NEUROSCIENCE
RESEARCH ARTICLE

Matthew Siegelman et al. / Neuroscience 413 (2019) 219–229

An Attempt to Conceptually Replicate the Dissociation between Syntax and Semantics during Sentence Comprehension

Matthew Siegelman,[ab] Idan A. Blank,[ac] Zachary Mineroff[a] and Evelina Fedorenko[ade]*

[a]*MIT, Department of Brain and Cognitive Sciences*
[b]*Columbia University, Department of Psychology*
[c]*UCLA, Department of Psychology*
[d]*MIT, McGovern Institute for Brain Research*
[e]*MGH, Department of Psychiatry*

Abstract—Is sentence structure processed by the same neural and cognitive resources that are recruited for processing word meanings, or do structure and meaning rely on distinct resources? Linguistic theorizing and much behavioral evidence suggest tight integration between lexico-semantic and syntactic representations and processing. However, most current proposals of the neural architecture of language continue to postulate a distinction between the two. One of the earlier and most cited pieces of neuroimaging evidence in favor of this dissociation comes from a paper by Dapretto and Bookheimer (1999). Using a sentence-meaning judgment task, Dapretto & Bookheimer observed two distinct peaks within the left inferior frontal gyrus (LIFG): one was more active during a lexico-semantic manipulation, and the other during a syntactic manipulation. Although the paper is highly cited, no attempt has been made, to our knowledge, to replicate the original finding. We report an fMRI study that attempts to do so. Using a combination of whole-brain, group-level ROI, and participant-specific functional ROI approaches, we fail to replicate the original dissociation. In particular, whereas parts of LIFG respond reliably more strongly during lexico-semantic than syntactic processing, no part of LIFG (including in the region defined around the peak reported by Dapretto & Bookheimer) shows the opposite pattern. We speculate that the original result was a false positive, possibly driven by a small subset of participants or items that biased a fixed-effects analysis with low power.

Key words: fMRI, syntax, semantics, replication, LIFG, Broca's area.

INTRODUCTION

Sentence comprehension requires us to retrieve the word meanings from the mental lexicon (lexico-semantic processing), and infer how they relate to one another within the sentence (syntactic processing) — i.e., recover their dependency structure using a combination of lexico-semantic constraints, word order, and/or functional morphology (e.g., Dryer, 2002; Gibson et al., 2013). Together, individual word meanings and the way they combine determine the propositional content of the sentence (i.e., who is doing what to whom). Whether these two components of sentence comprehension rely on distinct pools of cognitive and neural resources has been long debated (e.g., Dick et al., 2001).

In order to search for a potential dissociation between lexico-semantic and syntactic processing, cognitive neuroscientists have tested whether some brain regions respond selectively, or at least preferentially, to one or the other. To this end, several manipulations contrasting the two kinds of processes have been used, including a) linguistically degraded materials like lists of unconnected words that require lexical-level understanding but not putting words together into complex representations, vs. "Jabberwocky" sentences that contain a coarse-level representation of the dependency structure but not lexical meanings (e.g., Friederici et al., 2000; Humphries et al., 2006; Fedorenko et al., 2010; see Bautista & Wilson, 2016 for a related approach), b) violations of lexico-semantic vs. syntactic expectations (e.g., Embick et al., 2000; Kuperberg et al., 2003; Cooke et al., 2006; Friederici et al., 2010; Herrmann et al., 2012), and c) adaptation to lexico-semantic content vs. syntactic structure (e.g., Noppeney and Price, 2004; Santi and Grodzinsky, 2010; Menenti et al., 2011; Segaert et al., 2012). These numerous studies have produced a complicated empirical picture filled with contradictions (e.g., see Fedorenko et al., 2018, for a discussion). Nevertheless, the dominant view among cognitive neuroscientists studying language remains that lexico-semantic and

*Corresponding author.
E-mail address: evelina9@mit.edu (Evelina Fedorenko).

https://doi.org/10.1016/j.neuroscience.2019.06.003
0306-4522/

syntactic processing rely on distinct pools of resources (e.g., Friederici, 2012; Baggio and Hagoort, 2011; Tyler et al., 2011; Duffau et al., 2014; Ullman, 2016; cf. Bates & Goodman, 1997; Fedorenko et al., 2012a; Blank et al., 2016; Bautista and Wilson, 2016).

One of the most cited studies that has argued for a dissociation between lexico-semantic and syntactic processing was conducted by Dapretto and Bookheimer and published in Neuron in 1999. The study used an original manipulation where participants made meaning judgments on pairs of sentences, which differed either in one word (replaced by a synonym, resulting in the same meaning, or by a non-synonym, leading to a change in meaning) or in the structure of the sentence (e.g., an Active/Passive alternation that either kept the thematic roles the same or switched them; see sample items in Methods). The key result was a double dissociation between the Semantics and Syntax conditions observed in the left inferior frontal gyrus (LIFG), i.e., two nearby peaks revealed by the Semantics > Syntax, and Syntax > Semantics contrast, respectively. This result, the authors argued, provided "unequivocal evidence that these functions [lexico-semantic and syntactic processing; *SBMF*] are […] subserved by distinct cortical areas".

Dapretto & Bookheimer's study has been cited 689 times (as of November 12, 2018; https://scholar.google.com/citations?view_op=view_citation&hl=en&user=fQ-cmN8AAAAJ&citation_for_view=fQ-cmN8AAAAJ:d1gkVwhDpl0C), and the pattern of citations over the years (Fig. 1) suggests that it is still being used by researchers as evidence for distinct brain regions supporting lexico-semantic vs. syntactic processing.

And yet, it appears that no replication of this study has ever been published either by one of the original author's labs, or by any other research group. Given a) the study's impact on the field, combined with b) recent studies that have argued for overlap between lexico-semantic and syntactic processing across the fronto-temporal language network (e.g., Fedorenko et al., 2012a; Bautista and Wilson, 2016; Blank et al., 2016; Fedorenko et al., 2018), and c) current emphasis on reproducibility in the fields of psychology (e.g., Ioannidis, 2005; Simmons et al., 2011; Button et al., 2013; Ioannidis et al., 2014) and cognitive neuroscience (e.g., Poldrack et al., 2017), we here attempted a conceptual replication of Dapretto & Bookheimer's findings.

EXPERIMENTAL PROCEDURES

Participants

Fifteen individuals (age 20–30 (25.3 ±4.1), five females), native speakers of English, participated for payment. Fourteen of the 15 participants were right-handed (as determined by the Edinburgh handedness inventory; Oldfield, 1971), but all 15 showed typical, left-lateralized, language activations (as assessed with an independent language "localizer" task conducted in the same session; Fedorenko et al., 2010). All participants had normal hearing and vision, and no history of neurological illness or language impairment. Participants gave written informed consent in accordance with the requirements of MIT's Committee on the Use of Humans as Experimental Subjects (COUHES).

Design, materials, and procedure

Each participant completed the critical task, as well as one or more additional tasks for unrelated studies. The entire scanning session lasted approximately 2 h.

