Ling 245 Project Presentation

Benjamin Sparkes

June 11, 2018

Motivation

> *Our approach was to test whether enrichments can be primed across expressions. If different sorts of enrichments can prime each other, there must be an abstract mechanism that is shared between them. By testing which enrichments prime each other and which don't, we can specify what the common mechanism might be.* *(2016, p. 118)*

☞ Are there are shared reasoning processes which apply to distinct instances of enrichment via alternatives, or whether each category of enrichment has its on specialised process?

- $\leadsto$ Bott and Chemla are interested in whether priming can occur at all *given* a prior assumption of enrichment via alternatives.

Thoughts on the Question

The question of whether or not there are shared reasoning processes which apply to distinct instances of enrichment via alternatives, or whether each category of enrichment has its on specialised process has a nice cognitive feel.

> Intuitively there's some positive upshot whichever way the data points.
>
> (i) If there is cross-category enrichment, then there is a need to posit shared reasonig processes.
>
> (ii) If there is no cross-category enrichment, then one should posit distinct reasoning processes.

However, it is important to note that each of these carries a presupposition that the data can/should/will support one of these resolutions. As we shall see, there is no guarantee that the data will be so clean.

Bott and Chemla's Experiments

Bott and Chemla ran three experiments.
Participants were presented with trials consisting of a setence and two picutures, and are asked to select the picture which best reflects the sentence.
Trials are split into *prime* and *response*, and every response trial is preceeded by two prime trials which are used to ensure the participant considers certain alternatives.

Bott and Chemla's Experiments

Prime trials:

Weak: One false picture, and one true but underinformative picture.

Strong: Two true pictues. One true informative picture, the other underinformative.

Response trials:

One picture true but underinformative picture.

A picture containg the words 'Better Picture?'

☞ Participants were instructed to click the 'Better Picture?' option if they felt the the other picture did not sufficiently capture the sentence meaning.

Details of the Experiment

The sentences were constructed using one of two frames:

(i) Some of the symbols are [symbol]
(ii) There are four [symbol]

Bott and Chemla included a third frame:

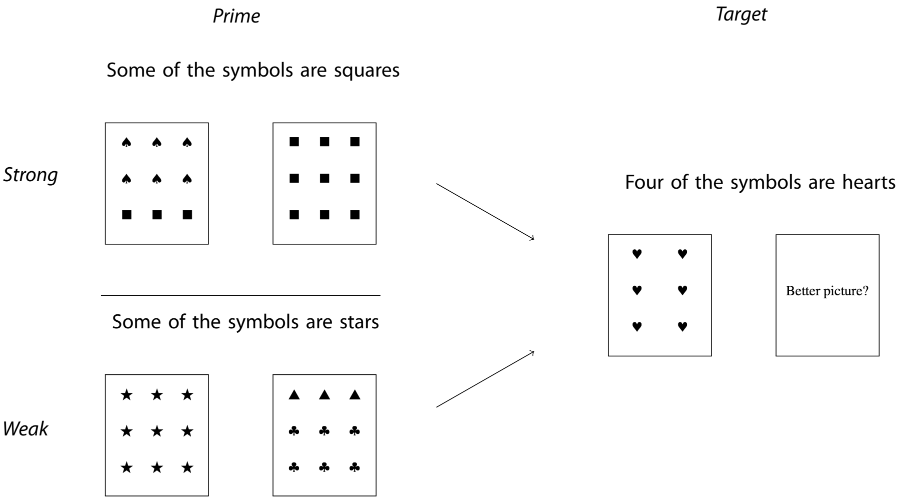
(iii) There is a [symbol].

The symbols were randomly chosen from:

♦ ♣ ✓ ♠ ♥ ■ ★ ● ♩ ▲

Sentences and pictures were generated for each participant at the start of each experiment.
Four instances of each prime strength, prime category, and reponse category combination were included in the trial.

Example


Prime
Target
Some of the symbols are squares
Strong
Four of the symbols are hearts
Better picture?
Some of the symbols are stars
Weak

‘Correct’ Responses

For each prime trials there was a ‘correct’ response, either due to the semantic content of the sentence in the case of weak trials, or due to pragmatics in the case of strong trials.

