Context and prediction in spoken word recognition: early left fronto-temporal effects of lexical uncertainty and semantic constraint

UNIVERSITY OF CAMBRIDGE

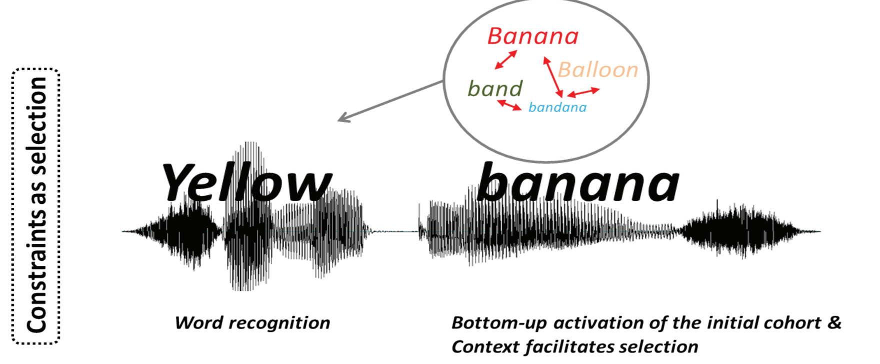
Anastasia Klimovich-Gray, Barry Devereux, Ece Kocagoncu, Billi Randall, William Marslen-Wilson, Lorraine K. Tyler (Centre for Speech, Language and the Brain, University of Cambridge)

European Research Council erc

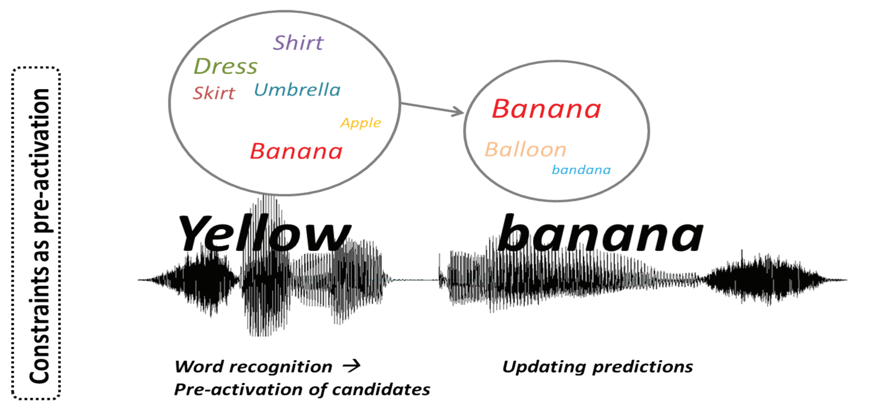
Motivation

The mechanisms and timing of semantic constraints on spoken word recognition are described by two strongly opposing views:

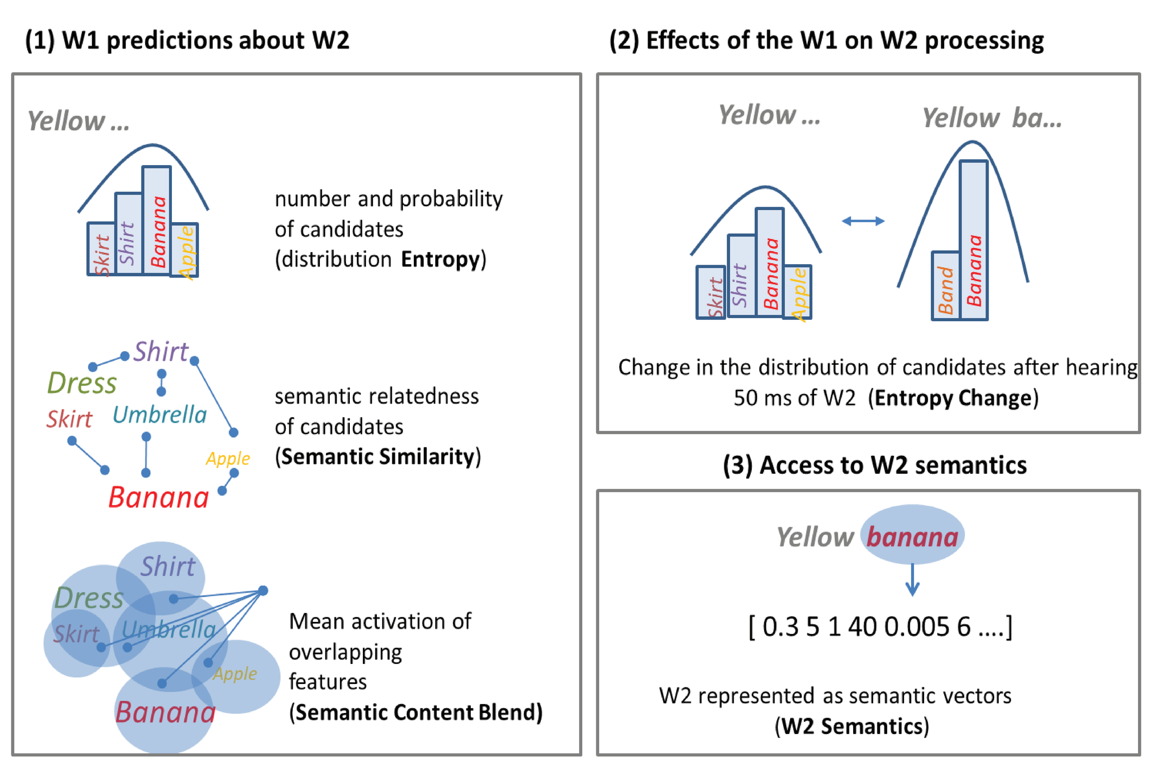
Bottom-up Priority : preceding context facilitates selection between multiple cohort candidates activated through bottom-up perceptual analysis - Cohort model (1,2).



Predictive Processing: context pre-activates compatible candidates at various levels of representation - word form, meaning & grammatical properties (3,4). Then predictions are updated based on perceptual input.



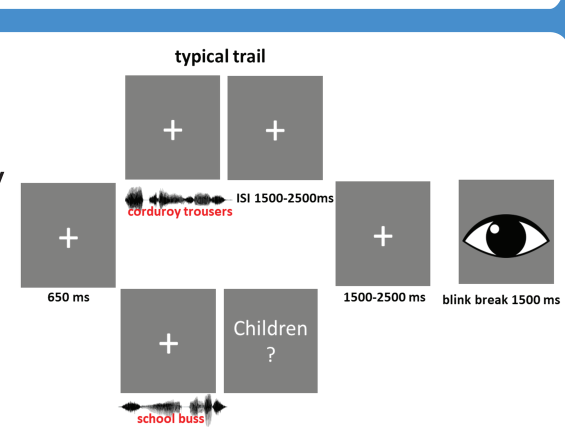
We recorded spatiotemporally resolved cortical activity (EMEG) during naturalistic processing of spoken 2-word phrases to ask:
(a) Is there evidence for pre-activation of context-compatible candidates? What kind of information is pre-activated and when? **(b)** How early do contextual constraints begin affecting word processing?
To answer these questions we used a gating paradigm to derive models capturing three distinct stages of spoken phrase processing:



We tested these measures against brain activity using ROI RSA (5) analysis, designed to explore whether and when specific information is encoded within the spatiotemporal activity patterns of specific ROIs.

Experimental Design

77 concrete nouns, each presented 3 times - (1) with a weakly (2) and with strongly constraining modifier and (3) in a nonsense phrase (filler).**We varied constraint** that W1 places on W2 by varying the conditional probability of the noun given the modifier (Google searches).