Design and materials

The basic design was the same as in Dapretto & Bookheimer's study. Participants were presented with pairs of sentences and asked to decide whether they meant roughly the same thing. The critical manipulation was whether the sentences in the pair differed in one of the words (the Semantics condition) or in the structure / word order (the Syntax condition). In particular, in the Semantics condition,

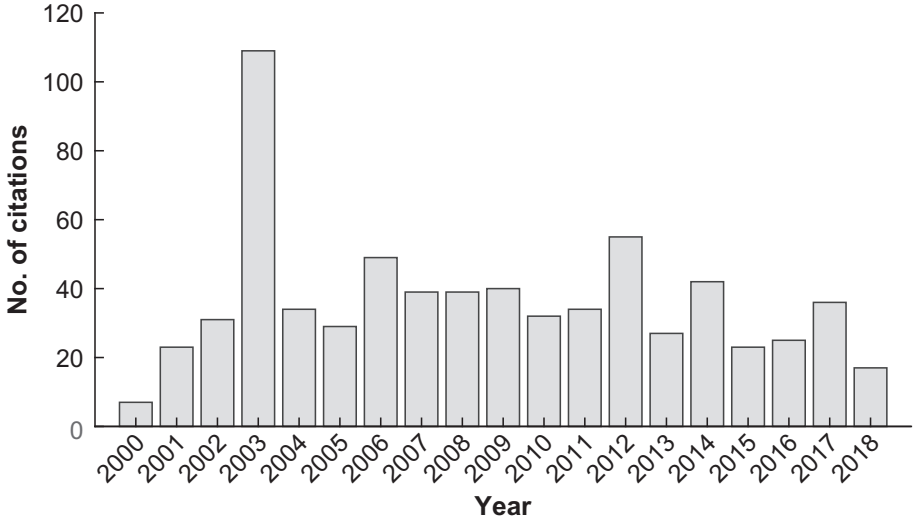

Figure 1. The citations for Dapretto and Bookheimer (1999) by year (total number = 689 as of November 12, 2018; the numbers come from Google Scholar).

Table 1. A summary of the key differences between the two studies.

	Current study	Dapretto & Bookheimer's study
Participants	$n = 15$	$n = 8$
fMRI design	Event-related	Blocked
Materials	40 pairs of sentences per condition, 40 events per condition	8 pairs of sentences per condition, 1 block per condition
Constructions used	Active/Passive alternation, Double Object/ Prepositional Phrase Object alternation	Active/Passive alternation, Locative Prepositional Phrase alternation
Presentation	Visual	Auditory
Acquisition device	A 3-T Siemens Trio scanner with a 32-channel head coil	A 3-T GE scanner; no coil information provided
Acquisition parameters	TR = 2000 ms, TE = 30 ms matrix size = 96×96 with 200-mm field of view	TR = 2500 ms, TE = 45 ms matrix size = 64×64 with 200-mm field of view
Preprocessing software	SPM5	SPM96
Preprocessing	4-mm FWHM Gaussian smoothing kernel; high-pass filtering	6-mm FWHM Gaussian smoothing kernel; no high-pass filtering
Statistical modeling	Random effects analysis	Fixed effects analysis? (not fully clear from the description)

one of the words in the first sentence was replaced by a synonym in the second sentence (roughly preserving the meaning) or by a word with a different meaning (leading to different meanings), as in (1a). In the Syntax condition, the sentences were either syntactic alternations with the same meaning, or the structure / word order was changed leading to a different meaning, as in (1b).

(1a) Semantics

Same: *Anna invited the composer. / Anna invited the songwriter.*

Different: *Anna invited the composer. / Anna invited the translator.*

(1b) Syntax

Same: *Anna invited the composer. / The composer was invited by Anna.*

Different: *Anna invited the composer. / The composer invited Anna.*

The materials consisted of 80 items (sentence pairs). Forty items used the Active / Passive constructions (as in Dapretto & Bookheimer's study), and 40 used the Double Object (DO) / Prepositional Phrase Object (PP) constructions. Each item had four versions, as in (1a–b), for a total of 320 trials. The full set of materials is available at https://osf.io/wtv9f/.

The 320 trials were divided into four experimental lists (80 trials each, 40 trials per condition) following a Latin Square Design so that each list contained only one version of any given item. Each participant saw the materials from just one experimental list, and each list was seen by three to four participants.

A number of features varied and were balanced across the materials. ***First***, the construction was always the *same* across the two sentences in a pair in the Semantics condition (balanced between active and passive for the Active / Passive trials, and between double object and prepositional phrase object for the DO / PP trials). In the Syntax condition, the construction was always *different* in the Same-meaning trials because this is how the propositional meaning was preserved. For the Different-meaning trials, the construction could either be the same (again, balanced beween active and passive for the Active / Passive trials, and between double object and prepositional phrase object for the DO / PP trials) or different, as follows:

(2a) Syntax–Different (active/passive):

Same construction: *Anna invited the composer. / The composer invited Anna.*

Different constructions: *Elizabeth disliked the proprietor. / Elizabeth was disliked by the proprietor.*

(2b) Syntax–Different (DO/PP):

Same construction: *Amanda lent the cook some money. / The cook lent Amanda some money.*

Different constructions: *Brenda read the expert a passage. / The expert read a passage to Brenda.*

For trials where the constructions differed between the two sentences in a pair, we balanced whether the first sentence was active vs. passive (for the Active / Passive trials), or whether it was DO vs. PP (for the DO / PP trials).

Second, all sentences (in both Active / Passive and DO / PP constructions) contained one occupation noun and one name. Whether the first noun in the first sentence in a pair was an occupation or a name was balanced across items.

And ***third***, for the Semantics condition, we varied how exactly the words in the second sentence in a pair differed from the words in the first. (This does not apply to the Syntax condition trials, where the content words are identical across the two sentences within each pair.) In particular, for the Active / Passive trials, either the occupation noun or the verb could be replaced (by a synonym or a word with a different meaning); and for the DO/PP trials, either the occupation noun or the direct object (inanimate) noun could be replaced.

Procedure

An event-related design was used. Each event (trial) consisted of an initial 300-ms fixation, 2000-ms presentation of the first sentence (presented all at once), 200-ms inter-sentence interval, 2000-ms presentation of the second sentence, and a 1500-ms window for participants to respond (by pressing one of two buttons on a button box), for a total of 6 s. The 80 trials in a list were divided into two runs, with each run consisting of 40 trials and additional 120 s of inter-trial fixation, for a total run duration of 360 s (6 min). Each

participant performed two runs. The optseq2 algorithm (Dale, 1999) was used to create condition orderings and to distribute fixation among the trials so as to optimize our ability to de-convolve neural responses to each condition. Eight orders were created, and order varied across runs and participants.

A summary of the key differences between the current study and Dapretto and Bookheimer's (1999) study:

Table 1 provides a summary of the key differences between the studies. The key improvements in the design of the current study concern *power*, with respect to both the number of participants tested (almost twice as many participants), and amount of data collected for each participant: we used five times as many trials per condition. The original study used a generally more powerful blocked design. However, given that the original study used only one block per condition, the current study is likely to be more powerful in spite of the use of an event-related design (with 40 events per condition) (e.g., Nee, 2019). (See Analyses below for a formal power calculation.)

We also deviated in one of the constructions used. Although we adopted the Active / Passive alternation, we replaced the Locative Prepositional Phrase alternation (e.g., *The pool is behind the gate. / Behind the gate is the pool.*) with a more commonly used Double Object / Prepositional Phrase Object (DO / PP) alternation (e.g., Allen et al., 2012; Gibson et al., 2013). The reason we chose not to use the Locative alternation from the original study is that fronted locative prepositional phrases (locative inversion) are rare in natural language (e.g., Gibson et al., 2013). Finally, we opted for the use of the visual presentation (cf. auditory presentation used by Dapretto & Bookheimer). Abundant prior evidence suggests that high-level language processing brain regions, including those in the frontal lobe, are robust to presentation modality (e.g., Buchweitz et al., 2009; Fedorenko et al., 2010, 2016; Braze et al., 2011; Bemis and Pylkkänen, 2012; Vagharchakian et al., 2012; Scott et al., 2016). Thus, the use of a different modality is not expected to matter.