> *[I]n the presence of both a weak picture and a strong picture, participants could not make a non-arbitrary choice solely based on the truth conditions of the weak interpretation which is true in both cases, hence the strong reading is a favored option in that it provides a non-arbitrary way to resolve the task.* *(2016, p. 124)*

Linking hypothesis:

☞ Prior trials will effect how participants evaluate sentences, and that in response trials participants click on ‘Better Picture?’ if they process the setence pragmatically, and the semantically adequate picture otherwise.

Picture Details

Some trials:

Strong pictures: involved three symbols matching the predicate in the sentence, and six of another type.
Weak pictures: Nine symbols matching the predicate in the sentence.
False pictures: Nine symbols of the same type which did not match the predicate.

Number4 trials:

Strong pictures: Symbols matching the number and predicate in the sentence, the number was always 'four'.
Weak pictures: A greater number of symbols than in the sentence which matched the predicate, this was always six.
False pictures: A smaller number of symbols than in the sentence which matched the predicate, this was always two.

Filler Trials

Filler trials were also included:

All sentences, an alternative to *some*.
*Number*6 sentences, an alternative to *number*4.

Each could occur in three forms:

(1) a weak picture with symbols that did not match the predicate in the sentence, and a “Better Picture?” option
(2) a weak picture with symbols that matched the predicate, and a “Better Picture?” option, and
(3) a weak picture with symbols that matched the predicate, and a strong picture.

Bott and Chemla used these to highlight alternatives to participants.
One filler triplet every six target triplets.

The ‘Replication’

Partial replication for two reasons:

1) Half the number of participants compared with the original experiment (100 and 200 participants, respectively)
2) Only two enrichment categories, as opposed to three in the original.
3) We included keyboard shortcuts
4) Areas where Bott and Chemla weren’t super clear on details

The basis for both modifications were straightforward cost considerations.

By uncommenting a few lines of code (and fixing any bugs that this may cause) the full experiment can be run.

`https://github.com/bsparkes/bottchemla2016`

`https://bsparkes.github.io/bottchemla2016/experiment/html/bottchemla2016.html`

Predictions

Bott and Chemla did not make predictions regarding the results of the experiment.

$\rightsquigarrow$ Core interest is in how the question about EVAs should be resolved.

They do note:

> *If enrichment can be primed at all, we would expect within-category priming …[and] …If the* numbers, some *and* ad hoc *EVAs share enrichment mechanisms we would expect them to prime each other, so that a strong some prime, for example, leads to a greater proportion of strong number responses.*
>
> *(2016, p. 122)*

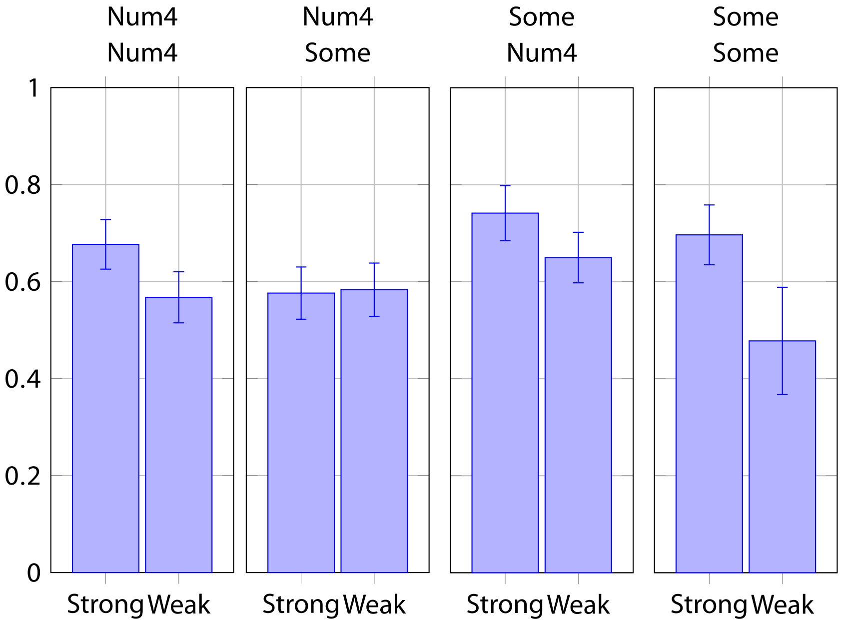
From the results of Bott and Chemla's experiment, one would expect to see a significant effect of priming, both within and between categories.

