Hi Ev and Melissa,

Here's the final (I hope) version of the PTB script along with the necessary input data (you just have to change the rootdir on line 19 to run it on your computer). More below on how it works. First, a few things about our first participant today:

- *timing error.* Ev, as I mentioned, the trial onsets drifted to 0.67 sec into the TR by the end of each run. I fixed that (should've noticed it earlier...). I don't think that's a big enough error to affect results by much (and certainly not big enough to exclude data);

- *responses.*For some reason, the button presses and RTs for the 4th run didn't record. I'm guessing the reason is that the button box cable got unplugged. I did several test runs after I saw that and everything works fine (also, the way the code works, I can't think of a reason why the responses for just one run wouldn't record). Anyway, no big deal: we still have most of the behavioral data.

- *behavioral results.*I looked at 4 things: 1) proportion female on target trials, per condition (missed trials not counted; i.e., prop.male + prop.female = 1); 2) proportion correct on catch trials; 3) RTs on target trials per condition; 4) RTs on catch trials. I'm attaching bar graphs. Error bars on the RTs are standard errors across items.

So, apparently, this participant thinks that women tend to respond in the negative more than men :) Either that, or, more problematically, there's a systematic difference between how female-like our "no" answers sound. Let's see what happens with a few more participants? Anyway, the important thing is that both proportion female and RTs even out for the literal and non-literal conditions when you collapse across yes/no, so we don't have a confound in this subject. Performance on catch trials was okay: 67% correct; note that's for 6 trials only (bc responses weren't recorded for the remaining 2).

Ev, I added 2 parts to the code that weren't there this morning: 1) format SPM inputs and 2) get descriptive stats & make bar graphs (at the end of run 4). Let's run the new version tomorrow.

The code has a bunch of comments in it, but I figured I'd briefly describe what each section does, in case you don't feel like going through it. I also thought that might help Melissa identify the fMRI-specific parts.

*Lines  What they do*

[50-60]: get a different random seed for the random number generator for each subject

[70-90]: generate lists that pick different halves of the 48 yes/no items to be assigned to the non-literal condition. I've used 16 such lists (e.g., [1 2 1 2...], [2 1 2 1...], [1 1 2 2...], etc.) ordered so that the same item occurs once as literal and once as non-literal across consecutive 2 subjects.

[92-103]: randomize item order within each of the 5 conditions (well-yes,well-no,lit-yes,lit-no, and catch);

[105-122]: get counterbalancing scheme: 6 4x4 latin squares concatenated to get 24 target items per run in 4 runs. This guarantees that each condition occurs once in each of 24 positions across runs. The squares are chosen from the possible 24 possible 4x4 latin squares and ordered in a way that preserves 1-st order counterbalancing (i.e., each condition precedes each different condition an equal number of times and precedes itself the same number of times as other conditions precede themselves). So we end up with a 4x24 matrix, where each row is a run. A different order of the 4 runs is selected for each participant (in 24 Ps, all orders will have occurred once).

Note for Melissa: you might know this, but in case you don't: besides the usual reasons why counterbalancing is good, I can think of (at least) 2 extra reasons why it's important in fMRI: 1) in fast event-related designs (like ours), there is some risk of spillover effects if the randomization of the rest durations isn't perfect; the n-1 counterbalancing takes care of that (if your trials are long enough); 2) there is low-frequency drift in fMRI, which changes mean signal levels over the course of a run, so you don't want many trials of 1 condition to be bunched together early on and others late on -- you might get spurious effects or real effects might go undetected.

[124-133]: insert 2 catch trials in random positions within each run.

[135-146]: load timing information generated by optseq2: consists of 4 lists (1 per run) with jittered rest period durations; the jitter is optimized to allow recovery of the hemodynamic response for each condition; together with the counterbalancing, this ensures no spillover effects.

[160-171]: set up PTB stuff

[173-184]: show instructions, wait for trigger

[186-278]: loop over trials, showing question, then answer, then response-prompt, then rest. Record key presses and RTs, as well as the \*actual\* onsets and durations, which should match the desired ones closely-enough (right now, our error is at ~10 msec per trial, which is negligible for fMRI).