Given these differences between the original study and the current one, this replication is not a direct replication, but a *conceptual* one, albeit a close one. Conceptual replications have been argued to be as important, if not more important in some cases, for establishing robust cumulative science (e.g., Schmidt, 2009).

fMRI data acquisition and preprocessing

Structural and functional data were collected on the whole-body 3-T Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 128 axial slices with 1-mm isotropic voxels (TR = 2530 ms, TE = 3.48 ms). Functional, blood oxygenation level dependent (BOLD) data were acquired using an EPI sequence (with a 90° flip angle and using GRAPPA with an acceleration factor of 2), with the following acquisition parameters: 31 4-mm-thick near-axial slices, acquired in an interleaved order with a 10% distance factor; 2.1 mm × 2.1 mm in-plane resolution; field of view of 200 mm in the phase encoding anterior to posterior (A > P) direction; matrix size of 96 × 96; TR of 2000 ms; and TE of 30 ms. Prospective acquisition correction (Thesen et al., 2000) was used to adjust the positions of the gradients based on the participant's motion one TR back. The first 10 s of each run was excluded to allow for steady-state magnetization.

MRI data were analyzed using SPM5 (using default parameters, unless specified otherwise) and supporting, custom MATLAB scripts. (The use of an older version of the SPM software should make the preprocessing and analysis more similar to those used by Dapretto & Bookheimer, who used SPM96.) Each participant's data were motion corrected and then normalized into a common brain space (the Montreal Neurological Institute, MNI, Brain Template) and resampled into 2-mm isotropic voxels. The data were then smoothed with a 4-mm FWHM Gaussian filter and high-pass filtered (at 200 s). The critical task's effects were estimated using a General Linear Model (GLM) in which each experimental condition was modeled with a boxcar function (corresponding to an event) convolved with the canonical hemodynamic response function (HRF).

Analyses

What counts as a replication in brain imaging studies is still debated (e.g., Hong et al., 2019. In an effort to be comprehensive, we performed three analyses to assess whether the dissociation reported by Dapretto and Bookheimer (1999) between syntactic and lexico-semantic processing holds in the current dataset.

First, we performed a traditional random-effects analysis (e.g., Holmes and Friston, 1998), where individual activation maps are overlaid in the common space, and a *t*-test is

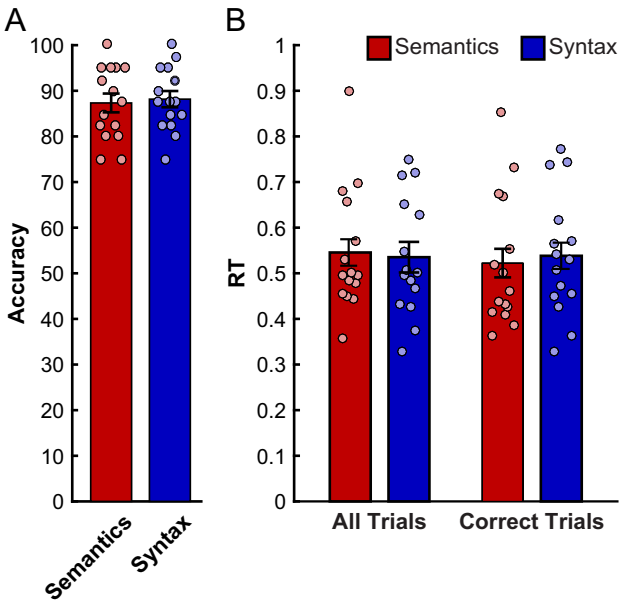

Figure 2. Behavioral performance (accuracies: left, RTs: right) during the Semantics and Syntax conditions. Dots correspond to individual participants; error bars represent standard errors of the mean across participants.

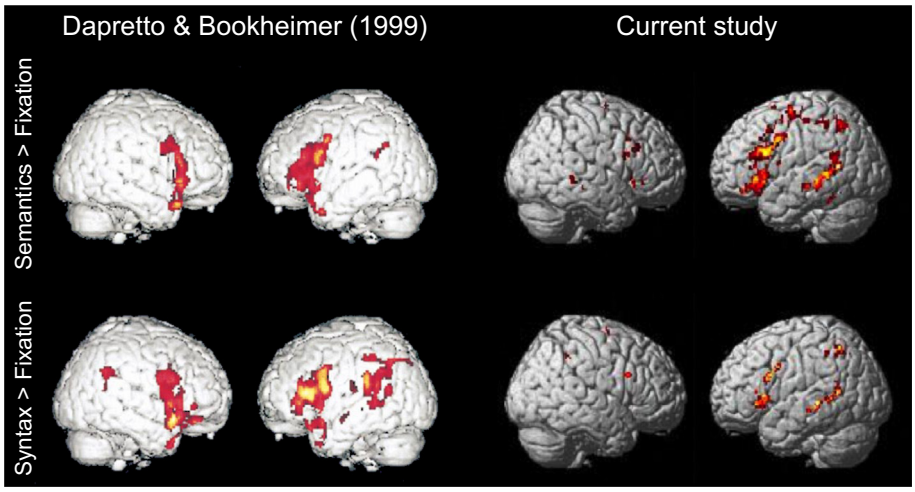

Figure 3. Whole-brain activation maps for the Semantics > Fixation (top) and for the Syntax > Fixation (bottom) contrasts in Dapretto & Bookheimer's study (left; $p < 0.0001$, uncorrected) and the current study (right; also $p < 0.0001$, uncorrected; see Table 2).

performed across participants in each voxel for each relevant contrast. In particular, following Dapretto & Bookheimer, we examined group-level effects for the following four contrasts: i) Semantics > Fixation, ii) Syntax > Fixation, iii) Semantics > Syntax, and iv) Syntax > Semantics.

Second, we performed a more targeted analysis of the activation peaks that emerged in Dapretto & Bookheimer's study for the direct contrasts of the Semantics and Syntax conditions: {−48, 20, −4} in Talairach space ({−48.5, 20.8, −3.6} in MNI space) for the Semantics > Syntax contrast, and {−44, 22, 10} ({−44.4, −23.2, 9.7} in MNI space) for the Syntax > Semantics contrast. To do so, we defined spherical regions of interest (ROIs) (of two different sizes: radius = 10 mm and 5 mm; available for download from https://osf.io/wtv9f/) around those activation peaks and extracted responses to the Semantics and Syntax conditions (including broken down by construction). We then performed one-tailed *t*-tests to evaluate whether the previously reported effects replicate in the current dataset.