Methods

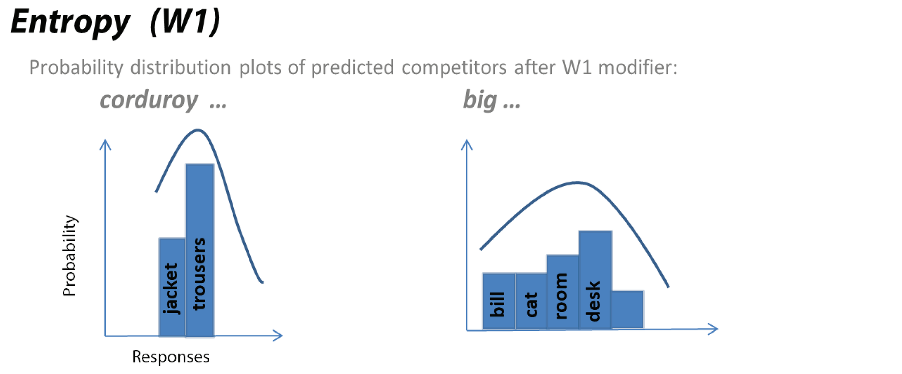
Procedure: natural listening + occasional (<10%) 1-back semantic association task. N = 16. 306 MEG +70 EEG.
Pre-processing: Maxfilter, band-pass filter: 0.1 to 40 Hz; ICA (blinks). **Source space**: Freesurfer (individual MRI), 3-layer BEM, minimum-norm solution, dSPM. **Alignment and epoching**: (1) W2 Onset, epoch [-200; +300 ms]; (2) W2 IP (identified through gating),epoch [-400; +400]. **Gating**: separate group of subjects (20) listened to W1 and produced guesses of W2, rating their confidence (gate zero); then W1 was played with 50 ms increments of W2 (subsequent gates), at each such gate recording further guesses and confidence scores; until W2 was correctly identified twice in a row (W2 IP point).
Cortical multivariate analysis: trial-level ROI RSA - see schematic overview of the RSA - (3) for details. RSA parameters used: 20 ms time-window, 5 ms time-step.

(a) Gating

Yellow ...	Responses	Confidence Scores
Big ...	dress	2
Corduroy ...	umbrella	1
	banana	3

(b) Deriving measures for RSA models

Entropy (W1)



If W1 triggers access to a distribution of W2 probable candidates we can capture the shape of this distribution using Shannon's entropy (H):

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

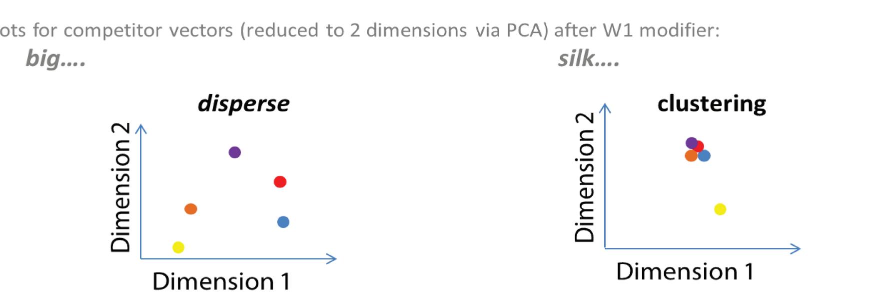
Where $P(x_i)$ is the summed confidence score for a given W2 predicted competitor across all participants divided by the sum of confidence scores for all W2 competitors for that item across all participants:

$$P(x_i) = \frac{\sum confidence_i}{\sum_{i=1}^{n} confidence_i}$$

Strong constraints (W1 on W2) produce only few highly probable candidates, while weak ones – a larger pool of less probable candidates

Semantic Similarity (W1)

Entropy captures the number and probability of the candidates. We can capture and summarise candidate semantics by calculating the average cosine similarity between the '**semantic vectors**' (Distributional Memory database - 6) of the competitors. Each semantic vector has 5000 dimensions, where entries are derived from summarising word-in-context co-occurrence data (random indexing method)