Melissa, a note on timing: since the code takes a bit of time to run, timing differs slightly from what you want it to be and small errors can add up over successive iterations (as they did today..). That's why it's good to build in a way to correct the timing on each trial. The way it's done here is that the rest period doesn't run for a fixed duration, but until the next trial is supposed to start. So each rest is a tiny bit shorted than its nominal duration, but the important thing is that trials start on the TR.

[280-306]: SPM stuff: 1) convert onsets and durations into TRs and format them the way SPM wants them; 2) specify contrast vectors: which conditions will you be comparing:  [1 1 -1 -1 0] &[-1 -1 1 1 0].

[318-397]: after collecting all the data for a participant (4 runs), process RTs and button presses and make figures.

Okay, that's it. Let me know if I should change something or if you have questions. If either of you get a chance to do a run or 2 and/or go over the code before we've scanned too many participants, please let me know if you have feedback, especially if you notice something suspicious :)

Alex

<nonlit\_fmri.zip>

<EV\_nonlit\_01\_behavior.jpg>

Hi Alex,  
Thanks!  Replies interleaved below! :)  
  
> On Feb 25, 2015, at 7:28 PM, Alexander M Paunov <[apaunov@mit.edu](mailto:apaunov@mit.edu)> wrote:  
>  
> Hi Ev and Melissa,  
>  
> So there was a mistake in the part of the nonlit code that determines accuracy on the reality-check trials. Sorry about that. Performance for all 3 subjects is actually 100%.  
  
Yeah, I figured that must be the case: all three subjects seemed pretty conscientious. :)  Good to hear!  
  
> For EV\_nonlit\_01 there are 2 missing trials (because responses didn't record in run 4); for EV\_nonlit\_03 there's 1 missing trial (because the subject didn't respond within the 3 sec window). I'm attaching the correct bar graphs.  
  
perfect  
  
> Looking at proportion female, 2 out of the 3 subjects judged the nonliteral answers to be more female-like (by 10% and by 20%, relative to the literal ones).  
  
And those subjects were female, no?  I think there are potentially interesting things to say about the behavioral data, regardless of what we discover in the fmri data..  
  
> It seems likely that this difference won't go away as we add more subjects (because the "well" is probably somewhat stereotypically female).  
  
yep  
  
> As we talked, that's probably not a big deal because it's hard to tell a plausible story that nonlit vs lit differences in ToM, language, or MD regions are really attributable to "female-ness" judgments. But, as Ev suggested, at some point (before we run more subjects) we should code the trials in terms of responses and see what happens (ideally, there won't be a main effect of female vs male in our regions nor a gender-judgment x literalness interaction).  
  
yep, all sounds good  
  
> As for RTs, it looks promising that there won't be a difference between the lit and nonlit conditions (although it's too early to tell).  
  
agreed, but looks pretty good anyway: no huge difference in any case.  
  
> Ev, I'm attaching the para files for your pipeline.  
  
awesome, thanks!  
  
> I haven't made different para files for EV\_nonlit\_01 because the timing error was always less than 1/2 a TR.  
  
you mean it wasn’t cumulative over the run, right?  if so, then that should indeed be ok.  
  
> Also, here's the correspondence of condition orders and runs:  
>             Run:   1     2     3     4  
> EV\_nonlit\_01: [2     1     3     4]  
> EV\_nonlit\_02: [3     4     1     2]  
> EV\_nonlit\_03: [2     4     1     3]  
>  
  
great!  i’ll try to take a look at the data when I am in Russia/Europe  
ev  
  
> That's it for now!  
>  
> Alex  
>  
> <nonlit\_1.para><nonlit\_2.para><nonlit\_3.para><nonlit\_4.para>  
> <EV\_nonlit\_01\_behavior.jpg><EV\_nonlit\_02\_behavior.jpg><EV\_nonlit\_03\_behavior.jpg>  
>

Hi Ev and Melissa,

I'm writing with the first-pass fMRI results for our 3 nonlit subjects. Sorry in advance for (another) long email. I'm attaching two folders with results (one with some whole-brain SPMs and one with ROI results).