And ***finally***, we gave the data the strongest chance to reveal a dissociation if such is present, using an individual-participant functional localization approach, which has been shown to benefit from higher sensitivity and functional resolution compared to group-based analyses (e.g., Saxe et al., 2006; Thirion et al., 2007; Nieto-Castanon and Fedorenko, 2012; see also the power calculation below). In particular, we searched, in each participant individually, for the most Semantics-preferring voxels (i.e., showing the strongest effect for the Semantics > Syntax contrast), and for the most Syntax-preferring voxels (i.e., showing the strongest effect for the Syntax > Semantics contrast) in the left inferior frontal gyrus (LIFG). To constrain the search, we used anatomical masks (Tzourio-Mazoyer et al., 2002) for the three sub-regions of LIFG: pars orbitalis (LIFGorb), pars triangularis (LIFGtri), and pars opercularis (LIFGop). To define individual functional regions of interest (fROIs), we divided the data in half, and using one half of the data we sorted the voxels within each mask by the *t*-value for the relevant contrast (i.e., Semantics > Syntax or Syntax > Semantics). We then chose the top 10% of voxels as the fROI. Thus, in each participant, we defined 6 fROIs: i) a Semantics-preferring fROI in LIFGorb, ii) a Semantics-preferring fROI in LIFGtri, iii) a Semantics-preferring fROI in LIFGop, iv) a Syntax-preferring fROI in LIFGorb, v) a Syntax-preferring fROI in LIFGtri, and vi) a Syntax-preferring fROI in LIFGop. We then extracted the responses to the Semantics and Syntax conditions (including broken down by construction) from the other half of the data and tested their difference using one-tailed *t*-tests. This analysis helps circumvent the high inter-individual variability that characterizes the human frontal lobes (e.g., Amunts et al., 1999; Tomaiuolo et al., 1999; Juch et al., 2005; Fedorenko et al., 2012b). Thus, even if the individual activation peaks for the Semantics > Syntax and Syntax > Semantics contrast are spatially variable enough so that group-level analyses (both whole-brain random-effects analysis, and ROI-based analysis) fail to detect them, this analysis would recover these effects if they hold across participants anywhere within the LIFG. The individual activation maps for the Semantics > Fixation and Syntax > Fixation contrasts are available for download at https://osf.io/wtv9f/.

To formally estimate power for this analysis, we first need to evaluate the expected effect size for the contrast between semantic and syntactic conditions in participant-specific fROIs. We do this in several steps: first, we note that, based on a sample of $n = 352$ participants (unpublished data from the Fedorenko lab), the average effect size for a robust contrast, between Sentences and Nonword-lists, is $d = 1.41$. Next, we note that in the left IFG, subtler contrasts, between different kinds of sentences, have been observed to elicit effect sizes that are about 60% of the Sentences > Nonwords effects (e.g., Blank et al., 2016). To err on the conservative side, we estimate the effect size in the current experiment to instead be 50% of the Sentences > Nonwords

effect, i.e., $d = 0.7$. This estimate is consistent with two other observations: first, a separate experiment in our lab (Fedorenko et al., 2018) estimated the difference between the reading of sentences with semantic (wrong word) vs. morpho-syntactic (wrong inflection) violations to be of similar magnitude ($d = 0.64$). And second, the effect sizes within participant-specific fROIs that we report for the current experiment are 1.25 (LIFGorb), 0.95 (LIFGtri), and 0.64 (LIFGop) (the effect sizes reported in the paper are smaller, because they were computed based on an independent-samples formula, which some claim is a more appropriate way to estimate effect sizes even for dependent-samples designs; in contrast, all estimates in the current paragraph are based on a dependent-samples formula, which is the estimate that is plugged into power calculations for dependent-samples designs). For an estimated effect size of $d = 0.64$–0.7, with $p = 0.05$, the power for our experiment is 75–82%.

RESULTS

Behavioral results

Dapretto and Bookheimer (1999) collected behavioral data from the scanned participants in a separate behavioral study (conducted at least 6 months after the fMRI session), and found that the two conditions were comparable in difficulty. We replicate similar across-condition accuracies and reaction times in our study (Fig. 2), although we collected the behavioral data during the scanning session. In particular, the accuracies for both conditions were close to 90% and not significantly different (Semantics condition: 87.3%

Table 2. Activation peaks in the random-effects group analyses for the contrasts of each condition against fixation (at $p < 0.0001$, uncorrected), and for the direct contrasts of the two conditions (at $p < 0.001$, uncorrected).

Comparison			Syntax Condition				Semantics Condition			
Region (Brodmann Area)			*x*	*y*	*z*	*t*	*x*	*y*	*z*	*t*
Inferior Parietal Lobule	(BA 1)	L					−46	−40	−58	6.06
Postcentral Gyrus	(BA 1)	L					−44	−24	54	6.30
Middle Frontal Gyrus	(BA 6)	L					−24	−10	52	7.91
Superior Frontal Gyrus	(BA 6)	L					−10	8	66	6.46
Middle Frontal Gyrus	(BA 6)	L					−34	8	46	6.43
Middle Frontal Gyrus	(BA 6)	L					−40	0	54	5.43
Inferior Parietal Lobule	(BA 7)	L					−38	−44	52	6.34
Inferior Frontal Gyrus	(BA 8)	L	−42	6	32	5.69				
Medial Frontal Gyrus	(BA 8)	L					−8	22	48	7.91
Middle Frontal Gyrus	(BA 8)	L					−46	14	46	7.57
Middle Frontal Gyrus	(BA 9)	R					52	26	30	6.85
Insula	(BA 13)	L					−34	24	2	9.20
Sub-Gyral	(BA 13)	R					32	22	8	6.21
Declive	(BA 19)	R					16	−64	−30	8.26
Middle Temporal Gyrus	(BA 21)	L	−54	−46	2	5.80	−56	−34	−4	8.86
Middle Temporal Gyrus	(BA 21)	L	−50	−26	−8	7.05	−50	−24	−8	12.15
Middle Temporal Gyrus	(BA 21)	R	−50	−34	0	8.29	62	−40	0	6.81
Temporal Lobe	(BA 21)	R					50	−28	−10	5.82
Fusiform Gyrus	(BA 37)	L					−36	−42	−24	7.97
Fusiform Gyrus	(BA 37)	R					38	−48	−16	6.44
Inferior Parietal Lobe	(BA 39)	L					−32	−58	48	7.38
Sub-Gyral	(BA 39)	L	−30	−52	42	7.60				
Superior Temporal Gyrus	(BA 39)	L	−48	−54	8	7.25				
Inferior Frontal Gyrus	(BA 44)	L	−50	12	26	6.61				
Inferior Frontal Gyrus	(BA 44)	R	46	14	26	5.81				
Inferior Frontal Gyrus	(BA 47)	L	−34	26	−2	8.90				
Caudate	(BA 48)	R					10	8	6	6.51
Thalamus	(BA 50)	L					−12	−12	10	7.61
Culmen	undefined	R					2	−54	−22	6.72
Sub-Gyral	undefined	R					40	−36	4	6.72
Extra-Nuclear	undefined	R	28	22	0	7.63				
Lentiform Nucleus	undefined	L	−16	−2	8	5.96				
Superior Frontal Gyrus	undefined	L	0	6	56	5.81				
Superior Frontal Gyrus	undefined	L	0	20	56	5.49				
undefined	undefined	L					−8	−58	−34	5.90
Comparison			**Syntax vs. Semantics**				**Semantics vs. Syntax**			
Region (Brodmann Area)			*x*	*y*	*z*	*t*	*x*	*y*	*z*	*t*
Precuneus	(BA 7)	R	10	−68	46	3.56				
Medial Frontal Gyrus	(BA 9)	L					−4	50	48	6.09
Superior Frontal Gyrus	(BA 9)	L					−12	56	32	7.53
Medial Frontal Gyrus	(BA 10)	L					−8	54	22	5.31
Inferior Frontal Gyrus	(BA 47)	L					−46	38	−12	6.10

