Balancing prediction and sensory input in speech comprehension: The spatiotemporal dynamics of word-recognition in context.

Anastasia Klimovich-Gray[a], Lorraine K. Tyler[a], Billi Randall[a], Ece Kocagoncu[a], Barry Devereux[a], and William D. Marslen-Wilson[a]

[a]Centre for Speech, Language and the Brain, Department of Psychology, University of Cambridge, CB2 3EB, UK

Corresponding Author: Lorraine K. Tyler

lktyler@csl.psychol.cam.ac.uk

Abstract

Spoken word recognition in context is remarkably fast and accurate, with recognition times of around 200ms, typically well before the end of the word. The neurocomputational mechanisms underlying these contextual effects are still poorly understood. This study combines source-localised electro- and magnetoencephalographic (EMEG) measures of real-time brain activity with multivariate Representational Similarity Analysis (RSA) to determine directly the timing and computational content of the processes evoked as spoken words are heard in context, and to evaluate the respective roles of *bottom-up* and *predictive* processing mechanisms in the integration of sensory and contextual constraints. Male and female human participants heard simple (modifier-noun) English phrases that varied in the degree of semantic constraint that the modifier (W1) exerted on the noun (W2), as in pairs like *yellow banana*. We used gating tasks to generate estimates of the probabilistic predictions generated by these constraints as well as measures of their interaction with the bottom-up perceptual input for W2. RSA models of these measures were tested against EMEG brain data across a bilateral fronto-temporo-parietal language network. Consistent with probabilistic predictive processing accounts we found early activation of semantic constraints in frontal cortex (LBA45) as W1 was heard. The effects of these constraints (at 100ms post W2 onset in L middle temporal gyrus and at 140ms in L Heschl's gyrus) were only detectable, however, after the initial phonemes of W2 had been heard. Within an overall predictive processing framework, bottom-up sensory inputs are still required to achieve early and robust spoken word recognition in context.

Significance

Human listeners recognise spoken words in natural speech contexts with remarkable speed and accuracy, often identifying a word well before all of it has been heard. In this study we investigate the brain systems that support this important capacity, using neuroimaging techniques that can track real-time brain activity during speech comprehension. This makes it possible to locate the brain areas that generate predictions about upcoming words and to show how these expectations are integrated with the evidence provided by the speech being heard. We use the timing and localisation of these effects to provide the most specific account to date of how the brain achieves an optimal balance between prediction and sensory input in the interpretation of spoken language.

Introduction

Processing spoken words involves the activation of multiple word candidates and competition between them until one candidate becomes uniquely compatible with the incremental speech input (Marslen-Wilson and Welsh, 1978; Kocagoncu et al., 2017; Zhuang et al., 2014). In everyday speech, despite the low predictability of successive words in natural discourse (Luke & Christianson, 2016) this recognition process is strongly affected by the prior context, with recognition times of around 200 ms from word onset, well before all of the word has been heard (Marslen-Wilson, 1975; Marslen-Wilson & Tyler, 1975; Sereno et al, 2003). The neurocomputational mechanisms underlying these powerful and early contextual effects remain unclear and in dispute.

Early bottom-up models assigned a primary role to the speech input in generating an initial cohort of word candidates, with context affecting the selection of the unique word candidate from among this set (Marslen-Wilson, 1987; Tyler and Wessels, 1983). A range of other models allowed contextual constraints to modulate directly the state of potential word-candidates before any of the word was heard (e.g., McClelland and Elman, 1986; Morton, 1969). More recently, the influential predictive coding framework views language comprehension as being driven by an internal generative model which reduces uncertainty about perceptual interpretation by generating probabilistic top-down hypotheses about potential upcoming words (Friston and Frith, 2015; Kuperberg and Jaeger, 2016). These hypotheses, generated in frontal cortex (Sohoglu and Davis, 2016), result in pre-activation of brain areas relevant to the processing of lexical form and content, so that top-down predictions can be compared with upcoming sensory data.

The strength and specificity of such predictions will vary as a function of preceding contextual constraints, with strong constraints generating lexically specific predictions and weaker constraints generating more graded semantic and syntactic predictions (Kuperberg and Jaeger, 2016; Luke & Christianson, 2016). To date, however, almost all research into the neural substrate for predictive processing has used constraining contexts where the target is highly predictable. It remains unclear

specifically where and when predictive constraints apply at more natural levels of constraint, and how these relate to incoming constraints provided by the sensory input.

Here we use source-localised EMEG data, combined with multivariate Representation Similarity Analysis (RSA), to determine directly the timing and the neurocomputational content of the processes evoked as spoken words are heard in variably constraining contexts, reflecting the levels of predictability seen in natural discourse. Participants listened to two word phrases (e.g., *yellow banana*) which varied the specificity of the semantic constraint imposed by the modifier (W1) on the noun (W2). To express these constraints as probability distributions of potential word-candidates we used a gating task in which participants listened to incremental fragments of the word-pairs and produced possible continuations (Grosjean, 1980). Based on these candidate sets, we devised several probes of the neural mechanisms whereby the constraints generated by hearing W1 could affect the processing of W2.

Three measures were based on the candidate sets generated when only W1 had been heard, asking whether different predictive representations of candidate properties were computed as W1 was heard, and how they modulated future processing events. An Entropy model captured the degree of uncertainty about the lexical identity of W2, given W1. A Semantic Similarity model captured the overall semantic dispersion of the predicted candidate sets, probing the representational framework in terms of which the Entropy measure was computed. Thirdly, a Semantic Blend model measured variations in the specific semantic content of the W1 candidate sets. These variations can potentially affect semantically related processing activity in lexical representation areas, both before and after the onset of W2. A fourth model, Entropy Change, reflects the shift in the distribution of predicted W2 candidates given the early perceptual input from W2. This measures the interaction between context-derived constraints and bottom-up perceptual input as W2 is heard. Using RSA, we evaluated these four models against EMEG source-localised activity estimates within an extended bilateral fronto-temporo-parietal language mask.

Methods

Participants

EMEG data acquisition was carried out on 20 healthy participants (9 males) with a mean age of 23.8 (range: 20-34 years). All were right-handed native British English speakers with normal hearing and normal or corrected-to-normal vision. The data for four subjects was discarded due to poor EEG quality. The experiment was approved by the Cambridge Psychology Research Ethics Committee.

