2.36	0.019
(ROI1 * mod1) * bottom-up	0.15	0.06	0.03 – 0.27	2.47	0.013
(ROI2 * mod1) * bottom-up	0.09	0.03	0.02 – 0.16	2.56	0.01
(ROI1 * mod1) * surprisal	0.01	0.02	-0.03 – 0.05	0.61	0.54
(ROI2 * mod1) * surprisal	-0.02	0.01	-0.04 – 0.00	-1.65	0.1
(ROI1 * mod1) * open nodes	0.02	0.01	0.00 – 0.03	2.54	0.011
(ROI2 * mod1) * open nodes	0.01	0	-0.00 – 0.02	1.58	0.114
N subj	52				
Observations	115743				
AIC	418491.644				

Table S4.2: Summary of model output of BOLD activity in production with top-down predictor. ROI1 refers to the contrast BA44 vs. BA45, ROI2 refers to the contrast BA44 & BA45 vs. pMTG. Mod stands for modality. AIC stands for Akaike Information Criterion, used for the production only models to determine model fit.

			BOLD		
Predictors	**Estimates**	**SE**	**CI**	**Statistic**	***p***
(Intercept)	-0.15	0.02	-0.18 – -0.11	-7.56	0.001
word rate	0.46	0.05	0.35 – 0.56	8.33	0.001
syllable	0.07	0.14	-0.21 – 0.35	0.5	0.62
frequency	0.08	0.07	-0.06 – 0.22	1.15	0.25
surprisal	0.15	0.03	0.09 – 0.22	4.78	0.001
ROI1	0	0.01	-0.02 – 0.02	0.12	0.905
ROI2	0	0.01	-0.01 – 0.01	0.05	0.958
top-down	0.51	0.19	0.14 – 0.88	2.73	0.006
bottom-up	-0.42	0.1	-0.61 – -0.23	-4.41	0.001
surprisal * ROI1	0.06	0.04	-0.01 – 0.14	1.75	0.08
surprisal * ROI2	0.03	0.02	-0.01 – 0.07	1.49	0.136
ROI1 * top-down	0.1	0.15	-0.20 – 0.40	0.67	0.501
ROI2 * top-down	-0.11	0.09	-0.28 – 0.07	-1.22	0.224
ROI1 * bottom-up	-0.1	0.1	-0.29 – 0.09	-1.03	0.303
ROI2 * bottom-up	0.03	0.06	-0.08 – 0.14	0.5	0.617
N subj	16				
Observations	45099				
AIC	170821.271				

Table S4.3: Summary of model output of BOLD activity in production with early top-down predictor. ROI1 refers to the contrast BA44 vs. BA45, ROI2 refers to the contrast BA44 & BA45 vs. pMTG. Mod stands for modality. AIC stands for Akaike Information Criterion, used for the production only models to determine model fit.

	BOLD				
Predictors	**Estimates**	**SE**	**CI**	**Statistic**	***p***
(Intercept)	-0.15	0.02	-0.19 – -0.11	-7.88	0.001
word rate	0.47	0.05	0.37 – 0.58	8.69	0.001
syllable	0.07	0.14	-0.21 – 0.35	0.49	0.622
frequency	0.06	0.07	-0.08 – 0.21	0.9	0.37
surprisal	0.16	0.03	0.09 – 0.22	4.83	0.001
ROI1	0	0.01	-0.02 – 0.02	0.12	0.906
ROI2	0	0.01	-0.01 – 0.01	0.06	0.956
early top-down	0.33	0.2	-0.06 – 0.72	1.64	0.1
bottom-up	-0.35	0.1	-0.55 – -0.16	-3.54	0.001
surprisal * ROI1	0.06	0.04	-0.01 – 0.14	1.78	0.076
surprisal * ROI2	0.03	0.02	-0.01 – 0.07	1.42	0.155
ROI1 * early top-down	0.1	0.12	-0.14 – 0.35	0.85	0.397
ROI2 * early top-down	-0.17	0.07	-0.31 – -0.03	-2.34	0.02
ROI1 * bottom-up	-0.09	0.09	-0.27 – 0.08	-1.04	0.296
ROI2 * bottom-up	0.04	0.05	-0.07 – 0.14	0.68	0.494
N subj	16				
Observations	45099				
AIC	170803.949				

Table S4.4: *Summary of model output of BOLD activity in production with chunked top-down predictor.* ROI1 refers to the contrast BA44 vs. BA45, ROI2 refers to the contrast BA44 & BA45 vs. pMTG. Mod stands for modality. AIC stands for Akaike Information Criterion, used for the production only models to determine model fit.

	BOLD				
Predictors	**Estimates**	**SE**	**CI**	**Statistic**	***p***
(Intercept)	-0.16	0.02	-0.19 – -0.12	-8.13	0.001
word rate	0.49	0.05	0.38 – 0.59	8.97	0.001
syllable	0.07	0.14	-0.21 – 0.35	0.52	0.603
frequency	0.05	0.07	-0.09 – 0.19	0.69	0.491
surprisal	0.15	0.03	0.09 – 0.21	4.62	0.001
ROI1	0	0.01	-0.02 – 0.02	0.12	0.906
ROI2	0	0.01	-0.01 – 0.01	0.05	0.957
chunked top-down	0.32	0.16	0.00 – 0.63	1.97	0.049
bottom-up	-0.31	0.08	-0.47 – -0.14	-3.65	0.001
surprisal * ROI1	0.06	0.04	-0.01 – 0.14	1.75	0.08
surprisal * ROI2	0.03	0.02	-0.01 – 0.07	1.49	0.135
ROI1 * chunked top-down	0.02	0.13	-0.23 – 0.27	0.17	0.866
ROI2 * chunked top-down	-0.04	0.07	-0.19 – 0.10	-0.56	0.574
ROI1 * bottom-up	-0.07	0.09	-0.25 – 0.10	-0.82	0.413
ROI2 * bottom-up	0	0.05	-0.10 – 0.10	0.02	0.984
N subj	16				
Observations	45099				
AIC	170834.489				

Table S4.5: *Summary of model output of BOLD activity in production and comprehension, including the temporal derivative (der) of all predictors of interest.* ROI1 refers to the contrast BA44 vs. BA45, ROI2 refers to the contrast BA44 & BA45 vs. pMTG. Mod stands for modality. AIC stands for Akaike Information Criterion, used for the production only models to determine model fit.

	BOLD				
Predictors	**Estimates**	**SE**	**CI**	**Statistic**	***p***
(Intercept)	-0.17	0.01	-0.20 – -0.15	-13.99	0.001
word rate	0.52	0.03	0.45 – 0.58	15.12	0.001
syll	0.22	0.12	-0.00 – 0.45	1.93	0.054
frequency	0.12	0.05	0.03 – 0.21	2.52	0.012
top-down der	0.02	0.35	-0.66 – 0.71	0.06	0.953
ROI1	0	0.01	-0.01 – 0.01	0.26	0.795
ROI2	0	0	-0.01 – 0.01	0	0.999
mod1	-0.01	0	-0.02 – 0.00	-1.72	0.085
bottom-up der	0.09	0.16	-0.23 – 0.41	0.56	0.576
surprisal der	0.09	0.12	-0.14 – 0.33	0.76	0.448
open nodes der	0.04	0.05	-0.06 – 0.15	0.84	0.399
top-down	-0.1	0.16	-0.42 – 0.21	-0.62	0.533
bottom-up	0.07	0.1	-0.13 – 0.27	0.65	0.515
surprisal	0.16	0.02	0.11 – 0.21	6.54	0.001
open nodes	0.03	0.02	-0.00 – 0.06	1.81	0.071
top-down der * ROI1	0.33	0.24	-0.14 – 0.79	1.38	0.166
top-down der * ROI2	-0.22	0.14	-0.49 – 0.05	-1.62	0.105
top-down der * mod1	-0.02	0.35	-0.70 – 0.67	-0.05	0.962
ROI1 * mod1	0	0.01	-0.01 – 0.01	0.06	0.952
ROI2 * mod1	0	0	-0.01 – 0.01	-0.12	0.906
ROI1 * bottom-up der	0.19	0.14	-0.08 – 0.45	1.37	0.171
ROI2 * bottom-up der	-0.12	0.08	-0.27 – 0.04	-1.5	0.134
mod1 * bottom-up der	-0.02	0.16	-0.33 – 0.30	-0.12	0.904
mod1 * surprisal der	-0.43	0.12	-0.66 – -0.19	-3.56	0.001
ROI1 * surprisal der	-0.02	0.06	-0.15 – 0.11	-0.29	0.775
ROI2 * surprisal der	-0.19	0.04	-0.26 – -0.12	-5.08	0.001
ROI1 * open nodes der	0.02	0.05	-0.07 – 0.11	0.39	0.7
ROI2 * open nodes der	0.14	0.03	0.09 – 0.19	5.22	0.001
mod1 * open nodes der	0.11	0.05	0.01 – 0.21	2.09	0.037
ROI1 * top-down	-0.13	0.13	-0.38 – 0.12	-1.05	0.295
ROI2 * top-down	-0.49	0.07	-0.63 – -0.34	-6.63	0.001
mod1 * top-down	-0.67	0.16	-0.98 – -0.35	-4.16	0.001
ROI1 * bottom-up	0.09	0.11	-0.13 – 0.30	0.8	0.427
ROI2 * bottom-up	0.38	0.06	0.26 – 0.50	6.08	0.001
mod1 * bottom-up	0.57	0.1	0.37 – 0.77	5.54	0.001
mod1 * surprisal	0.01	0.02	-0.04 – 0.05	0.25	0.803
ROI1 * surprisal	0.08	0.02	0.04 – 0.12	3.68	0.001
ROI2 * surprisal	0.01	0.01	-0.01 – 0.03	0.81	0.417
ROI1 * open nodes	0.01	0.01	-0.00 – 0.03	1.38	0.166
ROI2 * open nodes	0.01	0	0.00 – 0.02	1.98	0.048
mod1 * open nodes	0.03	0.02	0.00 – 0.06	2.04	0.041
top-down der * ROI1 * mod1	-0.27	0.24	-0.74 – 0.19	-1.16	0.248
top-down der * ROI2 * mod1	0.52	0.14	0.25 – 0.78	3.79	0.001
(ROI1 * mod1) * bottom-up der	-0.05	0.14	-0.32 – 0.22	-0.37	0.711
(ROI2 * mod1) * bottom-up der	-0.13	0.08	-0.28 – 0.03	-1.6	0.109
(ROI1 * mod1) * surprisal der	-0.24	0.06	-0.37 – -0.12	-3.79	0.001

continued ...

…continued

	BOLD				
Predictors	**Estimates**	**SE**	**CI**	**Statistic**	***p***
(ROI2 * mod1) * surprisal der	0.07	0.04	-0.00 – 0.14	1.9	0.057
(ROI1 * mod1) * open nodes der	0.04	0.05	-0.05 – 0.14	0.94	0.345
(ROI2 * mod1) * open nodes der	-0.04	0.03	-0.09 – 0.01	-1.48	0.14
(ROI1 * mod1) * top-down	-0.27	0.13	-0.52 – -0.02	-2.13	0.033
(ROI2 * mod1) * top-down	-0.07	0.07	-0.22 – 0.07	-1	0.318
(ROI1 * mod1) * bottom-up	0.22	0.11	0.01 – 0.43	2.07	0.039
(ROI2 * mod1) * bottom-up	0.02	0.06	-0.10 – 0.15	0.39	0.696
(ROI1 * mod1) * surprisal	0.01	0.02	-0.03 – 0.05	0.58	0.565
(ROI2 * mod1) * surprisal	-0.02	0.01	-0.04 – 0.01	-1.5	0.133
(ROI1 * mod1) * open nodes	0.02	0.01	0.00 – 0.04	2.3	0.021
(ROI2 * mod1) * open nodes	0.01	0	0.00 – 0.02	2.69	0.007
N subj	52				
Observations	115743				
AIC	418263.25				

Table S4.6: Summary of model output of the pause length preceding each word's production. AIC stands for Akaike Information Criterion, used for the production only models to determine model fit.

			pause length		
Predictors	**Estimates**	**SE**	**CI**	**Statistic**	***p***
(Intercept)	0.15	0.02	0.11 – 0.19	7.11	**<0.001**
frequency	0.00	0.00	-0.00 – 0.00	0.36	0.722
surprisal	0.02	0.00	0.02 – 0.03	11.22	**<0.001**
syllables	-0.03	0.01	-0.04 – -0.02	-5.36	**<0.001**
bottom-up	-0.02	0.00	-0.03 – -0.02	-10.36	**<0.001**
open nodes	-0.03	0.00	-0.03 – -0.02	-13.66	**<0.001**
top-down	0.10	0.01	0.08 – 0.11	10.66	**<0.001**
N subj	16				
Observations	45079				
AIC	82830.780				

Table S4.7: Summary of model output of word duration. AIC stands for Akaike Information Criterion, used for the production only models to determine model fit.

			word duration		
Predictors	**Estimates**	**SE**	**CI**	**Statistic**	***p***
(Intercept)	0.27	0.01	0.25 – 0.29	28.22	**<0.001**
frequency	-0.02	0.00	-0.02 – -0.01	-15.44	**<0.001**
surprisal	0.00	0.00	0.00 – 0.01	9.38	**<0.001**
syllables	0.08	0.00	0.08 – 0.09	22.50	**<0.001**
bottom-up	0.01	0.00	0.01 – 0.01	12.61	**<0.001**
open nodes	-0.00	0.00	-0.00 – -0.00	-5.35	**<0.001**
top-down	-0.05	0.00	-0.05 – -0.04	-20.05	**<0.001**
N subj	16				
Observations	45079				
AIC	-42830.283				

Chapter 5

5 | Distributed Neural Representations for Semantic Structures During Sentence Production

Abstract

The neural representations for compositional processing have so far mostly been studied during sentence comprehension. In an fMRI study of sentence production, we investigated the brain representations for compositional processing during speaking. We used a rapid serial visual presentation sentence recall paradigm to elicit sentence production from the conceptual memory of an event. With voxel-wise encoding models, we probed the specificity of the compositional structure built during the production of each sentence, comparing an unstructured model of word meaning without relational information (’bag-of-nouns’) with a model that encodes abstract thematic relations (’man-as-agent’) and a model encoding event-specific relational structure (’man-as-attacker’). We found that relational structure was encoded in a gradient of specificity in the superior temporal cortex (STC). The mid-anterior STC encoded sentences with specific event (verb) representations, while the posterior STC encoded sentences abstracting over events. This posterior cluster also encoded the ordinal and syntactic structure of the sentence, indicating that during production the posterior STC is involved in binding lexical items to semantic roles, as well as to subject and object roles. Whole-brain analyses revealed that the encoding of sentence meaning at different levels of specificity was by no means limited to the STC and instead highlighted a large left fronto-parieto-temporal network. Overall, we found evidence for the encoding of specific and abstract semantic structures during sentence production in distributed brain representations.

5.1 Introduction

One of the most characteristic features of human language is its infinite combinatorial potential. The same words can be used in different structures and combinations to compose different meaning. For example, the sentences “the surfer sees the violinist” and “the violinist sees the surfer” are formed by the same words, but have different meanings that depend on the relational structure of the event. Similarly, in “the surfer sees the violinist” and “the surfer is seen by the violinist”, the content words are the same and in the same order, but the voice of the sentence (active vs. passive structure) indicates a different relational structure and meaning. The sentence-level meaning is thus composed by a combination of word meanings based on syntactic rules that help form a semantic structure characterizing a conceptual event representation (Jackendoff, 1992). Investigations on how sentence-level compositional meaning is supported by brain structure and function have only just started.

At the phrasal level, extensive MEG evidence has found the left anterior temporal lobe (LATL) and the ventro-medial prefrontal cortex (mPFC) to sequentially activate for semantic composition (e.g. composing “red boat” vs. “red blue”) (e.g. Bemis & Pylkkänen, 2011; Blanco-Elorrieta et al., 2018; Flick et al., 2018; Pylkkänen, 2020; Pylkkänen et al., 2014; but see Kochari, Lewis, Schoffelen, & Schriefers, 2021). At the sentence-level, the verb argument structure indicates the thematic roles of entities participating in an event (e.g. who did what to whom). The angular gyrus, extending to posterior superior temporal gyrus (STG) and middle temporal gyrus (MTG), shows increased activation for verbs with more complex verb argument structures (Meltzer-Asscher, Mack, Barbieri, & Thompson, 2015; Meltzer-Asscher, Schuchard, Ouden, & Thompson, 2013), and is involved in event processing (Matchin, Liao, Gaston, & Lau, 2019). The temporal lobe was found to be involved in processing thematic roles. The mid left MTG was found to be involved in decoding the thematic roles of nouns in a video depicting an action event, where a rabbit “punching a monkey” was classified as an agent vs. a patient (Wang et al., 2016). An adjacent region in mid left STG was found to decode the identity of agents and patients (in this case determining if the agent was a rabbit or a monkey) (Frankland & Greene, 2015, 2020b). Lesions to these mid regions of temporal cortex were also found to lead to deficits in accessing thematic role knowledge at the linguistic and conceptual level (Wu, Waller, & Chatterjee, 2007).

There is an advantage for modelling compositional effects at the sentence level. Word meaning constrained by verb semantics provides a better fit to brain data relative to unconstrained meaning, indicating incremental composition in a parieto-temporo-frontal network (Lyu et al., 2019). Sentence-specific propositional meaning best models brain activity relative to unstructured models that use individual word meanings separately (bag-of-words) and is found to be distributed across the brain and language-network (Anderson et al., 2021). Regarding how compositional meaning may be supported by brain structure, Frankland and Greene (2015) suggested a computer-like architecture where regions of the cortex flexibly encode the meaning of sentence variables associated with a specific conceptual organization. In particular, they found that a region of the left mid superior temporal cortex (lmSTC) can distinguish between mirror sentences, which use the same words to create different meaning (e.g. "the surfer sees the violinist" and "the violinist sees the surfer") and carries information about the identity of the actors of an event to the amygdala. Additionally, adjacent subregions of the lmSTC separately carry information about the identity of the agents and patients of the events, showing that they can be encoded simultaneously to potentially form complex representations. The anterior medial prefrontal cortex (amPFC) and the hippocampus encode verb-specific information about thematic roles, i.e. thematic roles in combination with a specific event (Frankland & Greene, 2020b). The results of these studies were integrated to propose a functional architecture for dynamically encoding sentence meaning in the brain relying on amPFC, lmSTC and hippocampal representations (Frankland & Greene, 2020a).

Our goal was to test this functional architecture for compositional processing in sentence production. Sentence production diverges from comprehension in having as its goal a linear sequence that the speaker eventually utters. In comprehension, instead, the input is in linear order and must be combined into an event representation. In production, a conceptualizer system generates a preverbal (or prelinguistic) message in the form of a structured conceptual representation (Guhe, 2007; Jackendoff, 1992; Levelt, 1989). These conceptual structures are thought to be shared between production and comprehension (Levelt, 1989). The semantic structure is then formulated into a linear sequence that eventually takes the form of a phonetic plan for the articulators (Levelt, 1989). The formulator requires semantic concepts (entities and events), semantic roles between concepts (the argument structure) and features (e.g. number) to be encoded by the conceptualizer (Guhe, 2007). The semantic roles that take part in the event are mapped to lexical items assigned to syntactic roles (e.g. subject, object, verb). The syntactic roles are then ordered into a phrase-structure frame that includes morphological and phonological specification (Bock & Levelt, 1994).

One critical characteristic of sentence production is its incrementality (e.g. Bock & Ferreira, 2014; Levelt, 1989). Preverbal message generation is incremental, thus proceeding in increments that are fed to the formulator as they become available, rather than waiting for the complete semantic representation to be available (De Smedt & Kempen, 1987). Similarly, formulation proceeds incrementally and continues while words are being articulated. At the conceptual level, incrementality is argued to be either linear or hierarchical (e.g. Konopka & Brown-Schmidt, 2014). According to linear incrementality, speakers encode the message and plan the sentence in an element-by-element fashion. In hierarchical incrementality, instead, the gist of the event is encoded first, including the relations between the objects (which generally end up constituting the verb), before formulation and articulation of the first element.

The experimental study of message generation is hindered by the difficulty to control the content of messages a speaker may want to convey. Since the output is linguistic, the distinction between conceptualization and formulation is reduced, making it hard to understand what level of processing is affected by a task (Bock & Ferreira, 2014; Levelt, 1989; Papafragou & Grigoroglou, 2019). Evidence on speech onset times shows that longer utterances have slower onset latencies. For instance, for the sentence “The dog and the foot move up and the kite moves down” the onset is later than “the dog and the kite move up”, even though the subjects are equally long, suggesting that there is high-level processing of an event for the remainder of the sentence before speech onset (M. Smith & Wheeldon, 1999). Eye-tracking also provides a means to access event processing and its relationship with sentence formulation (Z. M. Griffin & Bock, 2000). Eye-tracking studies show that there is an initial central fixation of 300-400ms for higher-level coding of the scene before attention is moved to the first item to be uttered (Z. M. Griffin & Bock, 2000; Konopka, 2019). This again suggests that planning is not exclusively linear, otherwise attention would be directed directly to the subject. Konopka (2019) showed that the event is encoded before speech onset by finding that speech onset fixations are directed to the character that is most informative for action encoding, providing evidence for early event encoding during conceptualization, in support of hierarchical incrementality. In addition, during masked scene processing participants can rapidly identify the agents and patients of the depicted action as measured by questions following the rapid scene presentation (Hafri et al., 2013). Therefore, entities participating in an event are connected to their thematic role in very short processing timescales during scene processing (cf. Grimshaw, 1990; Jackendoff, 1992). Overall, evidence from scene description experiments suggests that the event is encoded early and that eye movements roughly follow identities in a scene in the order of formulation and articulation (see for a review Papafragou & Grigoroglou, 2019).

Therefore, psycholinguistic evidence based on scene description experiments suggests that compositional structures are built early and quickly in production. However, onset latencies and eye fixations in picture description paradigms are indirect indicators of the neuro-cognitive processes that unfold during language planning and production. Advances in neuroimaging, especially decoding and feature-based encoding models, have made it possible to probe them more directly. As such, further confirmation that entities are bound to their event role would be provided if composed sentence meaning (i.e. word meaning in association with the relational structure of the sentence) were found to be decodable from brain activity during sentence production, as it was for comprehension. A neuroimaging approach also allows for evidence on compositional processing with production elicited from the memory of an event rather than a picture description. Differently from comprehension, the incrementality of the message generation and sentence formulation may lead the ordering of entities in the event to become more prominent in production. The accessibility of lexical items and the codability of the events to be described have been found to affect the order in which entities are uttered during sentence production (Kuchinsky et al., 2011; van de Velde et al., 2014). Similarly, the verb is not always planned before sentence onset (Momma & Ferreira, 2019; Momma et al., 2016). As a consequence, compositional representations at different levels of abstraction may be built at different time points between production and comprehension. In comprehension, the compositional meaning represents the end goal, while in production it is the prerequisite, but its building dynamics are unclear. Frankland and Greene (2020b) found evidence for both abstract representations (across event types) and specific representations (event/verb-specific) to be encoded in the brain, in different regions. Here, we tested whether this proposed infrastructure for compositional processing was equally distinct for specific and abstract representations during sentence production.

Previous studies found that a similar network is engaged during sentence production and comprehension with some asymmetries in the level of engagement. In particular, both temporal and frontal areas are involved in syntactic processing (Giglio et al., 2022; Hu et al., 2022; Matchin & Wood, 2020; Segaert et al., 2012), but production and comprehension were found to differently load on them and on the connectivity between them (Chapters 2-3, Giglio et al., 2022). In addition, production and comprehension elicit different BOLD latencies, that are thought to reflect inverse processing dynamics during planning and parsing, as also reflected by temporal dynamics in MEG (Blanco-Elorrieta et al., 2018; Giglio et al., 2022; Pylkkänen, 2020; Pylkkänen et al., 2014, ; Chapter 4). These results thus suggest that a similar network can be expected for compositional processing in production and comprehension, since their neural resources seem to be shared.

Therefore, we investigated whether in production compositional structure at different levels of abstraction relied on the same brain areas as previously found in comprehension, with the anterior prefrontal cortex encoding specific representations, the hippocampus encoding specific representations via pattern separation, and the lmSTC encoding abstract relations (Frankland & Greene, 2020b). We investigated the specificity of relational encoding using voxel-wise encoding models trained to predict brain activity from sentence descriptors positing different levels of noun-role combinations. We focused on the lmSTC, amPFC and hippocampus ROIs, but also ran whole-brain analyses to determine whether other regions were involved in compositional processing in production. In addition, due to the incrementality of sentence production, we expected sentence meaning to be encoded in brain activity also in sequential and syntactic characterizations of sentence structure. We looked for clusters sensitive to these different aspects of sentence organization in the whole-brain, and averaged the performance of these encoding models in the lmSTC ROI that was previously found to encode abstract event structure (Frankland & Greene, 2020b). Finally, we tested for the separability of encoding of thematic roles such as agent and patient, as well as grammatical roles like subject and object, using the same encoding model procedure as well as multivoxel pattern classification analysis.

5.2 Materials and Methods

Participants

Forty right-handed native Dutch speakers participated in the study after giving written informed consent (27 females, mean age = 21.5 years, range 18-49). The study was approved by the ethical committee for Region Arnhem-Nijmegen.

Participants reported having no language-related or neurological disorders and normal or corrected-to-normal vision. Two participants were excluded due to poor performance during the task, which did not leave enough trials per run for the analysis.

Materials

The stimuli consisted of 256 Dutch sentences made of a combination of 16 nouns and 16 verbs; an additional 36 sentences of the same type followed by a question; and 72 filler sentences. Half of the sentences of interest were in active voice, while the other half was in passive voice, half of which in one word order and half in another word order (allowed in Dutch: "de bokser wordt door de muzikant herkend" or "de bokser wordt herkend door de muzikant"). The filler sentences were composed of several verbs and nouns to increase variability of content and structure in the production.

The nouns referred to individuals playing either sport or music, 8 for each category (see Table 5.1). The verbs were all transitive verbs either expressing an action that involves contact, or that requires perception. All perception verbs were experiencer-subject verbs, with the perceiver (i.e. experiencer) expressed in subject position and the perceived item (i.e. stimulus) as an object (e.g. "the athlete sees the musician"). The two categories of nouns and verbs were matched in length and frequency (see Table 5.1). We used this categorical semantic distinction to be able to differentiate sentences in their semantic representation at the single word level, as well as in the combination of words. We contrasted semantic categories rather than individual verbs and nouns (as done instead in Frankland & Greene, 2020b), to increase variability in the production output and to avoid too much word repetition. Both noun and verb categories were found to be decodable before (Arana, 2022; Xu et al., 2018). Each verb was used in 18 sentences and was combined with each subject and object at least once, with as few noun-noun combinations for subject and object as possible across all stimuli.