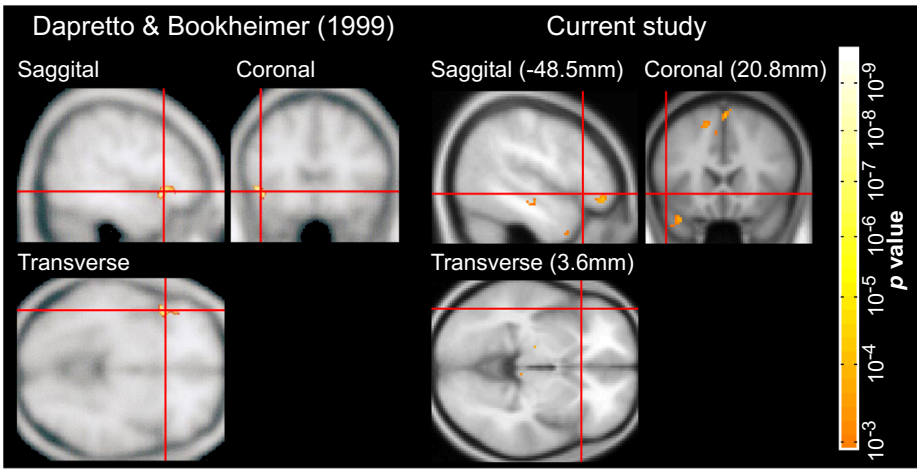

Figure 4. Whole-brain activation maps for the Semantics > Syntax contrast in Dapretto & Bookheimer's study (left; $p < 0.0001$, uncorrected) and the current study (right; $p < 0.001$, uncorrected). The crosshair for the map in the current study was centered on the peak in Dapretto & Bookheimer's study for ease of comparison.

±7.7 ($M \pm SD$); Syntax condition: 88.2% ±6.7; paired-samples $t_{(14)}=0.30$, $p>0.76$; Cohen's $d=0.11$, based on a conservative independent-samples test). Similarly, the RTs did not differ either when considering all trials (Semantics condition: 0.55 s ±0.13; Syntax condition: 0.53 s ±0.12; paired-samples $t_{(14)}=-0.39$, $p=0.7$; $|d| \ll 0.08$), or when considering correctly answered trials only (Semantics condition: 0.52 s ±0.14; Syntax condition: 0.54 s ±0.13; paired-samples $t_{(14)}=0.78$, $p>0.44$; $d=0.11$). Comparable behavioral performance suggests that whatever differences might be observed in neural responses between the two conditions would not be attributable to differences in cognitive effort.

fMRI results

Traditional random-effects analysis

Figure 3 shows whole-brain activation maps for each condition relative to the fixation baseline across the two studies. Visual examination of the maps suggests broad similarity between studies (see also Table 2 for a list of the activation peaks for each contrast in the current study), with, critically, robust responses detected for both contrasts in the left inferior frontal cortex.

Figure 4 shows the whole-brain activation map for the Semantics > Syntax contrast in Dapretto & Bookheimer's study and our study (centering the crosshair on the same stereotactic location). The Syntax > Semantics contrast did not reveal any significant peaks at a threshold of either 0.0001 or 0.001. The group-level (as well as individual) maps for all four contrasts (including versions of the maps smoothed with a larger, 8-mm, smoothing kernel) are available at https://osf.io/wtv9f/.

Activation-peak-based group-level ROI analysis

Figure 5 shows mean responses to the Semantics and Syntax conditions, including broken down by construction (Active / Passive vs. DO / PO), in each of the two activation peaks reported in Dapretto & Bookheimer's study. The peak reported as showing a Semantics > Syntax effect by Dapretto & Bookheimer showed a similar effect in our study (Semantics condition = 1.12±0.68 ($M \pm SD$), Syntax condition = 0.95±0.75, paired samples $t_{(14)}=1.97$, $p=0.03$; Cohen's $d=0.23$ based on an independent-samples test). However, the peak originally reported for the Syntax > Semantics effect also exhibited stronger activation in the semantic condition (Semantics condition = 0.82±0.51, Syntax condition = 0.63±0.48, $t_{(14)}=2.39$, two-tailed $p=0.03$; $d=0.37$). Reducing the size of the ROI from a 10-mm sphere to a 5-mm sphere did not affect these results (Figure 5B), which plausibly reflect the low sensitivity of group-level ROIs (e.g., Nieto-Castañon & Fedorenko, 2012). Further, this pattern of results was descriptively similar across the two constructions, with significant Semantics > Syntax effects in both ROIs for the Active / Passive alternation (which was shared between the current study and the original study), and non-significant effects in the same direction for the DO / PO alternation. These results argue against the idea that the failure to replicate the dissociation is due to the changes in the materials.

Individual-subject functional ROI analysis

Figure 6 shows responses to the critical conditions in individually defined functional ROIs. Here, half of the functional data was used to select the most Semantics- vs. Syntax-preferring voxels, in each participant separately and within each of the three sub-divisions of the LIFG. Then, responses in these voxels were independently estimated using the other half of the data. This analysis revealed reliable Semantics > Syntax effects in the Semantics > Syntax fROIs (i.e., fROIs consisting of most Semantics-preferring voxels) within LIFGorb (Semantics condition = 0.79±0.49, Syntax condition = 0.37±0.43, $t_{(14)}=4.85$, $p=10^{-4}$; $d=0.88$), LIFGtri (Semantics condition = 1.16±0.73, Syntax condition = 0.63±0.45, $t_{(14)}=3.70$, $p=0.001$; $d=0.83$), and LIFGop (Semantics condition = 0.94±0.64, Syntax