Stimuli

The test stimuli consisted of 154 spoken English noun phrases where an adjective modifier (W1) was followed by a noun (W2). All nouns were concrete objects selected from the CSLB norms database (Devereux et al., 2014) with average word form frequency of 14.9 per million, SD=18.3 (CELEX) and mean noun duration of 605 ms, SD=107 ms. In contrast to previous studies which emphasised highly constraining contexts, here we aimed for a moderate overall degree of constraint between W1 and W2. The modifier adjectives were selected to vary the strength of the semantic constraint they exerted on the noun. Constraint strength varied from relatively weak (e.g. '*yellow banana'*) to relatively strong (e.g. '*peeled banana'*) and was quantified by using frequency information from Google ngrams (2007-2008, British English corpus). This allowed us to calculate the log-transformed conditional probability of a phrase ('*yellow banana'*) given the modifier ('yellow'):

$$Constraint = -log(P(C_{ij}|C_i))) = -log\frac{P(C_i \cap C_j)}{P(C_i)}$$

Here C_{ij} is the frequency of a given two-word phrase ('*yellow banana'*) and C_i is the frequency of the modifier ('*yellow*') alone. The average constraint value for the test stimulus set was 1.89, SD 1.06, corresponding to a non-log transformed conditional probability average of 0.08, SD 0.19; min = 0; max = 0.8. This level and range of transitional probability between the word pairs is similar to the results of parallel computations for larger corpora including short narratives (e.g., Luke &

Christianson, 2016). This confirms that the balance of lower and higher constraint in the present stimulus set is consistent with naturally occurring degrees of predictability between successive words.

A separate gating study with 35 participants was also conducted on these word pairs (as described in the next section), from which we could calculate the cloze probability of W2 given W1. This was estimated at 0.17 (SD = 0.29), confirming that the probability of predicting the specific target word (W2) was generally low for this stimulus set.

To reduce the overall proportion of predictive modifiers in the set of word pairs heard by the participants, we also included modifier-noun pairs in which the noun was preceded by an unrelated word so that the 2 words did not form a meaningful phrase – e.g. '*lullaby banana'* (77 filler phrases). These filler items were not analysed. This manipulation brought down the relatedness proportion from 100% to a relatively high 67%, so that the majority of W1 items were still predictive (generally quite weakly) of the properties of the following W2 (but see Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R., 2017)

The presentation order of the phrases was pseudo-randomised and split into three blocks. To minimise the effects of block order, blocks were presented in 6 different orders across participants. The mean duration of phrases was 1213 ms and the duration of each block was approximately 10 minutes.

The stimuli were recorded onto digital audio tape at a sampling rate of 44100 Hz by a female native speaker of British English. All word-pairs were recorded as a single spoken phrase, with no added interval inserted between W1 onset and W2 offset. Recordings were transferred to computer and downsampled to 22050 Hz, 16 bits, mono-channel format using Cool Edit Software (Syntrillium Software Corporation, Phoenix, USA).

Experimental Design and Statistical Analysis

Behavioral pre-testing and Gating

A behavioural pre-test was conducted to rate the naturalness of the test stimulus phrases. 20 participants (mean age 24.8 years, 13 females), who did not participate in the MEG experiment, listened to the stimuli and rated the naturalness of each 2-word test-phrase on a scale of 1-5 (5=high naturalness). The final set of 154 phrases contained only phrases judged to be highly natural (mean=4.32, SD=0.42).

We also ran a behavioural gating study (Grosjean, 1980) with a separate group of 35 participants (mean age 20.6, range 18-35, 13 males) who were instructed (1) to listen to W1 followed by spoken segments of W2 presented incrementally in 50 ms intervals; (2) to type in their best guess of what that word might be after each segment and (3) to rate their confidence (1-7 from least to most confident). We used responses at the offset of W1 (gate 0) to derive measures of Entropy, Semantic Similarity and Semantic Blend. Responses at later gates were used to estimate an Entropy Change model, as well the W2 identification point (IP), defined as the average time (ms) from W2 onset when 80 % participants produced the correct W2 response twice in a row (Grosjean, 1980; Tyler and Wessels, 1983). The mean IP of W2 was 240 ms (SD 125 ms) from word onset.

Cognitive Models for RSA analyses

To determine how the probabilistic distribution of W2 candidates (derived from the gating experiment at the offset of W1) differed according to the strength of W1 constraints, and how these interacted with the processing of W2, we generated four cognitive models, as described below. The multivariate RSA technique then allows us to test for neural activity corresponding to these models against source-localised brain data covering an extended bilateral fronto-temporo-parietal language mask.

Entropy

To capture the shape of the probability distribution of potential W2 candidates for each word-pair, as estimated by gating data obtained at the offset of W1 (gate 0), we calculated a single metric of Entropy, using Shannon's entropy (H) formula:

$$H = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

Here $P(x_i)$ is the summed confidence score for a given W2 competitor at gate 0 across all participants divided by the sum of all confidence scores of all W2 competitors for that item across all participants.

High entropy indicates that there are many possible W2 candidates with low confidence scores, since W1 only weakly constrains upcoming W2 responses. In contrast, low entropy indicates that most participants in the gating task selected only one or two specific word continuations, reflecting stronger W1 constraints on W2 for that word-pair.

These computations of probability distributions are relevant to the predictive processing approach in two ways. First, this approach requires that such representations (viewed as hypotheses about the future properties of the sensory input) are computed incrementally as the input is interpreted, so that they can modulate expectations of the properties of this input at the relevant levels of neural description. Second, the pattern of variation in these probability distributions should modulate these patterns of neural activity before the onset of W2 as well as after. A bottom-up priority model, in contrast, predicts that such effects can only be seen after W2 onset, when an initial cohort has begun to be established.

Semantic Similarity

The Entropy measure reflects the probability distribution of word candidates produced at the offset of W1 for each phrasal modifier. This measure, however, simply tabulates the number (and

associated confidence) of the word candidates generated. It is likely, however, given the low average transitional probability of the modifier-noun pairs used here, that the words generated for most W1 modifiers reflected broad semantic and syntactic constraints on possible W2 continuations rather than specific hypotheses about lexical forms.

To measure the semantic distributional properties of these candidate sets we calculated the pair-wise similarity of the same sets of words, using a corpus-based Distributional Memory database (Baroni and Lenci, 2010). This represents words as vectors over 5000 semantic dimensions, where these dimensions are distilled from word co-occurrence data. The Semantic Similarity of the candidates for each phrasal modifier was defined as the average pairwise cosine similarity between vectors for all its gate 0 word candidates. The stronger the semantic constraint that the modifier places on the following word the less semantic dispersion there will be in the set of candidate words that it evokes. After the modifier 'cashmere', for example, gating participants tended to produce words from the semantic category 'clothing'– e.g. 'sweater', 'scarf'. For a less constraining modifier like 'massive' there was much more variability in the semantics of suggested words – 'car', 'tower', 'blow'.

