Noun and Verb Decoding Gives Insight into the Spatio-Temporal Dynamics of Sentential Meaning Construction

Hiroyoshi Yamasaki
hiroyoshi.yamasaki@etu.univ-amu.fr[1]

[1]Université Aix-Marseille, UFR ALLSH, Master 2 Science Cognitives, année universitaire 2021-2022

Travail d'Étude et de Recherche réalisé sous la direction de
Dr. F.-Xavier Alario, *Laboratoire de Psychologie Cognitive*
May 31, 2022

ILCB Institute of language, communication and the brain
Aix*Marseille université
Laboratoire de Psychologie Cognitive
cnrs

experiment. To minimise the effort we decided to use the stimulus set used in previous study (Strijkers et al. 2019) as the stimuli for this pilot. I owe much to Jean-Michel Badier for the preparation of the script used to present the stimuli at the MEG lab.

Our initial aim was to study typing using a special keyboard. Svetlana Pinet at BCBL has done pilot studies first on EEG and then MEG to this effect (personal communication). We intended to try something along those lines using fMRI compatible keyboard (thanks to Bruno Nazarian). Unfortunately, the keyboard generated too much artefact when moved. We then moved to speech production instead. This turned out to be more productive since much of the artefacts can be removed with ICA. This reorientation of the project was an important opportunity to learn how to manage in face of technical difficulties.

Last but not least, this internship has taught me how to cope with time pressure. It is clear that the project was far from complete and there are plenty more to be done (in particular, it was in last week that I began to have real results on the cluster). This has taught me to assess what's feasible within the time allotted (thanks to advice by my supervisor F.-Xavier Alario).

To conclude, this internship has taught me enough for me to attempt to embark on more scientific research in the near future.

Abstract

The final goal of language comprehension is to construct a holistic representation of the sentence. In this study we compared two models of sentence comprehension. Friederici's model (2011) predicts that this process takes place in the posterior part of the temporal lobe around 600 ms. Pylkkänen's model (Pylkkänen, 2020), by contrast, predicts the involvement of the ventromedial prefrontal cortex (vmPFC) at around 400 ms. We evaluated these predictions by performing multivariate pattern analysis (MVPA) on MEG data. The current study consists of two parts: a pilot study and a dataset analysis. In the pilot study a participant performed the overt reading task in MEG. The participant read either a noun phrase (e.g. 'ma méhode') or a verb phrase (e.g. 'je ferai'). The data was then classified using a machine learning algorithm into two conditions. The results revealed that both the activities predicted by Friederici and Pylkkänen can be observed. In addition, we observed a widely distributed pattern across temporo-parietal regions. We suggest that network based analysis in the future may be successful at explaining such observations. In the dataset analysis we observed a less clear picture. It became clear that there are several technical issues to solve before the analysis pipeline can be scaled up. We suggest possible paths to follow in the future studies.

Le but final de la compréhension du langage est de construire une représentation holistique de la phrase. Dans cette étude nous avons comparé deux modèles de compréhension du langage. Le modèle de Friederici (2011) prédit que cela se passe dans la partie postérieure du lobe temporal à environ 600 ms. Le modèle de Pylkkänen (2020), en contrast, prédit l'implication du cortex ventro-médial préfrontal à environ 400 ms. Nous avons évalué ces deux prédictions en faisant une analyse multivariée (multivariate pattern analysis, MVPA) sur les données (MEG). La présente étude comprends deux parties : une étude de pilote et une analyse d'un jeu de données. Dans l'étude de pilote un participant a lu à voix dans la MEG. Le participant a lu soit un syntagme nominal (ex. 'ma méthode') soit un syntagme verbal (ex. 'je ferai'). Les données étaient classifiées selon la condition avec un algorithme d'apprentissage automatique. Les résultats montrent que les deux activités prédites par Friederici et Pylkkänen peuvent être observées. En plus, nous avons observé une activité distribuée dans la région temporo-parietal. Nous suggérons que l'analyse basée sur les réseaux peut avoir plus de succès dans l'avenir. Dans l'analyse du jeu de données les résultats étaient moins clairs. Il apparaît que plusieurs problèmes techniques doivent être résolus pour appliquer le même pipeline aux données plus grandes. Nous suggérons les chemins à suivre pour les études dans l'avenir.

Keywords: sentence comprehension, MEG, MVPA, vmPFC, P600, semantic combinatorial effect

Contents

1 Introduction

The ability to process individual words and turn them into a coherent semantic representation is an important part of language processing. Decades of study by linguists, psychologists and cognitive neuroscientists have enabled us to amass a wealth of data in this domain. Neuroimaging studies have shown that in order to understand language a large perisylvian network is necessary (Fedorenko et al., 2011; Fedorenko & Thompson-Schill, 2014). The act of parsing a sentence involves many different process including but not restricted to lexical retrieval/memory (Snijders et al., 2009; Hagoort, 2005, 2013), local syntactic processes (Friederici, 2011), argument structure analysis ('who did what to whom', Pylkkänen, 2019, 2020; Williams et al., 2017), hierarchical structure building ('Merge' Zaccarella et al., 2017, Friederici, 2017 chapter 1, or 'Unification' Hagoort, 2005, 2013). But arguably all these processes exist so that we can finally reconstruct the meaning uttered by others. This final step in language comprehension will be the main topic of this study.

While there are many models of language comprehension, we will focus on the models by Friederici (2002, 2011) and Pylkkänen (2019, 2020) for two reasons. First, they both provide detailed spatio-temporal dynamics. This is essential especially if we wish to understand the 'cognitive algorithm' of language comprehension (Marr, 2010; King et al., 2018). Second, they provide two very different views on where in the brain the final representation is generated, thus comparing the two models will be most informative for understanding this particular process.

Friederici, who bases her model on combination of ERP studies and fMRI studies, proposes that the final stage of language comprehension (syntactico-semantic 'integration' process, Friederici, 2002, 2011) takes place around 600 ms (auditory) post-stimulus onset (Friederici, 2002, 2011, based on Kwon et al., 2005; Service et al., 2007). According to her, this process involves syntactic reanalysis and repair and is associated with centro-parietal positivity (P600, Friederici, 2011). This conclusion is motivated by studies in which the participants had to process temporarily ambiguous sentences (so called 'garden path' sentences, Osterhout et al., 1994) and sentences requiring repair due to syntactic violations (Osterhout & Holcomb, 1992; Osterhout et al., 1994). For our purposes, this stage is interesting because it is also implicated in sentence-level semantic violations (Hoeks et al., 2004; Kim & Osterhout, 2005; Kolk et al., 2003; Friederici, 2011) in which semantically incongruous sentences like '*The hearty meal was devouring*' had to be processed (Friederici, 2011). Processing of these sentences causes 'semantic P600' (Friederici, 2011, p. 1382). Within her model, the source of this P600 effect is localised somewhere in the middle temporal gyrus, posterior portion of the temporal cortex as well as basal ganglia (Friederici, 2011 based on Frisch et al., 2003. Thus the process is thought to take place in the posterior part of the brain (see fig. 1, left).

In contrast to Friederici, Pylkkänen considers *ventromedial prefrontal cortex* (vmPFC,

see fig. 1, right) to be the main seat of sentence level semantics (Pylkkänen, 2019, 2020). This area was initially implicated in a study of 'coercion' studies. Coercion is a process by which a sentence takes on additional information not explicitly expressed by syntax (Brennan & Pylkkänen, 2008; Pylkkänen et al., 2009). She gives the example '*The little boy finally finished his pasta*' in which the word 'finished' signals the end of an event while 'pasta' stands for 'act of eating pasta' (Pylkkänen, 2020, p. 2). In the study dealing with this coercion process (Pylkkänen et al., 2009; Brennan & Pylkkänen, 2008) she discovered an increased activity in vmPFC at around 400 ms post-stimulus onset. Subsequent studies (so called 'red-boat studies') in which participants had to construct two word phrases consisting of an adjective and a noun have implicated this vmPFC together with left anterior temporal lobe (LATL) at around 400 ms and 200 ms respectively (Pylkkänen, 2020). She thus speculates vmPFC to be a 'representational site for the end product of composition in comprehension' (Pylkkänen, 2020, p. 3).

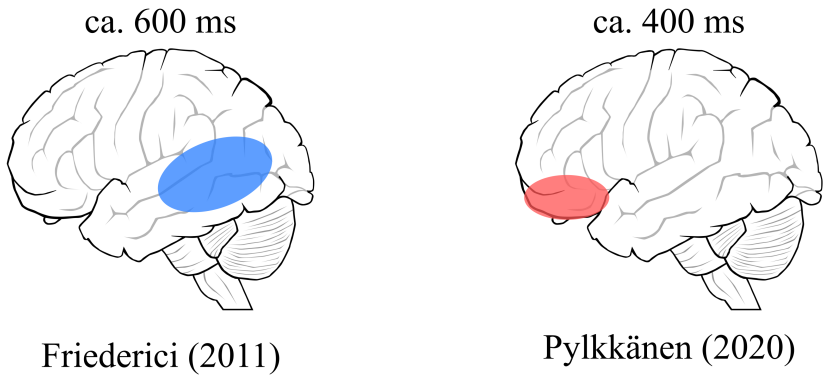

Figure 1: Two models considered in this study. (Left) Model by Friederici (2011). It predicts the involvement of the posterior portion of the temporal lobe at around 600 ms for the final 'integration' process. (Right) Model by Pylkkänen (2020). It predicts the involvement of the vmPFC at around 400 ms for the 'final semantic process'.

Thus we have two distinct views regarding the final stage of language comprehension. On the one hand Friederici considers the P600 effect localised to the posterior part of the brain to be responsible for the syntactico-semantic integration process (Friederici, 2011). On the other hand Pylkkänen considers vmPFC activation at around 400 ms post-stimulus onset to be related to sentence level holistic semantic process (Pylkkänen, 2019, 2020). While two positions are motivated by slightly different tasks (semantic incongruity vs. coercion/semantic combinatorial processes), nonetheless we would expect those tasks to involve sentence level semantic processes. Our aim in this study is to investigate this puzzle and conduct preliminary data collection and analysis to this end.

1.1 Current study

In order to study sentence level comprehension we will use magnetoencephalography (MEG) data. The motivation for using MEG is the fact that compared to other imaging

modalities (e.g. fMRI and EEG) MEG offers both high temporal and spatial resolutions (Baillet, 2017). In order to detect the final product of sentence level representation we will apply multivariate pattern analysis (MVPA) on this MEG data (for review see, e.g. Haxby et al., 2014, in MEG modality in particular King & Dehaene, 2014). More specifically, we will 'decode' the semantic content by classifying nouns against verbs using a machine learning algorithm. There are several motivations for using this approach. First, nouns and verbs are the most salient linguistic categories and are very wide spread across languages (Crepaldi et al., 2011, although an alternative view on their categorisation exists, see Croft, 2001, Chapter 2). Second, we know from a previous study that nouns and verbs are decodable in fMRI (Elli et al., 2019) and in MEG (Arana et al., 2021). Thus if we were to access the 'end product' of semantic composition our best bet is to use this distinction. It is, however, important to note that when we say 'noun versus verb' we actually mean 'noun versus verb and/or object versus action'. This is because these two grammatical categories are highly correlated with their prototypical semantic properties (Vigliocco et al., 2011). According to Vigliocco and colleagues (2011), for example, noun versus verb distinction is in fact less important than object versus action distinction. For our purposes, however, this ambiguity between syntactic (grammatical) and semantic aspects is less relevant. This is because (at least in our experiment) the stimuli are nouns and verbs and the the end product is some neural activity that correlates with either object or action. It is nonetheless worth keeping this potential confound in mind when we interpret the syntax-semantics interface.