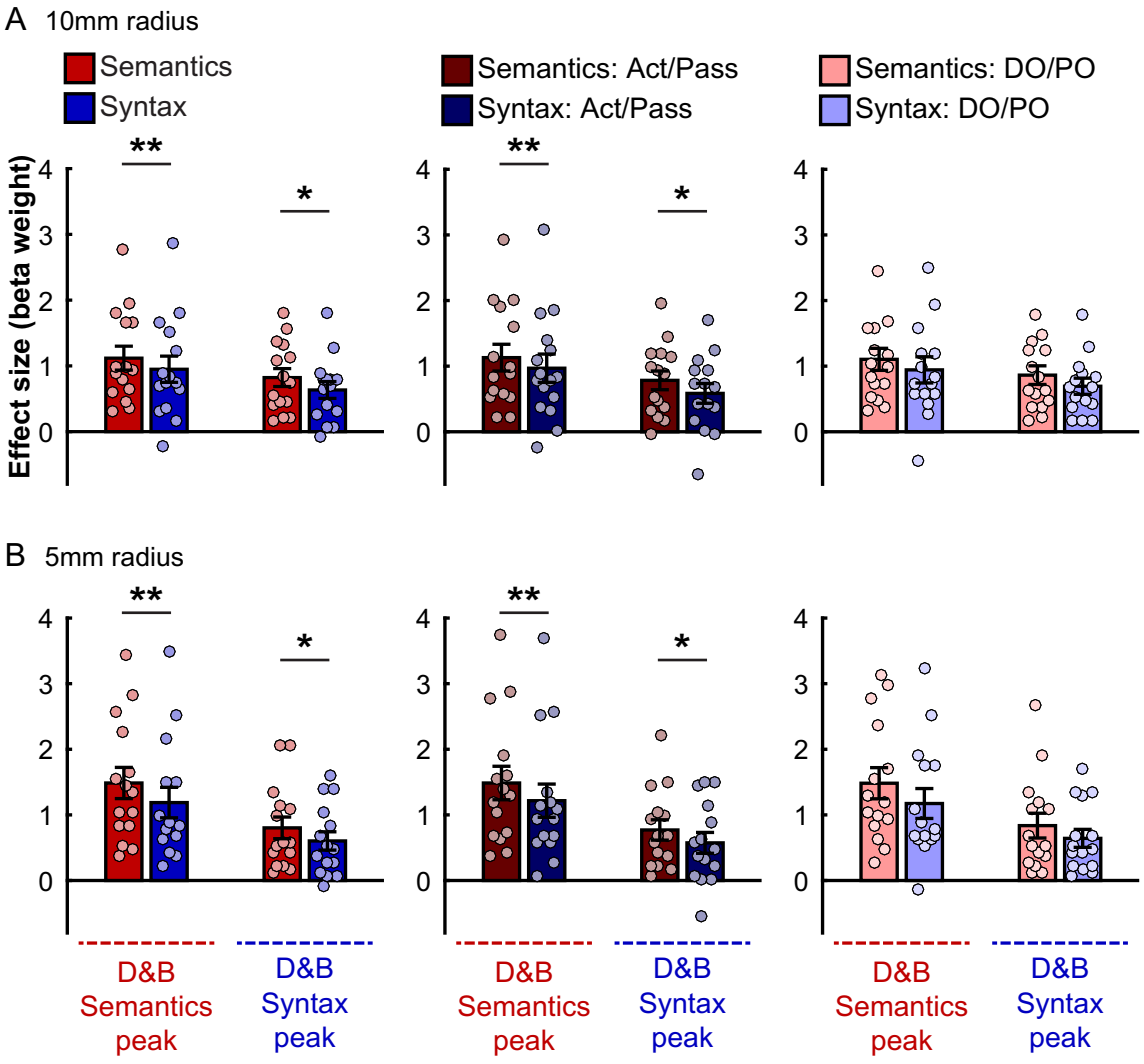

Figure 5. Responses to the critical, Semantics (red hues) and Syntax (blue hues), conditions in spherical ROIs defined around the peak coordinates reported by Dapretto & Bookheimer. (A) Spheres with a 10-mm radius. (B) Spheres with a 5-mm radius. Results are shown for all trials in each condition (left), or broken down by construction, i.e., only trials based on the Active / Passive (Act / Pass) alternation (middle), or only trials based on the Double Object / Prepositional Object (DO / PO) alternation (right). In each panel, the left pair of bars shows data for a sphere around Dapretto & Bookheimer's peak for the Semantics > Syntax contrast, and the right pair shows data for a sphere around Dapretto & Bookheimer's peak for the Syntax > Semantics contrast. Significant differences between conditions are marked with two stars for one-tailed tests, and a single star for two-tailed tests ($p<0.05$, uncorrected). Conventions are the same as in Figure 2.

condition = 0.69±0.55, $t_{(14)}$=2.47, p=0.014; d = 0.40). These effects suggest that the LIFG may contain areas that show robustly and replicably (across runs) greater engagement during the Semantics condition than the Syntax condition (although we note that the effect in LIFGop would not survive correction for multiple comparisons). Furthermore, these results seem stable across the two constructions: reliable for the Active/Passive alternation in all three fROIs (LIFGorb: $t_{(14)}$=4.69, p=0.0002, d=0.62; LIFGtri: $t_{(14)}$=3.46, p=0.002, d=0.66; LIFGop: $t_{(14)}$=2.18, p=0.02, d=0.34), and for the DO/PO alternation in the LIFGorb ($t_{(14)}$=3.92, p=0.0008, d=1.07) and LIFGtri ($t_{(14)}$=2.95, p=0.005, d=0.91), but not LIFGop ($t_{(14)}$=1.63, p=0.06, d=0.39). In contrast, the analysis of Syntax > Semantics fROIs (i.e., fROIs consisting of most Syntax-preferring voxels) did not reveal any replicable Syntax > Semantics effects in any of the three sub-divisions of the LIFG (ps>0.31). In fact, within the LIFGorb, the responses still showed a numerically stronger response to the Semantics than Syntax condition. This is striking (given that we specifically searched for most Syntax-preferring voxels) and suggests that no voxels within LIFG respond robustly and replicably (across runs) more strongly during the Syntax condition than the Semantics condition, at least in this paradigm. The fact that subject-specific fROI analyses are characterized by high sensitivity (e.g., Nieto-Castañon & Fedorenko, 2012), these results increase our confidence that the original Dapretto & Bookheimer finding was a false positive.

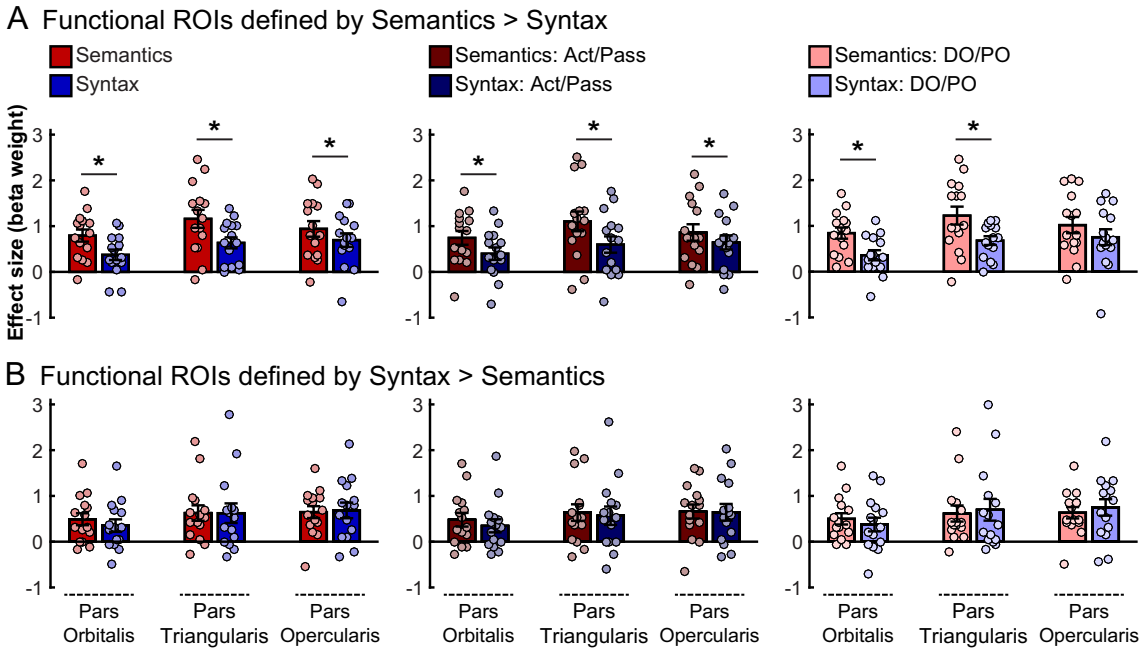

Figure 6. Responses to the critical, Semantics (red hues) and Syntax (blue hues), conditions in participant-specific fROIs defined by intersecting individual functional maps for either (A) the Semantics > Syntax contrast, or (B) the Syntax > Semantics contrast, with each of three anatomical masks for subdivisions of the inferior frontal gyrus (pars orbitalis: left pair of bars in each panel; pars triangularis: middle pair of bars; pars opercularis: right pair of bars). Half of the data was used to define the fROIs, and the other half was used to estimate the responses (using across-run cross-validation). Significant differences between conditions are marked with a star (one-tailed tests, $p<0.05$, uncorrected). Conventions are the same as in Figures 2 and 5.