If the probability distributions for W2 candidates are well captured by the semantic dispersion between candidates, then model fit should parallel the results for the Entropy measure, since they would both reflect the semantic constraints generated by the W1 modifier. Further, if the informational substrate for computing constraints is primarily semantic/distributional in nature and represented in middle temporal cortices, then any top-down modulation involving frontal regions (e.g. Musz and Thompson-Schill, 2017) may primarily affect lexical semantic representations rather than phonological representations in STG and HG (Binder et al., 2000;)

Semantic Blend

Entropy and Semantic Similarity models capture the overall properties of the distribution of expected word candidates – the shape of the candidate distribution and its dispersion in the semantic domain. However, they do not directly test the retrieval of the predicted candidate-specific semantic content of W2. To test when, with respect to W2 auditory onset, and in which brain regions such information is accessed, we derived the Semantic Blend model.

The Distributed Cohort Model of speech perception (Gaskell and Marslen-Wilson, 1997; 2002) assumes that lexico-semantic information associated with any given item involves a distributed activity pattern across the lexico-semantic representational space. When multiple word candidates are accessed the resulting pattern of activation should reflect this ambiguity and encode a 'blend' of overlapping representations, where lexico-semantic features shared across multiple candidates dominate the activation pattern. To derive Semantic blends for each stimulus ('e.g. *yellow banana'*) we first normalised the semantic vectors (Baroni and Lenci, 2010) obtained from the gating responses collected at gate 0. This was to prevent outlying values for a given candidate from dominating the blend. These normalised (by vector length) candidate vectors were averaged to produce a single 'blend' vector, weighting each vector by its associated confidence score. The resulting blend therefore mixed together the semantics of every W2 candidate, with a more coherent blend reflecting greater degrees of W1 constraint. In terms of evaluating the balance between predictive and bottom-up inputs under the relatively weak levels of constraint present in this study, we ask here whether these variations in the predictability of W2 candidate semantics will be reflected before W2 onset in the patterns of neural activity in temporal lobe areas responsible for the representation of lexical semantics, or whether they will require the additional constraints provided by the initial phonemes of W2.

Entropy Change

The above models capture potential predictive constraints that can be generated once W1 has been recognised and before W2 onset. To examine how these W1 constraints interact with the early

perceptual input for W2, we subtracted the entropy values calculated at gate 0 (as in the Entropy model) from the entropy values calculated from the candidates generated at gate 1, 50ms after the onset of W2. At Gate 1 the first phoneme of W2 will in most cases be identifiable, so that the candidates generated could potentially be constrained both by bottom-up phonological cues and by the constraints derived from W1. There was a significant drop ($p<0.001$, paired t-test) in Entropy at gate 1 (mean 1.78), compared to gate 0 (mean 2.14).

On a predictive processing approach, the bottom-up constraints available early in W2 should serve to update expectations about existing W2 lexical candidates, with corresponding reductions in lexical uncertainty. This in turn could lead to model fit in brain areas responsible both for computing candidate probabilities and for using these hypotheses to modulate incoming perceptual analyses. The timing of these activations is uncertain but may require more input than just the first 50ms of the word in question.

Procedure

The auditory stimuli were delivered binaurally through MEG-compatible ER3A insert earphones (Etymotic Research Inc., IL, USA). The instructions to the participants were visually presented on a monitor screen positioned 1m in front of them. They were told to listen attentively to the word pairs. To encourage them to do so, on 10% of the trials the spoken phrase (e.g. *school bus*) was followed by a single written word in the middle of the screen (e.g. *children?*). Participants were instructed to judge the semantic compatibility of the written word with the preceding phase and to give a yes/no answer via button presses. These semantically-related question trials were followed by a standard spoken phrase trial, which was treated as a dummy and not included in the analyses.

E-Prime Studio version 2 (Psychology Software Tools) was used to present the stimuli and record participants' responses. The experiment began with a short practice run, which was followed by three experimental blocks. The duration of the entire experiment was approximately 40 minutes.

Before each spoken stimulus presentation, a cross appeared in the middle of the screen for 650 ms prompting the participant to focus their eyes on the cross. The inter-trial interval (ITI), measured from the offset of the spoken phrase, was jittered between 3000 and 4000 ms. Note that the duration of the ITI for the infrequent 'question trials' was the same overall as for the standard trials. To reduce the potential contamination of the EMEG recordings by blink artefacts during the test phase, the last 1500 ms of each ITI was made up of a 'blink break', during which participants were encouraged to blink their eyes. The start of the blink break was indicated by an image of an eye that appeared in the middle of the screen. Participants were also asked to refrain from movement during the entire block of recording.

EMEG recording

Continuous MEG data was recorded in a magnetically shielded room (IMEDCO GMBH, Switzerland) at the MRC Cognition and Brain Sciences Unit in Cambridge, using the Vector View system (Elektra-Neuromag, Helsinki, Finland) containing 102 magnetometers and 204 planar gradiometers, arranged within a helmet covering the head of the subject. The position of the head relative to the sensors was monitored using the Head Position Indicator (HPI) coils attached to the subject's head. EEG was recorded simultaneously from 70 Ag-AgCl electrodes within an elastic cap (ESACYCAP GmbH, Herrsching-Breitbrunn, Germany) on the subject's head. Vertical and Horizontal EOGs were also recorded for blink detection. A 3D digitizer was used to record the position of the EEG electrodes and the HPI coils and approximately 100-150 head points along the participants' scalp, relative to the 3 anatomical positions (the nasion and the left and right preauricular points). Acquired data was sampled at 1kHz and band-pass filtered from 0.03 to 330 Hz.

Data Pre-Processing

The raw data were processed using MaxFilter 2.2 (Elekta Oy, Helsinki, Finland). Static magnetometer and gradiometer bad channels were identified and reconstructed using interpolation. Temporal

extension of the signal space separation technique (Taulu, Simola, and Kajola, 2005) was applied to separate the external noise from the head-internal signals. Finally, correction for head movement across the blocks was applied and each subject's data were transformed to a default head position. The EEG raw data was then manually inspected and noisy EEG channels were removed.

The raw data was low pass filtered to 200 Hz and notch filtered at 50 Hz (to remove the mainline frequency components). Independent component analysis was applied to the MEG and EEG data separately for further de-noising (MNE python – Gramfort et al., 2013). The ICA components most strongly correlating with the vertical EOG channel were identified and removed. For the current analysis raw data was then further band-pass filtered 0.1 to 40 Hz, all previously identified bad channels were interpolated.

The data file for each test trial was then epoched with respect to the auditory onset of W2. The W2 onset epoch was 500 ms in length (range: -200 to 300 ms, aligned to W2 auditory onset). During epoching the data was baseline corrected using 100 ms of pre-stimulus data that did not contain any speech signal. Trials with large artefacts were considered noisy and were removed (EEG>150µV, magnetometer>5e-12T gradiometer>2000e-13T/cm).

