Localizing incremental linguistic prediction in the mind

Cory Shain, 5/7/2019

From: Shain*, Blank*, van Schijndel, Schuler, & Fedorenko (under review). *fMRI reveals language-specific predictive coding during naturalistic sentence comprehension*.

*co-first author

Human sentence processing is incremental and predictive

- Visual world (Tanenhaus et al., 1995)
- Electrophysiological (Kutas & Hillyard, 1984)
- Reading (Smith & Levy, 2013)

Human sentence processing is incremental and predictive

- What is being predicted?
- What purpose does prediction serve?
- **What neural mechanisms support linguistic prediction?**

Is linguistic prediction domain-specific or domain-general?

- **(Primarily) domain-specific (DS):**
  - We know some predictive coding is local (Singer et al., 2018)
  - Predictive coding for language might also be implemented by domain-specific circuits

Is linguistic prediction domain-specific or domain-general?

- **(Primarily) domain-general (DG):**
    - Many have argued that linguistic prediction is carried out by domain-general executive resources (Smith & Levy, 2013; Huettig & Mani, 2016; Pickering & Gambi, 2018)
        - Prediction effects modulated by individual and group level differences in executive function (Federmeier et al., 2002; Martin et al., 2013, Gambi et al., 2018, *inter alia*)
            - Cf. Ryskin et al. (under review)
        - Domain-general executive involvement in language processing (Kaan & Swaab, 2002; January et al., 2009)
        - Prediction effects across tasks and species (Smith & Levy, 2013)

Is linguistic prediction domain-specific or domain-general?

- Both DS and DG hypotheses rely on notion of *generality*
  - DG: Predictive mechanism is domain-general
    - Unified mechanism predicts, specialized mechanisms query it
  - DS: Learning mechanism is domain-general
    - Specialized mechanisms predict, and learn to do so under general plasticity rules

Measuring predictive coding via *surprisal*

- Predictive coding should evoke a predictability response
    - Greater effort for less predictable stimuli
- Predictability can be quantified via *surprisal* (Shannon, 1948; Hale, 2001)
    - Negative log probability of events given context
- Search for networks where surprisal modulates neural response

Measuring predictive coding via *surprisal*

- Surprisal by what model?
- Previous fMRI studies have used “syntactic” surprisal (Henderson et al., 2016) or unlexicalized (PoS) *n*-gram surprisal (Brennan et al., 2016)
- Best-attested behavioral effects are for lexicalized *n*-gram surprisal (Frank & Bod, 2011; Smith & Levy, 2013)
  - Surprise broadly construed, abstracting away from structure
- **This study:** Lexicalized *n*-grams (5-grams)

Localizing surprisal effects in the brain

- Domain-specific:
  - **LANG:** Fronto-temporal language network (Fedorenko et al., 2010)
  - Prediction: Surprisal effects should primarily reside in LANG
- Domain-general:
  - **MD:** Fronto-parietal multiple-demand network (Duncan, 2010)
    - Supports top-down executive functions
    - Response modulated by cognitive effort (Duncan & Owen, 2000)
    - Argued to relay predictive signals to other regions (Strange et al., 2005)
  - Prediction: Surprisal effects should primarily reside in MD

Localizing surprisal effects in the brain

- Not possible with behavioral or EEG studies
- Subject to task artifacts from constructed stimuli (Miller & Cohen, 2001; Hasson & Honey, 2012; Campbell & Tyler, 2018)
- Best studied using **Naturalistic fMRI**
    - Few fMRI studies of naturalistic language processing
    - Even fewer that explore lexicalized surprisal (Brennan et al., 2016; Willems et al., 2015; Lopopolo et al., 2017)
    - Mixed evidence for (1) existence and (2) location of lexicalized *n*-gram surprisal