The rest of the paper is organised as follows. The first section describes the pilot study we conducted with MEG and MVP analysis of the data. The second section describes the MVP analysis we conducted on an existing large dataset. The final section is dedicated to the assessment of the results and discussion of approaches to take in the future explorations.

2 Pilot Experiment

The pilot experiment was conducted with two distinct aims in mind. First, we wished to collect data under controlled experimental settings in order to assess how well we can decode nouns against verbs. Second, we wished to assess the possibility of moving to production study later on using MEG which (due to various technical difficulties) is less studied than comprehension. To reconcile these two aims, we used a simple word reading paradigm of noun phrases (NPs) and verb phrases (VPs) presented visually. The reason for using NPs and VPs instead of isolated words is based on the previous studies which indicated that category specific effects tended to disappear once words are deprived of syntactic context (Vigliocco et al., 2011; Siri et al., 2008).

The collected MEG data was analysed using both regions of interest (ROIs) and whole brain analysis approaches. The analysis consisted of classifying the neural data in a

particular cortical area at a particular time as either a noun or a verb. Since the stimuli are simple NPs and VPs, whose meanings are either mostly static or dynamic respectively, we expect that once the phrasal structure has been parsed and its semantic content is constructed we would observe an increase in decodability for the two conditions. The two models we discussed above provided two distinct hypotheses. According to Friederici's model we should observe above chance decoding performance at around 600 ms post stimulus onset in the posterior part of temporal lobe and its vicinity (Friederici, 2011, see fig. 1, left). According to Pylkkänen's model we should observe an above chance decoding performance at around 400 ms in vmPFC (Pylkkänen, 2019, 2020, see fig. 1, right).

2.1 Materials and Methods

2.1.1 Participants

The pilot study was conducted with a single native speaker of French whose anatomical MRI was made available for our purposes.

2.1.2 Stimuli

The stimuli used in this study was adapted from Strijkers et al. (2019). The stimuli used in this study were grouped into noun phrases and verb phrases and they were already controlled for phonological length (i.e. number of phonemes), graphical length (i.e. number letters) and lexical frequency (Strijkers et al., 2019). Thus the stimuli served as a convenient starting point for this pilot study. The initial stimuli set contained noun-verb pairs which were assigned particular paired context words (possessive adjective for noun condition and pronouns for verb condition. There were 6 such possibilities ('ma'-'je', 'ta'-'tu', 'son'-'il', 'sa'-'elle', 'votre'-'vous', 'leurs'-'elles'). We selected 20 pairs per context word manually. This generated 6 context words $\times$ 2 conditions $\times$ 20 pairs $=$ 240 items. (see table 3).

Figure 2: Setup for the pilot experiment. The fixation cross was shown for 1000 ms followed by the (entire) stimulus phrase for 2000 ms. After the stimulus, a blank screen was shown for 1000 ms. The duration was fixed for all three steps

2.1.3 Procedure

The visual words were displayed using Neurobehavioural Systems' 'Presentation' software. The stimuli were displayed on the screen with white font with black background. The stimuli were preceded by a fixation cross that remained for 1000 ms. Following the fixation cross the entire phrase appeared and remained for 2000 ms. The stimuli were followed by a blank screen which lasted for 1000 ms (see fig. 2). The participant was instructed to read out loud the phrases as soon as they appeared on the screen. Thus the participant had 3s to read out loud the phrases which was more than enough. The participant was allowed to rest every 80 trials. The speech production was recorded in order to measure reaction time and duration later.

2.2 Data Analysis

2.2.1 Data Acquisition

MEG: MEG data was recorded with 4D Neuroimaging Magnes 3600 Whole Head 248 Channel Scanner (Timone Hospital, Marseille, France). The data were recorded at the sampling frequency of 2034.15 Hz. Head shape and position coil location were recorded using a Polhemus Fastrak 3-D digitising stylus. In order to facilitate the off-line artefact removal, electro-oculogram (EOG), electro-cardiogram (ECG) and EEG (for muscle artefact) were measured.

sMRI: Structural MRI image (needed for the source localisation) was made with SIEMENS 3T Prisma at the Centre IRM (Timone Hospital, Marseille, France). The slice-matrix size: $180 \times 256 \times 256$ mm.

2.2.2 Preprocessing

The MEG data was band-pass filtered at 0.2 Hz and 40 Hz using the MNE-Python package with the standard setting (Gramfort et al., 2013). In order to remove ocular, cardiac and muscular artefacts ICA algorithm provided by the MNE-Python package was

Poss.adj.	Noun	Pron.	Verb	Nb.phon.	Nb.lett.	Lex.freq.N	Lex.freq.V
ma	rivière	je	préfère	6	7	36.49	36.15
ta	boutique	tu	aimerais	5	8	36.01	36.01
sa	tribu	il	exige	5	5	13.58	14.53
son	serpent	elle	prétend	5	7	13.24	13.24
votre	tumulte	vous	devriez	6	7	12.09	11.89
leurs	châteaux	elles	arrêtent	4	8	10.74	10.81

Figure 3: Examples of the stimuli used. Each row shows a pair of noun phrase and verb phrase together with their (a) number of phonemes, (b) number of letters (c) lexical frequency of the noun (d) lexical frequency of the verb.

used. Number of components (`n_components`) was set to explain 99% of the variance in the data. Ocular, cardiac and some muscular artefacts were removed using built-in functions. Additional 3 components corresponding to muscular artefacts were manually selected and removed (see Appendix A for manually selected components).

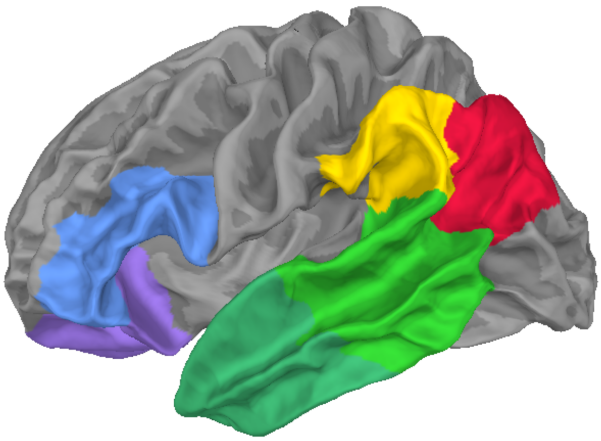

Figure 4: Regions of Interest used in the pilot study. The main ROIs are posterior temporal lobe (PTL) in light green, inferior parietal lobe (IPL) in red and ventro-medial prefrontal cortex (vmPFC) in purple. Additional ROIs are supramarginal gyrus (SMG) in yellow, left anterior temporal lobe (LATL) in dark green and Broca's area in blue.

The Boundary Element Model (BEM) was generated from the MRI with FreeSurfer's Watershed BEM algorithm through MNE-Python's command line tools (Gramfort et al., 2013). Three layer BEM with conductivity `(0.3, 0.006, 0.3)` was used with spacing `ico5` on FreeSurfer (this gives 10242 sources per hemisphere and source spacing of 3.1 mm). Parcellation of cortical areas was done by applying `aparc_sub` parcellation scheme (a subdivided version of `aparc` parcellation Desikan et al., 2006, used in Khan et al., 2018). This parcellation gives 448 labels (224 per hemisphere) and thus on average each cortical area contains 46 sources. These labels subdivide the averaged brain surface provided by FreeSurfer ('fsaverage'). The labels are then morphed to match the surface structure of the participant. Source estimation was performed with minimum norm estimate (Hämäläinen & Ilmoniemi, 1994) using `dSPM` method provided by MNE-Python's minimum norm estimate implementation using the default setting (Gramfort et al., 2013). Source estimate was performed on individual epoched data.

Behavioural Analysis: Praat (Boersma & Weenink 2021, version 6.1.38) was used to annotate the recorded speech in order to measure the reaction time and duration.

Event Related Field Analysis: For the event related analysis we downsampled the data to 250 Hz (for computational speed) and epoched the data from 200 ms before the stimulus onset to 1000 ms after the stimulus onset.

ROI Analysis: For the ROI analysis data was downsampled to 1000 Hz and epoched from 200 ms before the stimulus onset to 1000 ms after the stimulus onset. Regions of interest were constructed by combining `aparc_sub` parcellations manually (see Appendix C for the details). The ROIs are shown in fig. 4. The main regions of interest are posterior temporal lobe (PTL), inferior parietal lobe (IPL) and ventro-medial prefrontal

cortex (vmPFC). PTL and IPL were not combined together to avoid them becoming too big as handling high-dimensional data can be tricky when the dataset is small (see e.g. Hastie et al., 2009). Additionally, Broca and LATL were added as regions of interest since those areas were included in both Friederici's and Pylkkänen's models for earlier syntactico-semantic processes (Friederici, 2011; Pylkkänen, 2020).

Whole Brain Analysis: For whole brain analysis the data was downsampled to 250 Hz as there are many cortical areas to process. The data was epoched from 500 ms before the stimulus onset to 1500 ms post-stimulus onset. Data was subdivided into individual areas defined by `aparc_sub` parcellation (see Appendix D for the visualisation of `aparc_sub`).

2.2.3 MVPA Analysis

In order to be able to perform MVPA classification task, we extracted neural data in a particular cortical region (either ROI or small parcel defined in `aparc_sub`) at a particular time window (from time $t - \frac{w}{2}$ to $t + \frac{w}{2}$ where t is the centre of the time window and w is the width of the time window). This neural data was then flattened to create a vector. For individual trials we generated a vector. Thus, for the pilot study we had 240 vectors in total.