Source Reconstruction:

The pre-processed data was source localised by using the minimum norm estimate (MNE) procedure (MNE Python - Gramfort et al., 2013) based on distributed source modelling (Lin et al., 2006). MNE constrains the sources of currents by using *a priori* assumptions about their distributions (dipole orientation and location summarised in the lead field matrix, derived from the structural MRI scans) and the noise estimates covariance matrix.

The first step in source localisation was the acquisition of structural MRI images of each participant using the GRAPPA 3D MPRAGE sequence (time repetition = 2250 ms; time echo = 2.99 ms; flip angle = 9; acceleration factor = 2) on a 3-T Trio scanner (Siemens, Erlangen, Germany) with 1 mm isotropic

voxels. MRI structural images were processed with FreeSurfer software (Fischl, 2012) to parcellate brain volumes into inner and outer skin and scull, white and grey matter. Subsequent steps were performed using the MNE-Python environment. The source grid was set up on the white-grey matter boundary surface and downsampled to 4098 sources per hemisphere. The MRI and MEG coordinate systems were co-registered using the MNE analysis interface, with respect to the anatomical locations marked during acquisition (the nasion and the left and right preauricular points) and additional 100-150 head points.

Second, for each subject the forward model (lead field matrix) was created using a 3-layer boundary element model (BEM) that assigned different electrical conductivities to inner and outer surfaces (5120 triangles per surface) of the skull. A regularised covariance matrix was estimated from the epochs using the baseline period. The forward solution and the covariance matrix were used to estimate the linear inverse regularisation parameter (inverse operator) for every source across all channels. To improve the spatial accuracy of the localisation and correct for a bias towards assigning signals to superficial sources, a loose source orientation (0.2) constraint and a depth constraint (0.8) were applied (Lin et al., 2006). To derive the source estimates at every time point and for every trial the inverse operator was applied to pre-processed data by taking the norm of the dipole components. The estimated activations were normalised with respect to signal noise by dividing the estimates by their predicted standard error, thus producing unsigned dynamic statistical parametric maps (dSPM, Dale et al., 2000). To account for increased noise levels in the single-trial estimates the $lambda^2$ parameter was set to 1. The subject-specific estimates were then morphed to the averaged brain surface (produced with FreeSurfer) for further statistical analysis.

Representational Similarity Analysis (RSA) analysis

RSA analysis makes it possible to compare directly the correlational structure in patterns of brain activity with the correlational structures predicted by different cognitive models (Kriegeskorte et al., 2008). We used a sliding window RSA analysis, in which each subject's data Representational

Dissimilarity Matrices (RDMs) across the length of the epoch were compared to cognitive model RDMs within every ROI (see Figure 1 for a schematic overview of the RSA analysis and parameters used). Distinct cognitive model RDMs were derived for each cognitive measure (matrix size 154x154, reflecting the number of test phrases). The Entropy RDM was defined by taking the absolute pair-wise differences between the lexical Entropy scores for all word-pairs. The Semantic Similarity model RDM was produced by taking the absolute pair-wise differences of the Semantic Similarity scores associated with each word-pair. The Semantic Blend model RDM was computed by taking the pairwise cosine distances between every pair of 'blended' semantic vectors (of gate 0 predicted candidates for each word-pair). The Entropy Change RDM was calculated by taking the absolute pair-wise differences in the Entropy Change measure. Group-level data was derived by extracting subject-specific model-fit r-values and conducting a one-sample t-test across subjects at each ROI and time-point. Only t-value clusters that survived this initial $p<0.01$ threshold and then the correction for multiple comparisons (cluster-permutation permutation analysis, Maris and Oostenveld, 2007; Su et al., 2012) at $p<0.05$ threshold are reported. We did not apply an additional correction for the number of ROIs tested.

The cognitive models outlined above (converted into model RDMs) were tested against activity patterns (captured in subject-wise data RDMs) in a set of bilateral fronto-temporal ROIs [see inset brain in Fig 1, ROIs taken from the Desikan-Killiany Atlas, Freesurfer] consisting of BA44, BA45, BA47, Heschl's gyrus (HG), posterior superior temporal sulcus (pSTS), supramarginal gyrus (SMG), inferior-parietal area (IPA, that included the angular gyrus), temporal pole (TP). Temporal ROIs – superior temporal gyrus (STG), middle temporal gyrus (MTG) and inferior temporal gyrus (ITG) - were split into mid-posterior and anterior parts and the latter joined to form a functional anterior temporal lobe (ATL) ROI. These ROIs were selected on the basis of previous studies showing their involvement in processing spoken language (Hagoort, 2013; Hickok and Poeppel, 2007; Kocagoncu et al., 2017; Tyler et al, 2013).

Finally, to assess the independence of the model-fit results for each RSA model from the other three models, we ran a set of additional partial correlation analyses. For each model separately the model-fit (r) was estimated after a partial correlation analysis where the contribution of the other relevant models was partialled out. The rest of the procedure was identical to the main ROI RSA analysis described above. The results of these analyses are reported alongside the results for each model tested separately.

Results

The questions at issue here concern the role of the constraints generated by W1 in the perception and identification of W2: Whether or not the context (the W1 modifier) triggers access to a distribution of potential word (W2) candidates, what is the timing of this activation relative to W1 processing, whether and when this information affects the processing of W2, and how it interacts with the bottom-up constraints made available as W2 is heard. Three models – Entropy, Semantic Similarity and Semantic Blend – encoded different properties associated with the probabilistic distribution of W2 candidates elicited at W1 offset, while the Entropy Change model tested for potential interactions between top-down contextual constraints and the early analysis of the perceptual input for W2 (for details see Cognitive Models section above). For all 4 models the analyses were aligned to W2 onset, looking for potential model fit both before and after the start of the critical word (see Figure 2). While the average Identification Point for W2 (as estimated from the gating data) falls 240 ms after W2 onset, we expect to see much earlier effects on model fit as lexical candidates begin to be selectively activated by the incoming speech.

For all of the models reported below, the size of the model-fit r-values falls within the range of values obtained in comparable RSA analyses reported elsewhere (e.g., Devereux, Clarke, & Tyler, 2018; Nili et al., 2014).