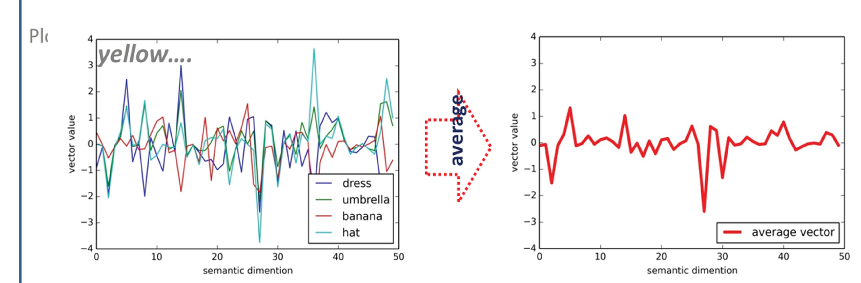
The greater the constraint from W1 on the semantics of W2 the more semantically similar and closer in the multidimensional semantic space are the W2 predicted competitors

Word 2 Semantics (W2)

After W2 has been recognised we expect that the semantic information unique to that word will be activated. We defined the W2 recognition point though a gating study - as the average time when all participants correctly guessed the target word twice in a row.
Semantic vectors encode distributional properties of a given word in a large corpus. To derive W2 semantics model we extracted and compared semantic vectors of every W2.

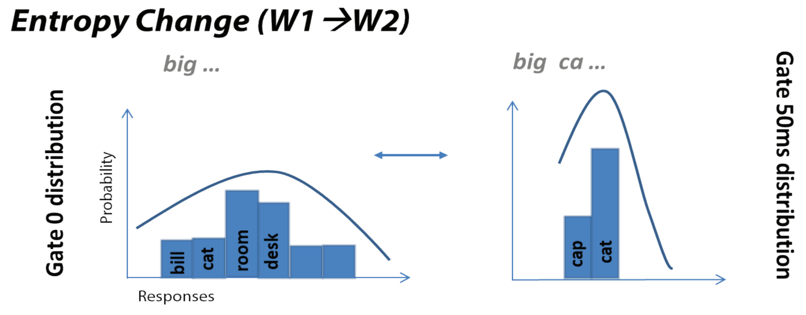
Semantic Content Blend (W1)

Information associated with a word can be represented as vector in the lexico-semantic feature space (6). After hearing W1 multiple predicted W2 candidates are activated simultaneously and the resulting pattern is a 'blend' of their overlapping representational features. Features shared across multiple candidates dominate the blend. For each W1 we can simulate the blend of W2 candidates by creating weighted averaging (by confidence) of all corresponding semantic vectors.



The greater the constraint that W1 places on semantics of W2 the closer the 'blend' will be to one specific highly likely candidate

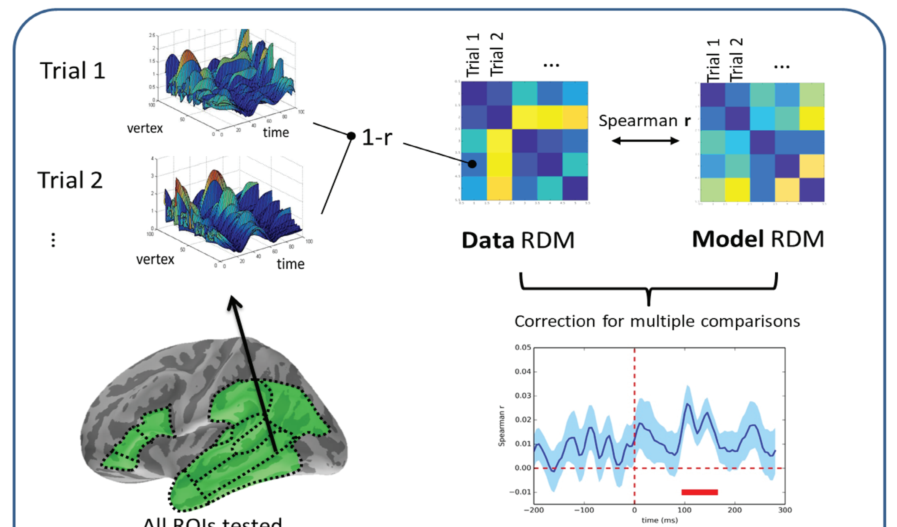
Entropy Change (W1 →W2)



If contextual constraints from W1 interact with the perceptual input to update the expectations formed about W2 lexical candidates we expect the language system to register the resulting shift in Entropy of the candidate distributions.