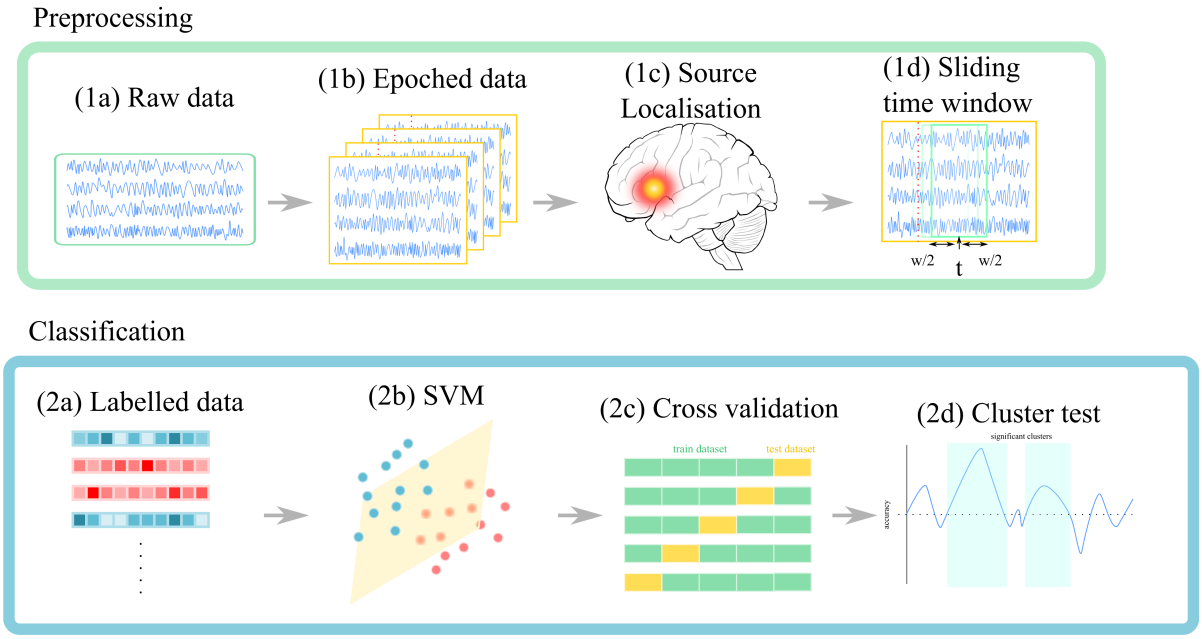

Figure 5: The visualisation of the whole pipeline used for the pilot as well as MOUS dataset analysis. (Top) (1a) Raw data collected from MEG, (1b) Epoched data created by time locking to an event (i.e. presentation of the stimulus) and taking a fixed time window, (1c) Source localisation applied to each epoch individually (1d) Sliding time window selects a slice within this time window. $\frac{w}{2}$ before and after each time point t is taken. (2a) Labelled set of arrays. Each array is labelled either noun condition or verb condition. (2b) Linear SVM classifier selects a single (flat) hyperplane that maximally separates trials corresponding to noun condition from those corresponding to verb condition. (Each dot corresponds to a trial). (2c) K-fold cross validation. The dataset is divided into k sections and one is selected as the test dataset and the rest is used to train the model. (2d) Permutation cluster test. It finds statistically significant cluster which exceeds the chance level (0.5) decoding performance

In order to classify the data we used sklearn (Pedregosa et al., 2011) implementation of linear SVM (Cortes & Vapnik, 1995). SVM is an algorithm which tries to find a hyperplane that separates data points corresponding to each class. A linear version of this algorithm is widely used in MVPA including for noun versus verb decoding (for fRMI Elli et al., 2019, for MEG Arana et al., 2021). A linear version of this algorithm is preferred as it is less likely to overfit (i.e. learn the noise) and has a relatively clear interpretation (i.e. data is distributed in a linearly separable manner, for discussion on use of linear vs. nonlinear algorithms see e.g. King et al., 2018; Naselaris et al., 2011; Ritchie & Carlson, 2016). The metric used to evaluate the model was simple accuracy, i.e. number of correct predictions (noun or verb) divided by total number of predictions. To cross-validate this metric we used 100-fold cross validation, that is, the dataset was divided into 100 sections and then 99 sections of data was used as training dataset and remaining dataset was used as test dataset. This process was repeated 100 times taking different sections as the test dataset each time. The entire training and testing cycle was repeated for each cortical area and each time window (see fig. 5 for the entire pipeline used).

ROI Analysis: For ROI analysis the time window used was 50 ms (25 ms on either side of the time point) and the time window was slid 1 ms at a time. The number of maximum iteration to perform for SVM (`max_iter`) was set to 200 to save computational time.

Whole Brain Analysis: For the whole brain analysis the time window used was 100 ms (50 ms on either side). This is much larger than for ROI. This decision was due to different sample frequencies used (higher the sample frequency, more time points available thus one can afford to use a smaller time window). However, we did not have the time and resources to evaluate this more systematically (see section for more discussion on this 4.3).

2.2.4 Statistical Significance Test

The significance for the evoked field data was tested by permutation cluster test for paired t-test provided by MNE-Python (Gramfort et al., 2013). The significance of the classification accuracy was tested with the temporal one-sample permutation cluster test (Maris & Oostenveld, 2007) provided by MNE-Python (Gramfort et al., 2013). This algorithm was used to test the statistical significance of the decoding accuracy against chance level (0.5) accuracy. Permutation cluster test allows us to find a significant cluster while taking into account the temporal adjacency and correct for multiple hypothesis testing (Maris & Oostenveld, 2007). The test was run with default settings and maximum number of iterations 10^6 times.

2.3 Results

2.3.1 Behavioural Analysis

The mean production onset was 489 s (SD=60) and the mean production duration was 597 s (SD=102).

2.3.2 Evoked Field Analysis

For evoked activity in the sensor space there was no significant cluster. In the source space one cluster almost reached the significance threshold (p=0.0517) and was localised to inferior frontal regions in the right hemisphere (see Appendix B for the results).

2.3.3 ROI Analysis

Posterior Regions (PTL, IPL and SMG):

There were 8 significant clusters in the PTL in total. They can be grouped into four groups. The first cluster appears around 80 ms ($80-87$ ms, p=1×10^{-6}). A couple of clusters with longer duration appear at around 200 ms ($177-195$ ms, p=1×10^{-6}, $199-202$ ms, p=0.000502). A short cluster appears around 350 ms ($341-346$ ms, p=2×10^{-6}). Finally several clusters appear between 550 and 650 ms ($561-563$ ms, p=0.010721, $609-611$ ms, p=0.010957, $641-651$ ms, p=1×10^{-6}, $653-656$ ms, p=0.000273).

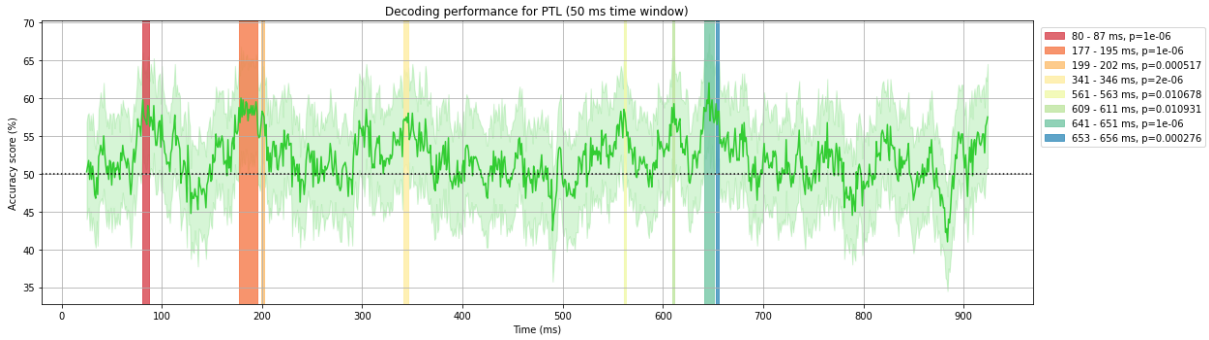

Figure 6: The figure shows the accuracy for PTL in percentage plotted against time. Shaded curves correspond to 95% confidence intervals. Vertical shaded areas are the clusters that reached the significance level.

There were 11 significant clusters in the IPL in total. They can be grouped into four groups. At around 400 ms there are two clusters ($402-410$ ms, p=1×10^{-6}, $437-439$ ms, p=0.00175). There are three closely spaced clusters around 520 ms ($514-518$ ms, p=7.7×10^{-5}, $522-524$ ms, p=0.002554, $526-528$ ms, p=0.019586, $668-674$ ms, p=1×10^{-6}). At around 850 ms there are two longer clusters ($835-841$ ms, p=1×10^{-6}, $835-841$, p=1×10^{-6}, $843-853$, p=1×10^{-6}).

There were 8 clusters in the SMG. They can be grouped into four broad groups of clusters. The first group of clusters appears very early ($29-32$ ms, p=0.002313, $34-36$

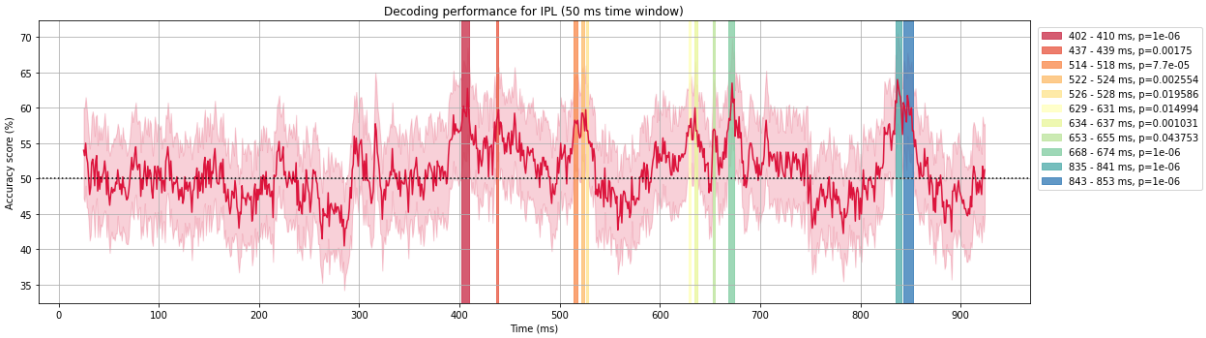

Figure 7: The figure shows the accuracy for IPL in percentage plotted against time. Shaded curves correspond to 95% confidence intervals. Vertical shaded areas are the clusters that reached the significance level.

ms, p=0.012189). A single cluster appears around 100 ms (105 − 114 ms, p=1×10^{-6}) Much longer clusters appear around 200 ms (191 − 215 ms, p=1×10^{-6}, 217 − 226 ms, p=1×10^{-6}). Three clusters appear around 600 ms (579 − 584 ms, p=3.2×10^{-5}, 603 − 616 ms, p=1×10^{-6}, 648 − 650, p=0.021276).

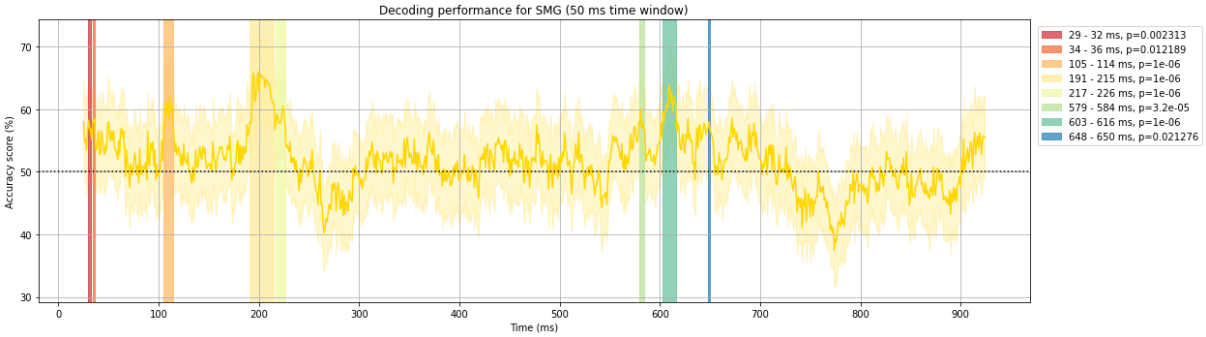

Figure 8: The figure shows the accuracy for SGM in percentage plotted against time. Shaded curves correspond to 95% confidence intervals. Vertical shaded areas are the clusters that reached the significance level.