The Entropy model (see Fig. 2A) showed significant model fit in L BA45, starting -70 ms before W1 offset and persisting until +165ms after W2 onset: (1) from -70 ms to +15 ms ($p<0.001$); (2) from +35 to +80 ms ($p<0.001$); and (3) from +120 ms to +165 ms ($p=0.007$). The Semantic Similarity model (Fig 2A) closely paralleled the Entropy model, also showing model fit only in L BA45, and also beginning at -70 ms before W1 offset though finishing slightly earlier at +95 ms after W2 onset. The significant clusters were: (1) from -70 ms to -45 ms ($p=0.02$); (2) from -5 ms to +15 ms ($p=0.03$); and (3) from +65 ms to +95 ms ($p=0.01$). The significant fit for both of these models starts after the W1 Identification Point - on average 194 ms before W2 onset - and ends before W2 is recognised – on average 240 ms after W2 onset. These results show that the number and probability (Entropy model) and the semantic dispersion (Semantic Similarity) of the W2 candidate distributions corresponded to patterns of neural activity that were activated after the modifier word had been recognised but before the onset of W2, and remained relevant to L inferior frontal processes well into the perceptual analysis of W2.

The close parallels between the timing and the location of the model fits elicited by the Entropy and Semantic Similarity models are consistent with the strong correlation (0.79) between their respective model RDMs, and with the claim that they both derive from the same underlying probabilistic source. The results of the partial correlation analysis confirm that the model fits for Entropy and Semantic Similarity are not independent – model fit in BA45 disappears across the board if Entropy is partialled out when estimating the effects of Semantic Similarity (and conversely for Semantic Similarity when Entropy is partialled out).

Turning to the Semantic Blend model, which tested for the processing relevance of the semantic content of the W2 candidates, we see a markedly different spatio-temporal pattern of model fit (Fig 2B). No effects were found prior to W2 onset, and only left temporal locations were significant, with model fit in L middle and posterior MTG (Fig 2B). The model fit was significant from 100 to 160 ms after W2 onset; initially from 100 to 120 ms ($p=0.02$) and then from 130 to 160 ms ($p=0.03$). These

results show that semantic information about potential W2 candidates was accessed in MTG soon after the onset of W2 and before the word was uniquely identified. Furthermore – and consistent with the weak correlations between the Semantic Blend model RDM and the other cognitive model RDMs - the model fit remains significant in the partial correlation analysis with unchanged locations and almost identical time-courses ($p=0.04$ for a first cluster from 100-115 ms and $p=0.02$ for a second cluster from 130-155 ms) after partialling out the other three cognitive models. This implies that the Semantic Blend model captures processes distinct from those of Entropy and Semantic Similarity.

Fourthly, the Entropy Change model (Fig 2C), which captures the update in the W1 Entropy model after 50 ms of W2 perceptual input had been heard, produced significant model fit in LH HG, from 140 ms to 180 ms ($p= 0.025$) after the perceptual onset of W2. These effects again proved to be independent from the other three models, as confirmed through partial correlation analysis where a similar significant cluster ($p=0.035$) was observed between 145 to 165 ms after W2 onset, consistent with the weak correlations between the Entropy Change model RDM and the other cognitive models.

A different set of partial correlation analyses were conducted to test the robustness of the model fit observed after W2 onset for the three 'prediction' models (Entropy, Semantic Similarity and Semantic Blend) against potentially uncontrolled variation in the bottom-up phonological and semantic properties of the W2 stimulus set. To address this issue, two additional models were constructed. To capture W2 semantics we calculated pair-wise cosine distances between the semantic vectors (Baroni and Lenci, 2010) for the W2 targets (as in Kocagoncu et al, 2017). For a phonological model we converted the first two phonemes of each W2 into a binary articulatory feature vector (Wingfield et al., 2017), calculating pair-wise cosine distances between each such vector to derive the model RDM. Separate partial correlation analyses were run for each cognitive model using the W2 semantic and phonological RDMs separately (the Entropy Change model was

not included since this was already based on the integration of W1 predictions and bottom-up W2 constraints). For the critical period of interest from the onset of W2, we see no evidence that the timing and significance of model fit for the cognitive models expressing constraints derived from W1 is significantly confounded with W2 stimulus properties. There is no general weakening of the significance of the model fits obtained, nor is there any change in the basic timing of the relevant effects in relation to W2 onset for the three models.

Finally, to provide more fine-grained information about the spatio-temporal distribution of the model fit for each cognitive model, we also conducted searchlight analyses across the entire language mask bilaterally. The results of these exploratory analyses confirm that the ROI analyses reported in Fig 2 correctly identified the time-periods and locations exhibiting consistent model fit for each model. No other spatially and temporally consistent clusters were visible in the searchlight data.

Discussion

This study addresses the specific neurobiological processes that underpin the dynamic and early integration of top-down and bottom-up constraints in the perceptual interpretation of spoken words heard in constraining contexts. We investigated these processes for a set of spoken two-word English phrases, presented in a minimal task 'attentive listening' environment, and sampling a naturalistic range of degrees of constraint. The average predictability of the second word in the phrase was low, so that the available constraints were in general lexically nonspecific and broadly semantic in nature.

Using gating data we generated four computational measures that captured specific properties of the probabilistic distribution of potential word candidates, where the combination of spatiotemporally well-resolved EMEG measures of dynamic brain-activity, together with RSA multivariate techniques, made it possible to use these models to probe the neurocomputational

content of processing activity as spoken words are heard. This provides a novel and revealing perspective on the dynamic functional architecture underpinning the integration of contextual prediction and sensory constraint in speech comprehension.

Bottom-up constraints in a predictive processing framework

Overall, the results suggest a system for dynamically combining bottom-up and top-down constraints in speech interpretation that shares key characteristics of predictive processing with a strong dependence on the incoming speech input – necessarily so under conditions of weak contextual constraint. Consistent with the predictive processing approach, as soon as listeners recognised the W1 context word they began to generate estimates of the probability distributions associated with potential W2 candidates. Significant RSA model fit for both Entropy and Semantic Similarity models was seen 70ms before W2 onset (see Figure 2A).

These probability estimates – viewed as potential hypotheses about the upcoming word (W2) – have several notable properties. The first is that model fit was seen only in L BA45 and not in any of the potential target regions for top-down modulation – such as auditory processing areas (HG and STG) or lexical content regions in posterior and middle MTG. While the processing role of BA45 is not fully understood, it is widely thought to be involved in semantic control processes (Novick et al., 2005; Musz and Thompson-Schill, 2017), and provides a plausible substrate for the computation of probabilistic semantic constraints (as indicated by the joint Entropy and Semantic Similarity model fit). We found no evidence, however, that the variations in constraint *per se* across word pairs being computed in BA45 were directly modulating neural patterns elsewhere in the language system.