Localizing surprisal effects in the brain

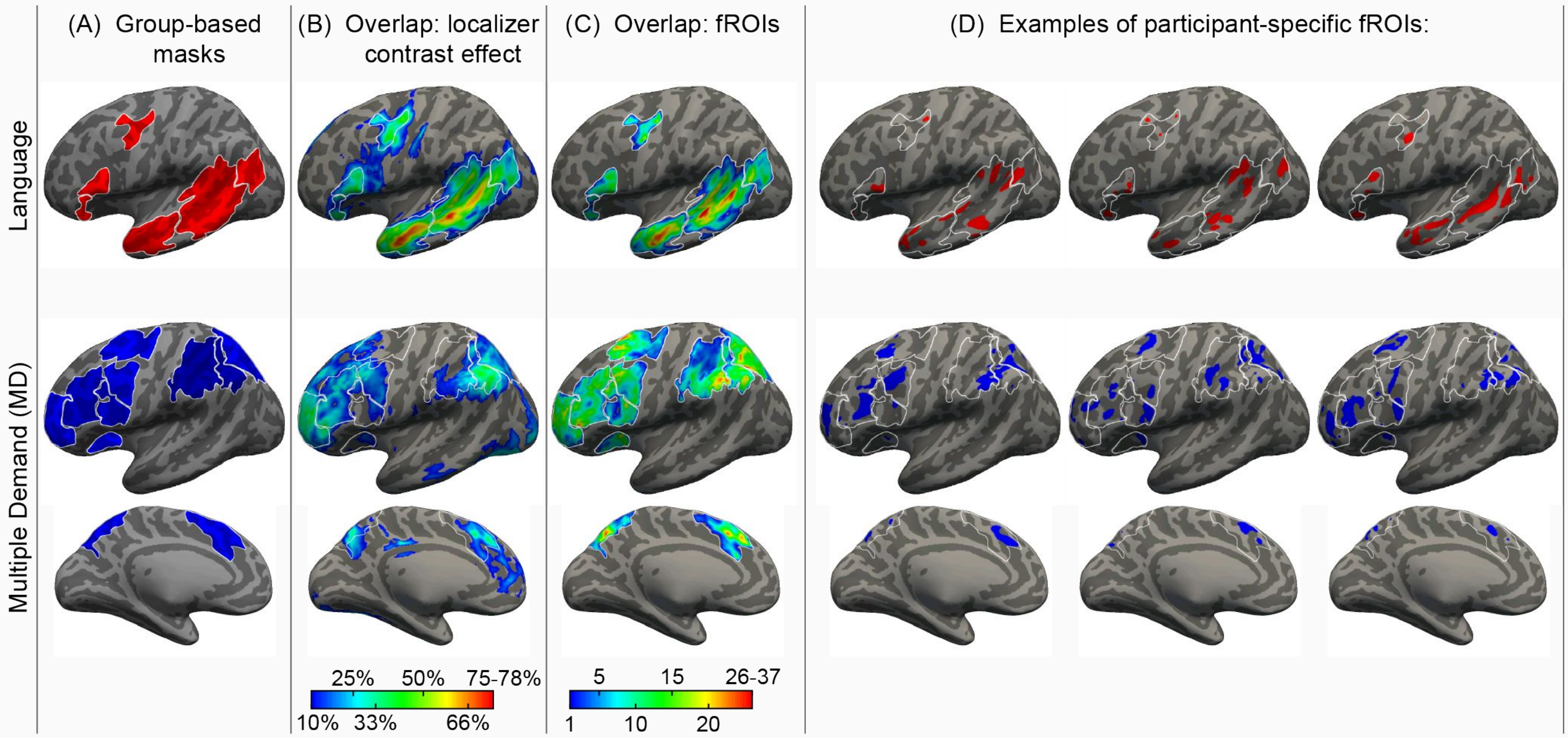
- **This study:** Test DS vs. DG by comparing surprisal effects in LANG vs. MD in fMRI measures of subjects listening to natural language.

Methods: Data

- Stimuli from the Natural Stories corpus (Futrell et al., 2018)
- Auditory presentation (1 female speaker, 1 male)
- 78 subjects (30 males)