DISCUSSION

To summarize, in a classic fMRI study, Dapretto and Bookheimer (1999) reported a dissociation between semantic and syntactic processing within the left inferior frontal gyrus. We here reported an fMRI study designed to conceptually replicate this early finding. We used the same two-condition design, but substantially expanded the set of experimental materials (five-fold), and included almost twice as many participants in order to increase statistical power.

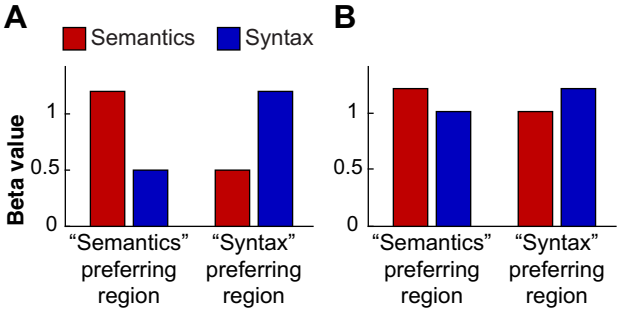

Figure 7. Hypothetical patterns of results in Dapretto & Bookheimer's study in two ROIs, one based on the Semantics > Syntax contrast, and the other on the Syntax > Semantics contrast. Semantics conditions are shown in red; syntax conditions are in blue. Left panel: a pattern with large effect sizes (strong selectivity for semantic vs. syntactic processing in the two ROIs, respectively); right panel: a pattern with small effect sizes (weak selectivity for semantic vs. syntactic processing).

Although the group-level whole-brain maps contrasting each condition to a low-level fixation baseline revealed broad similarity between the two studies (and between the two conditions), the direct contrasts of the Semantics and Syntax conditions did not replicate the originally reported dissociation. In particular, we found a number of reliable activation peaks for the Semantics > Syntax contrast, including within the LIFG, but the Syntax > Semantics contrast did not produce any reliable peaks within the LIFG. In line with this whole-brain analysis, we found a similar pattern in group-level ROIs defined around the original Semantics > Syntax, and Syntax > Semantics activation peaks from Dapretto & Bookheimer's study: the Semantics condition elicited reliably greater response in both the Semantics-peak ROIs, and the Syntax-peak ROIs. Finally, in an individual-participants functional localization analysis, which circumvents inter-individual anatomical and functional variability (rampant in the left frontal lobe, e.g., Amunts et al., 1999; Tomaiuolo et al., 1999; Juch et al., 2005; Fedorenko et al., 2012b), we were able to detect reliably greater responses to the Semantics than Syntax condition within the orbital and triangular sub-divisions of the LIFG. However, nowhere within the LIFG were there regions that responded reliably more strongly during the processing of the Syntax condition compared to the Semantics condition. Thus, the dissociation originally reported by Dapretto & Bookheimer does not appear to be robust to replication.

What can explain the non-replication of the original finding? The first, and perhaps most plausible, contributor

is the fact that Dapretto & Bookheimer appear to have relied on an analysis that treated participants as fixed effects rather than random effects. In a fixed-effects analysis, individual participants are not viewed as being randomly drawn from the population. Consequently, the results cannot be generalized beyond the sample tested, and the effects could be potentially driven by a small subset of participants (or even a single participant). The seminal publication about this significant limitation in many early brain-imaging studies had only come out a year earlier (Holmes and Friston, 1998), and thus it is possible that the authors had still relied on the fixed-effects analysis (it is difficult to determine this with certainty from the description provided in the Methods section).

Second, the original study used a small number of experimental items (eight per condition). This number is very low by the standards of language research, especially in cognitive neuroscience studies, where at least 20–30 items per condition are typically used to ensure generalizability. In the current study we used 40 unique trials per condition, and observed highly similar patterns across two distinct constructions. Thus, it is possible that in the original study, one or two of the items were driving the effects (see e.g., Bedny et al., 2007, for discussion).

It is also worth noting that Darpetto & Bookheimer, as is not uncommon in the fMRI literature, did not report the *magnitudes* of response to the Semantics and Syntax conditions. Thus, the effect size cannot be determined, only its significance (see Chen et al., 2017, for a discussion of this issue in fMRI research). More specifically, the significant peaks reported by Dapretto & Bookheimer are consistent with either of the hypothetical patterns shown in Figure 7. We suspect that the original result was more consistent with the possibility shown in the right panel of Figure 7, i.e., with small effect sizes. And small effects, especially observed in underpowered studies, are less likely to be real (e.g., Gelman and Carlin, 2014; Simonsohn, 2015; Open Science Collaboration, 2015).

To conclude, although the question of whether distinct pools of cognitive resources and cortical regions support lexico-semantic and syntactic processing is likely to keep generating controversy and further research (see also Fedorenko et al., 2018), we here found that at least one study that is commonly cited as evidence for this dissociation does not appear to replicate in an experiment with a similar design, materials, and greater statistical power. It may be important to ask, as researchers have recently done in the field of psychology (e.g., Open Science Collaboration, 2015), what proportion of fMRI studies is robust to replication (see Hong et al., 2019, for a discussion).

CONTRIBUTIONS

E.F. conceived the study and developed the experimental materials. Z.M. created the presentation script. All authors collected and analyzed the data. I.B. created the figures. M.S., I.B., and E.F. wrote the manuscript, with comments from Z.M.

ACKNOWLEDGMENTS

E.F. was supported by the NIH awards R00-HD057522, R01-DC016607, and R01-DC016950. We also acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, MIT. For technical support during scanning, the authors thank Atsushi Takahashi and Steve Shannon.

REFERENCES

Allen K, Pereira F, Botvinick M, Goldberg AE. (2012) Distinguishing grammatical constructions with fMRI pattern analysis. Brain Lang 123(3):174-182.

Amunts K, Schleicher A, Burgel U, Mohlberg H, Uylings HBM, Zilles K. (1999) Broca's region revisited: cytoarchitecture and intersubject variability. J Comp Neurol 412(2):319-341.

Baggio G, Hagoort P. (2011) The balance between memory and unification in semantics: A dynamic account of the N400. Lang Cogn Process 26(9):1338-1367.

Bates E, Goodman JC. (1997) On the inseparability of grammar and the lexicon: evidence from acquisition, aphasia and real-time processing. Language and Cognitive Processes. 12(5-6):507-584.

Bautista A, Wilson SM. (2016) Neural responses to grammatically and lexically degraded speech. Lang Cogn and Neurosci 31(4):567-574.

Bedny M, Aguirre GK, Thompson-Schill SL. (2007) Item analysis in functional magnetic resonance imaging. NeuroImage. 35(3):1093-1102.

Bemis DK, Pylkkänen L. (2012) Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. Cereb Cortex 23(8):1859-1873.

Blank I, Balewski Z, Mahowald K, Fedorenko E. (2016) Syntactic processing is distributed across the language system. NeuroImage. 127:307-323.

Braze D, Mencl WE, Tabor W, Pugh KR, Constable RT, Fulbright RK, Magnuson JS, Van Dyke JA, Shankweiler DP. (2011) Unification of sentence processing via ear and eye: an fMRI study. Cortex. 47 (4):416-431.