To test this we subtracted Entropy at gate 0 (distribution of predicted W2 candidates) from the Entropy at gate 50 ms (distribution of candidates after hearing 50 ms of W2).

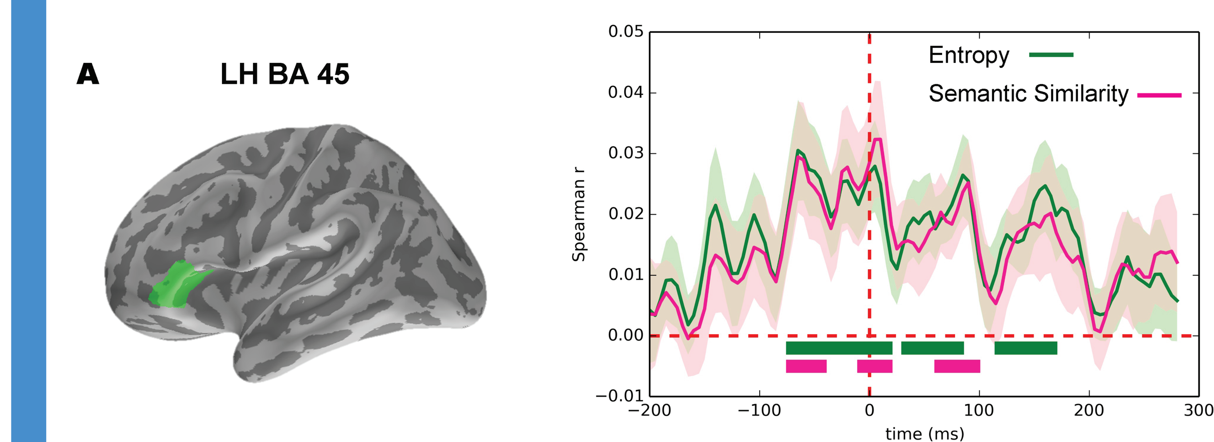
(c) Testing models with RSA



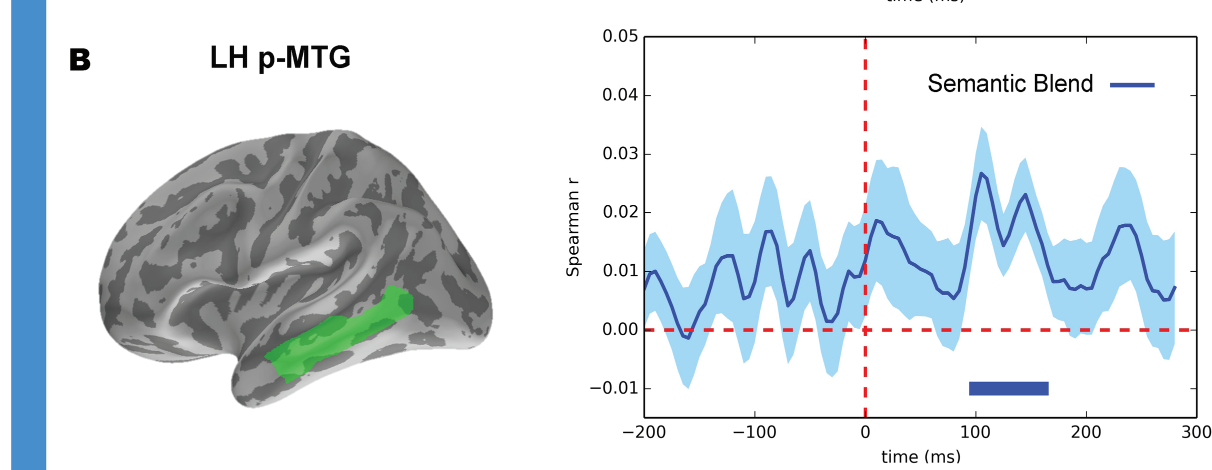
Deriving Model RDMs

To derive model RDMs we used different metrics of dissimilarity →**Entropy & Entropy Change** – absolute difference between trial values; **Semantic Similarity** – absolute difference between average cosine distance values; **Semantic Blend** – cosine distance between