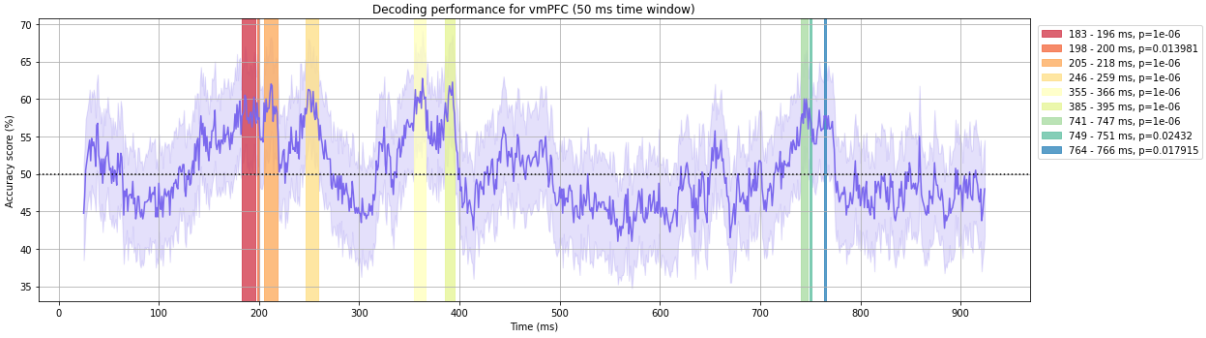

Figure 9: The figure shows the accuracy for vmPFC in percentage plotted against time. Shaded curves correspond to 95% confidence intervals. Vertical shaded areas are the clusters that reached the significance level.

Anterior Region (vmPFC): In the vmPFC there were 9 significant clusters in total. Four clusters appear around 200 ms (183 − 196 ms, p=1×10^{-6}, 198 − 200 ms, p=0.013981,

205 − 218 ms, p=1×10^{-6}, 246 − 259 ms, p=1×10^{-6}). Two clusters slightly before 400 ms (355 − 366 ms, p=1×10^{-6}, 385 − 395ms, p=1×10^{-6}). Three clusters appear around 750 ms (741 − 747 ms, p=1×10^{-6}, 749 − 751 ms, p=0.02432, 764 − 766 ms, p=0.017915)

Additional Regions (Broca, LATL):

There were four significant clusters for Broca's area. Two very early clusters (65 − 68 ms, p=0.001131, 87 − 89, p=0.017607). Second set of clusters appeared around 450 ms (453 − 455 ms, p=20.006706, 459 − 463, p=1×10^{-5}). The longest and most significant cluster appeared around 780 ms (782 − 802 ms, 1×10^{-6}).

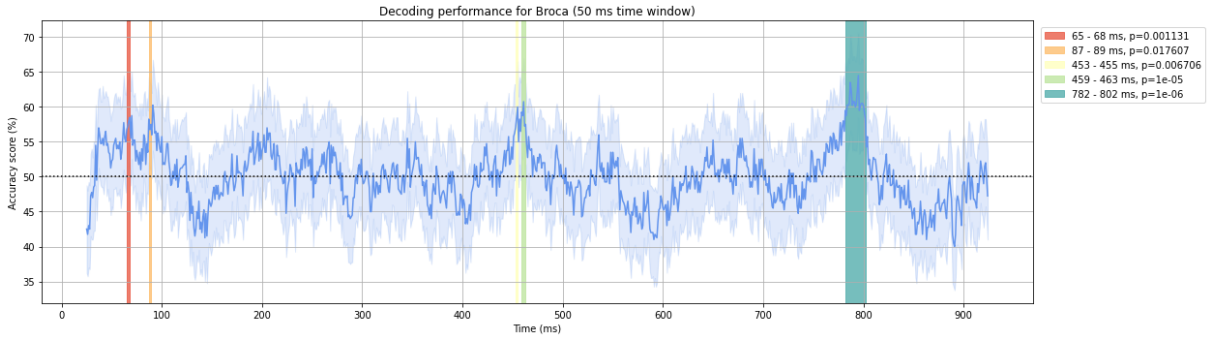

Figure 10: The figure shows the accuracy for Broca's area in percentage plotted against time. Shaded curves correspond to 95% confidence intervals. Vertical shaded areas are the clusters that reached the significance level.

In the LATL there were four significant clusters in total. The first two clusters appear around 200 ms (192−195, p=0.001405, 197−199, p=0.011241). The third cluster appears at around 350 ms (352−355, p=0.001602). The final cluster appears at 450 ms (445−447, p=0.010559).

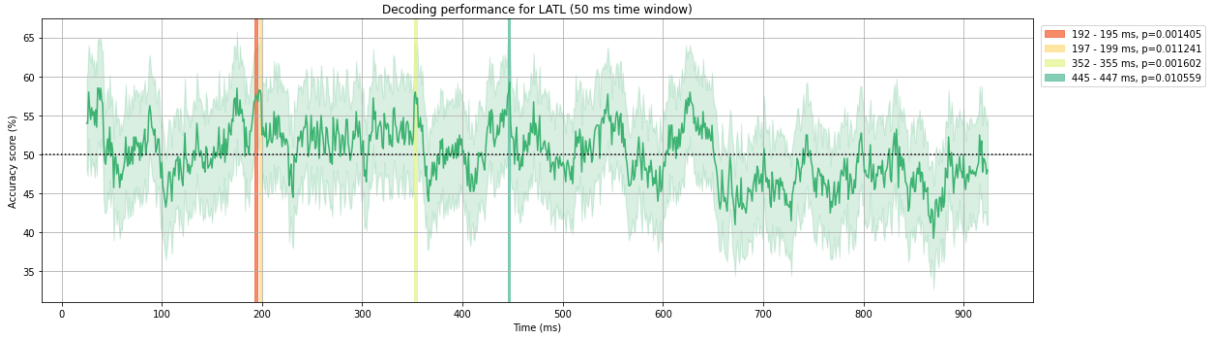

Figure 11: The figure shows the accuracy for LATL in percentage plotted against time. Shaded curves correspond to 95% confidence intervals. Vertical shaded areas are the clusters that reached the significance level.

2.3.4 Whole Brain Analysis

A large number of clusters were found for the whole brain analysis (12 clusters $p < 0.05$, 6 clusters $p < 0.01$ 1 cluster $p < 0.001$). In the following, the numbers after the name of the cortical areas are the indices given to the `aparc_sub` parcellation.

There was an early significant decoding performance in motor cortex (252 – 284 ms, p=0.0.035901, precentral 1-6, 10-16, posterior cingulate 3, 4) and superior frontal area (260 – 292 ms, p=0.017081, superior frontal 2-7, 16-18).

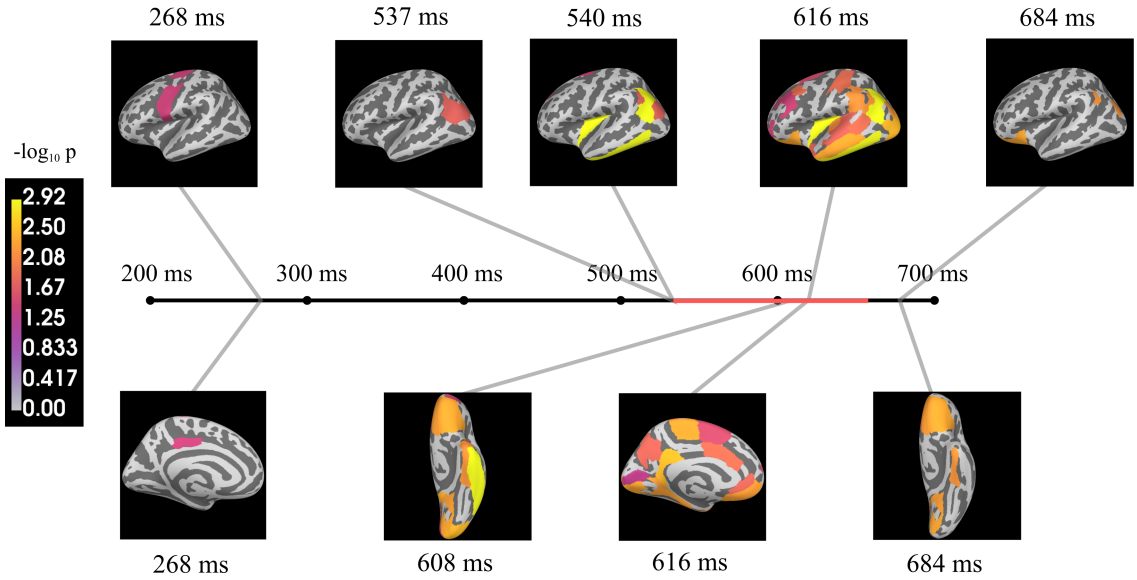

Figure 12: Whole brain analysis results. Colours represent $-\log_{10} p$ values (larger the value, more significant it is). The redline corresponds to the extent of the longest and most significant cluster found in inferior parietal, inferior temporal and insula. Only few characteristic patterns of cluster distributions are shown here. There are two main clusters of significant decoding performances. Of interest here is the second cluster of activities that appear post-production onset.

There were numerous clusters between ca. 500 ms and 700 ms. Those include: 536 – 596 ms, p=0.015 (inferior parietal 1-6, 10, fusiform 8), 540 – 668 ms, p=0.000576 (inferior parietal 4-9, inferior temporal 1-8, insula 1-4), 540 – 584 ms, p=0.003348 (superior frontal 2-8, 14-18), 544 – 680 ms, p=0.002 (isthmus cingulate 1-3, lateral occipital 1-8, 10-11), 552 – 688 ms, p=0.005267 (lingual 6-8, medial orbito-frontal 1-5), 552 – 596 ms p=0.047028 (superior temporal 1-5, 10, 11, superior parietal 9), 564 – 656 ms, p=0.00644 (supramarginal 5-9, temporal pole 1, transversal temporal 1-2), 576 – 668 ms, p=0.049044 (pericalcarine 1-3, postcentral 1), 584 – 652 ms, p=0.012909 (bankssts 1-3, caudal anterior cingulate 1-2, caudal middle frontal 1-2), 584 – 676 ms, p=0.012934 (postcentral 3-9, posterior cingulate 1), 596 – 692 ms, p=0.015505 (precuneus 5-9, rostral anterior cingulate 1), 600 – 692 ms, p=0.00605 (fusiform 7-8, inferior parietal 1-2, 10), 600 – 688 ms, p=0.004479 (lateral orbito-frontal 1-7, lingual 1-3), 600 – 652 ms, p=0.034614 (rostral middle frontal 1-4, 10-12), 604 – 668 ms, p=0.016482 (superior temporal 2-7, 11), 612 – 676 ms, p=0.004596 (middle temporal 3-7, paracentral 1-5, parahippocampal 1-3), 616 – 656 ms, p=0.0274 (superior frontal 2, 12-18). The results are summarised in the fig. 12.