Methods: Defining LANG and MD

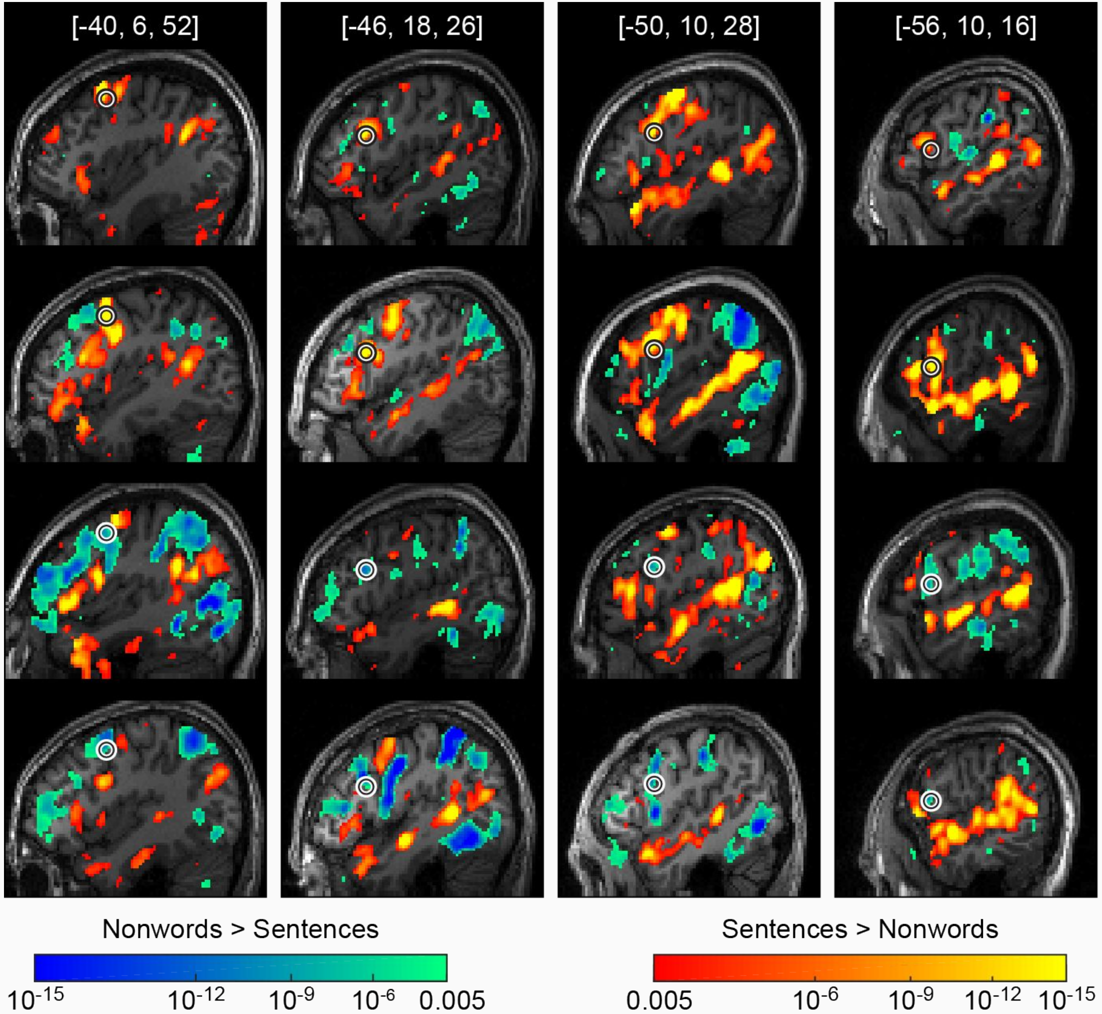
- LANG and MD defined with by-participant functional localization (Fedorenko et al., 2010)
- Independent localizer task (passive or probe)
- Sentence vs. non-word list conditions
- Functional regions of interest (fROIs) selected by
  - Masking
  - Selecting top 10% voxels within each mask



(A) Group-based masks
(B) Overlap: localizer contrast effect
(C) Overlap: fROIs
(D) Examples of participant-specific fROIs:
Language
Multiple Demand (MD)
25%
50%
75-78%
10%
33%
66%
5
15
26-37
1
10
20

Methods: Defining LANG and MD

- LANG contrast: Sent > Nonword (Fedorenko & Thompson-Schill, 2014)
- MD contrast: Nonword > Sent (Fedorenko et al., 2013; Mineroff et al., 2018)