A further significant property of the BA45 model fit is that it persisted well into the processing of W2, only dropping below significance at 165ms after W2 onset for the Entropy model, and at 95ms for the Semantic Similarity model. The probability estimates for W2 possible candidates based on W1 constraints, as computed by BA45, were apparently not modulated by bottom-up information

about actual W2 candidates until at least the first two phonemes of W2 had been identified. In a predictive processing framework, where top-down hypotheses modulate the perceptual interpretation of the incoming sensory input, any mismatch between hypothesis and input is fed back, in the form of prediction error, to refine top-down hypothesis formation. The relative lateness with which W1 hypothesis sets continue to fit patterns of neural activity in BA45, together with the evidence that BA45 is operating in semantic terms, leads to the inference that the required feedback to BA45 is in terms of the semantic properties of actual W2 candidates and that the availability of this information depends on bottom-up cues to the identity of these candidates.

The claim that the operations of BA45 are primarily semantic in nature is consistent with the observed close similarity between the Entropy and Semantic Similarity models, not only in the model fit they exhibit over time (Fig 2A), but also the close relationship of their respective RDMs in the partial correlation and cross-correlation measures. However, this lack of independence of the two models means that further research is still needed to confirm their specific shared underlying properties.

Supporting evidence for the role of W2 sensory input in enhancing access to candidate semantics comes from the Semantic Blend model (Fig 2B), which captures variations in the semantic coherence of the sets of W2 candidates generated at gate 0. This shows a very different pattern of model fit to the Entropy and Semantic similarity models. There is no model fit prior to W2 onset, and none is seen until 100ms into W2. The localisation of this to L middle and posterior MTG is consistent with evidence that these regions support the processing of lexical semantic information (Binder et al., 2000; Hagoort, 2013; Hickok and Poeppel, 2007), but begs the question of why these regions only generate model fit well after W2 onset.

The timing and location of the Semantic Blend model fit make several points. First, the significant fit for predicted W2 semantics in L MTG is indeed likely to reflect an interaction between semantic hypotheses generated in BA45 and the processing of potential lexical candidates in relevant brain

regions. However, this interaction only becomes neurocomputationally visible at 100ms after W2 onset, once the bottom-up input has started to generate lexical candidates whose semantic properties overlap with those of W1-generated constraints. This also shows that the bottom-up activation of lexical contents occurs very early in the word; as early as access to lexical form (Kocagoncu et al, 2017; Marslen-Wilson, 1975). Third, the interaction is short-lived, terminating at 160ms after onset, suggesting that the probabilistic constraints provided by W1 are rapidly superseded by new information from W2. The timing here is consistent with the termination of model fit at 165ms in BA45 for the W1 Entropy model, similarly displaced by new W2-based constraints.

Finally, the results for the Entropy Change model provide further evidence for the timing and the consequences of the interaction between bottom-up and contextual constraint. This model captures the reduction in lexical uncertainty generated by the integration of W1 constraints with the constraints provided by the initial 50ms of W2. The resulting model fit is seen – with some delay - at 140 ms after W2 onset, and is located in L Heschl's Gyrus - a brain region that supports the bottom-up perceptual analysis of the auditory input (Scott and Johnsrude, 2003; Uppenkamp et al, 2006; Warrier et al., 2012). The timing of this effect matches the time-course indicated by the Semantic Blend results for the integration of semantic and phonological constraints, and suggests that convergence on specific word candidates is largely complete within 150-200ms of word onset - consistent with behavioral evidence for the timing of word-recognition in context (Marslen-Wilson, 1973; 1975). The location of the Entropy Change model fit in auditory cortex suggests, finally, that the interaction of bottom-up phonological cues with contextual semantic constraints is able to modulate activity in primary sensory processing regions as candidate words are heard, consistent with predictive processing claims (Sohoglu & Davis, 2016). However, at the generally weaker levels of constraint provided by the word-pairs used here, the role of bottom-up perceptual evidence seems critical.

Conclusions

Research into the neural substrate for predictive processing in language comprehension has emphasized stimulus sequences where predictive constraints are strong and lexically specific, and delivered in 'prediction-friendly' experimental situations. In the current study, listeners heard words in less constraining contexts, arguably more representative of everyday language. In this situation, an 'all-or-nothing' predictive processing regime cannot hold – where contextual constraints predict the actual sensory forms of future inputs. Our word pairs were rarely sufficiently constrained to support such predictions.

We see instead a neurocognitive system where the basic components of a predictive processing framework are present, but where its forward predictions are tailored to the types of constraint that are at hand (largely semantic), and where bottom-up constraints are essential to the formation of phonologically and semantically specific perceptual hypotheses. There is evidence for early computation of probabilistic constraint representations, well before W2 onset, but we see no corresponding evidence for early modulation of neural activity in brain regions relevant to lexical form and content. It is only after information about the initial phonetic properties of W2 becomes available that we can detect interactions between W1 constraints and W2 interpretation – at 100ms for the Semantic Blend model and at 140ms for the form-specific predictions picked up by the Entropy Change model. The timing of the integration of these contextual and sensory constraints during W2 processing, while consistent with the rapid recognition of words in context, demonstrates that the sequential statistical properties of natural language require continuous contact with the sensory input to achieve the robust earliness of human real-time speech interpretation.

ACKNOWLEDGMENTS

This work was supported by a European Research Council Advanced Investigator grant to LKT under the European Community's Horizon 2020 Research and Innovation Programme (2014-2020 ERC Grant agreement no 669820), and an Isaac Newton Trust Research Grant (2017 Grant 15.40(k) to LKT).

Bibliography

Baroni M, Lenci A (2010) Distributional Memory: A General Framework for Corpus-Based Semantics. Comput Linguist 36:673–721.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex 10:512–28.

Dale AM, Liu AK, Fischl BR, Buckner RL, Belliveau JW, Lewine JD, Halgren E (2000) Dynamic Statistical Parametric Mapping. Neuron 26: 55–67.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2017) Comprehenders rationally adapt semantic predictions to the statistics of the local environment: a Bayesian model of trial-by-trial N400 amplitudes. 39th Annual Conference of the Cognitive Science Society, London, UK.

Devereux, B.J., Clarke, A.D., & Tyler, L.K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports, V*olume 8, Article number: 10636

Fischl, B. (2012). FreeSurfer. NeuroImage 62: 774-781

Friston KJ, Frith CD (2015). Active inference, communication and hermeneutics. Cortex 68: 129–143.

Gaskell MG, Marslen-Wilson WD (1997) Integrating Form and Meaning: A Distributed Model of Speech Perception. Language and Cognitive Processes 12: 613–656.

Gaskell MG, Marslen-Wilson WD (2002) Representation and competition in the perception of spoken words. Cognitive Psychology 45: 200-226.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Hämäläinen M (2013) MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience 7:1–13.