2.4 Discussion

2.4.1 ROI Analysis

ROI analysis revealed that classification accuracy for PTL and SMG first reached significance around 200 ms. The earlier significant cluster for SMG around 30 ms is probably a result of very poor SNR at the beginning of the epoch. Due to the normalisation process used before classification, the noise may have been amplified. IPL first reaches a significance threshold around 400 ms. Of interest is the significant decoding in PTL and SMG around 600 ms. This is in agreement with the predictions Friederici's model makes (Friederici, 2011). IPL also reaches significance ca. 70 ms later. There is, however, another cluster in IPL around 830 ms post-stimulus onset. This cluster is the longest and most significant one for IPL. This may be due to reactivation of representation during the overt production process.

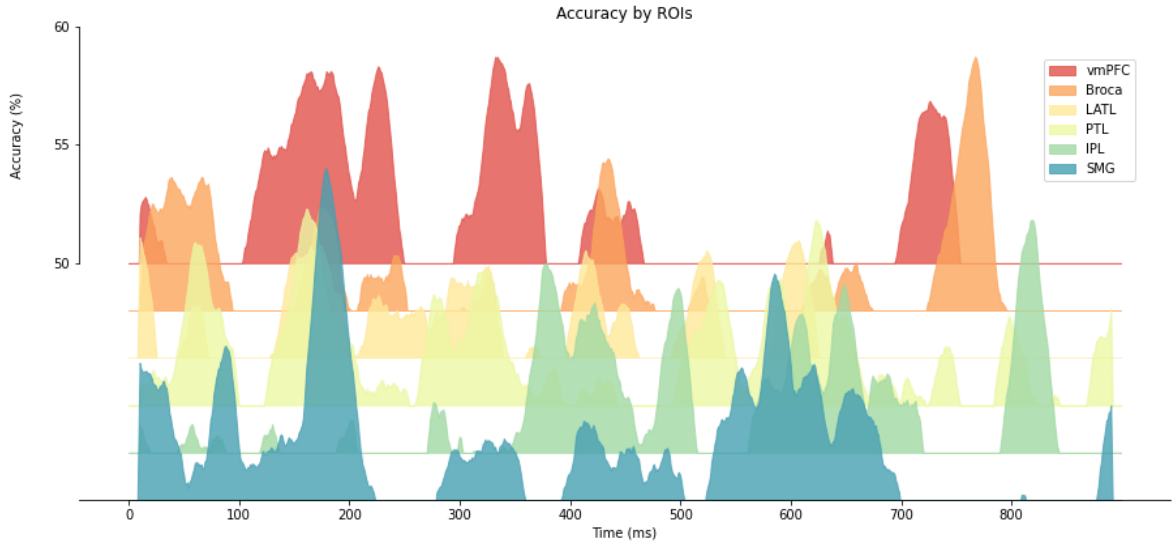

Figure 13: All ROIs plotted together. The data was smoothed by taking the running average by convolving with a kernel of size 20 filled with 0.05.

As for the anterior regions, apart from the very early significant clusters in Broca which may be artefact for the same reason as above, the first significant cluster is found around 200 ms in vmPFC. There are 4 clusters around this time window. This is much earlier than Pylkkänen's model suggests (Pylkkänen, 2019, 2020). There are, however, 2 clusters just before 400 ms which corresponds better to Pylkkänen's model. The first series of clusters may be a phenomenon missed by previous event related studies. Finally, around 750 ms post-stimulus onset we see 2 clusters in vmPFC, followed by 3 clusters around 800 ms in Broca. These may be reactivation of the representation mentioned above.

To summarise, our data suggest that both models capture the processes related to the scene construction. The process seems more complex than either models suggest,

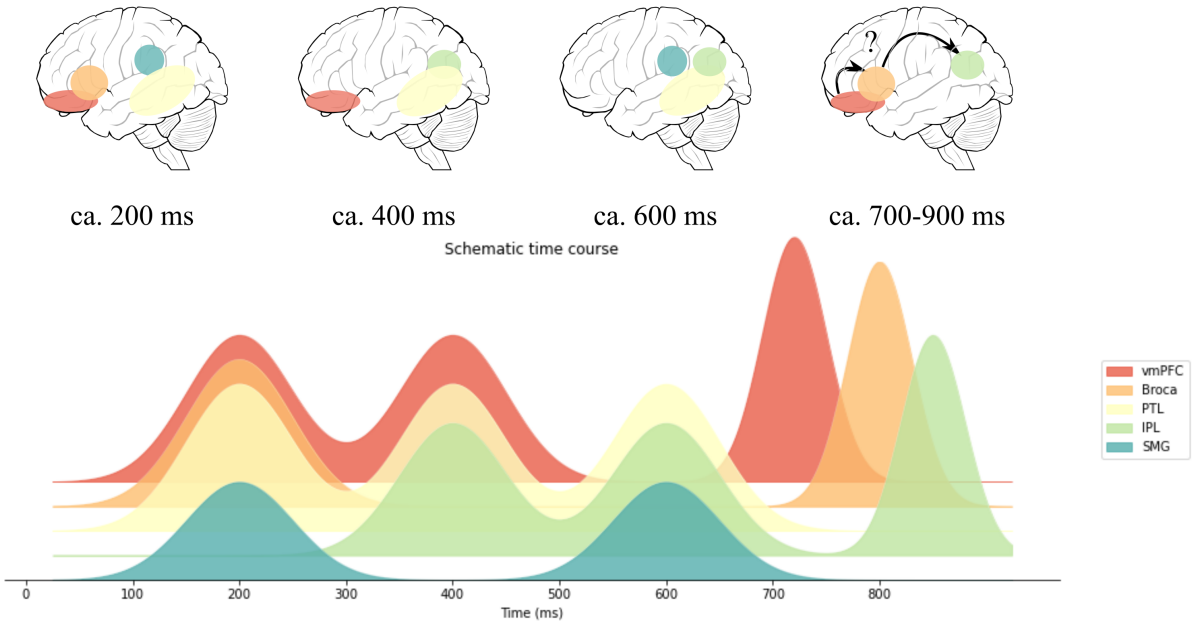

Figure 14: Schematic time course of the results. The clusters are grouped into four groups. (Left) At around 200 ms, activities found in vmPFC, Broca's area, SMG and PTL. (Second from left) At around 400 ms, activities found in vmPFC, PTL and IPL. (Second from right) At around 600 ms, activities found in PTL, IPL and SMG. (Right) Between 700 to 900 ms activities found in vmPFC, Broca's area and IPL in a cascading fashion. Note that our does not allow the conclusion that this is indeed a sequential process. For such inference Granger causality analysis is needed. (Seth et al. 2015)

however. In particular, we can observe 4 time windows where significant clusters can be grouped. First group is around 200 ms, which roughly corresponds to ELAN latency. In this time window vmPFC, Broca, PTL and SMG reach a significant threshold. This may correspond to initial grammatical processing predicted by Friederici (2011). Second group is around 300-400 ms and includes vmPFC, PTL and somewhat later, IPL. It is this time window that Pylkkänen's model (2019, 2020) predicts vmPFC to process semantic information. This may thus be the initial scene building process. The third group appears around 600 ms in PTL, IPL and SMG. These posterior clusters are predicted by Friederici's model. They may correspond to the repair or reanalysis process in her model (Friederici, 2011). The fourth, and unexpected clusters appear in vmPFC, Broca and IPL. These appear one after another and may correspond to sequential reactivation during the overt production process. With analyses we conducted we cannot conclude whether these are indeed sequential processes or mere coincidence. The schematic representation of this summary is shown in 14. Connectivity analyses such as Granger causality (Seth et al., 2015) is needed for more direct evidence. One study (Hauptman et al., 2022) has conducted G-causality analysis during production in MEG and shown that there is indeed anterior to posterior information flow. In the future such analyses is needed for our data as well. It must be said that the clusters we found were not very long overall. This is most likely due to the small sample size (single participant, 240 trials). Thus we cannot make any strong conclusions from this data.

2.4.2 Whole Brain Analysis

The whole brain analysis revealed a slightly different picture. Overall, a vast portion of the brain reached significance threshold for decoding accuracy at some point. Upon closer look, however, we may nonetheless speculate that inferior parietal regions and vmPFC play a crucial role. First, inferior parietal regions are the first to reach significance after the production onset (at around 530 ms after stimulus onset). Second, the inferior parietal region (together with the inferior temporal region and insula) is the longest and most significant cluster ($540 - 668$ ms, p=0.0006). Third, vmPFC and a portion of inferior parietal regions (numbers 1, 2, 10) are the last to drop below significance threshold. Thus while large areas of the brain were shown to contain information about nouns and verbs, inferior parietal regions and vmPFC may still play some key role. This point will be further discussed in the General Discussion 4.1.

We will briefly discuss the early involvement of the motor cortex around 270 ms. The most straightforward interpretation would be to say that because pronominal adjectives (e.g. 'mon') and pronouns (e.g. 'je') preceded nouns and verbs, and the articulation of those words would give information about the category of the following words. Thus this decoding performance could be explained in terms of articulation of the first word in the phrase. Although the syllabification process is predicted to be around 350-450 ms in Indefrey and Levelt model (Indefrey & Levelt, 2004; Indefrey, 2011) our task did not involve conceptualisation process, i.e. converting non-linguistic representation of the picture to linguistic representation. Thus this could mean that the syllabiciation took place around 270 ms. There is, however, an alternative explanation. This comes from the domain of embodied cognition. Pulvermüller, for example, observed activation in the motor cortex for words related to arm and legs between 150 and 250 ms (Boulenger et al., 2012; Pulvermüller, 2013). Since verbs are more likely to involve movements, it may be the case that this involvement of the motor cortex reflects embodied aspects of meaning. Our data are inconclusive, however.

Finally, the whole brain analysis revealed a surprisingly distributed picture. One possible explanation is that this was a participant specific analysis. As Stelzer et al. (2014) notes, group level analysis can give the false impression of blob-like activation due to only a few areas overlapping across participants. Thus it is possible that many areas are indeed active during scene construction but previous group level analysis have missed this aspect. (Issues relating to inter-individual variability will be discussed in 4.4) Additionally, MVPA may be more sensitive to distributed aspects than univariate analysis. Haxby and colleagues (2001) have shown, for example, that areas outside the fusiform area also contain information about faces. Whatever the reason, we cannot make conclusive statements based on a single participant. We are currently planning an additional pilot study with a second participant which may reveal more on this point.

3 MOUS Dataset Analysis

Alongside the pilot study we also conducted the analysis of a large publicly available dataset. There are several reasons for doing this. First, we wished to assess the effect of inter-individual differences but within the limits of Master internship it was unrealistic to obtain data from many individuals. Second, we wanted to know how the pipeline we used for the pilot would scale up to a larger cohort. Finally, we wished to assess variables beyond simple nouns versus verbs[1]. Thus our aim here was more technically oriented than the pilot study.

3.1 Dataset Description

The dataset we used is called the 'Mother of Unification Study' (MOUS) dataset described in Schoffelen et al. (2019). This dataset contains a total of 204 participants who took part in experiments with either auditorily presented linguistic stimuli (N=102) or visually presented linguistic stimuli (N=102). For each participant, four different imaging techniques were used (sMRI, fMRI, DWI, MEG). We only made use of sMRI and MEG. In the following only the relevant details of this dataset will be present. For further details we refer to the paper Schoffelen et al. (2019).