Ev, before I go into the results, I'm also attaching a folder with new para files. I just realized that we need individual para files for each subject because the onset-condition pairing varies from subject to subject. That's because the 4 timing schedules we're using are always in the same order (so, identical onsets for run 1 in subs 1, ..., n, for run 2 in subs 1, ..., n, etc.), whereas the condition lists are in a different order for each subject. I can easily change the PTB code so that in the future we'll only need the same 4 para files across subjects, but the advantage of the current scheme is that across subjects we will have an identical distribution of rest durations for each condition. Does all this make sense? I hope you haven't modeled the data with the nonsense para files yet -- sorry about that. Also, I added the code for writing the para files to the PTB script -- we'll get them automatically for each subject.

Now, about results. First I ran a standard HRF-GLM but found no nonlit > lit effects in ToM regions and decided to run an FIR model to get a sense of the time-course -- the conditions can only start diverging when the answer is presented, so if we model the entire trial (as we have to because we have no jitter between Q & A), small differences during the answer period might go undetected. Both the whole-volume and ROI figures are from the FIR model. I'll send figures with the HRF model when I get a chance to make them. Those should be useful at least for comparability with results from Ev's pipeline. Ev, would you like me to also make para files for FIR?

*Whole-brain maps*

The SPMs show nonlit > lit at p < 0.01, k > 10; the contrast includes all the modeled time points (10 per condition) with an equal weight; so, t-stats could be driven by a big condition difference on few time points or smaller differences on more time points. I also did this contrast for just the last 6 time points (excluding most of the signal related to the question), and things look similar overall. I've used a low significance threshold because effects are small: many activations aren't there at p < 0.001, and the maps aren't too noisy at p < 0.01. I get roughly similar results with the HRF model (so the effects aren't small because of loss of degrees of freedom with the FIR model).

So, the blobs we get look mostly like MD and language regions; they seem to be right-lateralized in subjects 2 & 3. I need to double check my design matrix for subject 1 -- there are lots of occipital activations, which doesn't make sense.

1) Overall, it doesn't look like there's much going on around ToM regions, sadly (though inconclusively at this stage).

2) The MD-like activations make sense on the story that nonlit sentences are harder, as well as that our activations are confounded with inhibition of the "female" choice on non-literal trials (haven't done the response-coded models yet).

3) The apparent RH dominance is consistent with literature on the role of the RH in non-literal language.

4) I don't know what to say about the LH language-y blobs; insofar as we're looking at a coherence effect of sorts, I've previously seen coherence preference in LIFG and some other LH language ROIs.

Could you remind me what you guys got for language and MD in your jokes study? (maybe send slides?)

Don't know what to make of the following yet, but seemed potentially interesting: in subject 3 (the last 2 pictures in the folder), we have blobs well within in the vicinity of RTPJ and PC (both in green), but they're mostly non-overlapping with the subject's ToM localizer (in red). There's a similar situation for subject 1's RTPJ; subject 2 has a blob that's right on the border of the RTPJ \*parcel\* (might be a parietal MD region). I'm curious if this turns out to be a reliable pattern with more subjects (and if so, what it might mean).

*ROIs*

The ROIs are top 100 voxels with no contiguity constraint (I didn't use the spm\_ss toolbox for the current version, but will switch over to it, and use top 10% voxels for consistency). I haven't worried about overlapping voxels between networks, but at some point we should look at that. The attached ROI folder includes subject-wise and average event-related time courses (separately for LH\_Lang, RH\_Lang, LH\_MD, RH\_MD, and ToM ROIs). The lines are nonlit (blue) and lit (green). The x axis is time in TRs (the trial is done at point 5). Note that although the y-axes are fixed for each ROI type, they do vary between types (i.e., for Lang vs MD vs ToM).

1) Looking at the AVG figures, the ToM ROIs aren't promising. The TPJs don't look like they're responding much at all, and, if anything, MMPFC and/or RSTS might prefer literal. There might be some non-literal preference in RTPJ during later time points (the answer part of the signal), but I wouldn't get my hopes up. I also looked at individual subject t-stats (but not fixed effects across subjects) for the main contrast (nonlit > lit; but not, say, LH vs RH, etc.). I find effects in RH\_STS (03) (but lit > nonlit) and RH\_TPJ (~01), if you contrast just the last 6 time-points. The numbers in parentheses are subject numbers; the "~" means marginal (p < 0.1).