[-40, 6, 52]
[-46, 18, 26]
[-50, 10, 28]
[-56, 10, 16]
Nonwords > Sentences
10^{-15} 10^{-12} 10^{-9} 10^{-6} 0.005
Sentences > Nonwords
0.005 10^{-6} 10^{-9} 10^{-12} 10^{-15}

Methods: Defining LANG and MD

- 6 LANG fROIs (left hemisphere only):
    - Inferior frontal gyrus (IFG)
    - Orbital part of inferior frontal gyrus (IFGorb)
    - Middle frontal gyrus (MFG)
    - Anterior temporal cortex (AntTemp)
    - Posterior temporal cortex (PostTemp)
    - Angular gyrus (AngG)

Methods: Defining LANG and MD

- 10 MD fROIs (each hemisphere):
  - Posterior parietal cortex (PostPar)
  - Middle parietal cortex (MidPar)
  - Anterior parietal cortex (AntPar)
  - Precentral gyrus (PrecG)
  - Superior frontal gyrus (SFG)
  - Middle frontal gyrus (MFG)
  - Orbital part of middle frontal gyrus (MFGorb)
  - Opercular part of inferior frontal gyrus (IFGop)
  - Anterior cingulate cortex and pre-supplementary motor cortex (ACC/pSMA)
  - Insula

Methods: Naturalistic fMRI modeling

- Naturalistic language stimuli are a problem for event-based stats methods in fMRI
  - Events (words) are variably spaced, don’t align with scan times

Methods: Naturalistic fMRI modeling

- Established solutions are problematic
  - Canonical HRF (Brennan et al., 2016)
    - Inflexible
    - Can’t account for regional variation (Handwerker et al., 2004)
  - Binned averaging (Wehbe et al., in prep)
    - Distorts event timestamps
    - Low-resolution filter
  - Interpolation (Huth et al., 2016)
    - Treats word properties as underlyingly continuous
    - Non-causal
    - Low-resolution filter

Methods: Naturalistic fMRI modeling

- **Our solution:** Deconvolutional time series regression (DTSR, Shain & Schuler, 2018)
- Uses ML to estimate continuous response shape
- Like a canonical HRF that adapts to the data
- No distortion of stimulus structure (temporal or featural)

Method	Train Mean Squared Error	Test Mean Squared Error
Canonical HRF	11.3548	11.8263
Binned Averaging	11.3478	11.9280
Linear Interpolation	11.4236	11.9888
Lanczos Interpolation	11.3536	11.9059
DTSR	**11.2749**	**11.6389**

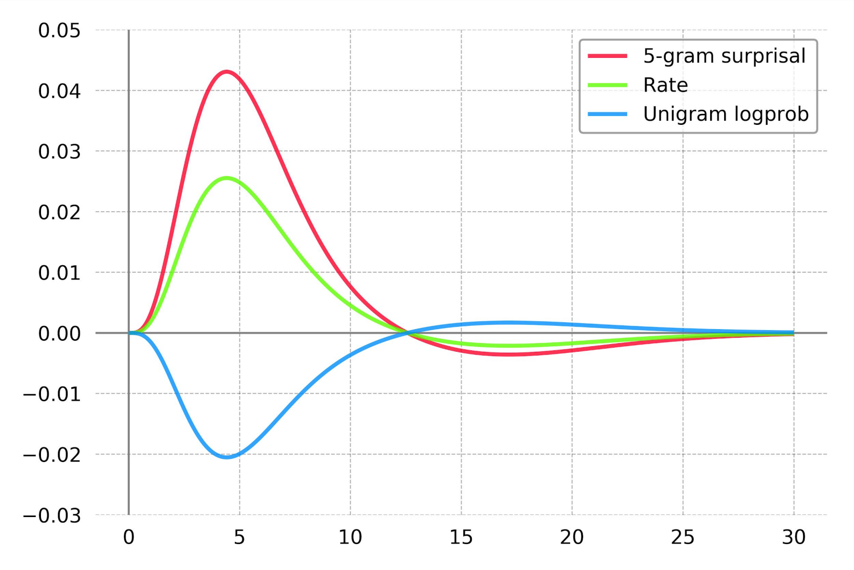
Methods: Naturalistic fMRI modeling

- Predictors:
  - Rate (convolved intercept)
  - Unigram logprob
    - KenLM (Heafield et al., 2013) on Gigaword 3 (Graff et al., 2007)
  - **5-gram surprisal**
    - Same as unigram
  - HRF params are tied between predictors within fROIs, by-predictor coefficients
  - Sound power (canonical HRF convolved)
  - TR number (linear)
- By-fROI random intercepts, slopes, HRF params
- By-participant random intercepts