3.1.1 Participants

102 native Dutch speakers took part in the visual condition of the experiment. They were aged between 18 and 32 (M=21.8, SD=2.8, 53 females, 49 males). 7 participants were dropped due to technical issues encountered during epoching (unknown bug causing the program to halt for that participant).

3.1.2 Stimuli

Original Annotation: The stimuli consisted of 360 sentences. Of those, 180 sentences contained relative clauses (complex condition) and 180 did not (simple condition). These 360 sentences were scrambled to generate 360 corresponding word lists. We, however, made use only of sentence conditions (both simple and complex) and dropped word list conditions, thus leaving 180 sentences per participant.

Part of Speech Annotation: In order to tailor this dataset for our study, the stimuli sentences were annotated for nouns and verbs as well as other grammatical categories. When annotating, certain choices were made with regard to what counts as nouns and what counts as verbs. First, all non-finite forms (i.e. infinitives, e.g. 'spreken' - to

[1]Unfortunately, this last aspect has not yet been attempted as of writing. This aspect will be continued in the future

speak, participles, e.g. 'gehad' - had, gerund, 'het spelen' - (act of) playing) as well as auxiliaries (i.e. 'hebben' - to have, 'zijn' - to be) were excluded from current analysis since such words are less likely to elicit verb specific activations (although this point could be investigated further). Second, verbs consisting of more than two parts, such as so called *verbs with separable prefixes* (e.g. 'opletten' - to pay attention, 'ik lett ... op' - I pay attention) and complex constructions consisting of participles and auxiliaries (e.g. present perfect). For nouns, out of extra precaution words with common gender or ambiguous gender, uncountable nouns, proper nouns and diminutives were dropped from the current analysis. There is in principle no reason to believe these choices should affect the end results but the dataset was large enough to allow us to stay on the safe side. The final annotation used in this study was validated by a native speaker who inspected 40 randomly sampled sentences (ca. 11%). An example of the annotation is given in fig. 15.

80 Ik **nam** een lekker **biertje** dat **gemaakt** **was** van **gerst**.
I took a nice beer that made was from barley
(eng.) I took a nice beer that was made from barley.

Figure 15: An example of the annotation added for this study. Each word was annotated if they were either noun or verb. (Blue) nouns (Red) verbs.

3.1.3 Procedure

The stimuli were visually presented on the LCD projector (refresh rate 60 Hz). Following a fixation cross (duration 300 ms) words were presented one by one with duration mean 351 ms (min. 300 ms, max. 1400 ms). Duration varied in proportion to the pronunciation duration of the corresponding auditory stimulus (Schoffelen et al., 2019).

3.1.4 Data Acquisition

MEG: Data were acquired with a 275-channel axial gradiometer system (CTF) with sampling frequency of 1200 Hz. 3 head localiser coils (nasion, left and right ear canals) were used to localise the head location relative to the sensors. Head surface was digitised using a Polhemus 3D-Space Fastrak scanner.

sMRI: Structural MRI image (needed for the source localisation) was made with SIEMENS Trio 3T scanner using a 32-channel head coil. T1-weighted magnetization-prepared rapid gradient-echo (MP-RAGE) pulse sequence was used (volume TR=2300 ms, TE=3.03 ms, 8 degree flip-angle, 1 slab, slice-matrix size = 256×256, slice thickness = 1 mm, field of view= 256 mm, isotropic voxel-size = $1.0 \times 1.0 \times 1.0$ mm).

3.2 Data Analysis

As the dataset size is far larger than usual dataset size (several Terabytes of data) we used a cluster of the Laboratoire de Psychologie Cognitive (St Charles, Marseille, France). The cluster has 128 CPUs and is managed by resource managing software SLURM (Yoo et al., 2003).

3.2.1 Preprocessing

The data was band-pass filtered between 0.1 and 45 Hz using the MNE-Python package with the standard setting (Gramfort et al., 2013). The artefact removal was skipped partly due to the technical issues, partly due to expectation that cardiac and ocular artefacts are less problematic than muscular artefacts related to production for decoding analysis (Honari-Jahromi, Chouinard, Blanco-Elorrieta, Pylkkänen, & Fyshe, 2021).

The Boundary Element Model (BEM) was generated from MRI of each participant with FreeSurfer's Watershed BEM algorithm provided by MNE-Python command line tools (Gramfort et al., 2013). There were issues when generating a three layer BEM model possibly due to sMRI image quality (software indicated that three layers intersect each other). As a temporary workaround we used a single layer BEM model with 0.3 conductivity. It is unclear how this would influence the final results. The spacing was `ico5` as before. Parcellation of cortical areas was done by applying the `aparc` parcellation scheme (Desikan et al., 2006). Due to time constraints we could not use the custom parcellation scheme used for the pilot study. In particular, this means that in the temporal lobe our regions of interest are subdivided along the dorso-ventral axis rather than antero-posterior axis. ROIs included areas around (a) temporo-parietal junction (TPJ), i.e. pSTS ('bankssts' in `aparc`), supramarginal gyrus, inferior parietal region, (b) vmPFC, i.e. lateral orbito-frontal region, medial orbito-frontal region, and few additional areas included in two models (c) Broca's area, i.e. pars opercularis, pars orbitalis,[2] (d) LATL, i.e. middle temporal gyrus, superior temporal gyrus, temporal pole, (e) other regions, inferior temporal gyrus, lateral occipital region, transverse temporal. The source space was morphed into a common space based on the averaged brain provided by FreeSurfer ('fsaverage').

The data were then converted in the manner described in the pilot study section. The only difference is the time window which was 100 ms in this analysis. This is because the sampling frequency is lower thus there are fewer data points within a given time window. Finally, in order to balance the number of items, noun items were randomly dropped (nouns were more numerous than verbs). The final dataset consisted of 45018 items per cortical area (22509 nouns and 22509 verbs).

[2] due to a glitch pars triangularis data is missing. There was unfortunately no time to correct this error before the deadline.

3.2.2 Classification

As with the pilot study the classification was done with sklearn (Pedregosa et al., 2011) implementation of linear SVM (Cortes & Vapnik, 1995). The metric used to evaluate the model was simple accuracy. As the data is balanced in the previous step, this would yield the chance level of 0.5. (with unbalanced data, e.g. with more nouns than verbs, this metric would be inappropriate). To cross-validate this metric we used 100-fold cross validation. The entire training and testing cycle was repeated for each cortical area and each time window.

3.2.3 Statistical Significance Test

Temporal one-sample permutation cluster test (Maris & Oostenveld, 2007) provided by MNE-Python (Gramfort et al., 2013) was used to test the statistical significance of the decoding accuracy against chance level (0.5) accuracy. The test was run with default settings and maximum number of iterations 10000.

3.3 Results

Main ROIs: From the main regions of interest pSTS and SMG reached significance. pSTS reached significance at $800-810$ ms (p=0.0209) and SMG reached significance in three clusters ($295-305$ ms, $295-305$ ms, p=0.0457, $785-795$ ms, p=0.0371, $895-905$ ms, p=0.0011). Two ROIs in vmFC (lateral orbito-frontal area and medial orbito-frontal area) did not reach significance.

Additional ROIs: From the additional ROIs ITG and MTG reached significance. ITG reached significance at $450-480$ ms (p=0.0004). MTG reached significance twice ($500-510$ ms, p=0.00062, $760-770$ ms p=0.00067). No other regions reached significance.

3.4 Discussion

The results for this dataset analysis shows that the posterior temporal regions reach significance around 800 ms. This time window is much later than that of the pilot study but this may be due to it involving complex sentences. However, the cluster is very short (mere 10 ms) and p-value is fairly large (p=0.02) thus we refrain from making further interpretations.

The vmPFC did not reach significance. This may be due to the fact that the vmPFC effect is known to be not very robust (Pylkkänen, 2020). It may also be due to inter-individual variability (see below 4.4).

Besides the two main regions of interest we also observed significant decoding in ITG $450-480$ ms and in MTG around 500 ms and 760 ms. These regions are linked to semantic

processing (Binder & Desai, 2011) and may be the same effect we observed in the whole brain analysis of the pilot data.

Overall the decoding accuracy is surprisingly poor. While the accuracy in MVP analysis for science (as opposed to engineering) does not necessarily need to be high (Hebart & Baker, 2018), given the performance we observed for the pilot study it is surprising to see such low performance with 95 participants. These results may be due to several factors. First, there may be some issues with preprocessing. The pipeline for MOUS dataset is still in the early stage of development as of writing and we cannot deny the possibility that there are issues with it. Second, the stimuli were less controlled than for our pilot study. Thus it is more naturalistic but also adds noise to the data. Third, this could reflect the inter-individual variabilities. We will discuss this issue in detail below 4.4. Finally, related to the second point is that we did not include position within the sentence as a factor. This is an important factor since verbs and nouns tend to take a particular position within the sentence. It also affects at what stage of comprehension it is. At the beginning of the sentence one may have only a vague idea of the semantic content while towards the end this would be much clearer. The position specific analysis was planned but due to time constraints we could not perform it.

To conclude this section, it has become evident that while scaling up the pipeline is possible there will be additional challenges along the way. These aspects should be taken into consideration in the planning of future experiments.

4 General Discussion

In this study we set out to compare the two models of language comprehension with respect to the final stage of scene building. To this end we performed a pilot study and dataset analysis. The pilot study used overt reading with controlled stimuli. The results of this pilot study revealed that both PTL, as predicted by Friederici's model (Friederici, 2002, 2011), and vmPFC, as predicted by Pylkkänen's model (Pylkkänen, 2019, 2020), are likely to be involved in this process. Importantly, it was shown that semantic information at the latest stage of comprehension involves more than those two regions. Instead a wide number of regions are shown to be involved. The MOUS dataset analysis, on the other hand, showed that there are some technical hurdles that deserve consideration. Implications of these points will be discussed below.

4.1 Reconciling the Two Models

The pilot study, and to a lesser extent MOUS dataset analysis, have revealed that both areas around TPJ and vmPFC are likely to be involved in the construction and/or representation of holistic scene information. This is surprising since those two areas are very distant from each other. However, whole brain analysis gives us some clues as to how to

solve this puzzle. First, even in the whole brain analysis, which showed large portions of the brain to contain relevant information, inferior parietal regions and vmPFC stand out as key regions. It is in the inferior parietal region that the earliest and most significant decoding is observed (around 536 ms). The vmPFC on the other hand is the latest region (together with parts of inferior parietal regions and some ventral temporal regions) to drop below significance level (around 688 ms). One possible interpretation is that they are the hubs for a larger network. Such interpretation would potentially reconcile two models.