Results

1. Significant replication of priming effects in general.
2. Almost significant replication of within category priming effect.
3. Failure to replicate between category priming effect.
4. Different kinds of priming effects, for within category priming.
5. Replication of no significant effects when splitting the data in half.

☞ Less support for shared reasoning processes between distinct instances of enrichment via alternatives?

A Bar Plot


Num4
Num4
Num4
Some
Some
Num4
Some
Some
1
0.8
0.6
0.4
0.2
0
Strong Weak
Strong Weak
Strong Weak
Strong Weak

Data Treatment

Bott and Chemla use the fact that each target trial was preceded by two prime trials this design to filter out target responses where they cannot be sure that the participant understood the correct interpretation of the prime sentence.

In terms of a comparison of relative target response removals, the numbers are 6.5% (B&C) and 7.1% (R) of all trials.

As Bott and Chemla do not include information about the categories the incorrect primes were removed from, so we don't have sufficient information to provide further insights.

Data Treatment

One hundred participants were recruited using Amazon Turk.

- We removed 7 participants who did not declare English as their native language.

We included keyboard shortcuts to help participants complete the experiment.
This meant that in principle the participants could complete the experiment very quickly.

- E.g., going through the experiment as fast as possible (using the keyboard, not reading the sentnces, etc.) takes around 40 seconds.

One and a half second seems a reasonable lower bound for time spent on a trail.

- 3 mins lower bound, average time of 9 mins.

One participant who feel below this, after the lang. restrictions.

Analysis Procedure

Following Bott and Chemla the response-type likelihood was modelled using logit mixed-effect models.
Analyses were conducted lme4 (Bates et al. 2014), languageR (Baayen 2011), and memisc (Elff 2012), libraries for the R statistics program (Team et al. 2013).
Treatment and sum coding were used as described by Bott and Chemla, with any factor not explicitly mentioned receiving treatment contrasts.
The analysis starts with an general model involving all of the data, in which the interaction of with and between-expression priming is assessed.
A more detailed analysis is then performed by restricting the analysis to within and between-expression trials only.
The dependent measure was the log odds of choosing a strong over a weak prime.

Analysis

Bott and Chemla report three analyses:

1. Whether EVAs can be primed at all,
2. Whether priming occurs at the within-category level, and
3. Whether priming occurs at the between-category level.

Without the βs, Avoiding the ps, and Forgetting the Zs

Bott and Chemla observed priming of EVAs at the within-category level and the between category level, we only observed priming of EVAs at the within-category level.

Fancy Results: Overview

Bott and Chemla:

		β	S.E.	Z	p-value
Overview	Prime $*$ WithBet $+$ (1 $+$ Prime $*$ WithBet $\mid$ subject)				
	(Intercept)	−0.594	0.198	−2.991	.003
	Prime	0.563	0.034	16.342	<.001
	WithBet	0.126	0.029	4.284	<.001
	Prime:WithBet	−0.430	0.033	−13.177	<.001
Within simple	Prime	0.993	0.059	16.950	<.001
Between Simple	Prime	0.133	0.033	4.082	<.001

Replication:

		β	S.E.	Z	p-value
Overview	Prime $*$ WithBet $+$ (1 $+$ Prime $*$ WithBet $\mid$ subject)				
	(Intercept)	0.962	0.346	2.778	<.010
	Prime	0.310	0.074	4.196	<.001
	WithBet	−0.006	0.067	−0.089	.929
	Prime:WithBet	0.294	0.071	4.135	<.001
Between Simple	Prime	0.016	0.089	0.181	.857
Within Simple	Prime	0.603	0.114	5.277	<.001

Details

For the first model, sum contrasts are used for both factors, and a significant effect of a strong prime increasing the rate of strong responses, $\beta = 0.56$, $p < .001$, a significant effect of strong responses happening in between category trials, rather than within category, $\beta = 0.126$, $p < .001$ and an interaction between the two $\beta = -0.43$, $p < .001$ showing that the effect of the prime was greater in between within trials.

We observed slightly different results. First, a significant effect of priming $\beta = 0.310$, $p < .001$ for within category trials, no significant effect of between category trials, $p = .929$, but a significant effect of the interaction between the two $\beta = .294$, $p < .001$ showing that the effect of prime was less for within category trails. Though, in all cases these effects were smaller than those observed by Bott and Chemla.