Procedure

The experiment used a rapid serial visual presentation (RSVP) sentence recall paradigm. Participants read sentences word-by-word, with each word presented for 150 ms and no blank screen in-between words. After sentence reading, there was a short distraction task, where a list of four numbers was presented, followed

Table 5.1: Nouns and verbs used in different combinations to create the sentences of interest, with their mean length and log frequency values ± standard deviation.

Nouns		**Verbs**	
Sport	**Music**	**Contact**	**Perception**
schaatser (skater)	cellist (cellist)	vasthouden (hold)	opmerken (notice)
wielrenner (cyclist)	saxofonist (saxophonist)	aanvallen (attack)	zien (see)
voetballer (footballer)	violist (violinist)	wegduwen (push away)	ontdekken (discover)
surfer (surfer)	gitarist (guitarist)	krabben (scratch)	begluren (peep)
zwemmer (swimmer)	pianist (pianist)	schoppen (kick)	herkennen (recognize)
atleet (athlete)	drummer (drummer)	grijpen (grab)	bekijken (look)
turner (gymnast)	muzikant (musician)	trappen (kick)	aanstaren (stare)
bokser (boxer)	zanger (singer)	knijpen (squeeze)	waarnemen (perceive)
Mean length: 7.5 ±1.73	7.5 ± 1.12	7.87 ± 1.05	8 ± 1.58
Mean log freq.: 0.22 ± 0.43	0.20 ± 0.40	0.98 ± 0.60	1.04 ± 1.04

by a single number written in letters. Participants had to decide whether the last number was present in the list. After their response, they repeated aloud the sentence they had just read as they remembered it (see Figure 5.1). This paradigm was shown to lead to sentence production from conceptual memory of the sentence just read, rather than from verbatim memory (Lombardi & Potter, 1992; Potter & Lombardi, 1990, 1998; van de Velde & Meyer, 2014). We thus considered it to be a suitable paradigm to elicit sentence production with constrained stimuli. Since the task was hard, participants sometimes made mistakes in the recall. They were asked to always utter full sentences, even if they did not remember one of the items they read, by using one of the other words they had been exposed to.

The experiment was divided in 6 blocks of 60 sentences each. In 6 trials per block, a question was asked in place of producing. The questions asked about the identity of the agent or patient of the sentence just read (e.g. “Who kicked?” and a choice between “boxer” and “musician”). The questions were meant to have a measure of participants’ understanding of what they were reading.

fMRI data acquisition

MR data were acquired in a 3T MAGNETOM PrismaFit MR scanner (Siemens AG, Healthcare Sector, Erlangen, Germany) using a 32-channel head coil. The MRI protocol included a T1-weighted MRI scan for anatomical reference and several fMRI scans. The T1-weighted scan was acquired in the sagittal orientation using a 3D MPRAGE sequence with the following parameters: repetition time (TR)/inversion time (TI) 2300/1100 ms, echo time (TE) 3 ms, 8° flip angle, field of view (FOV) 256 mm × 216 mm × 176 mm and a 1 mm isotropic resolu-

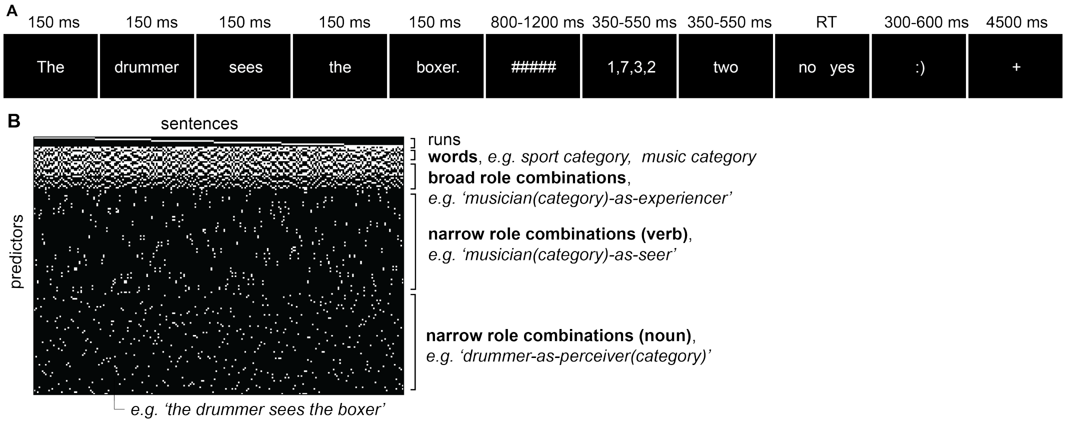


Figure 5.1: ***A****: Experimental procedure including presentation times and jittering between events. The inter-trial interval ranged from .5 to 3.5 sec (mean = 0.64 sec).* ***B****: All the predictors used for the encoding model of brain activity. Only a section of the predictors was used for each encoding model. The white spots indicate that a sentence feature (e.g. semantic category for the bag-of-nouns) was present in a given sentence.*

tion. Parallel imaging (iPAT = 2) was used to accelerate the acquisition resulting in an acquisition time of 5 min and 21 sec. Whole-brain functional images were acquired using a 3D EPI sequence following the implementation of Stirnberg et al. (Narsude, Gallichan, Zwaag, Gruetter, & Marques, 2016; Stirnberg et al., 2017). This choice was motivated by the need to have a short TR to help separability of the tasks, while keeping a good sensitivity to internal brain structures and reducing the sensitivity to motion (multi-band sequences are more sensitive to motion and with poorer resolution in medial structures). These were the parameters: TR 700 ms, TE 33 ms, flip angle 15°, FOV 210 mm x 210 mm x 150 mm, voxel size 2.5 mm isotropic. Fieldmap images were also acquired to correct for distortions. We acquired 6 fMRI runs per participant.

Behavioural analysis

The recording of each sentence was transcribed by a Dutch native speaker and rated for accuracy. We used two accuracy measures. One is standard accuracy for memory, where the sentence was rated to be correct if it was identical to the sentence that was just read. The other accuracy measure determined whether a sentence was suitable for the analysis, that is, whether the subject, object and verb were all part of the stimuli of interest, even if they did not match the sentence just presented. We decided to include these sentences in the analysis because we were primarily interested in sentence production, rather than correct memory. We also extracted the onset and offset times of the speech using Praat, after scanner noise removal. The onset and offset times were used for the timing of the production in fMRI analysis.

fMRI preprocessing

Preprocessing was performed using *fMRIPrep* 20.2.6 (Esteban, Blair, et al., 2018; Esteban, Markiewicz, et al., 2018), which is based on *Nipype* 1.7.0 (Gorgolewski et al., 2011, 2018).

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection, distributed with ANTs 2.3.3 (Avants et al., 2008) and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, Zhang et al., 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (A. Klein et al., 2017). Volume-based spatial normalization to MNI standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template.

Functional data preprocessing

For each of the 6 BOLD runs per subject, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. A B0-nonuniformity map (or fieldmap) was estimated based on a phase-difference map. The fieldmap was then co-registered to the target EPI (echo-planar imaging) reference run and converted to a displacements field map. Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al., 2002). The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD* run in *MNI152NLin2009cAsym* space. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Automatic removal of motion artifacts using independent component analysis (ICA-AROMA, Pruim et al., 2015) was performed on the *preprocessed BOLD* on MNI space time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). The "aggressive" noise-regressors were used for nuisance regression in first-level analysis. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), the derivative of the relative (frame-to-frame) bulk head motion variance (DVARS) and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al., 2014) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al., 2002). FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off). For anatomical CompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file.

First-level analysis for sentence *t* map extraction

To extract beta and *t* maps on which encoding and decoding procedures were run, we ran a first-level analysis in SPM12 in Matlab 2021a. We ran a GLM on the preprocessed files in MNI space and we included confound regressors from fMRIPrep for DVARS, Framewise Displacement, 6 aCompCor parameters and 6 motion parameters. We also added the AROMA noise components computed in fMRIPrep as additional nuisance regressors, to perform non-aggressive denoising, that better accounts for motion related noise. The design matrix included as condition regressors: one single regressor for all sentences read, one single regressor for the number task (from the presentation of the list of numbers to the response), one regressor for fillers, one regressor for question trials, and one regressors for all sentences that were erroneous. For the voxel-wise encoding model procedure, we included one regressor of interest per sentence, thus extracting about 40 beta maps per run (depending on the number of correct productions in each participant). For the multivoxel classification analysis, we instead included one regressor per combination of semantic categories across subject, verb and object, in each voice (active and passive, e.g. musician subject, sports object, perception verb, active voice, leading to 8 category combinations per voice). The onset of the sentence/combination conditions was set as the onset of speech, with duration until speech offset. These *t* maps representing one type of event in each voice were then fed to the classifier in different combinations for decoding of different dimensions of the sentence.

Encoding model analyses

Encoding models

We compared encoding models that hinged on different aspects of the sentence (Fig. 5.1). As introduced above, each sentence can be described in several ways: by its subject and object, that follow the order of the articulation in the sentence, or by its thematic roles. We used contact verbs, that take Agent and Patient thematic roles, and perception verbs, that take Experiencer and Stimulus thematic roles (experiencer-subject psych verbs). We focused on encoding models with two organizations: (a) models that represent the sentence structure at different levels of (compositional) specificity; (b) models that represent the sentence structure following different types of organization, i.e. ordinal, syntactic or based on the causal-temporal structure of the event.

Among the models of relational specificity, we used a bag-of-nouns model, which included the nature of the nouns present in the sentence independently of their role in the sentence (the 4 sentences in Table 5.2 would thus be deemed identical). This model thus had 2 parameters, one for each semantic category of the nouns. We also used a model encoding broad noun-role combinations across verbs (musician-as-agent), leading to 8 parameters (4 thematic roles x 2 categories). The agent/patient, stimulus/experiencer thematic roles were kept separate, meaning that implicitly we distinguished between event types (verb categories) as well, because agent/patient roles can only occur with contact verbs, and stimulus/experiencer roles with perception verbs, in our stimuli. Finally, a narrow model targeted noun-verb specific combinations, leading to 64 parameters (2 semantic categories for nouns x 16 verbs x 2 thematic roles). These models thus characterized each sentence at different levels of specificity in the relations between nouns in the event, that were previously proposed to be encoded in different brain regions. We tested this model of compositional meaning in a production study with different materials (semantic categories vs. word distinctions) and task (sentence recall vs. sentence reading; cf. Frankland & Greene, 2020b).

Table 5.2: Example sentence descriptors for 4 events in active and passive voice. These events are deemed equal by the bag-of-nouns model, but distinct in ordinal, syntactic and event structure.

Sentence	First noun	Second noun	Subject	Object	Agent/ stimulus	Patient/ experiencer
The musician hits the athlete.	Musician (agent)	Athlete (patient)	Musician (agent)	Athlete (patient)	Musician (agent)	Athlete (patient)
The musician is hit by the athlete.	Musician (patient)	Athlete (agent)	Athlete (agent)	Musician (patient)	Athlete (agent)	Musician (patient)
The musician sees the athlete.	Musician (experiencer)	Athlete (stimulus)	Musician (experiencer)	Athlete (stimulus)	Athlete (stimulus)	Musician (experiencer)
The musician is seen by the athlete.	Musician (stimulus)	Athlete (experiencer)	Athlete (experiencer)	Musician (stimulus)	Musician (stimulus)	Athlete (experiencer)

Among the structural models, the ordinal model focused on the semantic categories of the first and second noun, leading to 4 parameters (2 categories x 2 positions). This model also reflects the prominence that the first mentioned items receive, since they are the ones the sentence is about (Bornkessel-Schlesewsky, Grewe, & Schlesewsky, 2012). The syntactic model targeted deep syntactic roles, thus distinguishing external arguments from internal arguments, or grouping deep subjects, i.e. the surface subject of active sentences with the indirect object of passive sentences (see Table 5.2, Grimshaw, 1990). Since we only used experiencer-subject verbs, this meant grouping agents with experiencers, and patients with stimuli. This model thus also tested for an event organization that focuses on the mental state of event participants, where agents and experiencers are the more prominent arguments, because they are sentient (Levin & Hovav, 2005). This model had 4 parameters as well (2 categories x 2 syntactic roles/argument groupings). The final structural model focused on the causal-temporal structure of the event, where agents and stimuli are grouped because they are the “cause” of the event, which temporally is perceived to happen first, and patients and experiencers are the ones affected by the event (Frankland

& Greene, 2020b; Jackendoff, 1992). Again, this model had 4 parameters (2 categories x 2 thematic role groupings). By comparing these models, we could determine whether different areas of the brain were sensitive to these different structures. Additionally, we could learn whether ordinal and syntactic structure were encoded more strongly due to the production nature of the task, compared to previous comprehension results (Frankland & Greene, 2020b).

Voxel-wise encoding model procedure

The voxel-wise encoding model analyses followed the procedure in Frankland and Greene (2020b) as closely as possible. Some differences were unavoidable due to the different design and materials. The encoding models were run in *Nilearn* (Abraham et al., 2014; *nilearn/nilearn*, 2022) on the beta maps for each sentence in a grey matter mask that included the cortex and subcortical structures relevant in production (caudate and putamen, hippocampus and cerebellum).

The encoding models were trained to predict the activity associated with each sentence at each voxel as a linear combination of sentence dimensions. The models were trained on 5 out of 6 runs and tested on the held-out sentences from the remaining run in a k-fold cross-validation procedure. We thus extracted beta parameters per sentence descriptor per voxel per subject in a multiple regression of the sentence descriptors relevant for each encoding model. The model's performance was then evaluated by comparing predicted values generated from the learned beta parameters for each voxel for the held-out sentences with observed values. The comparison was performed with z-scored squared differences between the predictions and observation in each run (leading to a predictions x observations squared matrix). We then averaged the on-diagonal elements (which correspond to correct mappings between predicted and observed sentences, i.e. predicted activity for sentence$_i$ vs. observed activity for sentence$_i$) and the off-diagonal elements (corresponding to incorrect mappings, i.e. predicted activity for sentence$_i$ vs. observed activity for sentence$_j$). Finally, we compared these averages: if the voxel encoded information on the sentence descriptors, the average for the correct mappings (difference between predicted and observed values) should be lower than the average for the incorrect mappings. The difference between correct and incorrect mappings was then averaged across the 6 runs for each voxel and multiplied by -1, so that informative voxels were represented as greater than zero. We then smoothed the maps at 8 mm FWHM, since the encoding model procedure was run on unsmoothed preprocessed data on each voxel separately.

Group-level whole-brain and small-volume correction analysis

To run group-level whole-brain analysis, we used a permutation procedure in pyMVPA (Stelzer, Chen, & Turner, 2013). We created 100 permutations for each encoding model after permuting the sentence descriptors for each sentence. A cluster forming threshold was estimated via bootstrapping of the permutations (n=100 per model per participant) to estimate the group-average accuracy maps under the null hypothesis. For whole-brain analyses, we used a $p < 0.005$ voxelwise threshold, and a $p < 0.05$ cluster correction (False Discovery Rate, Benjamini-Hochberg procedure Benjamini & Hochberg, 1995).

For the exploratory analyses in the STG mask, we selected a mask of anterior and posterior superior temporal gyrus based on the Harvard Oxford atlas (Desikan et al., 2006). We then ran one-sample t tests against 0 for the generalization performance with a group-level analysis in SPM12 using the STG mask as explicit mask. We used a $p < 0.05$ voxelwise threshold, and a $p < 0.05$ FWE-corrected cluster threshold. Note that voxelwise p thresholds are selected based on the thresholds indicated by Frankland and Greene (2020b) for the different analyses.

ROI analysis

Following Frankland and Greene (2020b), we also inspected average generalization performance of the encoding models in 3 a priori regions of interest, left mid-STG, left hippocampus and left anterior medial PFC. The left hippocampus ROI was extracted from the Harvard Oxford atlas (Desikan et al., 2006) and masked with the grey matter mask used in the encoding models. The lmSTG and amPFC ROIs were extracted as grey matter voxels surrounding the peak voxels indicated in Frankland and Greene (2020b) transformed to MNI coordinates. For the models distinguishing sentence structure, we averaged encoding model performance in the same lmSTG region defined a priori. We then compared generalization performance of each model with *emmeans* package (Lenth et al., 2022) after running a linear mixed-effects model with *lme4* in R (Bates et al., 2015) with encoding model as a fixed effect and by-participant random effects.

For the models distinguishing relational specificity, we focused on voxels in all three ROIs that were found to have significant generalization performance with a full encoding model, that included all models at different levels of specificity. We could then determine whether it was one specific model that drove the generalization performance of the full model. We selected voxels independently for each participant, by selecting voxels with $t > 2.4$ (or $t > 1.68$ if no voxels survived the higher threshold) in all participants except for one held-out participant iteratively. We could then localize regions independently for each participant, and average each model's performance in these regions. We then ran a linear mixed-effects model with *lme4*, predicting model's generalization performance with factors ROI (lmSTC, L hippocampus and LamPFC) and model (bag-of-nouns, broad roles and narrow roles), and by-participant random slopes for ROI and class. We ran pairwise comparisons using *emmeans* (Lenth et al., 2022).

Multivariate pattern classification

The encoding analysis described so far had the advantage of considering multiple aspects of a sentence at once, thus identifying voxels sensitive to sentence-level meaning at different levels of specificity or structure. However, the encoding model procedure is voxel-wise and does not consider distributed patterns of activity. Therefore, to ask where the different dimensions of a sentence are decodable, we ran a multivoxel pattern classification of semantic categories that characterized each sentence. We classified the semantic category for subject, object, verb, agent, patient, experiencer and stimulus from different grouping of the t maps extracted for each possible category combination.

We ran a classification of semantic categories from t maps extracted after first-level analysis in SPM12. The classification was performed in pyMVPA with cross-validation using a linear SVM classifier, with a searchlight across the whole brain. We also created 100 searchlights after attribute permutations to be able to run multiple comparisons correction at the group level (following Stelzer et al., 2013). A cluster forming threshold is estimated via bootstrapping of the permutations to estimate the group-average accuracy maps under the null hypothesis. Since this analysis was not run in Frankland and Greene (2020b), we used $p < 0.001$ voxelwise (in line with other chapters), $p < 0.05$ FDR corrected clusterwise.

5.3 Results

Behavioural results

First, we determined whether participants suitably performed the distraction task, thus ensuring they were not actively rehearsing the sentence just read for the following production. All participants had above chance accuracy in the distraction task (mean = 0.91, SE = 0.01, with average reaction time 451.1 ms, SE = 16.7). On the main sentence recall task, performance varied consistently across participants (Fig. 5.2). Some participants exhibited relatively poor memory for the sentence just read, while others had very high accuracy (mean = 0.77, SE = 0.02). However, we were not just interested in sentences purely accurate in memory, but sentences that were constituted by words that are part of the stimulus set were also considered acceptable (Table 5.1). These sentences additionally increased the separability between stimuli read during comprehension and spoken during production (23% of acceptable sentences were different from previously read sentences). The number of 'acceptable' sentences was higher than 'accurate' sentences across participants (mean = 0.93, SE = 0.007). Incorrect sentences included words that were not part of the stimuli (e.g. uttering "man" instead of "musician"), or sentences with different structure or that did not include two participants and a verb. We also checked the performance during question trials, to make sure participants were correctly parsing the sentences. Performance was overall good (mean = 0.83, SE = 0.016), but one participant that had accuracy around chance level was excluded (accuracy = 0.53). Finally, we determined what parts of the sentence were more prone to errors and therefore possibly encoded less strongly. Verbs had the highest accuracy (mean = 0.94, SE = 0.006), followed by the first noun (mean = 0.93, SE = 0.007), while the second noun was the most likely to be forgotten (mean = 0.86, SE = 0.014; $ts > 7.7$, $p < 0.0001$; verb vs. first noun: $t = 2.35$, $p = 0.07$ corrected for multiple tests). Errors were more likely to happen in passive relative to active sentences (accuracy, active = 0.83, SE = 0.017; passive = 0.72, SE = 0.025; $t = 9.8$, $p < 0.0001$; acceptability, active = 0.94, SE = 0.006; passive = 0.92, SE = 0.009; $t = 3.8$, $p = 0.0005$).

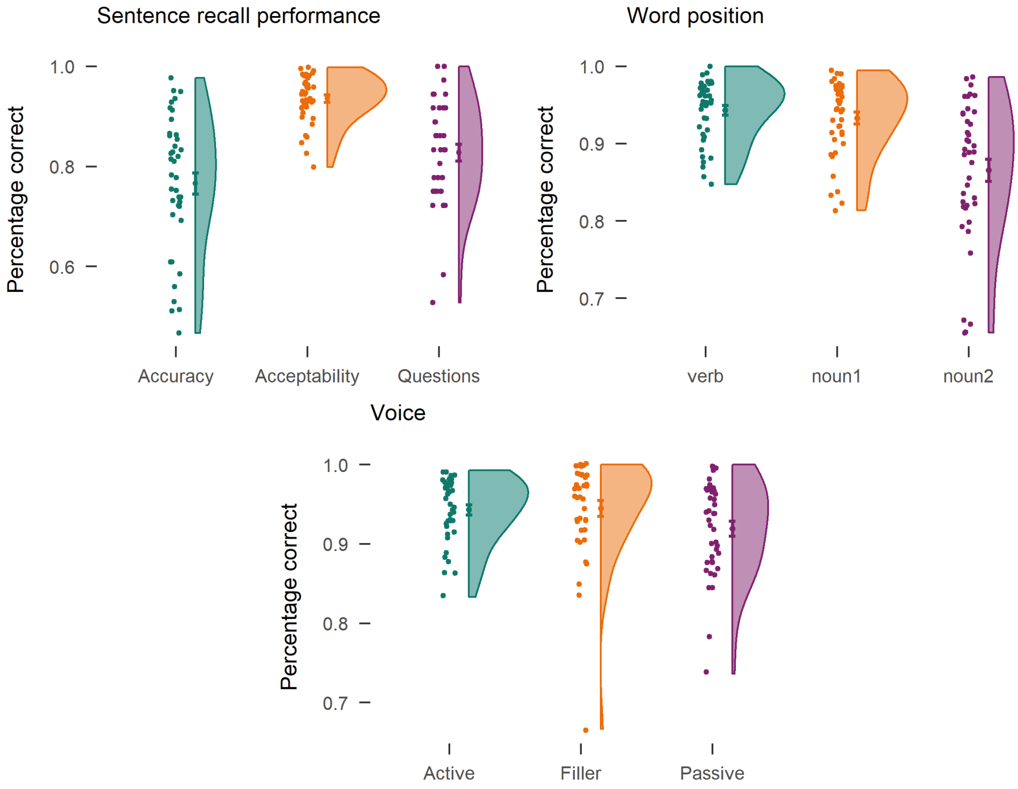


Figure 5.2: Summary of the behavioural performance. Points on the left of the violin plot represent individual participants' performance. Points on the violin plot represent the mean, error bars standard error of the mean. Voice performance relates to acceptability values.

Specificity of event encoding in lmSTC, hippocampus and amPFC

ROI analysis

First, we asked whether the compositional structure of sentences was decodable in production and we sought to replicate the ROI organization for the specificity of relational encoding (Frankland & Greene, 2020b). We reasoned that the conceptual organization of events should be similarly encoded in production and comprehension, but it may be less decodable in production due to the incrementality of message generation. We compared a bag-of-nouns model (presence of music and sport actors (by category) in the sentence, defining all words in italics independently of position and thematic role: "the *violinist* attacks the *surfer*", musician and athlete present) to a broad roles model (musician-as-agent/stimulus/patient/experiencer: "the violinist attacks the surfer" is encoded as musician-as-agent, athlete-as-patient), and a narrow roles model (musician-as-attacker and athlete-as-attackee, using the specific verbs instead of the thematic role associated to a type of event). Since thematic roles agent/patient and stimulus/experiencer are defined based on the type of verb, the broad roles model encoded thematic roles across verbs, but not independently of verb categories (i.e. contact or perception verbs). The narrow roles model instead included idiosyncrasies in verb meaning. Note that in these three models the event participants were not defined individually, but by their semantic categories.

First, we looked for clusters of voxels that had significant generalization performance within each ROI for each model. No model was significant in the amPFC and hippocampus ROIs, while all models led to significant clusters in the lmSTC (bag-of-nouns: k = 77, $p = 0.024$; broad roles: k = 62, $p = 0.016$; narrow roles: k = 46, $p = 0.021$; threshold voxelwise $p < 0.01$, clusterwise $p < 0.05$ FWE-corrected). To learn more about the specific representational profile of each region, we averaged the generalization performance of each encoding model in the left hippocampus, left amPFC and lmSTC ROIs (Fig. 5.3E). We selected voxels that performed above threshold in the full encoding model (including sentence dimensions at different levels of specificity) independently for each participant with a cross-validation procedure (see Methods for details). We then compared the average performance across the different ROIs and encoding models with a linear mixed-effects model with type of encoding model and ROI as factors (Fig. 5.3D). There was a main effect of ROI ($\chi^2 = 15.16$, $p < 0.0006$) and an interaction between model and ROI ($\chi^2 = 12.35$, $p < 0.015$). The effect of ROI indicated that the lmSTC had overall higher performance (lmSTC-amPFC: $t = 2.1$, $p = 0.09$; lmSTC-hippocampus: $t = 3.8$, $p = 0.0014$). All models showed above zero generalization performance in the lmSTC (bag-of-nouns: $t = 1.8$, $p = 0.07$; broad roles: $t = 2.6$, $p = 0.01$; narrow roles: $t = 5.7$, $p < 0.0001$), while in the amPFC only the broad roles model approached significance ($t = 1.88$, $p = 0.06$), and no model was significant in the left hippocampus. The interaction between ROI and model was due to the models being significantly different in lmSTC (bag-of-nouns and broad roles had lower performance than narrow roles: ts > 2.7, ps < 0.021), but not in the other ROIs. Therefore, we found almost opposite results to the ones found before (Frankland & Greene, 2020b). The lmSTC now was found to encode relational information at different levels of specificity, but most strongly in narrow roles, while it was sensitive only to broad roles in Frankland and Greene (2020b). Instead, the amPFC was only marginally significant for broad roles, whereas Frankland and Greene observed selective encoding of narrow roles in that region.