But what would this network look like? The answer may come from outside of language literature. Both vmPFC and inferior parietal regions are known to be part of what is called *core network* (Schacter & Addis, 2007; Benoit & Schacter, 2015). According to one hypothesis called *constructive episodic simulation hypothesis* (Schacter & Addis, 2007) this network is thought to process not only episodic memory but also simulating hypothetical situations. The idea behind this hypothesis is that memory is a constructive process rather than a pure retrieval process. And this constructive process can be applied not only to previously experienced episodes but also to hypothetical events that haven't happend (yet) (Schacter & Addis, 2007). The crucial point to note here is that language comprehension, particularly in its final stage, involves a very similar set of computations. In both episodic simulation and language comprehension one must 1) retrieve relevant information from past experience, 2) recombine those retrieved elements into a new holistic structure. Thus one may speculate that what we observed in this study is related to this domain's general capacity to integrate information from vast areas of the brain into a single coherent structure.

According to the meta-analysis by (Benoit & Schacter, 2015) the core network includes parahippocampal cortex (PHC), hippocampus (HC), anterior cingulate cortex (aCC), posterior cingulate cortex (pCC), ventral medial prefrontal cortex (vmPFC), dorsomedial prefrontal cortex (dmPFC), lateral surface of temporal cortex (LTC), posterior inferior parietal lobes (pIPL), superior temporal lobes (pSTL), precuneus (Prec) and postcentral gyrus (PCG). These areas (except for the hippocampus which is difficult to study with MEG due to its subcortical location, although see (López-Madrona et al., 2022)) also reached the significance threshold in our whole brain analysis. The figure 16 shows the areas shown to be activated both for episodic memory processing (blue) and episodic simulation tasks (red) in the meta-analysis by Benoit ant Schachter (2015). At the bottom of the same figure are the areas that reached significance at some point. They show certain resemblance to each other. Of course, one possible counter argument is that they pretty much include all of the brain so any extensive activation would resemble such a network. Thus this is merely a speculation at this point. However, it is clear that in order for the two models to reconcile we need to look beyond local structure.

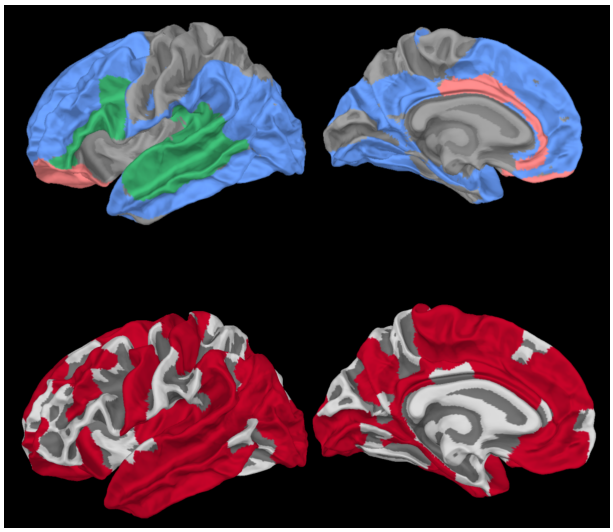

Figure 16: (top) Cortical areas found to be active during both episodic memory task and episodic simulation tasks by the meta-analysis and not more active for episodic simulation in red and areas found to be more active for episodic simulation task than for episodic memory task in blue (Benoit & Schachter 2015). And 'language network', here defined as BA 22, 41, 42, 44, 45, 47 in green (cf. e.g. Fedorenko & Blank 2020). (top left) lateral view, (top right) medial view. (bottom) Areas that reached significance threshold of $p < 0.05$ in our whole brain analysis in red. (bottom left) lateral view, (bottom right) medial view.

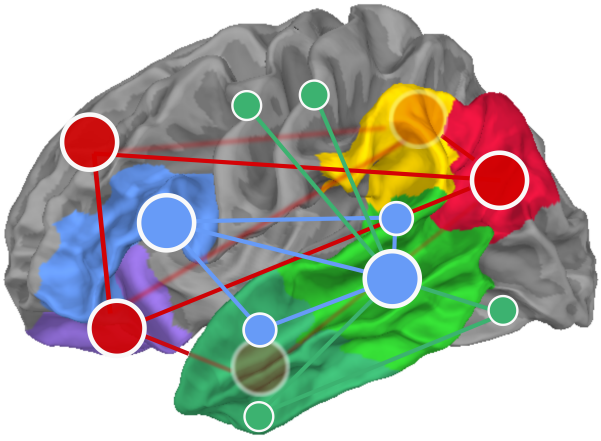

Figure 17: Possible 'networks of interest' (NOI) (Fedorenko & Thompson-Schill 2014). (Blue) Perisylvian language network. (Red) core network for episodic memory. (Green) Semantic network.

This brings us to a more general discussion about the limit of purely local analysis. It has been pointed out before (Fedorenko & Thompson-Schill, 2014) that in order for us to fully understand a cognitive system capable of handling language we must go beyond individual regions and look at networks. Apart from the core network we discussed we may have to look at semantic network (Binder & Desai, 2011), of which we also see some involvement in our study, multiple demand network (Fedorenko & Blank, 2020), and, last but certainly not least, the perisylvian 'language' network (Fedorenko & Thompson-Schill, 2014). In the future, it would be interesting to investigate how various networks as whole contain and process information by combining connectivity studies with MVPA approach.

4.2 Technical Challenges Encountered in MOUS Dataset Analysis

While analysing the MOUS data (and to a lesser extent also the pilot data) we encountered some technical difficulties. We will discuss these issues and what could be done in the

future experiments.

4.3 Classifiers are Sensitive to Various Parameters

Our classification pipeline requires several parameters to be specified, including sampling frequency, cortical parcellation to segment source estimation, size of the time window, whether or not to use dimensionality reduction (we used none), number of folds for k-fold cross validation and source localisation method and their parameters. During initial testing we observed that these factors could potentially influence the final results. Due to lack of time, the parameters were chosen in a somewhat *ad hoc* manner. This is not a big issue for pilot study such as this but for future studies we must plan ahead how to select these parameters or else we may end up in p-hacking (Wicherts et al., 2016). Grid-search approach in which all possible combinations of the parameters are tried is one option (Syarif et al., 2016). However due to the time consuming nature of the classification analysis this could only be applied to smaller dataset (for example, to pilot data).

4.4 Managing Inter-Individual Variability

We have already briefly mentioned the issue of inter-individual variability. This is a very general problem. It has been shown in the fMRI modality that while within-individual variability across sessions is minimal, inter-individual variability can be quite significant (Miller et al., 2012). Several researchers have warned that not explicitly taking these variabilities into account can lead to misleading conclusions such as identifying overlap when there is none (Fedorenko, 2021), or identifying blob-like activation when it is more distributed in reality (Stelzer et al., 2014).

In our case inter-individual variability may have contributed to relatively poor decoding performance. In order to aggregate the multi-participant data we morphed the data into a single common source space on the basis of structural information alone. However, MVPA is shown to be sensitive to participant specific patterns (Haxby et al., 2011) and between-participant decoding is known to be much harder (Haxby et al., 2011). Thus by aggregating multiple participant data without taking into account this *functional* variability may have added a large amount of noise. In the domain of (mass) univariate analysis, some researchers recommend functional localisers to identify subject specific ROIs before doing the group level analysis (Fedorenko, 2021). There are ongoing attempts to overcome this difficulty in the domain of MVPA (Haxby et al., 2011, 2020). These techniques, however, are designed for fMRI and it is unclear how well they apply to MEG modality. An alternative, and technically simpler, workaround is to use participants but more trials per participant. This will allow the dataset size to increase while minimising noise introduced by inter-individual variability. (For similar arguments in favour of studying fewer individuals see Naselaris et al., 2021; Fedorenko, 2021). In MEG studies, there is

an additional benefit that only single sMRI data is needed per participant, thus cost for sMRI imaging per trial decreases as we collect more data from a single participant.

5 Conclusion

In this study we investigated the final stage of language comprehension, namely the construction of holistic scene representation. To this end we performed multivariate analysis of MEG data. The two parallel components of this study, the pilot study and MOUS dataset analysis, have given us two different sets of insights.

The pilot study using overt reading task in MEG has shown that a) MVPA can by applied to higher cognition such as linguistic processes, b) overt articulation does not necessarily cause a major issue, and c) in order to have a better picture of the process we may need to go beyond analysis of local features to analysis of network-level activation patterns. The MOUS dataset analysis has shown that in order to scale up the pipeline we must account for the inter-individual variability. Finally, both studies have shown the need for a systematic approach to parameter selection to avoid p-hacking. Once these challenges are dealt with we will be in place to tackle the challenging theoretical questions with realistic participant group sizes.

6 Acknowledgements

I'd like to thank my supervisor F.-Xavier Alario for proposing this topic and guiding me through it. It was a pleasure to observe up close his scientific and pedagogical skills up close. I would also like to thank Jean-Michel Badier for helping me carry out the MEG experiment. I would like to thank Loic Bonnier and Jean Luc Blanc for assisting me in setting up the cluster for dataset analysis. I am also grateful for Floor Meewis for checking my annotation in Dutch. Last but not least, I'd like to thank my parents, who have supported me throughout. I would not be here without their support.

References

Arana, S., Schoffelen, J.-M., Mitchell, T., & Hagoort, P. (2021). Mvpa does not reveal neural representations of hierarchical linguistic structure in meg. *bioRxiv*.

Baillet, S. (2017). Magnetoencephalography for brain electrophysiology and imaging. *Nature neuroscience*, *20*(3), 327–339.

Benoit, R. G., & Schacter, D. L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia*, *75*, 450–457.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in cognitive sciences*, *15*(11), 527–536.

Boulenger, V., Shtyrov, Y., & Pulvermüller, F. (2012). When do you grasp the idea? meg evidence for instantaneous idiom understanding. *Neuroimage*, *59*(4), 3502–3513.

Brennan, J., & Pylkkänen, L. (2008). Processing events: Behavioral and neuromagnetic correlates of aspectual coercion. *Brain and language*, *106*(2), 132–143.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Crepaldi, D., Berlingeri, M., Paulesu, E., & Luzzatti, C. (2011). A place for nouns and a place for verbs? a critical review of neurocognitive data on grammatical-class effects. *Brain and language*, *116*(1), 33–49.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... others (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.

Elli, G. V., Lane, C., & Bedny, M. (2019). A double dissociation in sensitivity to verb and noun semantics across cortical networks. *Cerebral Cortex*, *29*(11), 4803–4817.

Fedorenko, E. (2021). The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Current Opinion in Behavioral Sciences*, *40*, 105–112.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433.

Fedorenko, E., & Blank, I. A. (2020). Broca's area is not a natural kind. *Trends in cognitive sciences*, *24*(4), 270–284.

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

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

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, *91*(4), 1357–1392.

Friederici, A. D. (2017). *Language in our brain: The origins of a uniquely human capacity*. MIT Press.

Frisch, S., Kotz, S. A., Von Cramon, D. Y., & Friederici, A. D. (2003). Why the p600 is not just a p300: The role of the basal ganglia. *Clinical Neurophysiology*, *114*(2), 336–340.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... others (2013). Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 267.

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

Hagoort, P. (2013). Muc (memory, unification, control) and beyond. *Frontiers in psychology*, *4*, 416.

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, *32*(1), 35–42.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hauptman, M., Blanco-Elorrieta, E., & Pylkkänen, L. (2022). Inflection across categories: Tracking abstract morphological processing in language production with meg. *Cerebral Cortex*, *32*(8), 1721–1736.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, *37*, 435–456.

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.