Grosjean F (1980) Spoken word recognition processes and the gating paradigm. Perception and Psychophysics 28: 267–283.

Hagoort P (2013) MUC (Memory, Unification, Control) and beyond. Frontiers in Psychology 4:1-13

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nature Neuroscience 8:393-402.

Kocagoncu E, Clarke A, Devereux B J, Tyler LK (2017) Decoding the Cortical Dynamics of Sound-Meaning Mapping. The Journal of Neuroscience 37: 1312–1319.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2:1-28.

Kuperberg GR, Jaeger TF (2016) What do we mean by prediction in language comprehension? Language Cognition and Neuroscience 31:32-59.

Lin FH, Witzel T, Ahlfors SP, Stufflebeam SM, Belliveau JW, Hämäläinen MS (2006) Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. NeuroImage 31:160-171.

Luke, SG, Christianson, K (2016) Limits on lexical prediction during reading. Cognitive Psychology 88: 22-60.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods 164:177–190.

Marslen-Wilson WD (1973) Linguistic structure and speech shadowing at very short latencies. Nature 244:522-523

Marslen-Wilson WD (1975) Sentence perception as an interactive parallel process. Science 189: 226-228

Marslen-Wilson WD (1987) Functional parallelism in spoken word-recognition. Cognition 25:71–102.

Marslen-Wilson WD, Welsh A. (1978) Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology 10: 29–63.

Marslen-Wilson WD, Tyler LK (1975) Processing structure of sentence perception. Nature 257: 784-786

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cognitive Psychology 18:1–86.

Morton J (1969) Interaction of information in word recognition. Psychological Review 76:165–178.

Musz E, Thompson-Schill SL (2017) Tracking competition and cognitive control during language comprehension with multi-voxel pattern analysis. Brain and Language 165:21–32.

Nili H., Wingfield C., Walther A., Su L., Marslen-Wilson WD., Kriegeskorte N. (2014). A toolbox for representational similarity analysis. PLoS Computational Biology, 10, e1003553.

Novick JM, Trueswell JC, Thompson-Schill S (2005) Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. Cognitive, Affective, & Behavioral Neuroscience 5:263-281.

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends in Neurosciences 26:100–7.

Sereno SC, Brewer CC, O'Donnell PJ (2003) Context effects in word recognition: evidence for early interactive processing. Psychol Sci 14: 328–333.

Sohoglu E, Davis MH (2016) Perceptual learning of degraded speech by minimizing prediction error. Proceedings of the National Academy of Sciences 113:1747–1756.

Su L, Fonteneau E, Marslen-Wilson WD, Kriegeskorte N (2012) Spatiotemporal searchlight

representational similarity analysis in EMEG source space. International Workshop on Pattern Recognition in NeuroImaging (PRNI), pp 97–100.

Taulu S, Simola J, Kajola M (2005) Applications of the signal space separation method. Signal Processing 53: 3359–3372.

Tyler LK, Cheung TPL, Devereux BJ, Clarke A (2013) Syntactic computations in the language network: characterizing dynamic network properties using representational similarity analysis. Frontiers in Psychology 4:1-19.

Tyler LK, Wessels J (1983) Quantifying contextual contributions to word-recognition processes. Perception and Psychophysics 34:409–420.

Uppenkamp S, Johnsrude IS, Norris D, Marslen-Wilson WD, Patterson RD (2006) Locating the initial stages of speech-sound processing in human temporal cortex. NeuroImage 31: 1284-1296.

Warrier C, Wong P, Penhune V, Zatorre R, Parrish T, Kraus N (2012) Relating structure to function: Heschl's Gyrus and acoustic processing. JN 29: 61–69.

Zhuang J, Tyler LK, Randall B, Stamatakis EA, Marslen-Wilson WD (2014) Optimally efficient neural systems for processing spoken language. Cerebral Cortex 24: 908–918.

Figure Legends

Figure 1: Schematic overview of RSA multivariate analysis procedures

(1) Representational Dissimilarity Matrixes (RDMs) are derived for each ROI and each subject across the length of the analysis epoch. These **Data** RDMs summarise the differences between the activation patterns (1-r) between trials in a given time window (20 ms in width) every 5 ms across all vertices of a given ROI. (2) Each data RDM is compared with **Model** RDMs (Spearman r), derived separately for each cognitive measure (see text). To produce group-level statistics a one-sample t-test is taken for each ROI and each time point across subjects. The resulting t-maps were thresholded at the $p<0.01$ level. To correct for multiple comparisons, values surviving the primary threshold were entered into a cluster-permutation analysis (1000 permutations, Maris and Oostenveld, 2007; Su et al., 2012). Only clusters that survived the $p<0.05$ cluster-correction threshold are reported (marked here by the red bar).

Figure 2. Spatio-temporal coordinates of RSA model fits for the W1 modifier-based constraints.

2A, 2B, 2C: Left: The ROIs (in green on the inflated MNE brain) producing significant model fit for the four models. Right: (**2A**) Subject-averaged model fit r-value in L BA45 across the analysis epoch (-200 ms to + 300ms relative to W2 onset) for the Entropy (green) and the Semantic Similarity (pink) models; (**2B**) Semantic Blend (blue) model fit in L middle and posterior MTG; (**2C**) Entropy Change (orange) model fit in L HG. Shaded areas indicate the SE of the model-fit means and the thick lines below the r-curves show the significant t-value model-fit clusters (corrected for multiple comparisons at $p<0.05$). The vertical broken red line marks W2 onset; (**2D**) (Left) Figure Key specifying the mean and SD of the critical time-points overlaid (Right) on the auditory waveform of a sample test phrase.