One critical difference between the current design and the one used by Frankland and Greene (2020b) was the use of semantic categories for characterizing nouns and verbs as opposed to individual words. Our decision to use categories was motivated by the preference for more varied production output and less repetition. However, characterizing a sentence's compositional meaning in terms of semantic categories rather than specific nouns leads to differences in the hypotheses about relational specificity. All the analyses described so far characterized nouns in terms of their thematic category (music vs. sport), while the

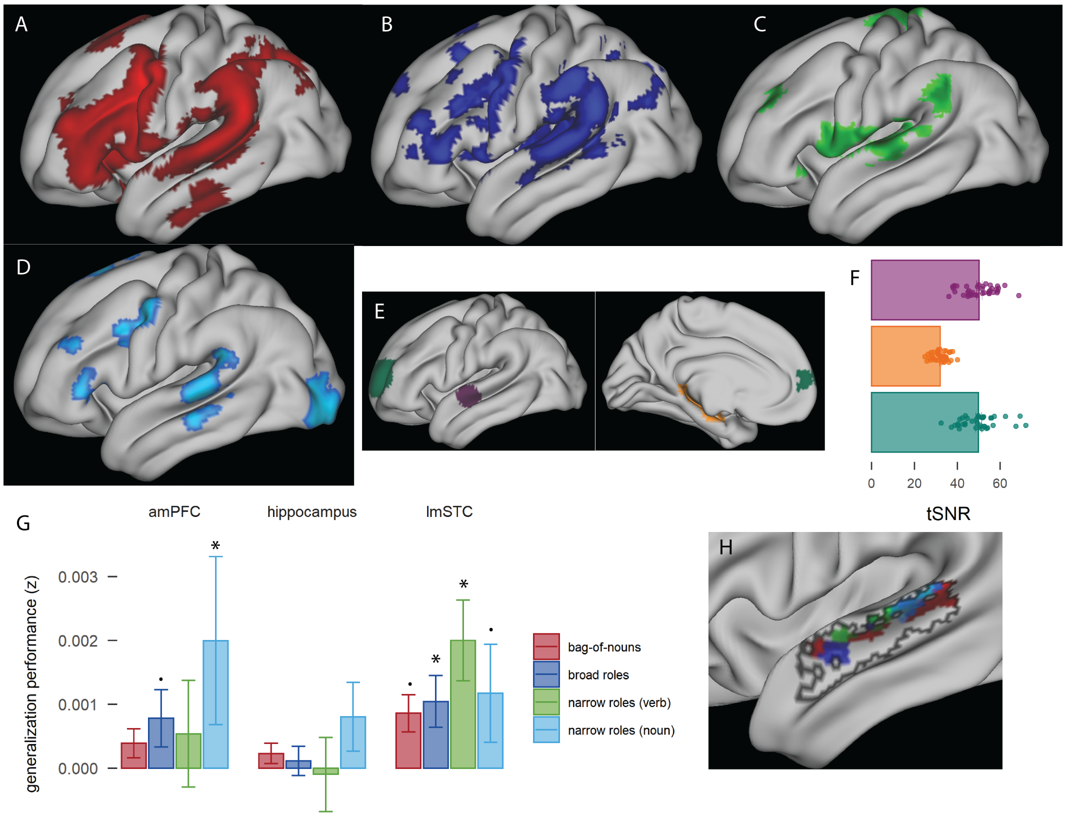


Figure 5.3: Specificity of event encoding. **A**: Whole-brain results for the generalization performance of the bag-of-nouns model. **B**: Whole-brain results for the broad roles model. **C**: Whole-brain results for the narrow roles model (specific for verb). **D**: Whole-brain results for the narrow roles model specific for nouns. All results are thresholded at $p < 0.005$, $p < 0.05$ FDR cluster corrected. **E**: ROI masks used for the ROI analysis. lmSTC: purple. amPFC: blue. Hippocampus: orange. **F**: average tSNR in each ROI, colors related to ROI colors. **G**: Average generalization performance in 3 ROIs, selected independently for each participant, shown in **E**. * indicates $p < 0.05$, while . is p < 0.1. **H**: Significant clusters in STG mask outlined in black. Colors respect model colors shown in whole-brain and ROI average results.

specific verb meaning was investigated with the encoding model of narrow noun-role combinations. Therefore, we ran a follow-up analysis to determine whether the specific noun meaning affected the way the event was encoded in the ROIs. We used an encoding model where each noun was associated to a thematic role and verb category (contact or perception) (e.g. “the violinist attacks the surfer” is encoded as violinist-as-agent, surfer-as-patient). We then determined its generalization performance in significant voxels of a full model that included both specific nouns and verbs (i.e. both narrow nouns and narrow verb combinations). This narrow role model for nouns was significantly different from zero in the amPFC (estimate = 0.002, $t = 2.28$, $p = 0.026$), and marginally different from zero in lmSTC (estimate = 0.0012, $t = 1.67$, $p = 0.098$). In the hippocampus it was not significant (estimate = 0.0008, $t = 1.25$, $p = 0.22$).

These results thus suggest a pattern of encoding where the amPFC is sensitive to noun-specific information in relation with a broad event type, while the lmSTC is most sensitive to the specific event with general semantic roles, but possibly also encodes the individual roles in the event. Frankland and Greene (2020b) had instead found that the amPFC and the hippocampus were sensitive to the specific event, while the lmSTC was sensitive to broad event roles. Note that here it was not possible to test a model where both nouns and verbs were specific, because the noun-verb combinations exceeded the number of sentences produced (16 nouns x 16 verbs x 2 thematic roles = 512 predictors for an average of 240 ± 10 sentences). The hippocampus was not found to be involved in the encoding of relational information. The temporal signal-to-noise ratio was much lower in this region relative to the amPFC and lmSTC (Fig. 5.3F), suggesting that the poor signal quality may have affected generalization performance.

Whole-brain results and STC small-volume correction

We also searched the whole-brain for clusters encoding roles at different levels of specificity, since other areas may have been sensitive to encoding relational roles. All encoding models, including the narrow nouns model, were significant in many fronto-temporal clusters (Fig. 5.3A-D, see Supplementary Figures S5.1-S5.4 for whole-brain figures, and Supplementary Table S5.1 for peaks). The bag-of-nouns model led to the most widespread clusters, in large parts of left posterior and inferior temporal lobe, extending into inferior parietal lobe, and inferior-middle frontal gyri, as well as precentral gyrus. The broad roles model had largely overlapping clusters. The narrow roles model for verbs instead was more limited, with a cluster in posterior STG extending into temporo-parietal junction, and a cluster in inferior frontal cortex. The narrow roles model for nouns was significant in a cluster in posterior STG and mid MTG, in inferior frontal gyrus and precentral gyrus and inferior occipital cortex. All models thus seemed to be localized to language-relevant areas, and to largely overlap in their distribution. However, none of the models was significant in the anterior medial frontal cortex.

The whole-brain analysis indicated that the areas of the superior temporal cortex that were sensitive to relational information were localized more posteriorly than reported previously (Frankland & Greene, 2020b) and used in the ROI analyses here. Therefore, we looked for significant clusters for each encoding model in a larger STG mask to understand where each model performed best (Fig. 5.3H). The bag-of-nouns model led to the largest cluster spanning most of the gyrus and peaking in the posterior part (peak: [-64 -40 16], $p < 0.001$ FWE-cluster corrected). The broad roles model cluster extended along the posterior part of the gyrus overlapping with the bag-of-nouns model and had a more mid cluster along the sulcus, partly overlapping with the lmSTC ROI (peak: [-66 -18 6], $p < 0.001$ FWE-cluster corrected). The narrow roles model (with verbs) had a cluster in the mid-STG, partly overlapping with the lmSTC ROI (peak: [-64 -20 6], $p = 0.082$ FWE-cluster corrected, $p = 0.007$ FWE corrected at the peak level). The narrow model with nouns had a cluster in posterior STG, overlapping with the posterior parts of the bag-of-nouns and broad roles model (peak: [-64 -35 9], $p = 0.002$ FWE-cluster corrected). These results, therefore, suggest that a more posterior section of the superior temporal cortex is involved in coding for noun-role combinations that are specific for nouns but abstract for verbs (events). The mid-anterior STC section instead codes for specific verb (event) representations, but abstract noun identity, as it is relatively less sensitive to the specific identity of the event participants.

Ordinal, syntactic and event structure are encoded in the posterior STC

Next, we investigated to what extent the sequential and syntactic structure of the sentence are encoded in the superior temporal cortex in production. Frankland and Greene (2020b) found that the lmSTC region, that encoded abstract relational structures, encoded sentences most strongly following event organization, whereas syntactic and especially ordinal structure were not reliably encoded. Because of the incrementality of sentence production, we expected sequential information to be encoded in the brain more strongly, and potentially to affect the representations in the same ROIs that represent relational structure. Therefore, we tested for sequential, syntactic and event structure in the whole-brain and in the lmSTC ROI.

The ordinal model groups event participants by the order in which they are uttered (e.g. focusing on similarities between "the *musician* attacks the athlete" and "the *musician* is attacked by the athlete", i.e. the noun in italics is given an identical state, see Table 5.2), regardless of their thematic role, which depends on the voice and verb of the sentence. The syntactic model considers the subject of active sentences identical to the indirect object of passive sentences, independent of the verb (e.g. focusing on similarities between "the *musician* attacks the athlete" and "the athlete is seen by the *musician*"). This approach hinges on the similarity between items that are syntactically connected to the verb, distinguishing external arguments from internal arguments, independently of their thematic role in the event (Grimshaw, 1990). However, it should be noted that because of the nature of the verbs agents and experiencers were contrasted to patients and stimuli, which could nevertheless reflect a semantic clustering where agents and experiencers are grouped because their mental state receives more focus than that of patients and stimuli (Levin & Hovav, 2005). Finally, the event model groups agents with stimuli and patients with experiencers following a causal-temporal reading of the event ("the *musician* attacks the athlete" and "the athlete sees the *musician*"), where patients and experiencers are the ones affected by the event (Jackendoff, 1992). Preference for this interpretation of the event was found in the lmSTC by Frankland and Greene (2020b).

We looked for clusters that encoded actors in a sentence in ordinal, syntactic and event structure in the whole-brain, and established generalization performance of an encoding model trained with these three types of structure in the lmSTC. We found that a mostly left-lateralized fronto-parieto-temporal network had above chance generalization performance for all three models (Fig. 5.4A-C). The clusters for all three models centred around the left posterior STG extending into inferior and superior parietal cortex and IFG, precentral gyrus, and supplementary motor cortex and posterior cingulate cortex. There were also a few clusters in the right hemisphere (see Supplementary Figs. S5.5-S5.7 for right hemisphere figures and Supplementary Table S5.1 for cluster peaks).

We also looked at the average generalization performance in the lmSTC, that was found to be sensitive only to event structure before (Frankland & Greene, 2020b). Generalization performance was found to be positive for all models (ordinal: $t = 2.9$, $p = 0.005$; syntactic: $t = 2.3$, $p = 0.023$; event: $t = 3.9$; $p = 0.0002$), but was not significantly different between structure models ($p > 0.1$) (Fig. 5.4D). From the whole-brain analysis, it appeared that all models had above chance generalization performance in clusters peaking more posteriorly than the lmSTC. We thus also explored where the encoding of the identity of the sentence roles was most strongly encoded in the posterior STG. We looked for clusters that had significant generalization performance in the STG mask for each model separately. We found that they mostly overlapped in a cluster just posterior to the lmSTC (Fig. 5.4E).

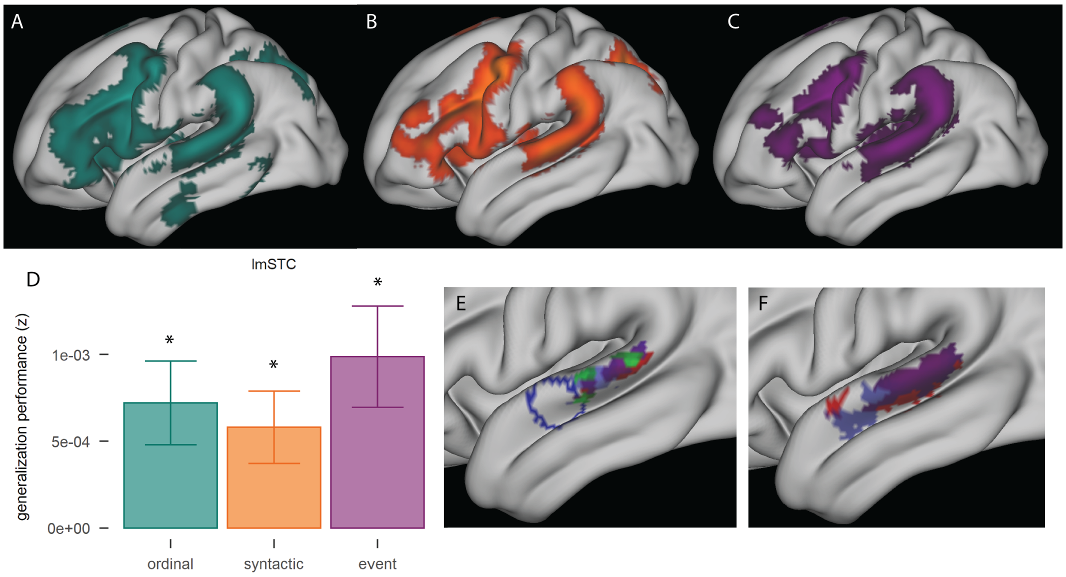


Figure 5.4: Ordinal, syntactic and event models. **A**. Whole-brain results for the generalization performance of the ordinal model. **B**. Whole-brain results for the syntactic model. **C**. Whole-brain results for the event model. All results are thresholded at $p < 0.005$, $p < 0.05$ FDR cluster corrected. **D**. Average generalization performance in a priori lmSTC ROI shown in blue in E. **E**. Significant clusters for the three encoding models in the STG mask. Ordinal model in red, event model in blue superimposed on the ordinal model (purple indicates overlap), syntactic model in green overlapping with other models. lmSTC ROI outline in blue. **F**. Clusters for agent/stimulus (blue) and patient/experiencer (red) encoding models in the STG mask ($p < 0.05$ voxelwise).

Therefore, the ROI results and the whole-brain results provide evidence that all three models of encoding are represented, in partly overlapping regions. These results suggest that there are different dimensions to the encoding of events and sentences, that are all functionally active during sentence production. In addition, these models of sentence structure all peaked in clusters that were close to the posterior ROI found to encode the specific relational structure at the level of event participants, rather than event types. This set of results tentatively suggests that a posterior STG cluster that encodes a broad event structure with the identity of the event participants is also sensitive to ordinal and syntactic representations of sentence structure.

Decoding the elements of a sentence

Finally, we sought to replicate the separate encoding for the identity of the agent and patient that was found by Frankland and Greene (2015, 2020b). In two studies, they had found that a lateral region of the lmSTC encoded only the identity of the patient, and a medial region of the lmSTC encoded the identity of the agent. To do this, we ran a searchlight classification analysis similar to Frankland and Greene (2015) in the whole-brain to understand how replicable the separate encoding in the superior temporal cortex was. We also ran this classification for subject, object and verb roles, to determine whether a similar segregation could be found for other sentence features. Note that this analysis differs from the encoding model analysis in two aspects: (1) we are now looking at the semantic value for one word in the sentence at a time, while the encoding models were characterizing all participants together; (2) this analysis is multivariate, thus looking for distributed patterns of encoding across voxels, while the encoding model was voxelwise (finding one value per voxel independently of the other voxels).

We found that the semantic category for a single element of a sentence was not always decodable (Supplementary Fig. S5.8). The category for the agent was decodable in a few clusters in the left fronto-temporal network, in posterior MTG and STG, left precentral gyrus and superior parietal lobule. The experiencer instead was decodable in a cluster in the visual cortex. The patient, stimulus and subject categories were not decodable. Given the potentially reduced power due to analyzing half of the sentences for each of the verb-arguments, we also tested for the causal thematic role combination (similar to the event structure encoding model) and found a cluster for the patient/experiencer in posterior MTG, precentral gyrus and superior parietal lobule. Instead, the decoding for the object was very widespread, and was very similar to the clusters of the encoding model (e.g. bag-of-nouns, ordinal), suggesting distributed rather than separate encoding of sentence roles. The verb decoding led to one cluster in posterior MTG. This pattern of results seems to indicate that there is less power in focusing on one feature of each sentence separately, while ignoring all other features, which may lead to more noisy estimates. In addition, there may be some washing out of activity related to items encoded and articulated early, due to the encoding of later words. Indeed the object, which appeared in last position in 75% of trials, was the one with the strongest decoding performance.

Therefore, we also took Frankland and Greene's (2020b) approach to use encoding models to probe the separation of agent and patient categories. We used encoding models that separately focused on the agent or patient category, using the causal-temporal clustering in line with Frankland and Greene's (2020b) results (agent/stimulus and patient/experiencer). Within the lmSTC ROI, there were no significant clusters for the agent or patient ($p < 0.05$ voxelwise, $p > 0.07$ clusterwise), but there was a significant peak for the agent ([-49 -8 -6], $p = 0.031$ peak-level). Therefore, we extended the search to the superior temporal cortex and found a marginally significant cluster for the agent in most of the ROI ($p < 0.05$ voxelwise, $p = 0.081$ clusterwise, peak [-52 -5 11]). There was a cluster for the patient that peaked more posteriorly but encompassing most of the ROI as well ($p < 0.05$ voxelwise, $p = 0.03$ clusterwise, peak [-62 -32 9]). Therefore, we found evidence for the agent being encoded in the same ROI previously indicated by Frankland and Greene (2015, 2020b), while the patient was encoded in the posterior STC (Fig. 5.4F). At their voxelwise threshold ($p < 0.05$), both thematic roles were encoded along the superior temporal mask, thus finding no evidence for segregation.

5.4 Discussion

In an fMRI study of compositional processing in sentence production, we used encoding and decoding techniques to investigate the representation of the compositional meaning of sentences during a sentence production task. The goals were to (i) replicate the organization of relational structure at different levels of abstraction found in a previous study (Frankland & Greene, 2020b), and (ii) determine whether the syntactic and ordinal structure of sentences is represented more strongly in production, due to its incrementality. Among the ROIs previously used by Frankland and Greene (2020b), we found that the superior temporal cortex encoded sentence meaning at different levels of specificity and in different structural organizations. The amPFC coded events broadly with specific representations for event participants. The hippocampus, instead, was not found to significantly encode sentence meaning at any level of abstraction. Whole-brain analyses revealed that a left fronto-parieto-temporal network encoded semantic structures at different levels of specificity. Overall, our results partly deviate from the representational structure that emerged in Frankland and Greene's study, possibly due to differences in the task, in the materials and in modality (production vs. comprehension). In addition, they suggest a distributed type of representational structure.

Compositional processing in production

First, we found evidence for the compositional representation of sentences in production, despite incremental message generation. Neural representations for compositional structure were found for a sentence recall task that highly differs from the scene description experiments with which message generation in production is usually studied (Konopka & Brown-Schmidt, 2014). This task is thought to resemble sentence production starting from a conceptual representation built during the comprehension of the sentence[1] (Lombardi & Potter, 1992; Potter & Lombardi, 1990, 1998), and is thus more similar to the conditions with which compositional processing is usually studied in sentence comprehension (e.g. Arana, 2022; Frankland & Greene, 2015, 2020b). Therefore, even though the conceptual representation of a sentence can be formed incrementally, with an event type possibly selected after speech onset (e.g. Momma et al., 2016), the words of the sentence are successfully composed into a specific relational structure that is decodable from brain activity. Although our results are not informative about timing, they show that a rich semantic structure can be decoded in production despite its incrementality, possibly because the rapid and spontaneous binding of thematic roles to event participants applies not only to scene processing but also to conceptualization in message generation (Hafri et al., 2013; Hafri, Trueswell, & Strickland, 2018).

Spatially, we found that the superior temporal cortex successfully encoded noun-role combinations in an event both abstracting over events ('musician-as-agent') and with specific event representations ('musician-as-attacker'). A bag-of-nouns model that did not encode relational information was also significant across the superior temporal cortex, but it performed less well than the event-specific encoding model ('musician-as-attacker'). A mid-anterior section of the STC was more sensitive to specific event information ('musician-as-attacker'), while a more posterior section encoded sentence information at different levels.

---

[1] Note that it is unlikely that the representation we decoded was in effect from the comprehension of the sentence, since production followed a distraction task, and the produced sentences differed from the sentences comprehended on 23% of trials.

In particular, the ordinal and syntactic structure of sentences was also encoded preferentially in the more posterior cluster. During speaking, the lexical items for event participants bound to a thematic role are functionally linked to a syntactic role (e.g. subject, object). Functional roles can then be bound into an ordered constituent structure (Bock & Levelt, 1994). Our results delineate a gradient of processing along the superior temporal cortex with compositional representations encoded more anteriorly and more abstract event representations as well as ordinal structure encoded more posteriorly to then be processed in inferior frontal areas for linearization and eventually articulation (Bornkessel-Schlesewsky & Schlesewsky, 2013; Chapter 3; see Leshinskaya & Thompson-Schill, 2020, for a similar gradient in perceptual-to-relational memory in visual processing in the RMTG).

In addition to the lmSTC encoding abstract relational structure, Frankland and Greene (2020b) suggested that the amPFC reuses structured knowledge to encode new specific relational combinations. In contrast to Frankland and Greene's results, the amPFC was found here to encode the specific identity of the participants of a scene, but not in combination with a specific event. Together, this supports the evidence for specific relational combinations that abstract over events. Therefore, the role of the amPFC in compositional processing may not be as specific as suggested. Modality may have affected the representational patterns. In production, events may be encoded with less specificity in frontal cortex, possibly because they are encoded to generate a linguistic output and do not need to be modified for conceptual integration to the same extent as during comprehension (Frankland & Greene, 2020b). However, MEG evidence shows that the ventral mPFC is involved in compositional semantic processing for adjective-noun combinations (e.g. "red boat") in both production and comprehension (Bemis & Pylkkänen, 2011; Blanco-Elorrieta et al., 2018; Pylkkänen, 2020; Pylkkänen et al., 2014). A pure modality explanation, therefore, seems unlikely.

An alternative is that our stimulus set limited the encoding potential we could explore. In fact, we could use either noun-specific or verb-specific role combinations, but not noun-verb specific combinations (e.g. 'violinist-as-attacker'), since the stimuli each participant produced did not span all of the combinations. It is therefore possible that specific noun-verb combinations may have improved generalization performance, under the assumption that the verb specific combination could also benefit the encoding model if bound to specific nouns. Tentative evidence for this explanation is provided by the generalization performance of a full model that includes both specific nouns in combination with their thematic role (verb-category dependent, 'violinist-as-agent') and specific verbs in combination with the semantic category for each of their thematic roles ('musician(category)-as-attacker'), but not their combination ('violinist-as-attacker'). This full model performed even better than the narrow nouns model (Supplementary Fig. S5.9). This explanation would suggest that the amPFC can only encode fully specific event representations where both the event (verb) and its entities are specific. Future studies addressing the specificity of compositional processing may be able to clarify this issue. A final alternative is that we were underpowered to reliably detect effects in the amPFC with n = 38. Frankland and Greene (2020b) mentioned that with 40 participants there was 66% power to detect reliable effects in amPFC at the whole-brain. However, since we ran a planned ROI analysis in the amPFC, it is unlikely that our low generalization performance was simply due to lack of power.

Another result that was not confirmed in the current study was the pattern separation encoding in the hippocampus for specific relational structure (Frankland & Greene, 2020b). Here, none of the models could significantly predict hippocampal activity in new sentences, possibly due to poorer signal-to-noise ratio in this region. Future studies may be able to determine whether the pattern separation for conceptual representations is replicable in sentence production and comprehension.

Multiple ways to build a thought

Our results, therefore, differ from Frankland & Greene's, in several ways. Frankland and Greene (2020b) found that the lmSTC encoded events only in abstract terms. Here, we replicated the abstract structure encoding, and additionally found evidence for specific relational structure encoding only in this ROI. Also, their lmSTC region encoding abstract relational structure was not found to encode ordinal and syntactic structure, which they took as evidence that the lmSTC is involved in building abstract compositional meaning regardless of sequential aspects of a sentence. However, we found significant encoding performance for all characterizations of sentence structure along the superior temporal cortex, in overlap with the relational structure encoding areas. The ordinal model reflected the surface structure of the sentence ('musician-as-subject' where the musician is the first noun in both active and passive sentences). This model also grouped elements based on information structure (i.e. what the sentence was about, see for discussion Bornkessel-Schlesewsky et al., 2012; Ferreira & Lowder, 2016). Instead the syntactic and event models grouped event participants following different assumptions. The syntactic model grouped the deep-subjects, which were external arguments in semantic structure, and sentient participants in the event (Grimshaw, 1990; Levin & Hovav, 2005). The event model instead grouped event participants based on the causal-temporal structure of the events (i.e. agents and stimuli need to exist for the event to occur) (Frankland & Greene, 2020b). These event characterizations were all encoded in the superior temporal cortex (but also in a more distributed way), but no specific event type was best. This pattern of results suggests that the brain tracks different types of abstract event structure simultaneously. The ordinal and syntactic structures could be encoded more strongly in production because of their importance for sentence formulation, but, as introduced, they capture other aspects of sentence meaning (‘aboutness’, sentience, causal-temporal structure) that are all potentially relevant for sentence interpretation. We are not able to distinguish the ordinal and syntactic interpretation from the semantic/pragmatic interpretation in this study, but in either case we find evidence for ‘multiple’ rather than ‘two ways to build a thought’ (cf. title Frankland & Greene, 2020b). We tentatively suggest this could form the basis for capturing sentence meaning at different levels of subtlety, e.g. a single sentence encoding meaning, speakers’ intentions, and focus based on prosody, information structure and pragmatic constraints.

Distributed representations for meaning

Frankland and Greene (2015, 2020b) also found that a medial region of the lmSTC encoded the identity of the agent, while a lateral region encoded the identity of the patient. We found that the identity of the agent was encoded in the medial lmSTC, but the patient was encoded in a more posterior cluster, where also the identity of the agent could be decoded. In addition, we ran a multivariate decoding analysis similar to Frankland and Greene (2015), and again we were not able to replicate the agent-patient organization. Frankland and Greene (2020b) proposed a computer-like organization of the lmSTC, thought to support reusable semantic representations bound to thematic roles. The STC, however, was now found to encode sentence meaning and structure at different levels of specificity and in different organizations in overlapping areas. In addition, the widespread whole-brain results are against any interpretation of neural specificity for compositional processing. Rather, we find a distributed nature of brain representations. Other studies found similar distributed representations for several aspects of sentence comprehension. For example, the language network is sensitive to both syntactic and lexical structure (Blank et al., 2016; Fedorenko, Blank, Siegelman, & Mineroff, 2020; Fedorenko, Nieto-Castañon, & Kanwisher, 2012; Hu et al., 2022; Shain et al., 2021). Similarly, the meaning of words in multiple positions in a sentence (e.g. subject, object) is encoded in frontal, inferior parietal and temporal regions (Anderson et al., 2019).