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

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *Elife*, *9*, e56601.

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, *180*, 4–18.

Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive brain research*, *19*(1), 59–73.

Honari-Jahromi, M., Chouinard, B., Blanco-Elorrieta, E., Pylkkänen, L., & Fyshe, A. (2021). Neural representation of words within phrases: Temporal evolution of color-adjectives and object-nouns during simple composition. *PloS one*, *16*(3), e0242754.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. *Frontiers in psychology*, *2*, 255.

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1-2), 101–144.

Khan, S., Hashmi, J. A., Mamashli, F., Michmizos, K., Kitzbichler, M. G., Bharadwaj, H., ... others (2018). Maturation trajectories of cortical resting-state networks depend on the mediating frequency band. *NeuroImage*, *174*, 57–68.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of memory and language*, *52*(2), 205–225.

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, *18*(4), 203–210.

King, J.-R., Gwilliams, L., Holdgraf, C., Sassenhagen, J., Barachant, A., Engemann, D., ... Gramfort, A. (2018). Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition.

Kolk, H. H., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and language*, *85*(1), 1–36.

Kwon, H., Kuriki, S., Kim, J. M., Lee, Y. H., Kim, K., & Nam, K. (2005). Meg study on neural activities associated with syntactic and semantic violations in spoken korean sentences. *Neuroscience research*, *51*(4), 349–357.

López-Madrona, V. J., Villalon, S. M., Badier, J.-M., Trébuchon, A., Jayabal, V., Bartolomei, F., ... others (2022). Magnetoencephalography can reveal deep brain network activities linked to memory processes. *bioRxiv*.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177–190.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

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

Naselaris, T., Allen, E., & Kay, K. (2021). Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences*, *40*, 45–51.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage*, *56*(2), 400–410.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785–806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *Journal of experimental psychology: Learning, memory, and cognition*, *20*(4), 786.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in cognitive sciences*, *17*(9), 458–470.

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66.

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*, *375*(1791), 20190299.

Pylkkänen, L., Martin, A. E., McElree, B., & Smart, A. (2009). The anterior midline field: Coercion or decision making? *Brain and language*, *108*(3), 184–190.

Ritchie, J. B., & Carlson, T. A. (2016). Neural decoding and "inner" psychophysics: A distance-to-bound approach for linking mind, brain, and behavior. *Frontiers in neuroscience*, *10*, 190.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 773–786.

Schoffelen, J.-M., Oostenveld, R., Lam, N. H., Uddén, J., Hultén, A., & Hagoort, P. (2019). A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific data*, *6*(1), 1–13.

Service, E., Helenius, P., Maury, S., & Salmelin, R. (2007). Localization of syntactic and semantic brain responses using magnetoencephalography. *Journal of Cognitive Neuroscience*, *19*(7), 1193–1205.

Seth, A. K., Barrett, A. B., & Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, *35*(8), 3293–3297.

Siri, S., Tettamanti, M., Cappa, S. F., Rosa, P. D., Saccuman, C., Scifo, P., & Vigliocco, G. (2008). The neural substrate of naming events: effects of processing demands but not of grammatical class. *Cerebral Cortex*, *18*(1), 171–177.

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J., 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.

Stelzer, J., Lohmann, G., Mueller, K., Buschmann, T., & Turner, R. (2014). Deficient approaches to human neuroimaging. *Frontiers in Human Neuroscience*, *8*, 462.

Strijkers, K., Chanoine, V., Munding, D., Dubarry, A.-S., Trébuchon, A., Badier, J.-M., ... others (2019). Grammatical class modulates the (left) inferior frontal gyrus within 100 milliseconds when syntactic context is predictive. *Scientific reports*, *9*(1), 1–13.

Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). Svm parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika*, *14*(4), 1502.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(3), 407–426.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, 1832.

Williams, A., Reddigari, S., & Pylkkänen, L. (2017). Early sensitivity of left perisylvian cortex to relationality in nouns and verbs. *Neuropsychologia*, *100*, 131–143.

Yoo, A. B., Jette, M. A., & Grondona, M. (2003). Slurm: Simple linux utility for resource management. In *Workshop on job scheduling strategies for parallel processing* (pp. 44–60).

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.

Appendices

A ICA Components of the Articulation Artefacts

As stated in the main text, 3 ICA components were manually selected.

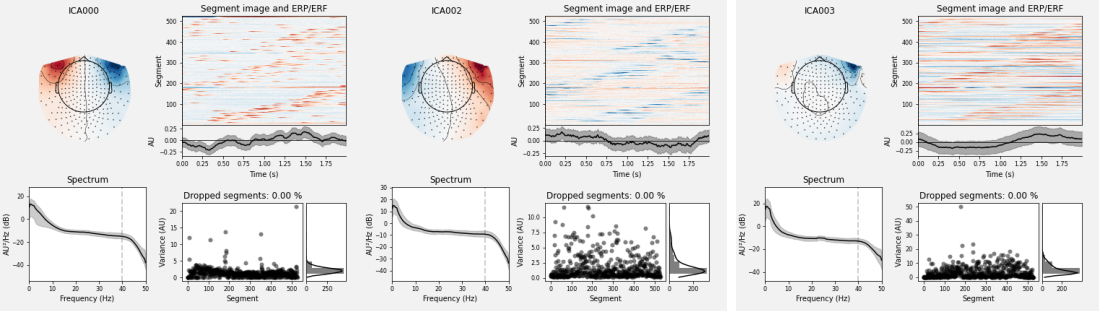

Figure 18: Three ICA components selected manually.

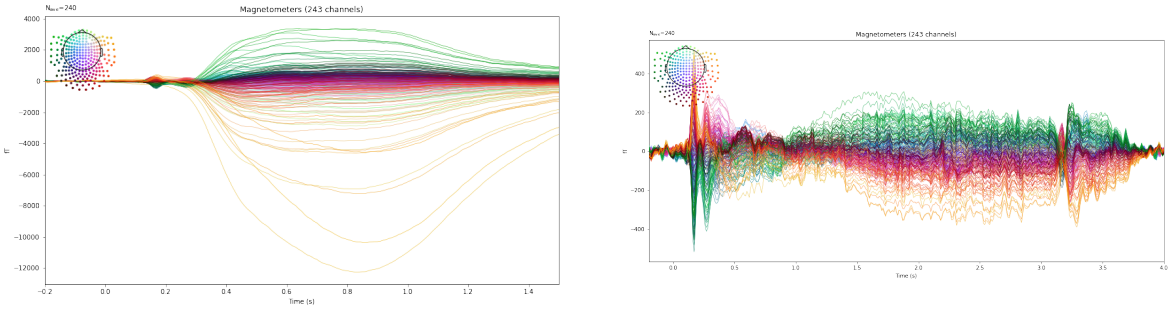

Figure 19: (Left) Evoked activity before ICA repair plotted by sensors. The colours represent the corresponding sensor. (Right) Same evoked after ICA repair.

B Evoked Field Analysis Results

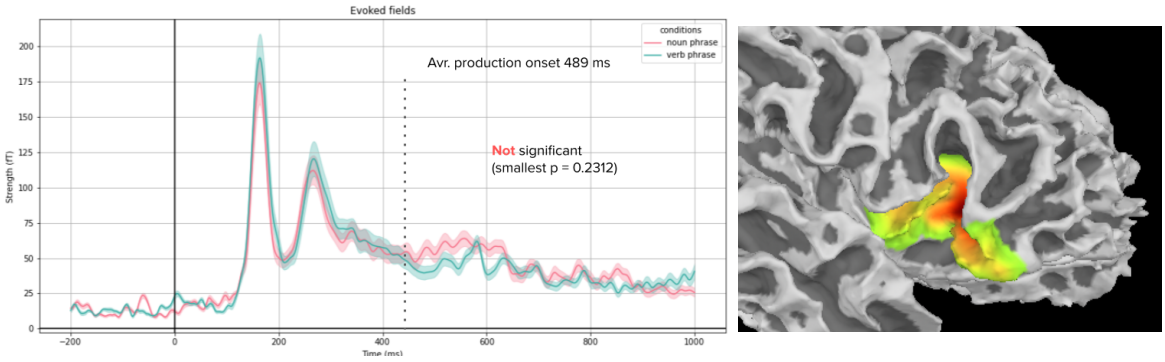

Figure 20: (Left) Direct comparison of averaged evoked activities in the sensor space. The activities are calculated by taking the absolute value and then averaged. (Right) Results of cluster test pairwise t-test on the source space. Single cluster reached close to the the significance threshold in the inferior prefrontal region in the right hemisphere.

C ROI Generation

The regions of interest are generated by selecting parcels from `aparc_sub` parcellation (see section D below). Indices used to select are given here (Note, indices given here are for bihemispheric parcellation. For single hemisphere the indices will differ.)

ROI	Indices for aparc_sub parcellation used
vmPFC	126, 128, 130, 132, 134, 136, 138, 155, 157, 159, 161, 163
Broca	198, 200, 202, 204, 206, 208, 210, 212, 214, 315, 332
LATL	45, 47, 71, 73, 75, 77, 81, 173, 175, 177, 404, 406, 418, 420, 422, 443
PTL	0, 2, 4, 79, 81, 83, 85, 165, 167, 169, 171, 402, 408, 410, 412, 414, 416
IPL	51, 55, 57, 61, 59, 63, 65, 69, 426, 437
SMG	424, 426, 429, 433, 435, 437, 439

D The Parcellation Used for the Whole Brain Analysis

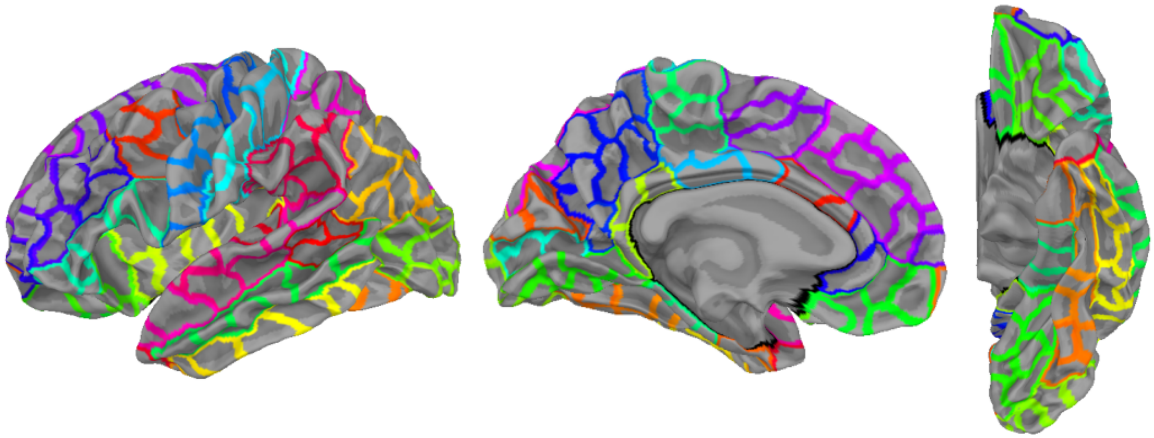

Figure 21: Parcellation used in for the whole brain analysis (Left) Lateral view. (Middle) Medial view. (Right) Ventral view.