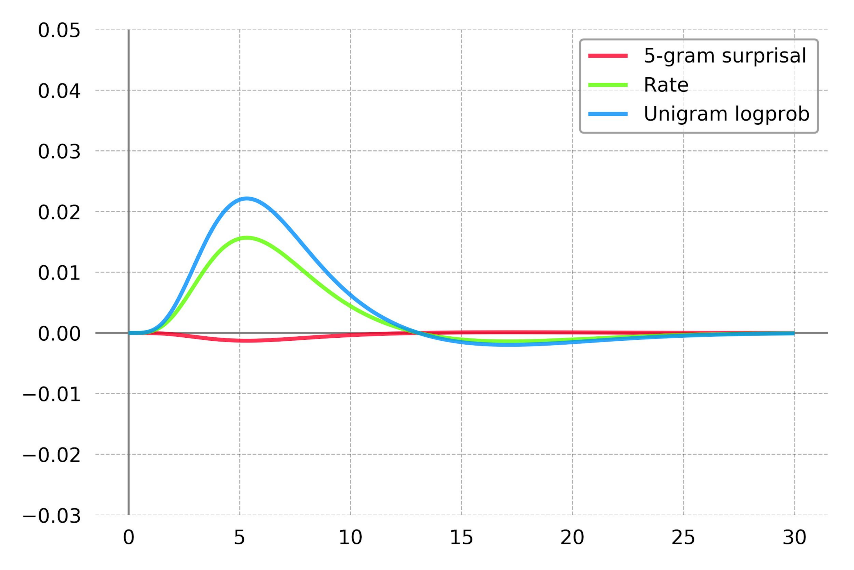
Methods: Naturalistic fMRI modeling

- Ablative non-parametric out-of-sample hypothesis tests
    - Common in ML
- 50% train, 50% test
- Separate models for LANG and MD test surprisal effects in each
- Combined model tests *difference* in surprisal between LANG and MD
    - Ablation: Surprisal:Network (0 = MD, 1 = LANG)

Results



LANG



MD

Results

Comparison	*p*	LL Improvement	Coefficient
Surprisal (LANG)	**0.0001*****	108.33	0.256
Surprisal (MD)	1.0	-3.23	-0.008
Surprisal by Network (combined)	**0.0001*****	86.69	0.231

Hypothesis tests

Surp in LANG, no surp in MD, significant difference between networks

Results

	LANG		MD		COMBINED	
	% Tot	**% Rel**	**% Tot**	**% Rel**	**% Tot**	**% Rel**
Ceiling	6.18%	100%	1.34%	100%	2.63%	100%
Model (train)	3.21%	51.9%	0.68%	50.7%	1.06%	40.3%
Model (test)	1.66%	26.9%	0.00%	0.00%	0.52%	19.8%

% variance explained

Results

- LANG surprisal effects
  - Large magnitude
  - Positive
  - Significant
  - Generalize well (large out-of-sample relative % variance explained)
- MD surprisal effects
  - Small magnitude
  - Negative
  - Non-significant
  - Generalize poorly (no out-of-sample variance explained)
- Significant difference in effect size

Conclusion

- Results support a domain-specific implementation of prediction:
  - Predictive coding for language, locally implemented in language-specialized circuits
- Prediction effect is over and above lexical frequency
- In line with patterns found in low-level sensory circuits (Singer et al., 2018)

Future directions

- What is the structure of the predictive model?
- Is there functional differentation *within* LANG wrt linguistic prediction?
- What is the relationship between predictive and integrative computation?

Thank you!

Thanks to:

Funders: NIH awards R00-HD-057522 and R01-DC-016607 (E.F.), a grant from the Simons Foundation via the Simons Center for the Social Brain at MIT (E.F.), and NSF grant #1816891 (W.S.)

Reviewers and participants at CUNY 2019

All of you!