This conclusion is supported by the whole-brain results, which showed that the relational structure of the sentence was encoded in several areas of the perisylvian language network. The different models had differences in their responses, but there was considerable overlap. This network is in line with the brain regions representing distinctions for thematic categories (i.e. music vs. sport categories) in a previous study, which peaked in the temporoparietal junction and included precentral gyrus and parietal areas (Xu et al., 2018). This raises the question whether our encoding models were only using the semantic information to predict sentence activity. While this is a possibility, the better performance of models with increasing relational specificity suggests instead that these areas simultaneously encode information at multiple levels, including word meanings as well as the relational structures they form.

Conclusions

Overall, we found that an extensive fronto-parieto-temporal network encoded abstract compositional sentence meaning in several organizational structures during sentence production. This distributed network was not in line with previous results on compositional processing in comprehension, that instead identified a localized network of superior temporal cortex, prefrontal cortex and hippocampus encoding relational meaning in either abstract or specific terms (Frankland & Greene, 2020b). It is unlikely that this difference in neural organization is due to modality (production vs. comprehension), as it is generally assumed that conceptual representations are shared between production and comprehension (Guhe, 2007; Levelt, 1989). Rather, the nature of the stimuli, organized in semantic categories, and our analysis approach, that was not limited to predefined ROIs but explored whole-brain representations, showed a more complex picture than previously suggested (Frankland & Greene, 2020b), with distributed representations for relational structures with different structural grouping.

In sum, we investigated the neural architecture supporting compositional processing during sentence production using voxel-wise encoding models. We found that the superior temporal cortex encodes sentence meaning in a posterior-to-anterior gradient of specificity. A mid-anterior cluster supports specific event representations, while a posterior cluster abstracts away from event-specific information and additionally encodes ordinal and syntactic aspects of sentence meaning, possibly due to the incremental nature of sentence production. Overall, we found evidence for distributed representations of relational structure across the perisylvian network.

S5 Supplementary Information

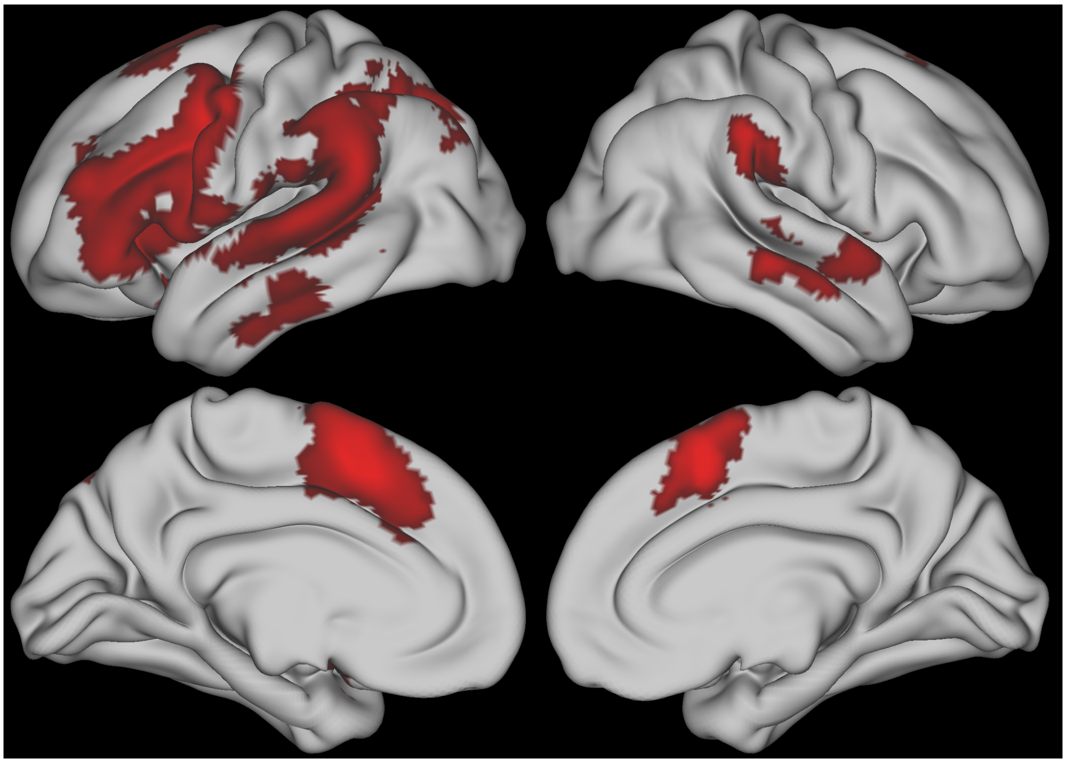


Figure S5.1: Whole-brain results of bag-of-nouns encoding model. Voxel-wise $p < 0.005$, cluster-wise $p < 0.05$ FDR corrected.

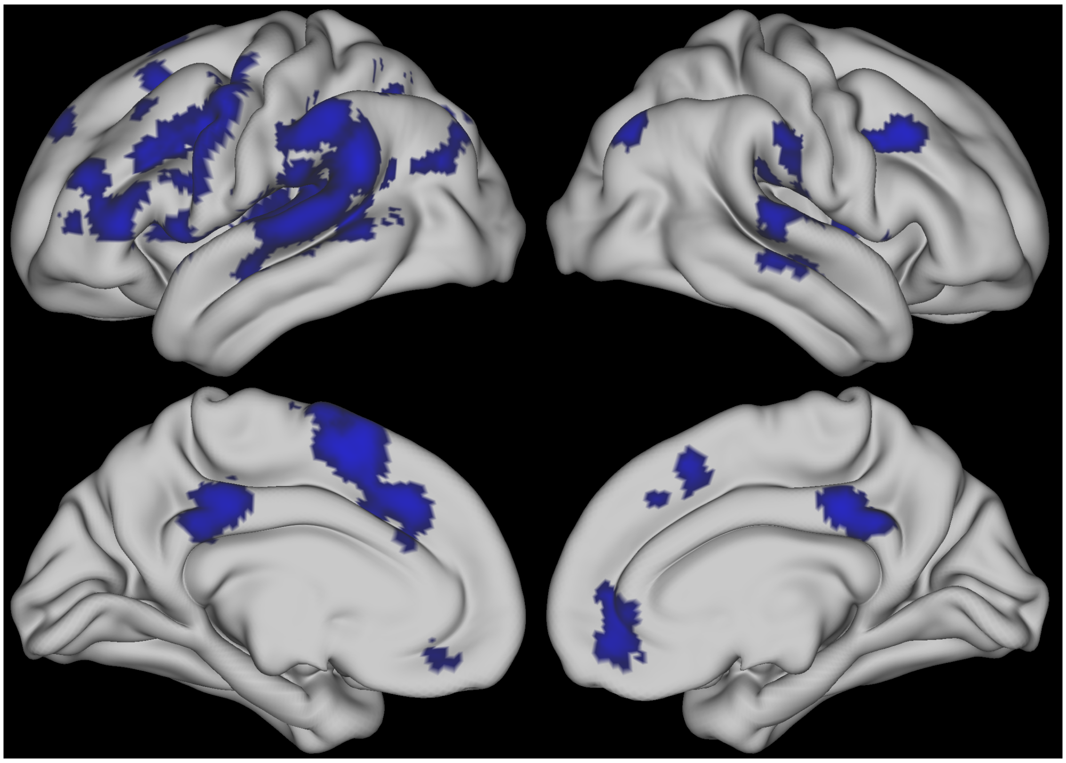


Figure S5.2: Whole-brain results of broad-roles encoding model. Voxel-wise $p < 0.005$, cluster-wise $p < 0.05$ FDR corrected.

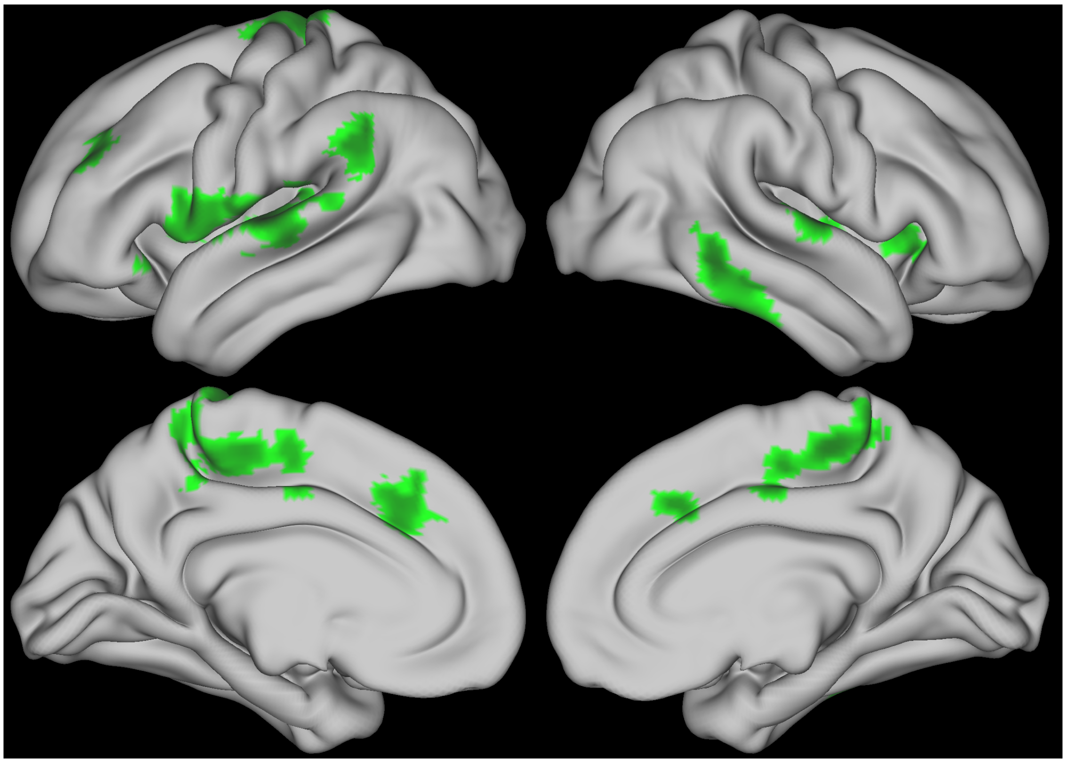


Figure S5.3: Whole-brain results of narrow roles (verbs) encoding model. Voxel-wise $p < 0.005$, cluster-wise $p < 0.05$ FDR corrected.

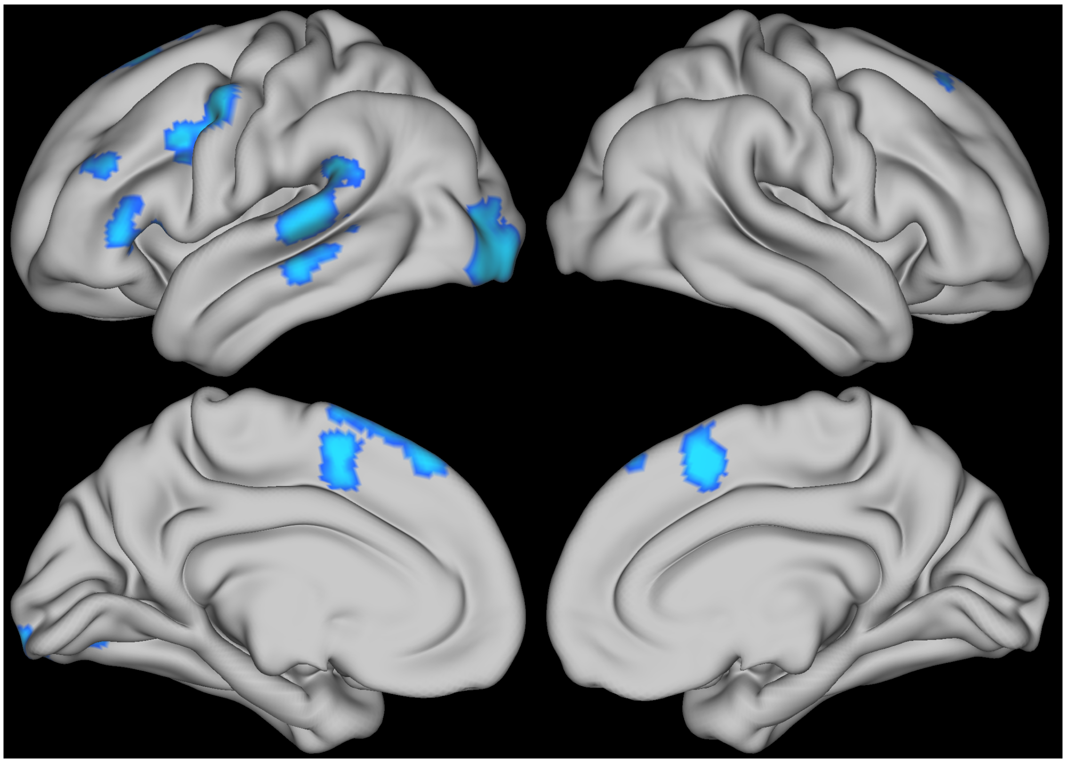


Figure S5.4: Whole-brain results of narrow roles (nouns) encoding model. Voxel-wise $p < 0.005$, cluster-wise $p < 0.05$ FDR corrected.

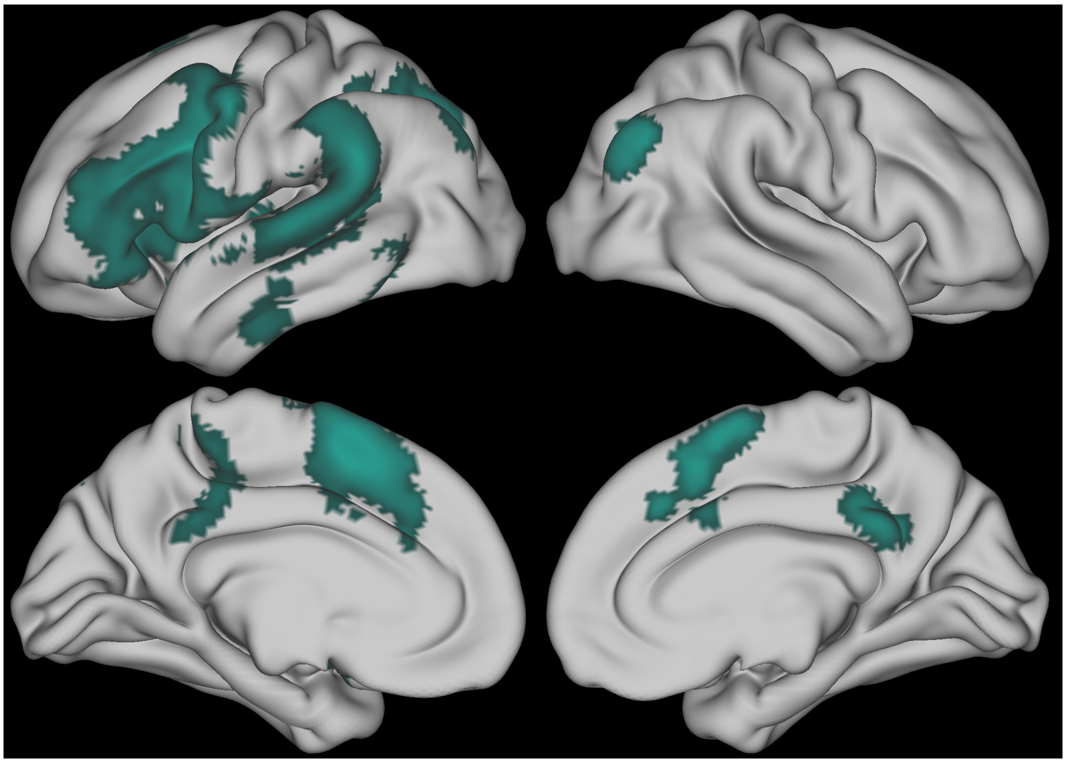


Figure S5.5: Whole-brain results of ordinal encoding model. Voxel-wise $p < 0.005$, cluster-wise $p < 0.05$ FDR corrected.

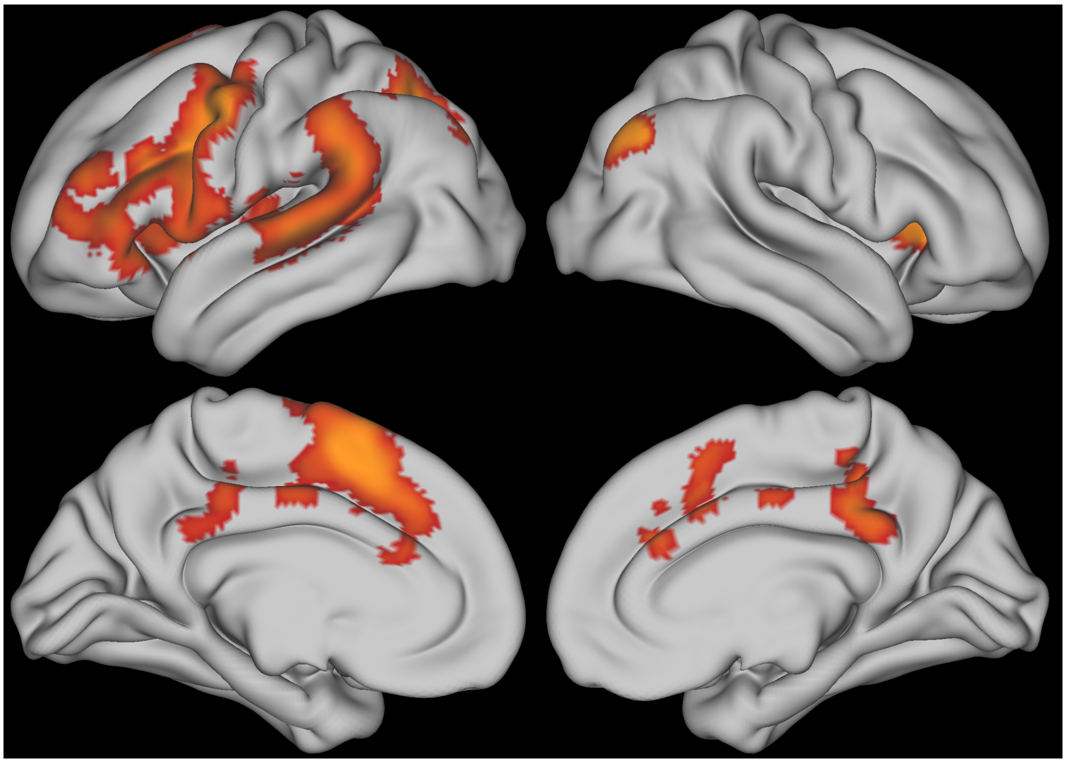


Figure S5.6: Whole-brain results of syntactic encoding model. Voxel-wise $p < 0.005$, cluster-wise $p < 0.05$ FDR corrected.

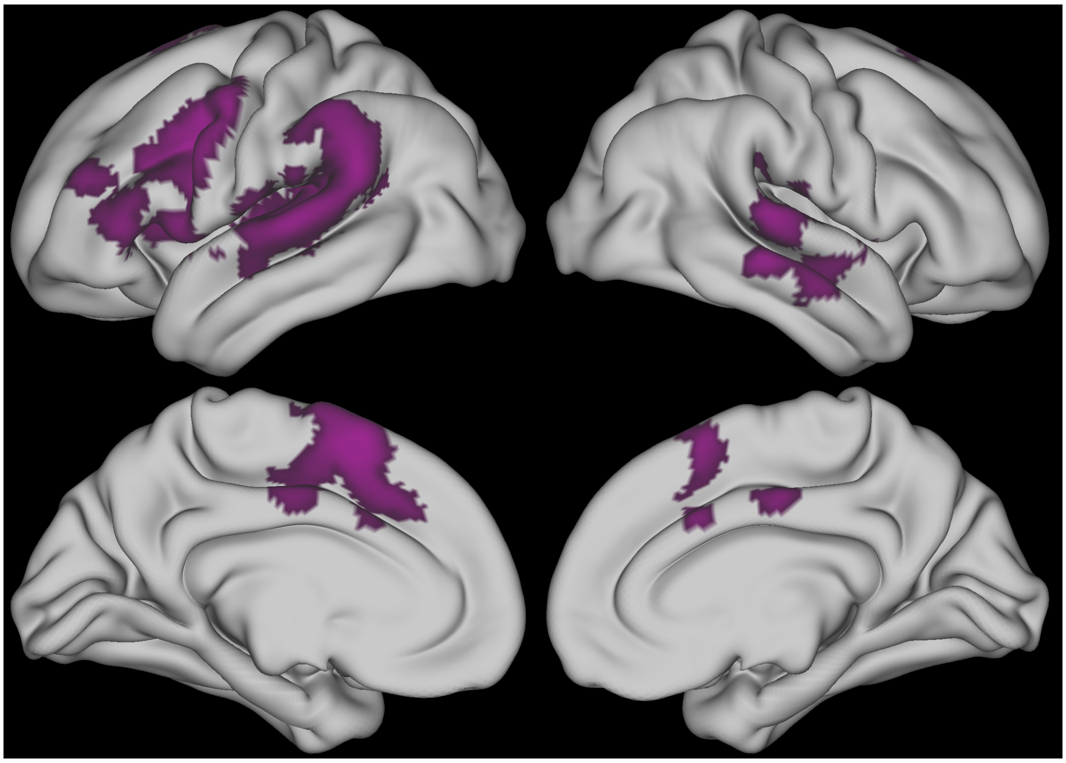


Figure S5.7: Whole-brain results of event encoding model. Voxel-wise $p < 0.005$, cluster-wise $p < 0.05$ FDR corrected.

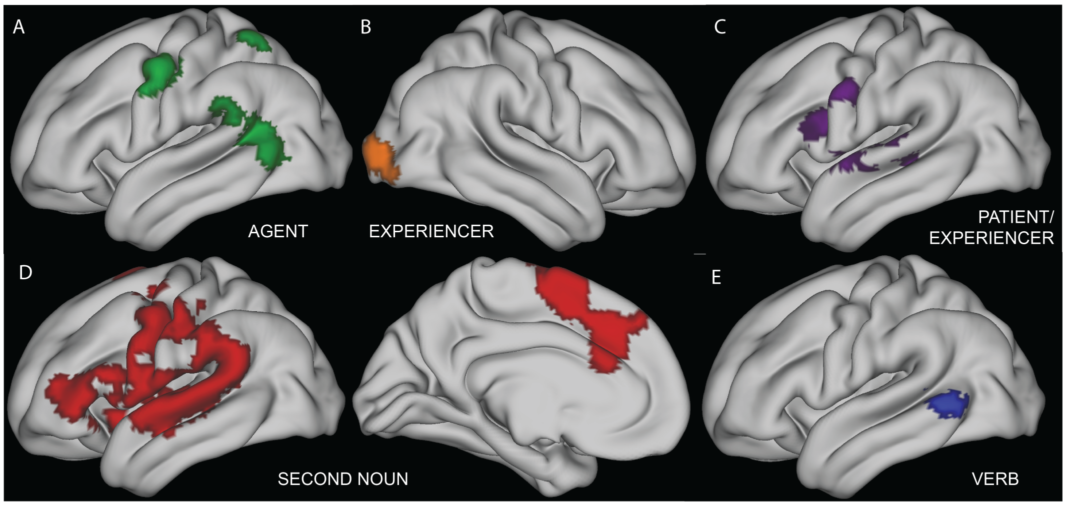


Figure S5.8: Whole-brain searchlights of above chance classification accuracy for sentence dimensions ($p < 0.001$ voxelwise, $p < 0.05$ FDR corrected clusterwise). **A**: classification of the semantic category of the agent. **B**: classification of the semantic category of the experiencer. **C**: classification of the semantic category of the patient/experiencer grouping. **D**: classification of the semantic category of the second noun. **E**: classification of the semantic category of verbs.

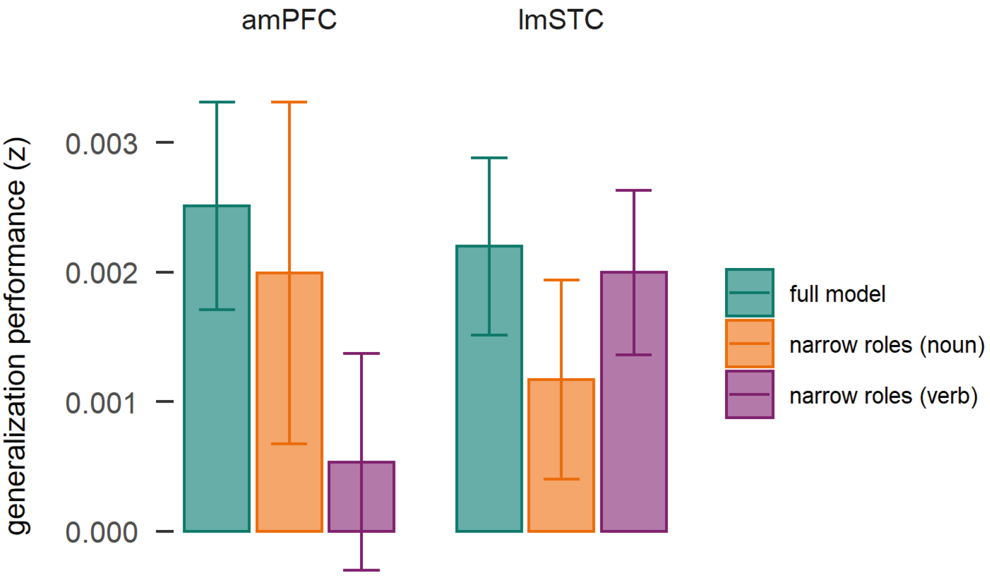
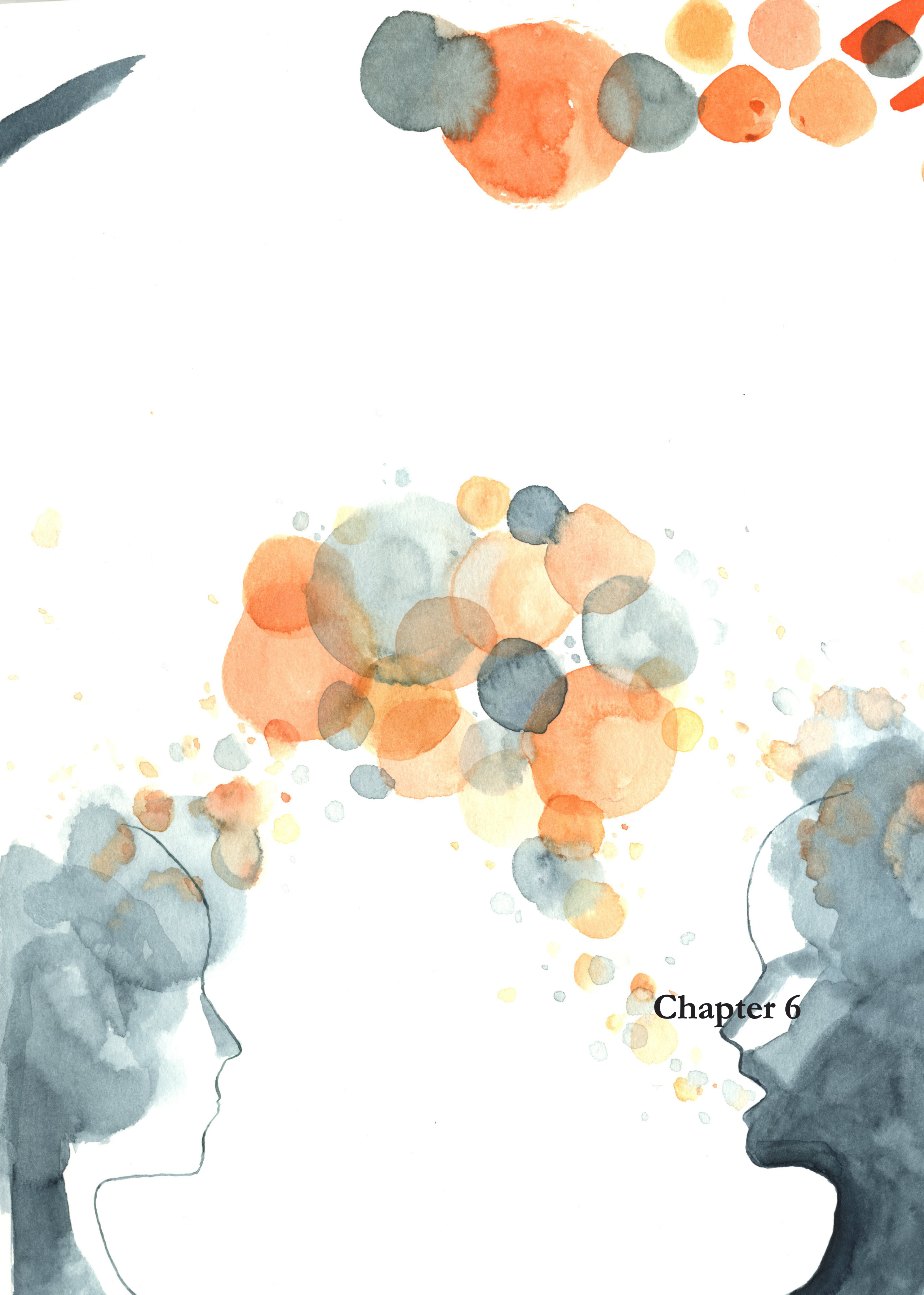


Figure S5.9: Average generalization performance in amPFC and lmSTC ROIs, selected independently for each participant, for a full model that includes all sentence predictors, and the two narrow roles models.