Buchweitz A, Mason RA, Tomitch L, Just MA. (2009) Brain activation for reading and listening comprehension: an fMRI study of modality effects and individual differences in language comprehension. Psychol Neurosci 2(2):111-123.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14(5):365-376.

Chen G, Taylor PA, Cox RW. (2017) Is the statistic value all we should care about in neuroimaging? Neuroimage 147:952-959.

Cooke A, Grossman M, DeVita C, Gonzalez-Atavales J, Moore P, Chen W, Gee J, Detre J. (2006) Large-scale neural network for sentence processing. Brain Lang 96(1):14-36.

Dale AM. (1999) Optimal experimental design for event-related fMRI. Hum Brain Mapp 8(2-3):109-114.

Dapretto M, Bookheimer SY. (1999) Form and content: dissociating syntax and semantics in sentence comprehension. Neuron. 24 (2):427-432.

Dick F, Bates E, Wulfeck B, Utman JA, Dronkers N, Gernsbacher MA. (2001) Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. Psychol Rev 108(4):759.

Dryer MS. (2002) Case distinctions, rich verb agreement, and word order type (comments on Hawkins' paper). Theoretical Linguistics 28:151-158.

Duffau H, Moritz-Gasser S, Mandonnet E. (2014) A re-examination of neural basis of language processing: Proposal of a dynamic hodotopical model from data provided by brain stimulation mapping during picture naming. Brain Lang 131:1-10, https://doi.org/10.1016/j.bandl.2013.05.011.

Embick D, Marantz A, Miyashita Y, O'Neil W, Sakai KL. (2000) A syntactic specialization for Broca's area. Proc Natl Acad Sci U S A 97 (11):6150-6154.
Fedorenko E, Hsieh PJ, Nieto-Castanon A, Whitfield-Gabrieli S, Kanwisher N. (2010) New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J Neurophysiol 104(2):1177-1194.
Fedorenko E, Nieto-Castanon A, Kanwisher N. (2012) Lexical and syntactic representations in the brain: an fMRI investigation with multivoxel pattern analyses. Neuropsychologia. 50(4):499-513.
Fedorenko E, Duncan J, Kanwisher N. (2012) Language-selective and domain-general regions lie side by side within Broca's area. Curr Biol 22(21):2059-2062.
Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, Schalk G, Kanwisher N. (2016) Neural correlate of the construction of sentence meaning. Proc Natl Acad Sci 113(41):E6256-E6262.
Fedorenko E, Mineroff Z, Siegelman M, Blank I. (2018 Jan) Word meanings and sentence structure recruit the same set of fronto-temporal regions during comprehension. bioRxiv 1:477851.
Friederici AD, Meyer M, von Cramon DY. (2000) Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. Brain Lang 74(2):289-300.
Friederici AD, Kotz SA, Scott SK, Obleser J. (2010) Disentangling syntax and intelligibility in auditory language comprehension. Hum Brain Mapp 31(3):448-457.
Friederici AD. (2012) The cortical language circuit: from auditory perception to sentence comprehension. Trends Cogn Sci 16(5):262-268.
Gelman A, Carlin J. (2014) Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect Psychol Sci 9 (6):641-651.
Gibson E, Bergen L, Piantadosi ST. (2013) Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. Proc Natl Acad Sci U S A 110(20):8051-8056.
Herrmann B, Obleser J, Kalberlah C, Haynes JD, Friederici AD. (2012) Dissociable neural imprints of perception and grammar in auditory functional imaging. Hum Brain Mapp 33(3):584-595.
Holmes A, Friston K. (1998) Generalisability, random effects and population inference. NeuroImage 7(4):S754.
Hong Y, Yoo Y, Wager T, Woo CW. (2019) False-positive neuroimaging: undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. bioRxiv :514-521.
Humphries C, Binder JR, Medler DA, Liebenthal E. (2006) Syntactic and semantic modulation of neural activity during auditory sentence comprehension. J Cogn Neurosci 18(4):665-679.
Ioannidis JPA. (2005) Why most published research findings are false. PLoS Med 2(8):696-701.
Ioannidis JPA, Munafo MR, Fusar-Poli P, Nosek BA, David SP. (2014) Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends Cogn Sci 18(5):235-241.
Juch H, Zimine I, Seghier ML, Lazeyras F, Fasel JHD. (2005) Anatomical variability of the lateral frontal lobe surface: implication for intersubject variability in language neuroimaging. NeuroImage. 24 (2):504-514.
Kuperberg GR, Holcomb PJ, Sitnikova T, Greve D, Dale AM, Caplan D. (2003 Feb 15) Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. J Cogn Neurosci 15 (2):272-293.
Menenti L, Gierhan SM, Segaert K, Hagoort P. (2011) Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. Psychol Sci 22(9):1173-1182.
Nee DE. (2019 Apr 12) fMRI replicability depends upon sufficient individual-level data. Commu Biol 2(1):130.
Nieto-Castanon A, Fedorenko E. (2012) Subject-specific functional localizers increase sensitivity and functional resolution of multisubject analyses. NeuroImage 63(3):1646-1669.
Noppeney U, Price CJ. (2004) An fMRI study of syntactic adaptation. J Cogn Neurosci 16(4):702-713.
Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Yarkoni T.. (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18(2):115.
Santi A, Grodzinsky Y. (2010) fMRI adaptation dissociates syntactic complexity dimensions. Neuroimage 51(4):1285-1293.
Saxe R, Brett M, Kanwisher N. (2006) Divide and conquer: a defense of functional localizers. Neuroimage 30(4):1088-1096.
Schmidt S. (2009 Jun) Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Rev Gen Psychol 13(2):90.
Scott TL, Gallee J, Fedorenko E. (2016) A new fun and robust version of an fMRI localizer for the frontotemporal language system. Cogn Neurosci :1-10.
Segaert K, Menenti L, Weber K, Petersson KM, Hagoort P. (2012) Shared syntax in language production and language comprehension — an fMRI Study. Cereb Cortex 22(7):1662-1670.
Simmons JP, Nelson LD, Simonsohn U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22(11):1359-1366.
Simonsohn U. (2015) Small telescopes: detectability and the evaluation of replication results. Psychol Sci 26(5):559-569.
The Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science. 2015;349(6251): aac4716.
Thesen S, Heid O, Mueller E, Schad LR. (2000) Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. Magn Reson Med 44(3):457-463.
Thirion B, Pinel P, Mériaux S, Roche A, Dehaene S, Poline JB. (2007 Mar 1) Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. Neuroimage 35(1):105-120.
Tomaiuolo F, MacDonald JD, Caramanos Z, Posner G, Chiavaras M, Evans AC, Petrides M. (1999) Morphology, morphometry and probability mapping of the pars opercularis of the inferior frontal gyrus: an in vivo MRI analysis. Eur J Neurosci 11(9):3033-3046.
Tyler LK, Marslen-Wilson WD, Randall B, Wright P, Devereux BJ, Zhuang J, Stamatakis EA. (2011) Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. Brain 134:415-431, https://doi.org/10.1093/brain/awq369.
Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15 (1):273-289.
Ullman MT. (2016) The Declarative/Procedural Model : A Neurobiological Model of Language Learning, Knowledge, and Use, 2016. In Neurobiology of Language: Academic Press.
Vagharchakian L, Dehaene-Lambertz G, Pallier C, Dehaene S. (2012) A temporal bottleneck in the language comprehension network. J Neurosci 32(26):9089-9102.

(Received 18 December 2018, Accepted 3 June 2019)
(Available online 11 June 2019)