the 'blend' semantic vectors; **W2 semantics** - cosine distance between the W2 semantic vectors.

Top-down effects (W1)

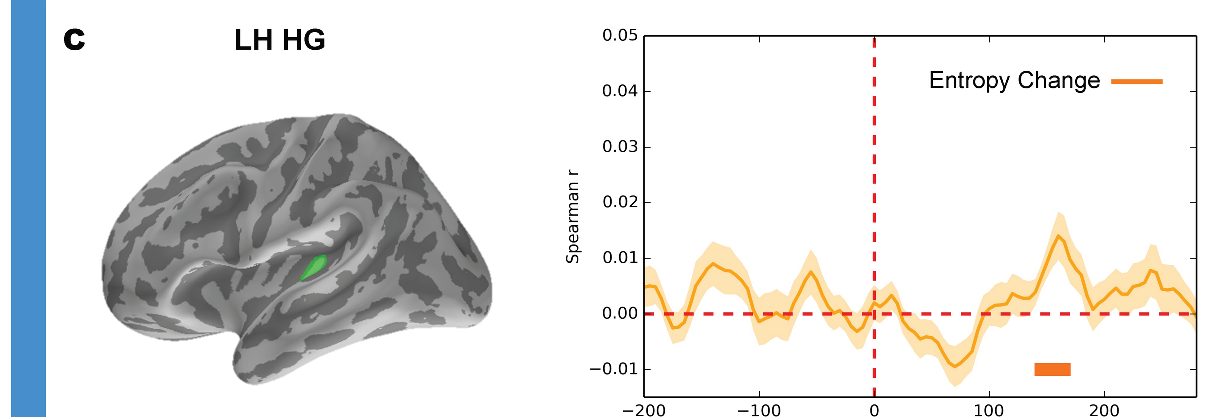


- W1 is recognised, general features (uncertainty & sparceness) of possible W2 continuations are accessed



- Access to the average semantic content of predicted W2 candidates is delayed [100 ms into W2]

Interaction of top-down and bottom-up effects (W1 -> W2)

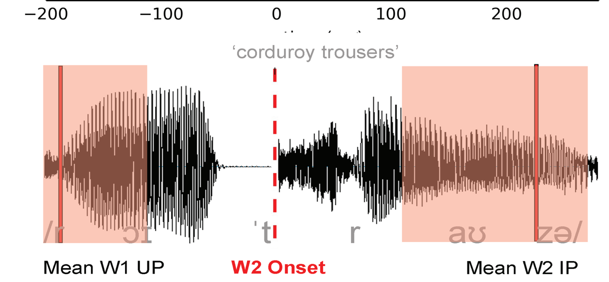


- 150 ms into W2, HG registers the change in lexical Entropy caused by 50 ms of W2 perceptual input