Table S5.1: Statistics and peaks for whole-brain analysis for each encoding model. Voxelwise $p < 0.005$, $p < 0.05$ FDR cluster corrected. Note that p values are not corrected for multiple tests: $p < 0.007$ is significant with Bonferroni correction for 7 encoding models.

Contrast	**Cluster**		**Peak Voxel (MNI)**				**Peak Anatomical Location**
	p	**Size**	**mean *z***	**x**	**y**	**z**	
bag-of-nouns	2.60E-05	6217	0.001391	-49	2	41	L Precentral Gyrus
	0.000286	1188	0.001447	2	12	51	Supplementary Motor Area
	0.011216	427	0.000847	26	-68	-51	R Cerebellum (VIIb)
	0.017631	330	0.000673	-22	10	-8	L Putamen
	0.017631	307	0.000793	68	-25	-6	R MTG
	0.035058	223	0.001185	61	-30	29	R Supramarginal Gyrus
broad roles	2.24E-05	1984	0.002009	-61	-45	24	L Supramarginal Gyrus
	4.48E-05	1259	0.002311	-51	2	41	L Precentral Gyrus
	0.002552	451	0.002305	-36	-60	56	L Superior Parietal Lobule
	0.002552	416	0.001809	56	-15	11	R Planum Temporale
	0.002552	414	0.002134	-2	12	54	Supplementary Motor Area
	0.021229	194	0.001937	1	32	36	Posterior Cingulate Cortex
	0.036216	147	0.002132	46	20	36	R Middle Frontal Gyrus
	0.036216	143	0.00152	6	45	-6	R Paracingulate Gyrus
	0.037371	135	0.00234	41	-72	36	R Superior Parietal Lobule
	0.038549	128	0.001733	-24	17	51	L Superior Frontal Sulcus
	0.03869	123	0.001432	46	-5	-1	R Opercular Cortex
	0.040269	117	0.001267	26	67	-51	R Cerebellum (VIIb)
narrow roles (verbs)	0.000528	950	0.003602	-59	-17	6	L Planum Temporal
	0.005237	415	0.003266	3	27	56	Supplementary Motor Area
	0.018807	250	0.002346	61	-45	-11	R MTG
	0.025291	204	0.003363	56	-12	4	R Planum Temporale
	0.025291	179	0.003387	-2	25	39	L Paracingulate Gyrus
	0.025291	177	0.003133	26	10	2	R Putamen
	0.031269	155	0.003693	-31	42	36	L Anterior Middle Frontal Gyrus
	0.032171	146	0.003239	-14	-30	74	L dorsal Prefrontal Gyrus
narrow roles (nouns)	0.003534	548	0.004156	-34	-90	-8	L Lateral Inferior Occipital Cortex
	0.003534	490	0.004079	3	15	51	L Superior Frontal Gyrus
	0.004663	402	0.004164	-61	30	2	L pSTG
	0.008314	305	0.004121	-44	47	16	L dorsal Prefrontal Cortex
	0.012133	248	0.005614	-49	2	39	L Precentral Gyrus
ordinal	2.69E-05	4716	0.001399	-51	2	42	L Precentral Gyrus
	0.000994	942	0.001522	2	12	54	Supplementary Motor Area
	0.002731	635	0.001461	-29	-72	44	L Superior Parietal Lobule
	0.017993	327	0.001331	46	-70	34	R Superior Parietal Lobule
	0.022674	278	0.001075	6	-47	29	R Posterior Cingulate Cortex
	0.035761	208	0.000639	-24	10	2	L Putamen
	0.035761	207	0.000885	-49	-60	-13	L Inferior Occipito-temporal gyurs
	0.049761	172	0.000788	23	-67	-51	R Cerebellum (VIIb)
syntactic	2.21E-05	3770	0.001341	-51	2	41	L Precentral Gyrus
	0.000663	971	0.001352	-2	12	54	Supplementary Motor Area
	0.003725	569	0.001541	-26	-70	46	L Superior Parietal Lobule
	0.046727	186	0.000918	3	-35	36	R Posterior Cingulate Cortex
	0.046727	184	0.000761	28	-67	-51	R Cerebellum (VIIb)
	0.046727	175	0.001116	36	27	2	R Frontal Orbital Cortex
	0.046727	174	0.00135	43	-70	34	R Superior Parietal Lobule
event	2.56E-05	1917	0.001565	-61	-45	24	L Supramarginal Gyrus
	5.12E-05	1696	0.00185	-49	2	41	L Precentral Gyrus
	0.001594	721	0.001554	2	12	54	Supplementary Motor Area
	0.001594	674	0.001304	66	-25	-4	L pMTG



Chapter 6

6 | General discussion

6.1 Summary of the results

In this thesis, I investigated the neural infrastructure for sentence production and integrated it with the current understanding of the neurobiology of language, which is mostly based on language comprehension. I ran four fMRI studies of sentence production and comprehension to uncover and understand similarities and differences in their neural correlates. The four chapters focused on different aspects of sentence-level linguistic processing and used different paradigms and analysis approaches to characterize their neural responses.

In Chapter 2, I identified the network of regions whose activity increases with the constituent size of spoken and heard word sequences. The same words were spoken in all constituent size conditions, but they were organized in constituents of increasing size. The same regions responded to syntactic and compositional processing during both production and comprehension. The response to constituent size was thus mostly shared between production and comprehension. The results also indicated effects of modality. Inferior frontal regions responded more during production and were more sensitive to constituent size in production than comprehension. Temporal regions instead were more active during comprehension. Therefore, I found a shared network for constituent structure building in production and comprehension, that was however differentially modulated by modality. The BOLD response additionally peaked later in comprehension with larger constituent structures, but it peaked earlier in production for the same structures, suggesting opposite dynamics of constituent structure building in production and comprehension.

In Chapter 3, I further probed this network to understand how modality affects the connectivity between some of the regions that were seen to respond to constituent size in Chapter 2. The inferior frontal gyrus and the posterior temporal lobe were functionally connected to the same regions that were sensitive to constituent size in Chapter 2. Connectivity among inferior frontal and anterior and posterior temporal regions was upregulated with increasing constituent size, suggesting increased network integration when building more complex constituent structure. Posterior temporal regions and the inferior frontal gyrus were functionally more connected to each other than to the anterior temporal lobe. This dorsal connectivity was further upregulated during production relative to comprehension, indicating the stronger need to prepare an articulatory plan in production. Therefore, connectivity between fronto-temporal regions is shared during constituent structure building in production and comprehension, and it is upregulated for production-specific processes.

In Chapter 4, I further addressed the dynamics of syntactic structure building during spontaneous production and comprehension using continuous metrics of syntactic complexity. Anticipatory phrase-structure building in production led to an increase in BOLD activity in the inferior frontal gyrus. In comprehension, integratory phrase-structure building led to a BOLD increase in the left inferior frontal gyrus and the left posterior temporal lobe. These results suggest that syntactic structure building was dominant at phrase-opening in production and at phrase-closing in comprehension. Additionally, in production a highly incremental parser, which posited structure building to occur word-by-word, led to a better model fit than a less incremental parser that predicted structure building to occur for chunks of words at a time. This study showed the incremental nature of structure building in both production and comprehension and highlighted their different dynamics.

In Chapter 5, I focused on the semantic structures represented during message generation in production. Encoding models were used to predict brain activity during the production of individual sentences based on their thematic role structure. A distributed perisylvian network was found to encode the compositional meaning of active and passive sentences. This network encoded sentence meaning following different organizational principles, including ordinal, syntactic and event structure. The superior temporal cortex was organized in a gradient of relational specificity, where an anterior region encoded the specific event structure, while a more posterior region simultaneously encoded abstract event structure in different configurations. Overall, I found evidence for relational structure being encoded during sentence production in a distributed brain network.

In summary, I probed the left hemispheric network for language and found that it was mostly shared in speaking and listening, although with differences in their dynamics of structure building. I will now discuss how these studies can inform the debate about the commonality of linguistic representations and processes in production and comprehension, and what they add to our understanding of the brain organization for language.

6.2 The neural infrastructure for sentence production and comprehension

The four empirical chapters just summarised identified a network of regions that are involved in sentence production and comprehension. Conceptual structures were seen to be built in the superior temporal gyrus in Chapter 5. While we only showed it for production, a comparable comprehension study also confirmed the involvement of the STG in compositional processing in comprehension (Frankland & Greene, 2020b). The anterior part was involved in encoding the specific relations within an event, while the posterior section better encoded abstract event structure. This posterior section was also seen to encode different aspects of sentence structure (ordinal, syntactic). This gradient, therefore, is in line with Bornkessel-Schlesewsky and Schlesewsky's (2013) account of a ventral stream that processes time-independent conceptual information, in contrast to a dorsal stream that processes sequential information. These results support evidence that the anterior temporal lobe is involved in processing semantic information (e.g. Lambon Ralph et al., 2017; Pylkkänen, 2020). Semantic structures were also encoded in the inferior parietal lobe, which has been reported before to encode information along thematic divisions (e.g. doctors vs. musicians) and to be involved in event processing (e.g. Matchin et al., 2019; Xu et al., 2018).

The results of Chapter 5 add production evidence to previous comprehension results (Frankland & Greene, 2020b). The common encoding of semantic structures in the superior temporal cortex provides the empirical basis for positing shared conceptual processing in production and comprehension as initially suggested by Levelt (1989). The comprehension results indicated that the amPFC and the hippocampus also take part in compositional processing of specific relational structures (Frankland & Greene, 2020b). This was not replicated in production. Differences in the design, such as using more varied stimuli grouped in semantic categories, may have affected the responses of these regions. In addition, our results indicate a distributed network encoding abstract relational structure. Therefore, the results of Chapter 5 suggest a different, more distributed, principle of encoding of semantic structures, relative to the one proposed in Frankland and Greene (2020b) that instead hinges on a few regions. This difference is unlikely to be a consequence of modality, since other comprehension studies also found distributed semantic representations (e.g. Anderson et al., 2021, 2019). Therefore, the neural differences between production and comprehension that emerged in Chapter 5 do not provide a basis for the segregation of semantic processes and representations in production and comprehension.

The posterior superior temporal sulcus and posterior middle temporal gyrus were seen to be modulated by syntactic structure building in both production and comprehension in Chapter 2, in line with previous studies in comprehension (e.g. Pallier et al., 2011; Zaccarella, Schell, & Friederici, 2017). The posterior MTG was also involved in phrase-structure building in Chapter 4, most directly in comprehension, similarly to other studies of continuous structure building (e.g. Brennan et al., 2016; Lopopolo et al., 2021). The LIFG was also consistently found to be involved in syntactic structure building in both production and comprehension (Chapters 2, 4, 5). In addition, the connectivity between posterior temporal and inferior frontal regions was seen to increase for constituent structure building in production and comprehension (Chapter 3), confirming and extending previous evidence for a dorsal stream involved in syntactic processing (e.g. Griffiths et al., 2013; Papoutsi et al., 2011).

Our tasks cannot directly distinguish between different accounts of the function of the LIFG (i.e. unification or control-based accounts) and the PTL (i.e. retrieval of lexico-syntactic information or hierarchical structure building) (Bornkessel-Schlesewsky & Schlesewsky, 2013; Hagoort, 2005, 2013; Matchin & Hickok, 2020; Zaccarella, Schell, & Friederici, 2017). However, some studies suggest that syntactic processing and lexical retrieval cannot be effectively distinguished within the language network (e.g. Fedorenko et al., 2020). The present findings, therefore, strengthened by the connectivity results, support theories of hierarchical structure building being processed by the interaction of PTL and LIFG, as highlighted by many models despite attempts at assigning specific computations to regions (e.g. Friederici, 2012; Hagoort, 2013; Tyler & Marslen-Wilson, 2008).

Therefore, the same regions were seen to respond to syntactic structure building in production and comprehension. These results confirm the previous adaptation studies that found shared resources for sentence production and comprehension (Menenti et al., 2011; Segaert et al., 2012). However, we also found differences in the modulation of this fronto-temporal network by production and comprehension. First, constituent size elicited a stronger response in production in the LIFG and in the LpMTG (Chapter 2). Second, the dynamics of structure building were seen to differ between production and comprehension (Chapters 2 and 4). In Chapter 2, we proposed the increased response in production to be due to the increased requirements of speaking relative to listening. Production requires a message to be encoded linguistically and results in an action (involving articulation, a motor plan), while listening can be more passive. In production, it is not possible to bypass computations by relying on heuristics (Bock, 1982; e.g. good-enough processing, Ferreira, 2003; Ferreira et al., 2002; but see Goldberg & Ferreira, 2022, for a discussion of good-enough production), due to the importance of 'getting it right' (Garrett, 1980, 1982). Another study that investigated sentence production and comprehension also found an increased response for sentence production across the language network, which was explained with production having a larger cost on the language network (Hu et al., 2022).

Interestingly, in Chapter 4 we did not find an increased response for structure building in production vs. comprehension in this fronto-temporal network. Taken together, fronto-temporal regions were similarly affected by structure building in production and comprehension, although in different dynamics. Therefore, it seems that when production is not elicited in the context of a task, it does not necessarily lead to increased brain activity relative to comprehension. In this study, the production was not constrained by a task, but it was spontaneous. Participants were asked to recall the events of a TV series they had just watched. They thus produced monologues in the scanner without the pressure to speak cleanly, but with the goal to remember as many events as possible. Their speech thus contained disfluencies, hesitations and self-corrections. Together with Chapter 2, these results may suggest that, when the pressure to 'get it right' is reduced, the production load on brain responses matches the comprehension load. Although disfluent speech is not usually studied in language experiments, it is common in conversation (average of 6 disfluencies per 100 words) (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Tree, 1995), and it can reflect the natural constraints of the production system. Chapter 4 thus makes use of more naturalistic stimuli for both production and comprehension, and as a consequence delineates a more realistic account of brain activity during speaking and listening.

6.3 Phrase-structure building in production and comprehension

The results of Chapters 2 and 4 showed that phrase-structure building proceeds with different dynamics in production and comprehension. In Chapter 2, the latency of the BOLD response to sentence processing was modulated by the interaction of constituent size with modality. In comprehension, the BOLD response peaked later with constituents of increasing size (see also Pallier et al., 2011). In production, instead, the BOLD response peaked earlier with larger constituents. In Chapter 4, I modelled the incremental (word-by-word) building of phrase structure and thus addressed the dynamics of structure building directly. The results were the same: BOLD activity increased at phrase-opening in production, while it increased at phrase-closing in comprehension. Phrase-structure is thus built at different sentence processing stages in production and comprehension. This is likely due to directional differences in the mapping of semantic, lexical and phonological representations in production and comprehension. Production needs to map semantic structures to a linear articulatory code, while comprehension needs to map sounds to semantic structures.

This directional difference in the mapping of representations does not imply that the syntactic processor differs between modalities. Kempen (2000) argues that the processes of grammatical encoding in production and parsing in comprehension form a single processing mechanism used for constructing syntactic structure. The differences between parsing and encoding are due to different processing contexts. In production, structures are built on the basis of lexico-syntactic information extracted from conceptual structure, while in comprehension the lexico-syntactic information is derived from word strings. Similarly, Momma and Phillips (2018) argue that differences between encoding and parsing may arise when tasks affect which piece of information becomes available to the processor and when. Therefore, processes mapping from higher to lower linguistic levels (called ‘production’ processes by Pickering & Garrod, 2013), and processes mapping from lower to higher linguistic levels (i.e. ‘comprehension’ processes according to Pickering & Garrod, 2013) can take place during both production and comprehension depending on different tasks or linguistic contexts.

In Chapter 4, the dynamics of structure building were modelled with two parsers, a top-down anticipatory parser and a bottom-up integratory parser. The two parsers make hypotheses about the structure building operations for the same syntactic tree. Therefore, in practice the parsers only differed in the timing of operations (Hale, 2014). The top-down predictive parser assumes structure building to take place at phrase-opening, i.e. before word processing. The bottom-up integratory parser instead builds the structure after encountering each word. Previous work suggests that different structure building dynamics can occur during both production and comprehension. For example, top-down as well as bottom-up operations have been found to successfully predict brain activity in comprehension, with fMRI, MEG and intracranial EEG (Bhattasali et al., 2019; Brennan et al., 2016; Coopmans, 2023; Lopopolo et al., 2021; Nelson et al., 2017). The use of a more or less predictive strategy of structure building may depend on the context of language use. For example, there may be situations where it is easier to predict, such as during audiobook listening, where the input is at a constant speed and is free of disfluencies or corrections (Huettig & Mani, 2016). Listening to spontaneous production like in Chapter 4, or reading sentences out of context, may instead discourage predictive structure building due to the higher likelihood of having to revise the initially hypothesised structure.

In addition, psycholinguistic studies have highlighted the flexibility of planning in scope in production, suggesting that structure building can be driven by lexical access (Bock & Ferreira, 2014). For example, in conversational settings, a speaker may be driven to start speaking before engaging in structure building to claim their turn. It is unlikely that the structure is built fully following a bottom-up strategy in production, since it would mean that the linear sequence is decided before the words are integrated in the hierarchical structure. However, the amount of incrementality and anticipatory structure building, as described in Chapter 4, may be more flexible. For example, when the pressure to speak rapidly is reduced, relative to the importance of speaking without corrections, such as during a presentation, a less incremental (e.g. *chunked*, see Chapter 4) strategy may be beneficial.

6.4 The brain vs. the language network

Until now, I have focused on the brain regions that are commonly denoted as the ‘language network’, since they have been most consistently found in neuroimaging studies, and are found to lead to aphasia if lesioned. However, there are other regions that are seen contribute to linguistic processing both in this thesis and in other studies. First, the right hemisphere was engaged in Chapter 2, especially for modality specific processes, but also for constituent size. For example, the right precentral gyrus was found to be involved in production more than comprehension, due to articulation needs. Conversely, the right superior temporal gyrus was engaged more in comprehension, possibly due to auditory processing. The modality effects therefore are partly due to peripheral processes (e.g. motor control, acoustic processing) that are not expected to be lateralized. Other right hemispheric areas were also found across the chapters.

For example, in Chapter 5, the right temporal lobe was seen to encode semantic structures. The right hemisphere has been found to be important for theory of mind processing (R. Griffin et al., 2006; Happé, Brownell, & Winner, 1999). In particular, right temporo-parietal regions and inferior frontal areas have been found for inferring speaker meaning and in discourse comprehension based on world knowledge (e.g. Bašnáková, Weber, Petersson, van Berkum, & Hagoort, 2014; Menenti, Petersson, Scheeringa, & Hagoort, 2009). Regions of the right hemisphere may therefore have been involved in building situation models in both language production and comprehension.

The cerebellum and the basal ganglia were also often found to be active, especially in production. In particular, in Chapter 2 the right cerebellum responded to constituent size in both modalities, but more strongly in production. It was also seen to encode sentence meaning in Chapter 5. The role of the cerebellum in language processing is only recently being uncovered. Generally, it is known that the cerebellum is important for precise motor control, but it can also represent cognitive operations that are processed in cortical areas, including language (Diedrichsen, King, Hernandez-Castillo, Sereno, & Ivry, 2019; Ito, 2008; LeBel, Jain, & Huth, 2021). Disruption to the cerebellum with transcranial magnetic stimulation was found to lead to a delay in predictive eye movements in language processing and to impaired language production, possibly due to impaired monitoring (Lesage, Morgan, Olson, Meyer, & Miall, 2012; Runnqvist et al., 2016). In Chapters 2 and 5, therefore, the cerebellum may have been involved in computing forward models of action (or action perception in comprehension, Chapter 2) that are functional for smooth production and monitoring, and prediction in comprehension.

The basal ganglia have been proposed to be involved in language via two loops (Bohsali & Crosson, 2016; Watkins & Jenkinson, 2016). A pre-supplementary motor area-basal ganglia loop is suggested to support lexical selection. A LIFG-basal ganglia loop is proposed to support phonological and articulatory processing. In Chapter 2, the basal ganglia were more active in production, and in Chapter 5 they were sometimes seen to encode sentence meaning, together with IFG and supplementary motor areas. Their proposed role in lexical retrieval and articulation is therefore in line with our results.

6.5 The study of language with fMRI

Although I have attempted to integrate the results of the empirical chapters within the current understanding of the neurobiology of language, one final point of caution should be mentioned. Overall, the results in Chapters 2 and 5 highlighted the distributed nature of linguistic representations in the brain. Recently, several studies have been reporting a lack of segregation in neural activity for many language-related tasks (e.g. Blank et al., 2016; Fedorenko et al., 2020; Fedorenko, Nieto-Castañon, & Kanwisher, 2012; Hu et al., 2022). The distributed representations relate both to the fact that multiple levels of linguistic representations are seen to be encoded within the same brain region (against a specific interpretation of functional specialization, see Friston, 2002), and to the fact that the same linguistic representation is encoded in many brain regions (against functional localization but supporting functional integration, e.g. Anderson et al., 2019). This pattern of results is found in univariate studies, as well as multivariate studies (Box 1.1).

How can it be that many brain regions are sensitive to many linguistic processes? The ‘localizationist’ (or neo-phrenological, see Embick & Poeppel, 2015; Friston, 2002; Tyler & Marslen-Wilson, 2008) attempts at characterizing brain function originated from the early approach of lesion studies. Early fMRI studies aimed to assign brain function to specific regions. While this approach was extremely productive, increasing our understanding of coarse brain organization (e.g. language processing in fronto-temporal areas, visual processing in the occipital lobe, confirming functional specialization in a broad sense), it seems to fail when attempting to assign a specific linguistic function (e.g. syntactic processing) to a particular brain region (Poeppel, 2012). There are two reasons why this may be.

First, the improvement in magnetic resonance methods over the last decades led to increased power (e.g. shorter repetition times and smaller voxels, increasing the signal-to-noise ratio) to detect changes in haemodynamic activity relative to what was possible before the early 2000s. As a consequence, it is now possible to detect subtler and broader changes in brain activity, which may be related to correlational, but perhaps not necessary, activity during a task. This correlation among brain regions means that regions often act together in a distributed network, as also suggested by resting-state studies that find the same regions activated by a task to be functionally connected at rest (Buckner et al., 2013). It is unlikely that all connected regions are necessary for a certain process (which would result in a disruption in processing if lesioned), but they may instead be involved in less specific processing, such as building a situation model as suggested for the right hemisphere. Based on these considerations, fMRI studies can hardly inform localizationist interpretations of brain function: the brain organization for cognition is more complex than the studies that focus on localization and segregation can indicate, because of the correlational activity found with fMRI (e.g. Hall, Howarth, Kurth-Nelson, & Mishra, 2016).

A second consideration relates to the isolation of linguistic processes that is afforded by tasks. For example, comparing brain activity during the reading of sentences and word sequences can show a network that is involved in high-level sentence processing, but it cannot distinguish between compositional processing, syntactic structure building, morphological processing, etc. Processing word sequences additionally is not ecologically valid, which might influence the response of cognitive control regions. The use of coarse tasks, therefore, may be another reason for distributed patterns of brain activity. MVPA and encoding models (see Box 1.1) aim to distinguish very similar conditions (with the same task loading) that only differ in one dimension ideally (e.g. semantic category). They are thus able to isolate the linguistic factor of interest by reducing task effects, but they may still be confounded in the specific linguistic characterisation. For example, as explained in Chapter 5, the ordinal, syntactic and event models capture differences both at pragmatic and syntactic levels. This inadequate characterization of linguistic function allows for only coarse mapping to brain activity, which as discussed is also characterized in large scales. As a consequence, this approach allows for at best correlational neurolinguistics (Embick & Poeppel, 2015). Instead, fMRI can be useful to address processing questions that cannot be directly answered with other methods (i.e. using neurobiology to adjudicate among computational theories of language, see integrated neurolinguistics, Embick & Poeppel, 2015). I took this approach in Chapter 4 by targeting the scope of incrementality in production.