Figure 1

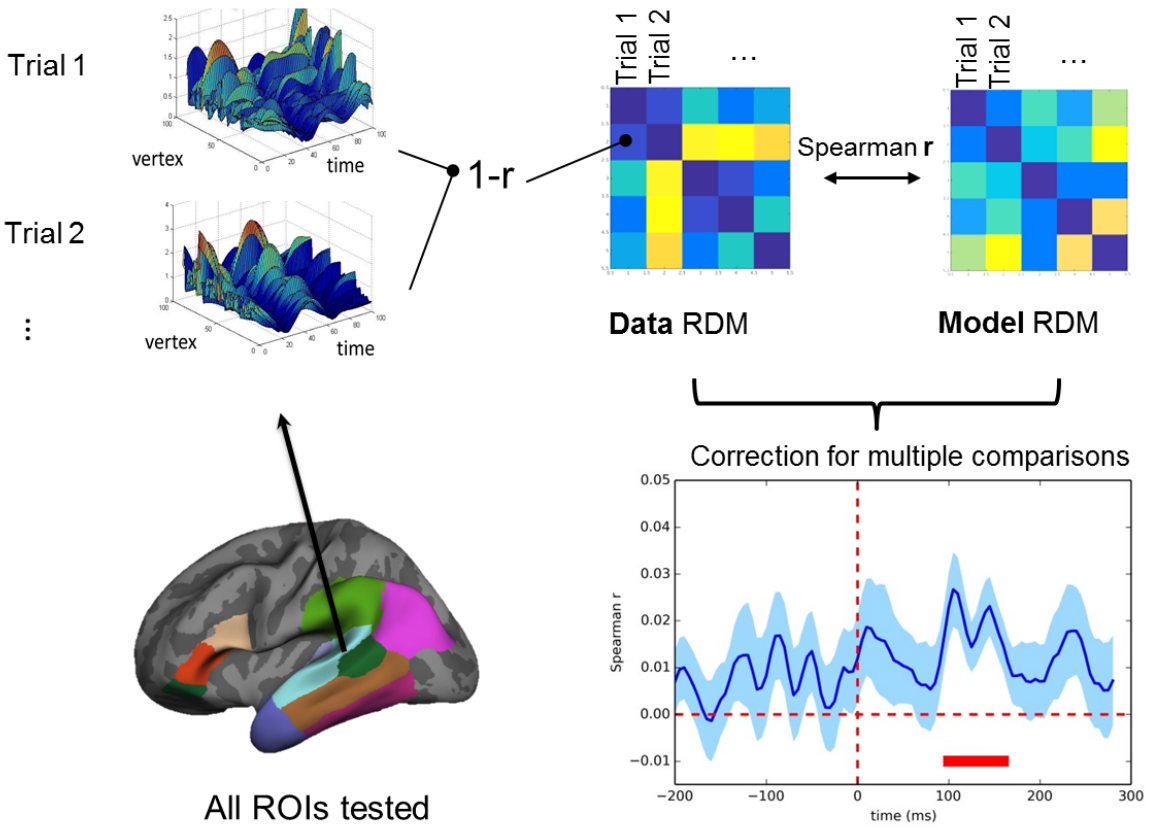
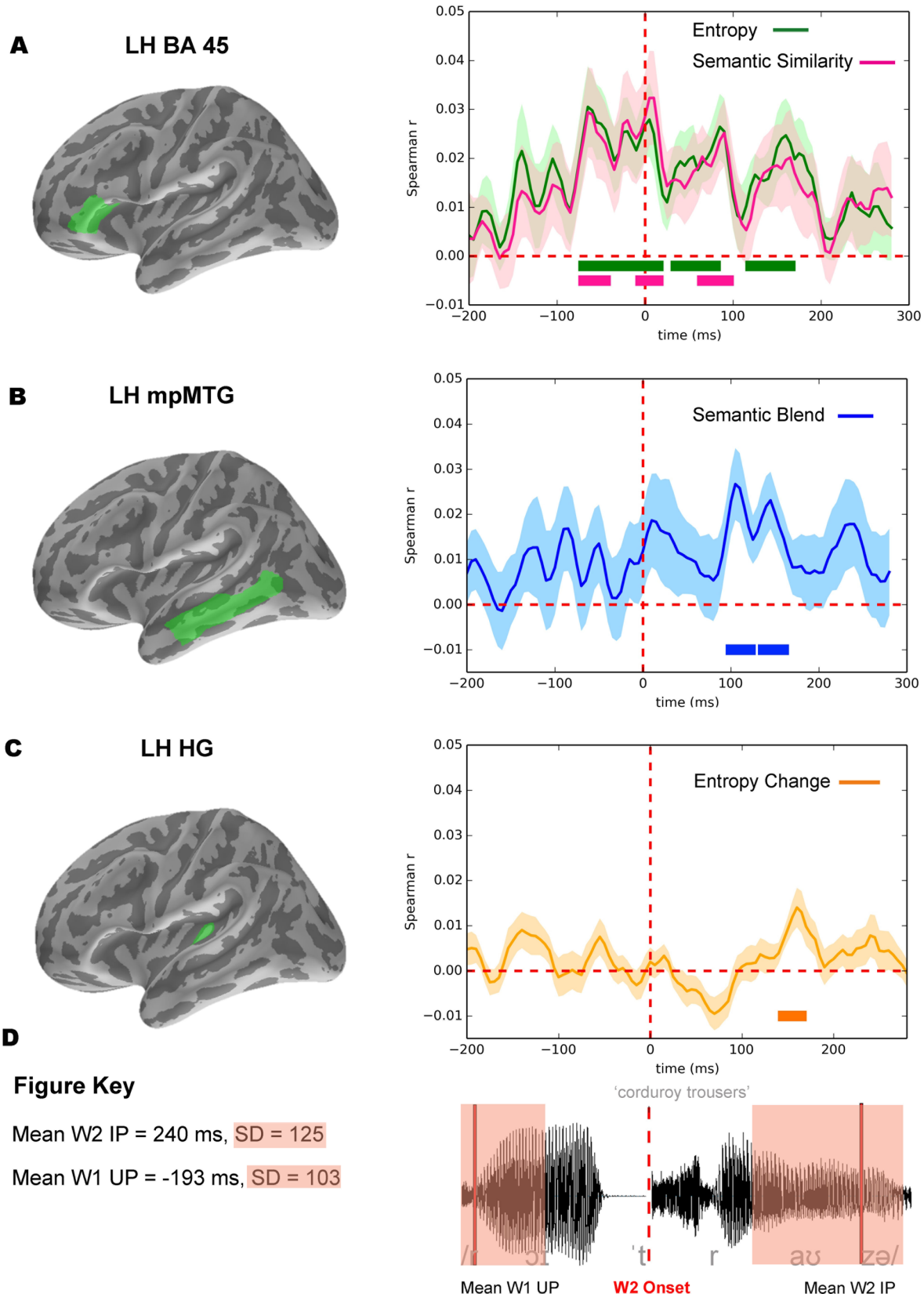

Trial 1
Trial 2
vertex
time
vertex
time
1-r
Trial 1
Trial 2
...
Spearman r
Trial 1
Trial 2
...
Data RDM
Model RDM
All ROIs tested
Correction for multiple comparisons
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
Spearman r
−200
−100
0
100
200
300
time (ms)

Figure 2


A
LH BA 45
B
LH mpMTG
C
LH HG
D
Entropy
Semantic Similarity
Semantic Blend
Entropy Change
Spearman r
time (ms)
'corduroy trousers'
Mean W1 UP
W2 Onset
Mean W2 IP
Figure Key
Mean W2 IP = 240 ms, SD = 125
Mean W1 UP = -193 ms, SD = 103