2) For MD, it does seem that the preference for nonlit is somewhat right-lateralized -- most ROIs show it, qualitatively. The significant effects are LH\_MFGorb (03), RH\_IFGop (02), RH\_MFG (~01,02), RH\_MFGorb (~01,~02), RH\_PrecG (02), and RH\_ParInf (02).

3) In the language AVG figures, it looks like 2 ROIs prefer non-literal: IFG and IFGorb, bilaterally. The sig. ROIs within subjects are LH\_IFGorb (03), LH\_AngG (~01), RH\_IFG (~02), and RH\_MFG (~02). In the RH AVG figure, there's an early preference for literal in several ROIs, which is too early and must be noise.

Notes on analysis:

1) Spikes in the signal aren't accounted for in the current version. The code I used adds an artifact regressor for each spike directly in the model; this uses up too many degrees of freedom in combination with the number of FIR predictors, so doesn't work. One way around that is to NaN-out the spikes beforehand. Another is to use 1 regressor per condition across runs (currently, each run is modeled with a separate regressors; then betas are averaged across runs); this would reduce the number of regressors 4-fold and shouldn't be a problem if the data are preprocessed well.

2) More generally, seeing as we might be dealing with small effects, it's worth considering preprocessing with Wandell et al's GLM-denoising toolbox ([http://kendrickkay.net/GLMdenoise/](http://kendrickkay.net/GLMdenoise/" \t "_blank)). I've used it once before and it made a huge difference. Basically, what it does is find optimized HRF for each subject, find voxels that don't respond to the task at all, find noise components among these voxels with PCA, and regress out the noise components from the task-responsive voxels. It would take me a day or two to get the results with this preprocessing, but it seems worth it. Should I go for it?

3) Should we try MVPA and/or WSC?

Notes on design:

It seems too early to consider changing the design (get another ~3 subjects first?) but if we want to consider more drastic measures, a few possibilities that might be relevant:

1) Add context? ToM regions, esp. the TPJs and PC don't like short language stimuli; they still prefer mental over non-mental but the signal changes from baseline are much smaller. Relevant data: a) Rebecca and I have data with ToM stories of different lengths (from 1 to ~10 sentences); b) Hasson et al only find ToM ISCs for unscrambled stories and ones scrambled at the paragraph level (not for sentences, words, or backward stories); and c) a number of studies on irony, metaphor, and sarcasm don't find ToM regions unless there's context (cf. Spotorno et al). So it might be worth adding a sentence or two of context?

2) Add jitter between Q & A (and maybe reduce between-trial rest to save time; it's a bit long now anyway)? That way we can cleanly recover the HRFs for the answer alone, and not rely on washing out the Q-A spillover effect across subjects.

3) Revert to old task? Insofar as one subject is any indication, we seem to get stronger overall responses with the previous task (which was to judge whether answer is yes/no/incoherent in nonlit, lit\_coh, and lit\_incoh trials). This might not be relevant for ToM activations, specifically, either because it boosts responses across the board (task engagement?) or because it boosts responses in particular networks (MD, if task is harder?) Besides that, the main downside I can think of is that the task is essentially different in the literal and nonliteral conditions: on nonlit trials, it's inferring whether the answer is yes/no; on lit trials, it's inferring whether the Q-A pair is coherent or incoherent; this might be a confound, at least in the MD network. The upside is that we get to compare responses for incoherent vs non-literal (e.g., in my coherence study, the language voxels in IFGorb preferred incoherent, and now they seem to prefer non-literal, whereas IFG voxels preferred coherent and now they prefer non-literal).

That's it for now! Looking forward to your questions and thoughts and figuring out what to do next.

Alex

PS: Ev, I've been working on the ISC/WSC data too. I have to pause for a bit to get ready for my meeting with Ted, but I'll get back to it ASAP and will hopefully have results this week. Haven't done much on the ISC-mega stimuli, but Idan and I talked about finalizing the set this week as well. We'd still need to edit (matching length, cutting out text from movie clips) and write the PTB script. Realistically, we won't be scanner-ready before the middle of next week. More soon!