6.6 Future directions

Overall, in this thesis I targeted the nature of the brain organization for sentence production and comprehension, and attempted to unify our understanding of the neural infrastructure for language with knowledge from both modalities. We are one small step ahead but we are still far from a complete understanding of production and comprehension mechanisms and shared infrastructure. There are a few areas that could benefit from theoretical, methodological or ecological advances.

As suggested in Chapter 4, parsing models that are inspired by processing in production could increase our understanding of the incrementality of structure building and lexical selection. In Chapter 4, I initially used comprehension-based parsers to address phrase-structure building in production. I then developed two parsers that are better suited to production. These were initial attempts that assumed fixed dynamics of structure building during the speakers' monologues. By taking into account the flexibility in planning scope, and the difficulty of lexical retrieval, more dynamic and complete parsers could be developed that better model planning and neural dynamics in production. This enterprise would further benefit from studies with better temporal resolution, such as electrophysiological studies. The electrophysiological literature on sentence production is limited because of the sensitivity to motor artefacts. The dynamics of word production have been studied with M/EEG methods, by focusing on the time just preceding the onset of speech (e.g. Carota, Schoffelen, Oostenveld, & Indefrey, 2022; Conner et al., 2019; Fairs, Michelas, Dufour, & Strijkers, 2021; Piai, Roelofs, Rommers, & Maris, 2015). The incrementality of sentence production, however, prevents a full account of the dynamics of sentence generation in the time before speech onset, since sentence encoding is interleaved with articulation. A few studies that attempted to study sentence production with EEG focused on planning stages (i.e. before speech onset) or used subvocal articulation (e.g. Mascelloni, Zamparelli, Vespignani, Gruber, & Mueller, 2019; Sauppe et al., 2021).

Additionally, cross-linguistic studies can answer questions about the dynamics of structure building by taking advantage of different features of languages (Bornkessel-Schlesewsky & Schlesewsky, 2016). For example, Sarvasy et al. (2022) found that the scope of planning can be very extended by using a language where the verb needs to mark if the subject of the next clause differs from the subject of the current clause. Momma et al. (2016) found properties of planning scope in Japanese, which could not be uncovered with English due to its different word order. Similarly, studying the spontaneous production of different languages could help elucidate whether structure building is equally incremental in head-final languages.

Finally, studying language processing in dialogue would allow for uncovering the live interaction of production and comprehension processes in the brain, with the added value of naturalistic language processing (see for psycholinguistic evidence in dialogue e.g. Corps, Knudsen, & Meyer, 2022; Meyer, Alday, Decuyper, & Knudsen, 2018).

6.7 Conclusions

In this thesis, I investigated the neural correlates of sentence production and comprehension with different experimental paradigms and analysis approaches. I used a picture description experiment, a sentence recall paradigm and spontaneous production to learn more about the brain organization of language production in different contexts. I also used (i) standard analysis techniques to uncover brain regions involved in syntactic processing load; (ii) functional connectivity analysis to understand the connectivity of regions for structure building; (iii) continuous regression of structure building measures against average activity in key regions for language processing; (iv) encoding models to characterize semantic structures used during sentence production. I found a reliable fronto-temporal network that was involved in both sentence production and comprehension. In particular, compositional processing in production engaged a broad network that encoded meaning in different relational structures. Syntactic structure building was processed in a dynamic fronto-temporal network in both production and comprehension. Therefore, I found evidence for the same brain regions to process sentence-level structures in production and comprehension. The different processing contexts and goals of production and comprehension were seen to lead to different dynamics of structure building. Overall, studying sentence production improved our understanding of the neural correlates of sentence processing. In addition, it raised questions about the commonality of neural resources, representations and processes between modalities that can drive efforts for a deeper and broader understanding of the neurobiology of language processing. Finally, the use of spontaneous production uncovered aspects of the incrementality of production with novel techniques. Future studies taking advantage of more naturalistic approaches for both speaking and listening may be able to further characterize the complexities of the moment in the brain between listening, thinking and speaking.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., . . . Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*.

Alday, P. M., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2017). Electrophysiology Reveals the Neural Dynamics of Naturalistic Auditory Language Processing: Event-Related Potentials Reflect Continuous Model Updates. *eNeuro*, *4*(6). (Publisher: Society for Neuroscience Section: New Research) doi: 10.1523/ENEURO.0311-16.2017

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., . . . Lalor, E. C. (2021). Deep Artificial Neural Networks Reveal a Distributed Cortical Network Encoding Propositional Sentence-Level Meaning. *Journal of Neuroscience*, *41*(18), 4100–4119. (Publisher: Society for Neuroscience Section: Research Articles) doi: 10.1523/JNEUROSCI.1152-20.2021

Anderson, A. J., Lalor, E. C., Lin, F., Binder, J. R., Fernandino, L., Humphries, C. J., . . . Wang, X. (2019). Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences. *Cerebral Cortex*, *29*(6), 2396–2411. doi: 10.1093/cercor/bhy110

Andric, M., & Small, S. L. (2015). fMRI methods for studying the neurobiology of language under naturalistic conditions. In R. M. Willems (Ed.), *Cognitive Neuroscience of Natural Language Use* (pp. 8–28). Cambridge: Cambridge University Press. doi: 10.1017/CBO9781107323667.002

Arana, S. (2022). *Abstract neural representations of language during sentence comprehension: Evidence from MEG and Behaviour.* PhD Thesis, Radboud University Nijmegen, Nijmegen..

Avants, B., Epstein, C., Grossman, M., & Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. doi: 10.1016/j.media.2007.06.004

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, *47*(1), 31–56. (Publisher: SAGE Publications Ltd) doi: 10.1177/00238309040470010201

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*(9), 1338–1367. (Publisher: Routledge _eprint: https://doi.org/10.1080/01690965.2010.542671) doi: 10.1080/01690965.2010.542671

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Basilakos, A., Smith, K. G., Fillmore, P., Fridriksson, J., & Fedorenko, E. (2018). Functional Characterization of the Human Speech Articulation Network. *Cerebral Cortex*, *28*(5), 1816–1830. (Publisher: Oxford Academic) doi: 10.1093/cercor/bhx100

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. (Number: 1) doi: 10.18637/jss.v067.i01

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the Language Given: The Neural Correlates of Inferring Speaker Meaning. *Cerebral Cortex*, *24*(10), 2572–2578. doi: 10.1093/cercor/bht112

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. doi: 10.1016/j.neuroimage.2007.04.042

Bemis, D. K., & Pylkkänen, L. (2011). Simple Composition: A Magnetoencephalography Investigation into the Comprehension of Minimal Linguistic Phrases. *Journal of Neuroscience*, *31*(8), 2801–2814. (Publisher: Society for Neuroscience Section: Articles) doi: 10.1523/JNEUROSCI.5003-10.2011

Bemis, D. K., & Pylkkänen, L. (2013). Basic Linguistic Composition Recruits the Left Anterior Temporal Lobe and Left Angular Gyrus During Both Listening and Reading. *Cerebral Cortex*, *23*(8), 1859–1873. (Publisher: Oxford Academic) doi: 10.1093/cercor/bhs170

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal* *of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x) doi: 10.1111/j.2517-6161.1995.tb02031.x

Bhattasali, S., Fabre, M., Luh, W.-M., Saied, H. A., Constant, M., Pallier, C., … Hale, J. (2019). Localising memory retrieval and syntactic composition: an fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, *34*(4), 491–510. doi: 10.1080/23273798.2018.1518533

Binder, J. R. (2017). Current Controversies on Wernicke's Area and its Role in Language. *Current Neurology and Neuroscience Reports*, *17*(8), 58. doi: 10.1007/s11910-017-0764-8

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, *34*(4), 537–541. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.1910340409) doi: 10.1002/mrm.1910340409

Blanco-Elorrieta, E., Kastner, I., Emmorey, K., & Pylkkänen, L. (2018). Shared neural correlates for building phrases in signed and spoken language. *Scientific Reports*, *8*(1), 5492. (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41598-018-23915-0

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. doi: 10.1016/j.neuroimage.2015.11.069

Bock, K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*(1), 1–47. (Place: US Publisher: American Psychological Association) doi: 10.1037/0033-295X.89.1.1

Bock, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin & Review*, *3*(4), 395–421. doi: 10.3758/BF03214545

Bock, K., & Ferreira, V. (2014). Syntactically speaking. In *The Oxford handbook of language production* (pp. 21–46). New York, NY, US: Oxford University Press. doi: 10.1093/oxfordhb/9780199735471.013.008

Bock, K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In *Handbook of psycholinguistics. ed. by Morton A. Gernsbacher,* (pp. 945–984). San Diego, CA: Academic Press.

Bohsali, A., & Crosson, B. (2016). The Basal Ganglia and Language: A Tale of Two Loops. In J.-J. Soghomonian (Ed.), *The Basal Ganglia: Novel Per-*

spectives on Motor and Cognitive Functions (pp. 217–242). Cham: Springer International Publishing. doi: 10.1007/978-3-319-42743-0_10

Bornkessel-Schlesewsky, I., Grewe, T., & Schlesewsky, M. (2012). Prominence vs. aboutness in sequencing: A functional distinction within the left inferior frontal gyrus. *Brain and Language*, *120*(2), 96–107. doi: 10.1016/j.bandl.2010.06.004

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Reconciling time, space and function: A new dorsal–ventral stream model of sentence comprehension. *Brain and Language*, *125*(1), 60–76. doi: 10.1016/j.bandl.2013.01.010

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2016). The importance of linguistic typology for the neurobiology of language. *Linguistic Typology*, *20*(3), 615–621. (Publisher: De Gruyter Mouton) doi: 10.1515/lingty-2016-0032

Bornkessel-Schlesewsky, I., Schlesewsky, M., & Cramon, D. Y. v. (2009). Word order and Broca's region: Evidence for a supra-syntactic perspective. *Brain and Language*, *111*(3), 125–139. doi: 10.1016/j.bandl.2009.09.004

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, *44*(2), 123–147. (Publisher: SAGE Publications Ltd) doi: 10.1177/00238309010440020101

Boylan, C., Trueswell, J. C., & Thompson-Schill, S. L. (2017). Relational vs. attributive interpretation of nominal compounds differentially engages angular gyrus and anterior temporal lobe. *Brain and Language*, *169*, 8–21. doi: 10.1016/j.bandl.2017.01.008

Brennan, J. R. (2016). Naturalistic Sentence Comprehension in the Brain. *Language and Linguistics Compass*, *10*(7), 299–313. doi: 10.1111/lnc3.12198

Brennan, J. R., Dyer, C., Kuncoro, A., & Hale, J. T. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, *146*, 107479. doi: 10.1016/j.neuropsychologia.2020.107479

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, *14*(1), e0207741. doi: 10.1371/journal.pone.0207741

Brennan, J. R., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163–173. doi: 10.1016/j.bandl.2010.04.002

Brennan, J. R., & Pylkkänen, L. (2017). MEG Evidence for Incremental Sentence Composition in the Anterior Temporal Lobe. *Cognitive Science*, *41*(S6), 1515–1531. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12445) doi: 10.1111/cogs.12445

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157-158*, 81–94. doi: 10.1016/j.bandl.2016.04.008

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis using an SPM toolbox. Sendai, Japan.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216. doi: 10.1016/j.jml.2016.10.002

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. doi: 10.1016/j.actpsy.2014.04.010

Buckner, R. L., Krienen, F. M., & Yeo, B. T. T. (2013). Opportunities and limitations of intrinsic functional connectivity MRI. *Nature Neuroscience*, *16*(7), 832–837. (Number: 7 Publisher: Nature Publishing Group) doi: 10.1038/nn.3423

Bulut, T. (2022). Meta-analytic connectivity modeling of the left and right inferior frontal gyri. *Cortex*. doi: 10.1016/j.cortex.2022.07.003

Carota, F., Schoffelen, J.-M., Oostenveld, R., & Indefrey, P. (2022). The Time Course of Language Production as Revealed by Pattern Classification of MEG Sensor Data. *Journal of Neuroscience*, *42*(29), 5745–5754. (Publisher: Society for Neuroscience Section: Research Articles) doi: 10.1523/JNEUROSCI.1923-21.2022

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, *11*(3), e1460. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1460) doi: 10.1002/wics.1460

Chang, C. H. C., Dehaene, S., Wu, D. H., Kuo, W.-J., & Pallier, C. (2020). Cortical encoding of linguistic constituent with and without morphosyntactic cues. *Cortex*. doi: 10.1016/j.cortex.2020.04.024

Chen, Leong, Y.C., Honey, C.J., Yong C.H., Norman, K.A., & Hasson, U. (2018).

Sherlock. OpenNeuro. [Dataset].

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/nn.4450

Clark, E. V., & Hecht, B. F. (1983). Comprehension, Production, and Language Acquisition. *Annual Review of Psychology*, *34*(1), 325–349. (_eprint: https://doi.org/10.1146/annurev.ps.34.020183.001545) doi: 10.1146/annurev.ps.34.020183.001545

Collina, S., Seurinck, R., & Hartsuiker, R. J. (2014). Inside the Syntactic Box: The Neural Correlates of the Functional and Positional Level in Covert Sentence Production. *PLOS ONE*, *9*(9), e106122. doi: 10.1371/journal.pone.0106122

Conner, C. R., Kadipasaoglu, C. M., Shouval, H. Z., Hickok, G., & Tandon, N. (2019). Network dynamics of Broca's area during word selection. *PLOS ONE*, *14*(12), e0225756. (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0225756

Coopmans, C. (2023). *Triangles in the brain: The role of hierarchical structure in language use*. PhD Thesis, Radboud University Nijmegen, Nijmegen..

Corps, R. E., Knudsen, B., & Meyer, A. S. (2022). Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. *Cognition*, *223*, 105037. doi: 10.1016/j.cognition.2022.105037

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. doi: 10.1006/cbmr.1996.0014

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, *8*(2-3), 109–114.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, *9*(2), 179–194. doi: 10.1006/nimg.1998.0395

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394. (Publisher: Royal Society) doi: 10.1098/rstb.2012.0394

den Ouden, D.-B., Hoogduin, H., Stowe, L. A., & Bastiaanse, R. (2008). Neural correlates of Dutch Verb Second in speech production. *Brain and Language*, *104*(2), 122–131. doi: 10.1016/j.bandl.2007.05.001

den Ouden, D.-B., Saur, D., Mader, W., Schelter, B., Lukic, S., Wali, E., … Thompson, C. K. (2012). Network modulation during complex syntactic processing. *NeuroImage*, *59*(1), 815–823. doi: 10.1016/j.neuroimage.2011.07.057

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. doi: 10.1016/j.neuroimage.2006.01.021

De Smedt, K., & Kempen, G. (1987). Incremental Sentence Production, Self-Correction and Coordination. In G. Kempen (Ed.), *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics* (pp. 365–376). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-009-3645-4_23

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. doi: 10.1016/j.neuroimage.2010.06.010

Dick, A. S., Bernal, B., & Tremblay, P. (2014). The Language Connectome: New Pathways, New Concepts. *The Neuroscientist*, *20*(5), 453–467. (Publisher: SAGE Publications Inc STM) doi: 10.1177/1073858413513502

Diedrichsen, J., King, M., Hernandez-Castillo, C., Sereno, M., & Ivry, R. B. (2019). Universal Transform or Multiple Functionality? Understanding the Contribution of the Human Cerebellum across Task Domains. *Neuron*, *102*(5), 918–928. doi: 10.1016/j.neuron.2019.04.021

Embick, D., & Poeppel, D. (2015). Towards a computational(ist) neurobiology of language: correlational, integrated and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, *30*(4), 357–366. (Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2014.980750) doi: 10.1080/23273798.2014.980750

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., … Gorgolewski, K. J. (2018). *fMRIPrep*. (Publisher: Zenodo)

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2018). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 1. doi: 10.1038/s41592-018-0235-4

Europa, E., Gitelman, D. R., Kiran, S., & Thompson, C. K. (2019). Neural Connectivity in Syntactic Movement Processing. *Frontiers in Human Neuroscience*, *13*. (Publisher: Frontiers) doi: 10.3389/fnhum.2019.00027

Fairs, A., Michelas, A., Dufour, S., & Strijkers, K. (2021). The Same Ultra-Rapid Parallel Brain Dynamics Underpin the Production and Perception of Speech. *Cerebral Cortex Communications*, *2*(3), tgab040. doi: 10.1093/texcom/tgab040

Fedorenko, E., & Blank, I. A. (2020). Broca's Area Is Not a Natural Kind. *Trends in Cognitive Sciences*, *24*(4), 270–284. doi: 10.1016/j.tics.2020.01.001

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348. doi: 10.1016/j.cognition.2020.104348

Fedorenko, E., Duncan, J., & Kanwisher, N. (2012). Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area. *Current Biology*, *22*(21), 2059–2062. doi: 10.1016/j.cub.2012.09.011

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. (Publisher: American Physiological Society) doi: 10.1152/jn.00032.2010

Fedorenko, E., Nieto-Castañon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513. doi: 10.1016/j.neuropsychologia.2011.09.014

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, *30*(2), 210–233. doi: 10.1016/0749-596X(91)90004-4

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203. doi: 10.1016/S0010-0285(03)00005-7

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. (Publisher: SAGE Publications Inc) doi: 10.1111/1467-8721.00158

Ferreira, F., & Lowder, M. W. (2016). Chapter Six - Prediction, Information Structure, and Good-Enough Language Processing. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 65, pp. 217–247). Academic Press. doi: 10.1016/bs.plm.2016.04.002

Flick, G., Oseki, Y., Kaczmarek, A. R., Al Kaabi, M., Marantz, A., & Pylkkänen, L. (2018). Building words and phrases in the left temporal lobe. *Cortex*, *106*, 213–236. doi: 10.1016/j.cortex.2018.06.004

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., & Crone, N. E. (2015). Redefining the role of Broca's area in speech. *Proceedings of the National Academy of Sciences*, *112*(9), 2871–2875. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1414491112

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47, Supplement 1*, S102. doi: 10.1016/S1053-8119(09)70884-5

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., … R-Core (2020). *car: Companion to Applied Regression.*

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., … R-Core (2021). *car: Companion to Applied Regression.* Retrieved 2022-04-26, from `https://CRAN.R-project.org/package=car`

Frankland, S. M., & Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, *112*(37), 11732–11737. doi: 10.1073/pnas.1421236112

Frankland, S. M., & Greene, J. D. (2020a). Concepts and Compositionality: In Search of the Brain's Language of Thought. *Annual Review of Psychology*, *71*(1), 273–303. (_eprint: https://doi.org/10.1146/annurev-psych-122216-011829) doi: 10.1146/annurev-psych-122216-011829

Frankland, S. M., & Greene, J. D. (2020b). Two Ways to Build a Thought: Distinct Forms of Compositional Semantic Representation across Brain Regions. *Cerebral Cortex*, *30*(6), 3838–3855. (Publisher: Oxford Academic) doi: 10.1093/cercor/bhaa001

Friederici, A. D. (2009). Pathways to language: fiber tracts in the human brain. *Trends in Cognitive Sciences*, *13*(4), 175–181. doi: 10.1016/j.tics.2009.01.001

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, *91*(4), 1357–1392. (Publisher: American Physiological Society) doi: 10.1152/physrev.00006.2011

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, *16*(5), 262–268. doi: 10.1016/j.tics.2012.04.001

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254. doi: 10.1016/j.conb.2012.10

.002

Friston, K. J. (2002). Beyond Phrenology: What Can Neuroimaging Tell Us About Distributed Circuitry? *Annual review of neuroscience*, *25*, 221–50. doi: 10.1146/annurev.neuro.25.112701.142846

Friston, K. J. (2011). Functional and Effective Connectivity: A Review. *Brain Connectivity*, *1*(1), 13–36. (Publisher: Mary Ann Liebert, Inc., publishers) doi: 10.1089/brain.2011.0008

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and Modulatory Interactions in Neuroimaging. *NeuroImage*, *6*(3), 218–229. doi: 10.1006/nimg.1997.0291

Friston, K. J., Frith, C. D., Frackowiak, R. S. J., & Turner, R. (1995). Characterizing Dynamic Brain Responses with fMRI: A Multivariate Approach. *NeuroImage*, *2*(2, Part A), 166–172. doi: 10.1006/nimg.1995.1019

Friston, K. J., Penny, W. D., & Glaser, D. E. (2005). Conjunction revisited. *NeuroImage*, *25*(3), 661–667. doi: 10.1016/j.neuroimage.2005.01.013

Gambi, C., & Pickering, M. J. (2017). Models linking production and comprehension. In *The handbook of psycholinguistics* (pp. 157–181). Wiley Online Library.

Garrett, M. F. (1980). Levels of processing in sentence production. In *Language production Vol. 1: Speech and talk* (pp. 177–220). Academic Press.

Garrett, M. F. (1982). Remarks on the relation between language production and language comprehension systems. In *Neural models of language processes* (pp. 209–224). Elsevier.

Giglio, L., Ostarek, M., Weber, K., & Hagoort, P. (2022). Commonalities and Asymmetries in the Neurobiological Infrastructure for Language Production and Comprehension. *Cerebral Cortex*, *32*(7), 1405–1418. doi: 10.1093/cercor/bhab287

Glasser, M. F., & Rilling, J. K. (2008). DTI Tractography of the Human Brain's Language Pathways. *Cerebral Cortex*, *18*(11), 2471–2482. doi: 10.1093/cercor/bhn011

Goldberg, A. E., & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, *26*(4), 300–311. doi: 10.1016/j.tics.2022.01.005

Golestani, N., Alario, F. X., Meriaux, S., Le Bihan, D., Dehaene, S., & Pallier, C. (2006). Syntax production in bilinguals. *Neuropsychologia*, *44*(7), 1029–1040. doi: 10.1016/j.neuropsychologia.2005.11.009

Gorgolewski, K. J., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, *5*, 13. doi: 10.3389/fninf.2011.00013

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., … Ghosh, S. (2018). Nipype. *Software*. (Publisher: Zenodo) doi: 10.5281/zenodo.596855

Graessner, A., Zaccarella, E., & Hartwigsen, G. (2021). Differential contributions of left-hemispheric language regions to basic semantic composition. *Brain Structure and Function*, *226*(2), 501–518. doi: 10.1007/s00429-020-02196-2

Grande, M., Meffert, E., Schoenberger, E., Jung, S., Frauenrath, T., Huber, W., … Heim, S. (2012). From a concept to a word in a syntactically complete sentence: An fMRI study on spontaneous language production in an overt picture description task. *NeuroImage*, *61*(3), 702–714. doi: 10.1016/j.neuroimage.2012.03.087

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. doi: 10.1016/j.neuroimage.2009.06.060

Griffin, R., Friedman, O., Ween, J., Winner, E., Happé, F., & Brownell, H. (2006). Theory of mind and the right cerebral hemisphere: Refining the scope of impairment. *Laterality*, *11*(3), 195–225. (Publisher: Routledge _eprint: https://doi.org/10.1080/13576500500450552) doi: 10.1080/13576500500450552

Griffin, Z. M., & Bock, K. (2000). What the Eyes Say About Speaking:. *Psychological Science*. (Publisher: SAGE PublicationsSage CA: Los Angeles, CA)

Griffiths, J. D., Marslen-Wilson, W. D., Stamatakis, E. A., & Tyler, L. K. (2013). Functional Organization of the Neural Language System: Dorsal and Ventral Pathways Are Critical for Syntax. *Cerebral Cortex*, *23*(1), 139–147. doi: 10.1093/cercor/bhr386

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. doi: 10.1016/j.tics.2005.11.006

Grimshaw, J. (1990). *Argument structure*. Cambridge, MA, US: The MIT Press. (Pages: x, 202)

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences*, *23*(1), 1–21. doi: 10.1017/S0140525X00002399

Guhe, M. (2007). *Incremental conceptualization for language production.* Lawrence Erlbaum Associates Publishers. (Publisher: Lawrence Erlbaum Associates Publishers)

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, *142*(3), 880–905. (Place: US Publisher: American Psychological Association) doi: 10.1037/a0030045

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, *175*, 36–52. doi: 10.1016/j.cognition.2018.02.011

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, *9*(9), 416–423. doi: 10.1016/j.tics.2005.07.004

Hagoort, P. (2013). MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology*, *4*, 416. doi: 10.3389/fpsyg.2013.00416

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current Opinion in Neurobiology*, *28*, 136–141. doi: 10.1016/j.conb.2014.07.013

Hagoort, P. (2017). The core and beyond in the language-ready brain. *Neuroscience & Biobehavioral Reviews*, *81*, 194–204. doi: 10.1016/j.neubiorev.2017.01.048

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, *366*(6461), 55–58. (Publisher: American Association for the Advancement of Science Section: Review) doi: 10.1126/science.aax0289

Hagoort, P., & Beckmann, C. F. (2019). Key issues and future directions: The neural architecture for language. In *Human language: From genes and brains to behavior* (pp. 527–532). MIT Press.

Hagoort, P., & Indefrey, P. (2014). The Neurobiology of Language Beyond Single Words. *Annual Review of Neuroscience*, *37*(1), 347–362. doi: 10.1146/annurev-neuro-071013-013847

Hale, J. T. (2014). *Automaton theories of human sentence comprehension*. Center for the Study of Language and Information.

Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. R. (2022). Neurocomputational Models of Language Processing. *Annual Review of Linguistics*, *8*(1), 427–446. (_eprint: https://doi.org/10.1146/annurev-linguistics-051421-020803) doi: 10.1146/annurev-linguistics-051421-020803

Hall, C. N., Howarth, C., Kurth-Nelson, Z., & Mishra, A. (2016). Interpreting BOLD: towards a dialogue between cognitive and cellular neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1705), 20150348. (Publisher: Royal Society) doi: 10.1098/rstb.2015.0348

Haller, S., Radue, E. W., Erb, M., Grodd, W., & Kircher, T. (2005). Overt sentence production in event-related fMRI. *Neuropsychologia*, *43*(5), 807–814. doi: 10.1016/j.neuropsychologia.2004.09.007

Happé, F., Brownell, H., & Winner, E. (1999). Acquired 'theory of mind' impairments following stroke. *Cognition*, *70*(3), 211–240. doi: 10.1016/S0010-0277(99)00005-0

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. doi: 10.1126/science.1063736

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2201968119

Henderson, J. M., Choi, W., Lowder, M. W., & Ferreira, F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage*, *132*, 293–300. doi: 10.1016/j.neuroimage.2016.02.050

Hendriks, P., & Koster, C. (2010). Production/comprehension asymmetries in language acquisition. *Lingua*, *120*(8), 1887–1897. doi: 10.1016/j.lingua.2010.02.002

Henson, R. N. (2015). Analysis of variance (ANOVA). *Brain Mapping: an encyclopedic reference. Elsevier*, 477–481.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. doi: 10.1038/nrn2113

Hirotani, M., Makuuchi, M., Rüschemeyer, S.-A., & Friederici, A. D. (2011). Who was the agent? The neural correlates of reanalysis processes during sentence comprehension. *Human Brain Mapping*, *32*(11), 1775–1787. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.21146) doi: 10.1002/hbm.21146

Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., … Fedorenko, E. (2022). Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, bhac350. doi: 10.1093/cercor/bhac350

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31. (Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2015.1072223) doi: 10.1080/23273798.2015.1072223

Huizeling, E., Peeters, D., & Hagoort, P. (2021). Prediction of upcoming speech under fluent and disfluent conditions: eye tracking evidence from immersive virtual reality. *Language, Cognition and Neuroscience*, *0*(0), 1–28. (Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2021.1994621) doi: 10.1080/23273798.2021.1994621

Hultén, A., Schoffelen, J.-M., Uddén, J., Lam, N. H. L., & Hagoort, P. (2019). How the brain makes sense beyond the processing of single words – An MEG study. *NeuroImage*, *186*, 586–594. doi: 10.1016/j.neuroimage.2018.11.035

Humphreys, G. F., & Gennari, S. P. (2014). Competitive mechanisms in sentence processing: Common and distinct production and reading comprehension networks linked to the prefrontal cortex. *NeuroImage*, *84*, 354–366. doi: 10.1016/j.neuroimage.2013.08.059

Indefrey, P. (2018). The Relationship Between Syntactic Production and Comprehension. *The Oxford Handbook of Psycholinguistics*. doi: 10.1093/oxfordhb/9780198786825.013.20

Indefrey, P., Brown, C. M., Hellwig, F., Amunts, K., Herzog, H., Seitz, R. J., & Hagoort, P. (2001). A neural correlate of syntactic encoding during speech production. *Proceedings of the National Academy of Sciences*, *98*(10), 5933–5936. doi: 10.1073/pnas.101118098

Indefrey, P., Hellwig, F., Herzog, H., Seitz, R. J., & Hagoort, P. (2004). Neural responses to the production and comprehension of syntax in identical utterances. *Brain and Language*, *89*(2), 312–319. doi: 10.1016/S0093-934X(03)00352-3

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144. doi: 10.1016/j.cognition.2002.06.001

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience*, *9*(4), 304–313. (Number: 4 Publisher: Nature Publishing Group) doi: 10.1038/nrn2332

Jackendoff, R. S. (1992). *Semantic structures* (Vol. 18). MIT press.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi: 10.1016/j.jml.2007.11.007

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. doi: 10.1016/j.cogpsych.2010.02.002

Jaeger, T. F., & Levy, R. (2007). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems* (Vol. 19). MIT Press.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. doi: 10.1006/nimg.2002.1132

Karimi, H. (2022). Greater entropy leads to more explicit referential forms during language production. *Cognition*, *225*, 105093. doi: 10.1016/j.cognition.2022.105093

Kearney, E., & Guenther, F. H. (2019). Articulating: the neural mechanisms of speech production. *Language, Cognition and Neuroscience*, *34*(9), 1214–1229. (Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2019.1589541) doi: 10.1080/23273798.2019.1589541

Kempen, G. (2000). Could grammatical encoding and grammatical decoding be subserved by the same processing module? *Behavioral and Brain Sciences*, *23*(1), 38–39. (Publisher: Cambridge University Press) doi: 10.1017/S0140525X00402396

Kempen, G., Olsthoorn, N., & Sprenger, S. (2012). Grammatical workspace sharing during language production and language comprehension: Evidence from grammatical multitasking. *Language and Cognitive Processes*, *27*(3), 345–380. (Publisher: Routledge _eprint: https://doi.org/10.1080/01690965.2010.544583) doi: 10.1080/01690965.2010.544583

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. doi: 10.3758/BRM.42.3.643

Kircher, T. T. J., Oh, T. M., Brammer, M. J., & McGuire, P. K. (2005). Neural correlates of syntax production in schizophrenia. *The British Journal of Psychiatry*, *186*(3), 209–214. doi: 10.1192/bjp.186.3.209

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., … Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*(2), e1005350. doi: 10.1371/journal.pcbi.1005350

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03* (Vol. 1, pp. 423–430). Sapporo, Japan: Association for Computational Linguistics. doi: 10.3115/1075096.1075150

Kochari, A. R., Lewis, A. G., Schoffelen, J.-M., & Schriefers, H. (2021). Semantic and syntactic composition of minimal adjective-noun phrases in Dutch: An MEG study. *Neuropsychologia*, *155*, 107754. doi: 10.1016/j.neuropsychologia.2021.107754

Konopka, A. E. (2019). Encoding actions and verbs: Tracking the time-course of relational encoding during message and sentence formulation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. (ISSN: 1939-1285 Issue: 8 Pages: 1486 Publisher: US: American Psychological Association Volume: 45) doi: 10.1037/xlm0000650

Konopka, A. E., & Brown-Schmidt, S. (2014). Message encoding. In *The Oxford handbook of language production* (pp. 3–20). New York, NY, US: Oxford University Press. doi: 10.1093/oxfordhb/9780199735471.013.010

Konopka, A. E., & Meyer, A. S. (2014). Priming sentence planning. *Cognitive Psychology*, *73*, 1–40. doi: 10.1016/j.cogpsych.2014.04.001

Krause, F., Benjamins, C., Eck, J., Lührs, M., Hoof, R. v., & Goebel, R. (2019). Active head motion reduction in magnetic resonance imaging using tactile feedback. *Human Brain Mapping*, *40*(14), 4026–4037. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.24683) doi: https://doi.org/10.1002/hbm.24683

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. doi: 10.1073/pnas.0600244103

Kuchinsky, S. E., Bock, K., & Irwin, D. E. (2011). Reversing the hands of time: Changing the mapping from seeing to saying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 748–756. (Place: US Publisher: American Psychological Association) doi: 10.1037/a0022637

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/nrn.2016.150

LeBel, A., Jain, S., & Huth, A. G. (2021). Voxelwise Encoding Models Show That Cerebellar Language Representations Are Highly Conceptual. *Journal of Neuroscience*, *41*(50), 10341–10355. (Publisher: Society for Neuroscience Section: Research Articles) doi: 10.1523/JNEUROSCI.0118-21.2021

Lee, D. K., Fedorenko, E., Simon, M. V., Curry, W. T., Nahed, B. V., Cahill, D. P., & Williams, Z. M. (2018). Neural encoding and production of functional morphemes in the posterior temporal lobe. *Nature Communications*, *9*(1), 1877. doi: 10.1038/s41467-018-04235-3

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* Retrieved 2022-04-26, from `https://CRAN.R-project.org/package=emmeans`

Lenth, R. V., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means.*

Lesage, E., Morgan, B. E., Olson, A. C., Meyer, A. S., & Miall, R. C. (2012). Cerebellar rTMS disrupts predictive language processing. *Current Biology*, *22*(18), R794–R795. doi: 10.1016/j.cub.2012.07.006

Leshinskaya, A., & Thompson-Schill, S. L. (2020). Transformation of Event Representations along Middle Temporal Gyrus. *Cerebral Cortex*, *30*(5), 3148–3166. (Publisher: Oxford Academic) doi: 10.1093/cercor/bhz300

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* MIT Press Series in Natural-Language Processing. MIT Press, Cambridge, Massachusetts.

Levelt, W. J. M. (2013). *A history of psycholinguistics: The pre-Chomskyan era*. Oxford University Press.

Levin, B., & Hovav, M. R. (2005). *Argument realization*. Cambridge University Press Cambridge.

Li, J., & Hale, J. (2019). Grammatical predictors for fMRI timecourses. *Minimalist parsing*, 159–173.

Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short term memory. *Journal of Memory and Language*, *31*(6), 713–733. doi: 10.1016/0749-596X(92)90036-W

Lopopolo, A., van den Bosch, A., Petersson, K.-M., & Willems, R. M. (2021). Distinguishing Syntactic Operations in the Brain: Dependency and Phrase-

Structure Parsing. *Neurobiology of Language*, *2*(1), 152–175. doi: 10.1162/nol_a_00029

Lyu, B., Choi, H. S., Marslen-Wilson, W. D., Clarke, A., Randall, B., & Tyler, L. K. (2019). Neural dynamics of semantic composition. *Proceedings of the National Academy of Sciences*, *116*(42), 21318–21327. (Publisher: National Academy of Sciences Section: PNAS Plus) doi: 10.1073/pnas.1903402116

Mack, J. E., Meltzer-Asscher, A., Barbieri, E., & Thompson, C. K. (2013). Neural Correlates of Processing Passive Sentences. *Brain Sciences*, *3*(3), 1198–1214. (Number: 3 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/brainsci3031198

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, Maryland: Association for Computational Linguistics. doi: 10.3115/v1/P14-5010

Mascelloni, M., Zamparelli, R., Vespignani, F., Gruber, T., & Mueller, J. L. (2019). Distinct Neural Processes for Memorizing Form and Meaning Within Sentences. *Frontiers in Human Neuroscience*, *13*. doi: 10.3389/fnhum.2019.00412

Matchin, W. (2018). A neuronal retuning hypothesis of sentence-specificity in Broca's area. *Psychonomic Bulletin & Review*, *25*(5), 1682–1694. doi: 10.3758/s13423-017-1377-6

Matchin, W., Hammerly, C., & Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, *88*, 106–123. doi: 10.1016/j.cortex.2016.12.010

Matchin, W., & Hickok, G. (2016). 'Syntactic Perturbation' During Production Activates the Right IFG, but not Broca's Area or the ATL. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.00241

Matchin, W., & Hickok, G. (2020). The Cortical Organization of Syntax. *Cerebral Cortex*, *30*(3), 1481–1498. (Publisher: Oxford Academic) doi: 10.1093/cercor/bhz180

Matchin, W., Liao, C.-H., Gaston, P., & Lau, E. (2019). Same words, different structures: An fMRI investigation of argument relations and the angular gyrus. *Neuropsychologia*, *125*, 116–128. doi: 10.1016/j.neuropsychologia.2019.01.019

Matchin, W., & Wood, E. (2020). Syntax-sensitive regions of the posterior inferior frontal gyrus and the posterior temporal lobe are differentially recruited by production and perception. *Cerebral Cortex Communications*. doi: 10.1093/texcom/tgaa029

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *NeuroImage*, *61*(4), 1277–1286. doi: 10.1016/j.neuroimage.2012.03.068

Meltzer-Asscher, A., Mack, J. E., Barbieri, E., & Thompson, C. K. (2015). How the brain processes different dimensions of argument structure complexity: Evidence from fMRI. *Brain and Language*, *142*, 65–75. doi: 10.1016/j.bandl.2014.12.005

Meltzer-Asscher, A., Schuchard, J., Ouden, D.-B. d., & Thompson, C. K. (2013). The neural substrates of complex argument structure representations: Processing "alternating transitivity" verbs. *Language and Cognitive Processes*, *28*(8), 1154–1168. (Publisher: Routledge _eprint: https://doi.org/10.1080/01690965.2012.672754) doi: 10.1080/01690965.2012.672754

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared Language: Overlap and Segregation of the Neuronal Infrastructure for Speaking and Listening Revealed by Functional MRI. *Psychological Science*, *22*(9), 1173–1182. doi: 10.1177/0956797611418347

Menenti, L., Petersson, K. M., Scheeringa, R., & Hagoort, P. (2009). When Elephants Fly: Differential Sensitivity of Right and Left Inferior Frontal Gyri to Discourse and World Knowledge. *Journal of Cognitive Neuroscience*, *21*(12), 2358–2368. doi: 10.1162/jocn.2008.21163

Menenti, L., Segaert, K., & Hagoort, P. (2012). The neuronal infrastructure of speaking. *Brain and Language*, *122*(2), 71–80. doi: 10.1016/j.bandl.2012.04.012

Mesulam, M.-M., Rogalski, E. J., Wieneke, C., Hurley, R. S., Geula, C., Bigio, E. H., ... Weintraub, S. (2014). Primary progressive aphasia and the evolving neurology of the language network. *Nature Reviews Neurology*, *10*(10), 554–569. (Number: 10 Publisher: Nature Publishing Group) doi: 10.1038/nrneurol.2014.159

Meyer, A. S., Alday, P. M., Decuyper, C., & Knudsen, B. (2018). Working Together: Contributions of Corpus Analyses and Experimental Psycholinguistics to Understanding Conversation. *Frontiers in Psychology*, *9*. (Publisher:

Frontiers) doi: 10.3389/fpsyg.2018.00525

Meyer, A. S., Huettig, F., & Levelt, W. J. M. (2016). Same, different, or closely related: What is the relationship between language production and comprehension? *Journal of Memory and Language*, *89*, 1–7.

Momma, S., & Ferreira, V. S. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, *114*, 101228. doi: 10.1016/j.cogpsych.2019.101228

Momma, S., & Phillips, C. (2018). The Relationship Between Parsing and Generation. *Annual Review of Linguistics*, *4*(1), 233–254. doi: 10.1146/annurev-linguistics-011817-045719

Momma, S., Slevc, L. R., & Phillips, C. (2016). The timing of verb selection in Japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 813–824. (Place: US Publisher: American Psychological Association) doi: 10.1037/xlm0000195

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, *39*(1), 261–268. doi: 10.1016/j.neuroimage.2007.07.061

Narsude, M., Gallichan, D., Zwaag, W. v. d., Gruetter, R., & Marques, J. P. (2016). Three-dimensional echo planar imaging with controlled aliasing: A sequence for high temporal resolution functional MRI. *Magnetic Resonance in Medicine*, *75*(6), 2350–2361. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.25835) doi: https://doi.org/10.1002/mrm.25835

Nelson, M. J., Karoui, I. E., Giber, K., Yang, X., Cohen, L., Koopman, H., ... Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669–E3678. doi: 10.1073/pnas.1701590114

nilearn/nilearn. (2022). nilearn. Retrieved 2022-05-30, from `https://github.com/nilearn/nilearn` (original-date: 2011-01-09T19:02:23Z)

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. doi: 10.1073/pnas.1018711108

Papafragou, A., & Grigoroglou, M. (2019). The role of conceptualization during language production: evidence from event encoding. *Language, Cognition and Neuroscience*, *34*(9), 1117–1128. (Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2019.1589540) doi: 10.1080/23273798.2019.1589540

Papoutsi, M., Stamatakis, E. A., Griffiths, J., Marslen-Wilson, W. D., & Tyler, L. K. (2011). Is left fronto-temporal connectivity essential for syntax? Effective connectivity, tractography and performance in left-hemisphere damaged patients. *NeuroImage*, *58*(2), 656–664. doi: 10.1016/j.neuroimage.2011.06.036

Paul, H. (1970). The sentence as the expression of the combination of several ideas. In A. Blumenthal (Ed.), *Language and psychology: Historical aspects of psycholinguistics.* New York, NY: Wiley.

Peeters, D. (2019). Virtual reality: A game-changing method for the language sciences. *Psychonomic Bulletin & Review*, *26*(3), 894–900. doi: 10.3758/s13423-019-01571-3

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier. (Google-Books-ID: G_qdEsDlkp0C)

Phillips, C. (2013). Parser-grammar relations: We don't understand everything twice. In *Language Down the Garden Path: The Cognitive and Biological Basis for Linguistic Structures* (pp. 294–315). Oxford University Press Oxford, England. (Publisher: Oxford University Press Oxford, England)

Piai, V., Roelofs, A., Rommers, J., & Maris, E. (2015). Beta oscillations reflect memory and motor aspects of spoken word production. *Human Brain Mapping*, *36*(7), 2767–2780. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.22806) doi: 10.1002/hbm.22806

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1012551108

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. doi: 10.1017/S0140525X04000056

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. (Publisher: Cambridge University Press) doi: 10.1017/S0140525X12001495

Pinheiro, J., & Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag. doi: 10.1007/b98882

Poeppel, D. (2012). The maps problem and the mapping problem:

Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, *29*(1-2), 34–55. (Publisher: Routledge _eprint: https://doi.org/10.1080/02643294.2012.710600) doi: 10.1080/02643294.2012.710600

Posse, S., Wiese, S., Gembris, D., Mathiak, K., Kessler, C., Grosse-Ruyken, M.-L., … Kiselev, V. G. (1999). Enhancement of BOLD-contrast sensitivity by single-shot multi-echo functional MR imaging. *Magnetic Resonance in Medicine*, *42*(1), 87–97. doi: 10.1002/(SICI)1522-2594(199907)42:1<87::AID-MRM13>3.0.CO;2-O

Potter, M. C., & Lombardi, L. (1990). Regeneration in the Short-Term Recall of Sentences. *Journal of Memory and Language; New York*, *29*(6), 633–654.

Potter, M. C., & Lombardi, L. (1998). Syntactic Priming in Immediate Recall of Sentences. *Journal of Memory and Language*, *38*(3), 265–282. doi: 10.1006/jmla.1997.2546

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*(Supplement C), 320–341. doi: 10.1016/j.neuroimage.2013.08.048

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*(1), 62–88. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.2010.05444.x) doi: 10.1111/j.1749-6632.2010.05444.x

Price, C. J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. doi: 10.1016/j.neuroimage.2012.04.062

Pruim, R. H. R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*, *112*(Supplement C), 267–277. doi: 10.1016/j.neuroimage.2015.02.064

Przeździk, I., Haak, K. V., Beckman, C. F., & Bartsch, A. (2019). The human language connectome. *Human language: From genes and brains to behavior*, 467–480. (Publisher: The MIT Press London)

Pylkkänen, L. (2020). Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190299. Retrieved 2022-08-03, from `https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0299` (Publisher: Royal Society) doi:

10.1098/rstb.2019.0299

Pylkkänen, L., Bemis, D. K., & Blanco Elorrieta, E. (2014). Building phrases in language production: An MEG study of simple composition. *Cognition*, *133*(2), 371–384. doi: 10.1016/j.cognition.2014.07.001

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2), 676–682. (Publisher: National Acad Sciences)

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *NeuroImage*, *37*(4), 1083–1090. doi: 10.1016/j.neuroimage.2007.02.041

Ries, S. K., Piai, V., Perry, D., Griffin, S., Jordan, K., Henry, R., … Berger, M. S. (2019). Roles of ventral versus dorsal pathways in language production: An awake language mapping study. *Brain and Language*, *191*, 17–27. doi: 10.1016/j.bandl.2019.01.001

Rilling, J. K., Glasser, M. F., Preuss, T. M., Ma, X., Zhao, T., Hu, X., & Behrens, T. E. J. (2008). The evolution of the arcuate fasciculus revealed with comparative DTI. *Nature Neuroscience*, *11*(4), 426–428. doi: 10.1038/nn2072

Rogalsky, C., Matchin, W., & Hickok, G. (2008). Broca's area, sentence comprehension, and working memory: an fMRI study. *Frontiers in Human Neuroscience*, *2*. (Publisher: Frontiers) doi: 10.3389/neuro.09.014.2008

Runnqvist, E., Bonnard, M., Gauvin, H. S., Attarian, S., Trébuchon, A., Hartsuiker, R. J., & Alario, F. X. (2016). Internal modeling of upcoming speech: A causal role of the right posterior cerebellum in non-motor aspects of language production. *Cortex*, *81*, 203–214. doi: 10.1016/j.cortex.2016.05.008

Santi, A., & Grodzinsky, Y. (2010). fMRI adaptation dissociates syntactic complexity dimensions. *NeuroImage*, *51*(4), 1285–1293. doi: 10.1016/j.neuroimage.2010.03.034

Sarvasy, H. S., Morgan, A. M., Yu, J., Ferreira, V. S., & Momma, S. (2022). Cross-clause planning in Nungon (Papua New Guinea): Eye-tracking evidence. *Memory & Cognition*. doi: 10.3758/s13421-021-01253-3

Sauppe, S., Choudhary, K. K., Giroud, N., Blasi, D. E., Norcliffe, E., Bhattamishra, S., … Bickel, B. (2021). Neural signatures of syntactic variation in speech planning. *PLOS Biology*, *19*(1), e3001038. (Publisher: Public Library of Science) doi: 10.1371/journal.pbio.3001038

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., … Weiller, C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences*, *105*(46), 18035–18040. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.0805234105

Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fMRI study on two-word phrasal processing. *Cortex*, *96*, 105–120. doi: 10.1016/j.cortex.2017.09.002

Schoffelen, J.-M., Hultén, A., Lam, N., Marquand, A. F., Uddén, J., & Hagoort, P. (2017). Frequency-specific directed interactions in the human brain network for language. *Proceedings of the National Academy of Sciences*, *114*(30), 8083–8088. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1703155114

Schoot, L., Menenti, L., Hagoort, P., & Segaert, K. (2014). A little more conversation - The influence of communicative context on syntactic priming in brain and behavior. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00208

Segaert, K., Kempen, G., Petersson, K. M., & Hagoort, P. (2013). Syntactic priming and the lexical boost effect during sentence production and sentence comprehension: An fMRI study. *Brain and Language*, *124*(2), 174–183. doi: 10.1016/j.bandl.2012.12.003

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared Syntax in Language Production and Language Comprehension—An fMRI Study. *Cerebral Cortex (New York, NY)*, *22*(7), 1662–1670. doi: 10.1093/cercor/bhr249

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. doi: 10.1016/j.neuropsychologia.2019.107307

Shain, C., Kean, H., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., & Fedorenko, E. (2021). *'Constituent length' effects in fMRI do not provide evidence for abstract syntactic processing.* bioRxiv. (Pages: 2021.11.12.467812 Section: New Results) doi: 10.1101/2021.11.12.467812

Sharoh, D., Ruijters, L., Weber, K., Norris, D. G., & Hagoort, P. (submitted). Sentence-level meaning and compositionality in a left fronto-temporal network.

Sharoh, D., van Mourik, T., Bains, L. J., Segaert, K., Weber, K., Hagoort, P., & Norris, D. G. (2019). Laminar specific fMRI reveals directed interactions in distributed networks during language processing. *Proceedings of the National Academy of Sciences*, *116*(42), 21185–21190. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1907858116

Shekari, E., & Nozari, N. (2022). *A review of the anatomy and function of the white matter tracts in language production and comprehension.* PsyArXiv. doi: 10.31234/osf.io/kwf4r

Shetreet, E., Friedmann, N., & Hadar, U. (2009). An fMRI study of syntactic layers: Sentential and lexical aspects of embedding. *NeuroImage*, *48*(4), 707–716. doi: 10.1016/j.neuroimage.2009.07.001

Sierpowska, J., Bryant, K. L., Janssen, N., Blazquez Freches, G., Römkens, M., Mangnus, M., . . . Piai, V. (2022). Comparing human and chimpanzee temporal lobe neuroanatomy reveals modifications to human language hubs beyond the frontotemporal arcuate fasciculus. *Proceedings of the National Academy of Sciences*, *119*(28), e2118295119. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2118295119

Skeide, M. A., Brauer, J., & Friederici, A. D. (2016). Brain Functional and Structural Predictors of Language Performance. *Cerebral Cortex*, *26*(5), 2127–2139. doi: 10.1093/cercor/bhv042

Slobin, D. I. (1987). Thinking for speaking. In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society* (Vol. 13, pp. 435–445).

Smith, M., & Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition*, *73*(3), 205–246. doi: 10.1016/S0010-0277(99)00053-0

Smith, Z. (2001). *White teeth*. Penguin.

Snijders, T. M., Petersson, K. M., & Hagoort, P. (2010). Effective connectivity of cortical and subcortical regions during unification of sentence structure. *NeuroImage*, *52*(4), 1633–1644. doi: 10.1016/j.neuroimage.2010.05.035

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J. A., Petersson, K. M., & Hagoort, P. (2009). Retrieval and Unification of Syntactic Structure in Sentence Comprehension: an fMRI Study Using Word-Category Ambiguity. *Cerebral Cortex*, *19*(7), 1493–1503. doi: 10.1093/cercor/bhn187

Stanojević, M., Bhattasali, S., Dunagan, D., Campanelli, L., Steedman, M., Brennan, J., & Hale, J. (2021). Modeling Incremental Language Comprehension in the Brain with Combinatory Categorial Grammar. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 23–38). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.3

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, *65*, 69–82. doi: 10.1016/j.neuroimage.2012.09.063

Stirnberg, R., Huijbers, W., Brenner, D., Poser, B. A., Breteler, M., & Stöcker, T. (2017). Rapid whole-brain resting-state fMRI at 3 T: Efficiency-optimized three-dimensional EPI versus repetition time-matched simultaneous-multi-slice EPI. *NeuroImage*, *163*, 81–92. doi: 10.1016/j.neuroimage.2017.08.031

Takashima, A., Konopka, A., Meyer, A., Hagoort, P., & Weber, K. (2020). Speaking in the Brain: The Interaction between Words and Syntax in Sentence Production. *Journal of Cognitive Neuroscience*, *32*(8), 1466–1483. (Publisher: MIT Press) doi: 10.1162/jocn_a_01563

Thothathiri, M. (2018). Statistical experience and individual cognitive differences modulate neural activity during sentence production. *Brain and Language*, *183*, 47–53. doi: 10.1016/j.bandl.2018.06.005

Thothathiri, M., & Rattinger, M. (2015). Ventral and dorsal streams for choosing word order during sentence production. *Proceedings of the National Academy of Sciences*, *112*(50), 15456–15461. (Publisher: National Academy of Sciences Section: Biological Sciences) doi: 10.1073/pnas.1514711112

Tree, J. E. F. (1995). The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, *34*(6), 709–738. doi: 10.1006/jmla.1995.1032

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. doi: 10.1016/j.bandl.2016.08.004

Turken, A. U., & Dronkers, N. F. (2011). The Neural Architecture of the Language Comprehension Network: Converging Evidence from Lesion and Connectivity Analyses. *Frontiers in Systems Neuroscience*, *5*. (Publisher: Frontiers) doi: 10.3389/fnsys.2011.00001

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions* *on Medical Imaging*, *29*(6), 1310–1320. doi: 10.1109/TMI.2010.2046908

Tyler, L. K., & Marslen-Wilson, W. (2008). Fronto-temporal brain systems supporting spoken language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1037–1054. doi: 10.1098/rstb.2007.2158

Uddén, J., Hultén, A., Schoffelen, J.-M., Lam, N., Harbusch, K., van den Bosch, A., . . . Hagoort, P. (2022). Supramodal Sentence Processing in the Human Brain: fMRI Evidence for the Influence of Syntactic Complexity in More Than 200 Participants. *Neurobiology of Language*, 1–24. doi: 10.1162/nol_a_00076

van de Velde, M., & Meyer, A. S. (2014). Syntactic flexibility and planning scope: the effect of verb bias on advance planning during sentence recall. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.01174

van de Velde, M., Meyer, A. S., & Konopka, A. E. (2014). Message formulation and structural assembly: Describing "easy" and "hard" events with preferred and dispreferred syntactic structures. *Journal of Memory and Language*, *71*(1), 124–144. doi: 10.1016/j.jml.2013.11.001

Walenski, M., Europa, E., Caplan, D., & Thompson, C. K. (2019). Neural networks for sentence comprehension and production: An ALE-based meta-analysis of neuroimaging studies. *Human Brain Mapping*, *40*(8), 2275–2304. (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.24523) doi: 10.1002/hbm.24523

Wang, J., Cherkassky, V. L., Yang, Y., Chang, K.-m. K., Vargas, R., Diana, N., & Just, M. A. (2016). Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. *Cognitive Neuropsychology*, *33*(3-4), 257–264. (Publisher: Routledge _eprint: https://doi.org/10.1080/02643294.2016.1182480) doi: 10.1080/02643294.2016.1182480

Watkins, K. E., & Jenkinson, N. (2016). Chapter 8 - The Anatomy of the Basal Ganglia. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 85–94). San Diego: Academic Press. doi: 10.1016/B978-0-12-407794-2.00008-0

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE*, *9*(11), e112575. doi: 10.1371/journal.pone.0112575

Wernicke, C. (1892). *Gesammelte Aufsätze und kritische Referate zur Pathologie des Nervensystems.* (Vol. Facsimile reproduction (2006)). Saarbrücken: VDM Verlag.