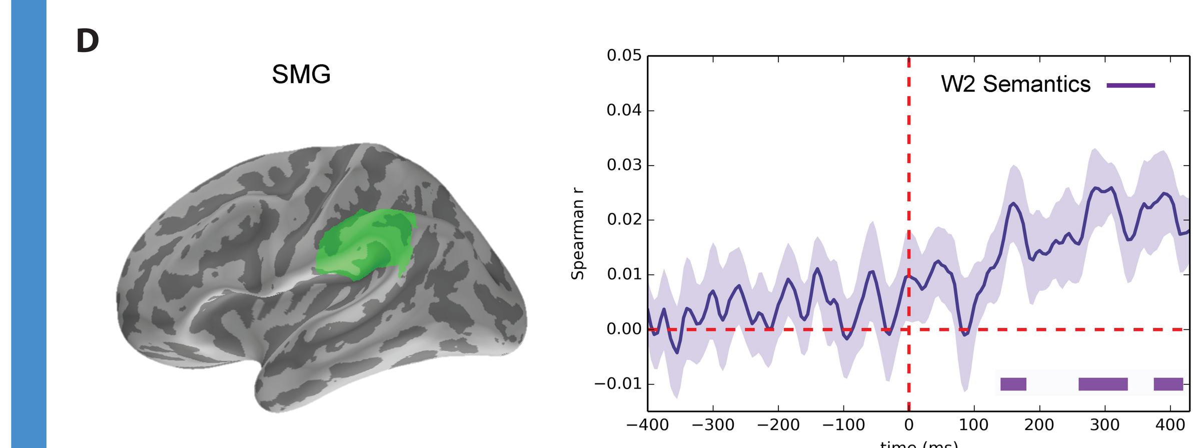
Figure Key

Mean W2 IP = 240 ms, SD = 125

Mean W1 UP = -193 ms, SD = 103



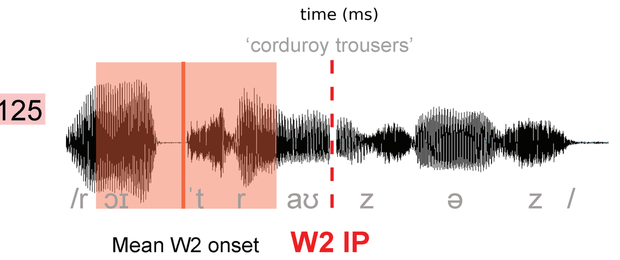
Access to W2 semantics



- W2 unique semantics are activated after the recognition point (IP, gating)

Figure Key

Mean W1 onset = -240ms, SD = 125



Conclusions

- Consistent with the predictive processing accounts, with minimal context the linguistic system can rapidly generate predictions about the overall properties (uncertainty & sparceness) of possible and not yet heard W2 candidates in BA 45. (Entropy, Semantic Similarity). However, such predictions are not realised as pre-activations of specific candidate's semantics (Semantic Blend model) but rather as limitations on the lexico-semantic space of possible candidates.
- Then at 150 ms into W2 top-down contextual constraints from W1 interact with W2's perceptual input to update these expectations (Entropy Change). HG is sensitive to the resulting shift in lexical uncertainty. The localisation of this effect is consistent with the predictive processing view; however its latency is more consistent with the Cohort model.
- Finally, after W2 recognition point its unique semantics is accessed in the SMG.

Our results show that predictive processing can provide a mechanistic account for how contextual constraints are generated and updated in spoken phrases. However, the timing with which these constraints interact with the bottom-up perceptual input seems to follow the Cohort model. Furthermore our results raise questions about the extent and nature of pre-activation routinely required in processing of spoken phrases with variable semantic constraint.

References

(1) Marslen-Wilson W., &. Tyler L. K. (1980). The temporal struc- ture of spoken language understanding.Cognition, 8, 1-71. (2) Tyler, L.K., Wessels, J., 1983. Quantifying contextual contributions to word recognition processes. Perception & Psychophysics 34, 409–420. (3) Kuperberg, G., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? Language, Cognition and Neuroscience, 31, 32–59. (4) Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and Brain Sciences, 36(4), 329-347. (5) Li, S., Fonteneau, E., Marslen-Wilson & Kriegeskorte, N. (2012). Spatiotemporal Searchlight Representational Similarity Analysis in EMEG Source Space. 2nd Int. Workshop on Pattern Recog. in NeuroImaging. (6) Baroni, M. & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4):673–721

This project has received funding to LKT from the European Research Council (ERC) under the European Union Horizon 2020 research andinnovation programme (grant agreement No 669820)

ak798@cam.ac.uk