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. doi: 10.1093/cercor/bhv075

Willems, R. M., & Gerven, M. A. J. v. (2018). New FMRI Methods for the Study of Language. *The Oxford Handbook of Psycholinguistics*. doi: 10.1093/oxfordhb/9780198786825.013.42

Wilson, S. M. (2017). Lesion-symptom mapping in the study of spoken language understanding. *Language, Cognition and Neuroscience*, *32*(7), 891–899. (Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2016.1248984) doi: 10.1080/23273798.2016.1248984

Wilson, S. M., DeMarco, A. T., Henry, M. L., Gesierich, B., Babiak, M., Mandelli, M. L., … Gorno-Tempini, M. L. (2013). What Role Does the Anterior Temporal Lobe Play in Sentence-level Processing? Neural Correlates of Syntactic Processing in Semantic Variant Primary Progressive Aphasia. *Journal of Cognitive Neuroscience*, *26*(5), 970–985. (Publisher: MIT Press) doi: 10.1162/jocn_a_00550

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6

Wu, D. H., Waller, S., & Chatterjee, A. (2007). The Functional Neuroanatomy of Thematic Role and Locative Relational Knowledge. *Journal of Cognitive Neuroscience*, *19*(9), 1542–1555. (Conference Name: Journal of Cognitive Neuroscience) doi: 10.1162/jocn.2007.19.9.1542

Wundt, W. (1970). The psychology of the sentence. In A. Blumenthal (Ed.), *Language and psychology: Historical aspects of psycholinguistics* (pp. 20–31). New York, NY: Wiley.

Xiang, H.-D., Fonteijn, H. M., Norris, D. G., & Hagoort, P. (2010). Topographical Functional Connectivity Pattern in the Perisylvian Language Networks. *Cerebral Cortex*, *20*(3), 549–560. doi: 10.1093/cercor/bhp119

Xu, Y., Wang, X., Wang, X., Men, W., Gao, J.-H., & Bi, Y. (2018). Doctor, Teacher, and Stethoscope: Neural Representation of Different Types of

Semantic Relations. *Journal of Neuroscience*, *38*(13), 3303–3317. (Publisher: Society for Neuroscience Section: Research Articles) doi: 10.1523/JNEUROSCI.2562-17.2018

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. (Number: 8 Publisher: Nature Publishing Group) doi: 10.1038/nmeth.1635

Zaccarella, E., Meyer, L., Makuuchi, M., & Friederici, A. D. (2017). Building by Syntax: The Neural Basis of Minimal Linguistic Structures. *Cerebral Cortex*, *27*(1), 411–421. (Publisher: Oxford Academic) doi: 10.1093/cercor/bhv234

Zaccarella, E., Schell, M., & Friederici, A. D. (2017). Reviewing the functional basis of the syntactic Merge mechanism for language: A coordinate-based activation likelihood estimation meta-analysis. *Neuroscience & Biobehavioral Reviews*, *80*, 646–656. doi: 10.1016/j.neubiorev.2017.06.011

Zadbood, Chen, Leong, Norman, & Hasson. (2018). *Sherlock_merlin. OpenNeuro. [Dataset].*

Zadbood, A., Chen, J., Leong, Y., Norman, K., & Hasson, U. (2017). How We Transmit Memories to Other Brains: Constructing Shared Neural Representations Via Communication. *Cerebral Cortex*, *27*(10), 4988–5000. doi: 10.1093/cercor/bhx202

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. doi: 10.1109/42.906424

Nederlandse Samenvatting

Wanneer we taal gebruiken, weten we dat onze hersenen aan het werk zijn. Het onderzoek naar de hersensystemen die taal mogelijk maken heeft zich lange tijd gericht op de relatie tussen herseninfarcten en taalstoornissen. De komst van functionele magnetic resonance imaging (fMRI) drie decennia geleden heeft het mogelijk gemaakt om hersenfuncties te bestuderen tijdens taalgebruik in gezonde mensen. Sindsdien hebben veel onderzoeken een verband gevonden tussen hersenactiviteit en specifieke onderdelen van taalgebruik. Echter, taalgebruik is vooral bestudeerd vanuit het oogpunt van receptieve verwerking (dat wil zeggen, tijdens het luisteren of lezen). Studies naar taalproductie zijn daarentegen zeldzamer, omdat fMRI erg gevoelig is voor beweging in de scanner, zoals de bewegingen van de kaak tijdens het spreken. Bovendien is het moeilijker om experimenten uit te voeren waarbij proefpersonen moeten praten, omdat het bijna onmogelijk is om experimentele controle te hebben over wat de proefpersonen gaan zeggen, zonder hen te vertellen wat ze precies moeten zeggen (bijvoorbeeld door hen te laten lezen). Het is belangrijk dat ons begrip van hersenfuncties tijdens taalverwerking ook gebaseerd is op het spreken, omdat taal ook in deze vorm wordt uitgewisseld: als je iets hoort, betekent dit dat iemand het heeft uitgesproken. Daarnaast maken taalproductie en taalbegrip gebruik van verschillende perifere systemen (zoals het spraakkanaal en de oren), terwijl ze wel dezelfde woorden gebruiken. Daarom focust dit proefschrift zich op hersenactiviteit tijdens de productieve kant van taal, om te begrijpen in hoeverre onze kennis over taalbegrip kan worden uitgebreid naar taalproductie. Alle onderzoeken die in dit proefschrift worden gepresenteerd zijn gericht op taalverwerking op zinsniveau, dat wil zeggen, taalverwerking tijdens het luisteren naar of produceren van zinnen.

Hoofdstuk 2 heeft zich gericht op het afbakenen van de set hersengebieden die actief zijn tijdens de productie en het begrip van aaneengesloten woordreeksen die variëren in lengte. We gingen ervanuit dat syntactische verwerking sterker betrokken is bij woordenreeksen van die langer zijn. Concreet werd hersenactiviteit tijdens het produceren van of luisteren naar “denken, springen, de jongen, het meisje” vergeleken met “het meisje denkt dat de jongen springt”. In deze reeksen worden dezelfde woorden geproduceerd, maar in verschillende syntactische structuren. We vonden dat hetzelfde netwerk van hersengebieden in de frontale en temporale kwabben van de linkerhersenhelft betrokken was bij het verwerken van de complexere zinnen in vergelijking met de eenvoudigere, voor zowel taalproductie als taalbegrip. Daarnaast vonden we dat sommige van deze hersengebieden gevoeliger waren voor deze vergelijking tijdens productie. Ten slotte waren er enkele gebieden die alleen reageerden op productie of begrip, zoals de motorische cortex bij productie en de auditieve cortex bij begrip. Deze studie toont aan dat taalverwerking tijdens het spreken en luisteren in dezelfde hersengebieden kan plaatsvinden.

In Hoofdstuk 3 onderzochten we of de hersenconnectiviteit verandert tijdens het produceren van en luisteren naar dezelfde woordreeksen als in Hoofdstuk 2. We vroegen ons met name af of twee relevante gebieden uit de vorige studie, de linker inferieure frontale gyrus en de linker posterieure temporale gyrus, op dezelfde manier aan elkaar verbonden waren tijdens productie en begrip. We ontdekten dat de connectiviteit tussen deze gebieden toenam voor de complexere zinnen, wat suggereert dat de interactie tussen deze hersengebieden toenam bij het samenvoegen van meerdere woorden. Er was ook een iets verhoogde connectiviteit tijdens spreken ten opzichte van luisteren, wat overeenkomt met het resultaat uit Hoofdstuk 2, waarin sommige gebieden sterker reageerden op productie dan op begrip.

Deze studies zijn gebaseerd op een experiment dat ver af staat van de manier waarop we normaal spreken. Proefpersonen werden gevraagd om de woordreeksen te produceren op basis van visuele displays die aangaven welke woorden ze moesten zeggen. Echter, als we spreken, zijn we normaal gesproken vrij om te beslissen wat we zeggen en hoe. Daarom wilden we begrijpen of hersenactiviteit tijdens syntactische verwerking zou verschillen in situaties waarin deelnemers vrijuit mogen spreken. In Hoofdstuk 4 hebben we een openlijk gedeelde fMRI-dataset gebruikt om deze vraag te beantwoorden. We ontdekten dat de hersengebieden in de frontale en temporale kwabben, waar we ons eerder ook op concentreerden, reageren op syntactische verwerking tijdens zowel productie als begrip. Daarnaast liet deze studie zien dat de timing van het opbouwen van syntactische structuur verschilt tussen productie en begrip. Tijdens het spreken hebben we de neiging om de structuur van de zin eerder op te bouwen, terwijl we tijdens het luisteren de neiging hebben om te wachten op de input.

Tot nu toe heeft het proefschrift zich gericht op het beschrijven van hersenactiviteit tijdens syntactische verwerking. Hoofdstuk 5, daarentegen, onderzocht hoe de hersenen betekenis coderen. Zo is de zin "de muzikant schopte de atleet"duidelijk anders dan de zin "de atleet schopte de muzikant", terwijl de woorden in de twee zinnen identiek zijn. In de eerste zin is de muzikant degene die de actie uitvoert (de agens), terwijl de atleet de actie ondergaat (de patiëns). Zinsbetekenis kan worden beschreven aan de hand van zowel de personen en handelingen in een gebeurtenis (hier zijn dat de muzikant en de atleet in een schoppende gebeurtenis) als hun rol in die gebeurtenis (agens, patiëns). In Hoofdstuk 5 produceerden proefpersonen zinnen die een vergelijkbare structuur hadden maar verschilden in de specifieke woorden waaruit de gebeurtenis bestond die door de zin wordt beschreven. Vervolgens bekeken we of de hersenactiviteit verschilde wanneer de gebeurtenis een muzikant of een atleet bevatte, wanneer de muzikant of de atleet de agens was, en wanneer het de muzikant of de atleet was die specifiek aan het schoppen was. We ontdekten dat de betekenis van de zin op deze verschillende niveaus gecodeerd werd in een netwerk van frontotemporale gebieden in de linkerhersenhelft. Deze studie toont aan dat de hersenen zinsbetekenis op verschillende manieren coderen tijdens productie, zoals ook al is aangetoond voor begrip in voorgaand onderzoek. Het wijst echter ook op een aantal verschillen in de patronen van hersenactiviteit tussen productie en begrip.

Samengevat, in dit proefschrift heb ik het patroon van hersenreacties op taalverwerking tijdens zinsproductie en -begrip beschreven. Ik heb me geconcentreerd op de hersenactiviteit tijdens syntactische verwerking, zowel bij en taak als bij spontane productie, en op de hersenpatronen van codering van zinsbetekenis tijdens het spreken. Hersenactiviteit werd gemeten op basis van de hoeveelheid bloedstroom in 3D pixels met zijden van 2.5 millimeter, met een vertraging van enkele seconden. Toekomstige studies zullen daarom moeten bepalen in hoeverre deze resultaten worden bevestigd in studies die gebruik maken van methodes met een betere temporele resolutie. Op deze schaal suggereren de resultaten in ieder geval dat er verschillen zijn in de hersengebieden die een rol spelen bij taalproductie en -begrip, zowel ruimtelijk als in de timing van reacties. Die verschillen kunnen te maken hebben met de verschillende doelen en vereisten van productie en begrip. Maar over het algemeen laten de studies in dit proefschrift zien dat dezelfde hersengebieden betrokken zijn bij het produceren en begrijpen van taal.

English Summary

When we use language, we know our brain is doing something about it. The investigation of the brain systems that allow for language focused for a long time on the relationship between stroke and linguistic impairment. The advent of functional magnetic resonance imaging (fMRI) three decades ago allowed for studying brain function during healthy language processing. Since then, many studies found a link between brain activity and specific linguistic processing. However, language processing has been mostly studied from a perceptive point of view (i.e. while listening or reading). Studies with speaking are instead more rare, because fMRI is very sensitive to movement in the scanner, such as the movement of the jaw elicited by speaking. In addition, it is harder to run experiments where participants speak, because it is almost impossible to have experimental control over what participants will say, without telling them exactly what to say (for example by reading). It is important that our understanding of brain function during language processing is also based on speaking, because language is exchanged in this form as well: if you're hearing something, it means someone spoke it. Additionally, language production and comprehension interface with different peripheral systems (vocal tract and ears), while using the same words. Therefore, the aim of this thesis was to focus on brain activity during the productive side of language, and to understand to what extent what we learn about language comprehension can be extended to language production. All the studies presented in this thesis focused on language processing at the sentence level, that is, language processing while hearing or saying sentences.

Chapter 2 focused on delineating the sets of brain regions that are active during the production and comprehension of connected strings of words of increasing length, which we assumed would engage syntactic processing to increasing extents. Concretely, brain activity while saying or hearing "to think, to jump, the boy, the girl" was compared to "the girl thinks that the boy jumps". In these sequences, the same words are said but in different syntactic structures. We found that the same network of regions in the frontal and temporal lobes was involved while processing the more complex sentences, relative to the easier ones, both in production and comprehension. Additionally, we found that in production a few of these brain regions were more sensitive to this comparison. Finally, there were some regions that responded only in production or comprehension, such as motor cortex in production, and auditory cortex in comprehension. Therefore, this study showed that linguistic processing may happen in the same brain regions while speaking and listening.

In Chapter 3, we asked whether brain connectivity changes during speaking and listening to the same word sequences as in Chapter 2. In particular, we asked whether two regions found in the previous study, the left inferior frontal gyrus and the left posterior temporal gyrus, were similarly connected during production and comprehension. Interestingly, we found that the connectivity between these regions increased with the more complex sentences, suggesting that the interaction between brain regions increased when composing multiple words together. There was also slightly increased connectivity during speaking relative to listening, which mirrors the result from Chapter 2 where a few regions responded more to production than comprehension.

These studies were based on an experiment that was distant from the way we normally speak. Participants were asked to produce the word sequences based on visual displays that indicated what words they had to say. When we speak, though, we are normally free to decide what to say and how. Therefore, we set out to understand whether brain responses to syntactic processing would differ in settings where participants are allowed to speak freely. In Chapter 4, we used an openly shared fMRI dataset to address this question. We found that the same brain regions we focused on before in the frontal and temporal lobes responded to syntactic processing during both production and comprehension. Additionally, this study showed that the timing of syntactic structure building differs between production and comprehension. During speaking we tend to build the structure of the sentence earlier on, while during listening we tend to wait for the input.

So far, the thesis focused on characterizing brain responses during syntactic processing. Chapter 5 instead explored how the brain encodes sentence meaning. For example, the sentence “the musician kicked the athlete” is clearly different from the sentence “the athlete kicked the musician”, although the words in the two sentences are identical. In the first sentence, the musician is the one doing the action (the agent), while the athlete is the undergoer (patient) of the action. Sentence meaning can be characterized by the entities that take place in an event (here, the musician and the athlete, in a kicking event), together with their role in the event (agent, patient). In Chapter 5, participants said sentences that had a very similar structure, but differed in the specific words that made up the event described by the sentence. We then asked whether brain activity differed when the event included a musician or an athlete, or when it was the musician or the athlete that was doing an action, or when it was the musician or the athlete that was specifically kicking. We found that a network of left fronto-temporal brain regions encoded the meaning of the sentence at these different levels. This study showed that the brain encodes sentence meaning in different ways during production, as was shown for comprehension in previous studies, but also indicated some differences in the patterns of brain responses between production and comprehension.

Overall, in this thesis I characterized the pattern of brain responses to linguistic processing during sentence production and comprehension. I focused on the brain responses to syntactic processing during a task and spontaneous production, and on the patterns of brain encoding of sentence meaning while speaking. Brain activity was measured based on the amount of blood flow in cubes of dimensions of 2.5mm, with a delay of a few seconds. Therefore, future studies will have to determine to what extent these results would be confirmed with better temporal resolution. At this scale, these results suggest that there are some differences in the brain responses to production and comprehension, both spatially and in the timing of responses, that may relate to the different goals and requirements of production and comprehension. In general, though, these studies indicate that the same brain regions are used for linguistic processing during speaking and listening.

Acknowledgements

These were four engaging and fun years. I am so grateful to all the people that were there throughout these years, and made it possible and exciting. I was lucky to be in the warm environment of the MPI and DCCN surrounded by brilliant people.

Peter, thank you for the opportunity to work with you in the inspiring environment of the NBL lab at the MPI. You believed in me and gave me the opportunity to learn independently, leaving freedom while not losing sight of what's important. I learnt a lot thanks to your piercing feedback on design, predictions, writing and anything work- and career related, and you've been a source of inspiration with your extensive knowledge of the field.

Markus, you came in at a moment of difficulty, both work-wise and world-wise. Your guidance during the transition to a new supervision team and new projects was invaluable. I looked forward to meeting with you every week to discuss ideas, analysis approaches, life, and of course bread. I learnt a lot from you, most importantly how to believe in myself.

Kirsten, I'm sad that we could not work together for longer. You made the start of the PhD way less stressful than it could have been by making fMRI easy and experimental design fun. Thank you also for showing me the unsweetened side of academic careers.

I would also like to thank the members of the **manuscript committee**, not only for taking the time to read and evaluate the thesis, but also for their individual contributions to this thesis given through their work.

Dan, I'm so glad we got to work together more recently. Our differences in work approaches always push me to do better, understand better, and make me appreciate the complexity of our work. I don't know if we have a career playing the accordion, but at least we are producing something together.

Joost, you showed me the way in the then unknown world of research. I first learnt the rigour and critical thinking necessary in research by working with you. Thank you so much for teaching me about EEG, prediction, language, memory and all there is to know about the academic world. I hope we'll have a chance to work together again in the future.

Michaela, it was great to meet you in the last phase. Your snacks and warm smiles were a great support in these months! **Carolin**, **Ina** and **Tildie** thank you for making the everyday business smoother with your undying efficiency.

Birgit, **Maarten**, **Iris**, **Eva**, **Marjolijn**, and **Janniek**: this thesis would not be here without all your help with coding hours of Dutch recordings and experiment set up. I really appreciate the time and accuracy with which you helped me carry out the repetitive part of these projects. **Ludy**, thank you for your help and advice with designing the cover and getting the thesis ready for print.

I would not have made it without the incredible support of the MRI team and the TG support at the DCCN. **Paul**, thanks for showing me how to use the scanners and for the fun chats during the long hours of testing. **Marcel**, you made fMRI preprocessing fun! Your work makes life easier for all of us fMRI buddies. **José**, thanks for advising on sequences and analysis, the difficulty of testing speaking participants was made easier by you.

Janice, **Savannah** and the rest of your lab, thank you for always being so ready to help and taking on hours of work to make science more open.

If these years were fun and engaging it is also thanks to all the people that were there, always ready to give feedback, ask the right questions and just share everyday difficulties and happiness. Thanks to all the NBL PhDs that were there biweekly to share the complexity of the work. **René**, **Marvin**, **Sophie**, **Julia**, **Ksenija**, **Micha**, **Rowan**, **Cas**, **Alessio**, **Nienke**, **Teun**, **Margot**, **Cecilia** and **Sara**, you were all individually a source of inspiration. I'm glad I could share this journey with so many brilliant people.

Cas, we have shared this journey from day one. You have been my linguistics guide, you were always ready to listen to whatever was going wrong that day, always there to give feedback and just chat our way into the afternoon. It was incredibly enriching to have the opportunity to discuss ideas and findings spontaneously in the office. It will be strange when sooner or later I won't be able to turn around and discuss the latest twitter storm with you. **Natascha**, we've been going through so much together, from the early days of master's courses to event organizations! It's been comforting to always have someone to talk to that was going through similar steps at work. And you got me to like dogs! **Adrian**, thanks for being there through all the struggles and the fun, for sharing the interest and the frustration with courses, deadlines and brains. **Alessio**, you never stop challenging us and forcing us to face our compromises. Cocco, il caos che porti nel corridoio dell'NBL e' una boccata d'aria fresca! **Ksenija**, I'm so glad there was at least one other person I could share all the fMRI scheduling, buddying, and analysis frustration with!

The NBL group was an incredible place where to grow. **Ashley**, by now you're my official Friday drinks buddy. I think we can proudly say we made the NBL lab an (almost?) social environment. Thanks for going along with my attempts at getting people together and for all the nice chats about science and life over the years. **Ellie**, you've been through it all! It's been fun to work together lately, and I hope we'll be able to do more conferencing together soon. **Maarten**, you always made sure I was running experiments in the best possible way. **Harmut**, I got to know you during the corona Friday drinks, and I learnt a lot from our never ending Friday night zoom discussions. **Ambra**, your motivation is contagious! **Jakub**, thanks for all the nice discussions. **Eva** and **Iris**, thanks for all the lunches! **Atsuko**, unfortunately it feels good to know that I am not the only one fighting with fmriprep. **Yingying**, can't wait for our next match! **Jana**, your passion for running almost makes me want to go for a run now. **Francesca**, thanks for sharing your experience working in so many different countries. **Anny**, it was about time someone stopped by offices for a chat! **Linda**, **Noor** and **Sara**, thanks for adopting me during SNL!

Xiaochen, thanks for always being interested and for the nice mentoring chats. **Ryan**, it was so much fun to discuss research and cooking with you. **Hugo**, I'll never stop admiring your patience in explaining the latest analytical approach. **Marco**, discussing academic careers with you has been a wake-up call.

The MPI is a special place not just for the incredible facilities and opportunities to come in touch with the latest research, but also for the PhD environment. Organising the IMPRS Career Event taught me so much about the non-academic world and about working in a team. Thank you **Kevin**, **Natascha**, **Adrian**, **Fenja** for this experience. If organising the IMPRS Conference was even more of a challenge, it was also incredibly rewarding. **Kevin**, **Cecilia**, **Koen**, **Giulio**, **Figen**, **Rong**, **Goekberk**, **Sara**, we did a great job!

Also huge thanks to all the IMPRS members that made working at the MPI and Friday evenings so great: **Candice**, **Teun**, **Orhun**, **Marlijn**, **Naomi**, **Marlou**, **Cecilia**, **Sophie**, **Melis**, **Filiz**, **Taylor**, **Rong**, **Ine**, **Christophoros**...

Kevin, you are the one that made this all possible.

If I was surrounded by amazing people at the MPI, the DCCN was no less. Thanks to **Flavia**, **Alessio P.**, **Nishant**, **Mahdad**, **Alejandro**, **Raimon**, for sharing on the daily work in the flex room on the second floor. I also know I might have starved already without Betty & Mora's food supply. **Betty**, ci manchi! **Mora**, come avrei fatto senza le nostre chiacchiere di inizio giornata. E **Lorenzo**, senza le chiacchierate pomeridiane e i tuoi gin tonic! **Max**, you're replacing both of them with your high wine and food knowledge! You all didn't just comfort me with good food, but also with the endless dinners and evenings together, and made these years that much more pleasant.

I also want to thank all the people that kept me sane over the years, and that made these years so cheerful! **Marti**, you were there from the start, sharing the tough parts and the happy parts. I'm so grateful we could do this journey together, with our walks, cocktails and sushi nights. **Michi**, thanks for painting the best cover I could have wished for. But mostly thanks for being a great friend over the years. **Michi** and **Marti**, you've been my home in the expat world.

Randi, I don't know how you could stand me over the years. Thank you for being so patient and for making me see reason always, thank you for the company in the pandemic times, and for sharing so much of this journey.

Jill, thankfully I found a book and yoga buddy in you!

Socia, you were always there even from afar.

Vale, **Vale** e **Agne**, thanks for being there every time I'm back home.

Mrudula, **Giulia**, **Fran**, **Giacomo**, **Cemre**, **Sofi**, you made me feel at home in the first few years in Nijmegen.

Silvia <3 finally a late lunch buddy at the MPI! **Larraitz**, **Elena**, **Nora**, **Ioanna**, **Elie**, **Achille**, **Natalie**, **Tiziana**, **Simen**, **Meri**, **Noor**, thanks for making the post-corona world so much fun!

I also wouldn't be the person I am today without the constant support of the climbing crew. **Harry**, **Floris**, **Jeroen**, **Lenno**, and all the other climbers I met along the way, you helped me go home at night with an empty mind and a smile.

I wouldn't have made it without the constant support of my parents and encouragement in undertaking new challenges. You always pushed me to follow my curiosity and it's made me who I am today. Thanks to my grandmothers, for being great role models in their individual ways. **Luca**, thank you for showing me alternative ways to look at the world. **Gabri**, you've been the lighthouse that shows me the way during tough and happy times.

Curriculum Vitae

Photo credit: Kevin Lam

Laura Giglio was born in Ivrea, a small town in the north of Italy, in 1993. After high school, she moved to Glasgow (United Kingdom), where she obtained a Bachelor's degree in Neuroscience from the University of Glasgow in 2016. After that, she moved to Nijmegen in the Netherlands, where she completed her Master's degree in Cognitive Neuroscience at Radboud University in 2018. For her Master's thesis, she worked on the relationship between predictability in language and memory. In September 2018, she started her PhD in the Neurobiology of Language Department of the Max Planck Institute for Psycholinguistics, in affiliation with the Donders Centre for Cognitive Neuroimaging. Under the supervision of Peter Hagoort, Kirsten Weber and Markus Ostarek, and in close collaboration with Daniel Sharoh, she investigated the brain networks for language production and comprehension. She is currently working as a postdoctoral researcher in the same lab.

Publications

Giglio, L., Ostarek, M., Weber, K. & Hagoort, P. (2022). Commonalities and Asymmetries in the Neurobiological Infrastructure for Language Production and Comprehension. *Cerebral Cortex*, *32*(7), 1405–1418. https://doi.org/10.1093/cercor/bhab287