Simple Effects

Bott and Chemla use a model with a similar structure, but using treatment contrasts for the within/between factor and sum contrasts for the prime factor to investigate simple effects.

They observe significant priming occurred at the within category level, $\beta = .99$, $p < .001$, and at the between category level $\beta = .13$, $p < .001$.

We observe significant priming at the within category level $\beta = .603$, $p < .001$, but no significant priming at the between category level $\beta = .016$, $p < .857$.

Now, I don't really understand all the details here, but from what I gather this means the effect of between category priming is significant *given* the effect of within category priming, but not really significant in and of itself.

Details: Further Observations

Bott and Chemla broke the data down into within-category trials and between-category trials to assess the observed effects in more detail, conducting separate analyses on each. In each model treatment contrasts were used for the categories, and sum contrasts for the prime.

Details: Within

Bott and Chemla observed a significant effect of prime, $\beta = 1.24$, *p* $<$ .001, for within category trials, showing an effect for *ad hoc* categories. In contrast to Bott and Chemla we found a significant effect for the *number*4 category, $\beta = 0.759$, $p <$ .001. And, in line with Bott and Chemla, we did not find a significant effect with respect for the *some* category.

Fancy Results: Within

Bott and Chemla:

		β	S.E.	Z	p-value
Within detail	Prime $*$ WithCat $+$ (1 $+$ Prime $*$ WithCat $\mid$ subject)				
	(Intercept)	−2.088	0.255	−8.185	<.001
	Prime	1.239	0.109	11.374	<.001
	WithCatNUM4	2.068	0.195	10.588	<.001
	WithCatSOME	1.823	0.157	11.598	<.001
	Prime:WithCatNUM4	0.174	0.166	1.046	.269
	Prime:WithCatSOME	−0.138	0.137	−1.007	.314

Replication:

		β	S.E.	Z	p-value
Within Detail	Prime $*$ WithCat $+$ (1 $+$ Prime $*$ WithCat $\mid$ subject)				
	(Intercept)	1.361	0.460	2.960	<.010
	Prime	0.759	0.206	3.678	<.001
	WithCatSOME	−0.784	0.432	−1.816	.069
	Prime:WithCatSOME	−0.164	0.265	−0.618	.536

Details: Between

Between categories, Bott and Chemla found a significant effect of priming $\beta = .15$, $p < .05$ for *number4/ad hoc* trials and for *some/number4* trials, but not when interaction with the strength of the prime was accounted for, and we found no significant interaction.

Bott and Chemla also perform further analysis of the data pooled from three experiments they conducted.

We followed a similar analysis, looking at the separate prime and target categories in the case of between category trials and with splits of the data into first and second halves of the experiment.

In contrast to Bott and Chemla analysing the separate prime and target categories in the case of between category trials revealed no significant effects, and in line with Bott and Chemla the split by halves did not reveal a difference in responses.

Fancy Results: Between

Bott and Chemla (directions pooled):

		β	S.E.	Z	p-value
Between detail	Prime $*$ BetCat $+$ (1 $+$ Prime $*$ BetCat $\mid$ subject)				
	(Intercept)	-0.691	0.204	-3.384	<.001
	Prime	0.145	0.058	0.058	.012
	BetCatSOMEADH	-0.054	0.089	-0.611	.540
	BetCatSOMENUM4	0.889	0.112	7.915	<.001
	Prime:BetCatSOMEADH	-0.069	0.079	-0.873	.383
	Prime:BetCatSOMENUM4	0.078	0.088	0.888	.374

Replciation (directions *not* pooled):

		β	S.E.	Z	p-value
Between detail	Prime $*$ BetCat $+$ (1 $+$ Prime $*$ BetCat $\mid$ subject)				
	(Intercept)	0.899	0.506	1.777	.076
	Prime	-0.086	0.160	-0.541	.589
	BetCatSOMENUM4	0.861	0.451	1.910	.056
	Prime:BetCatSOMENUM4	0.362	0.282	1.281	.200

Analysis of priming effect for direction between category trials

Bott and Chemla:

		β	S.E.	Z	p-value
	Prime + (1 + Prime | Subject)				
some $\rightarrow$ *number4*	Prime	0.345	0.069	5.140	<.001
number4 $\rightarrow$ *some*	Prime	0.221	0.065	3.412	<.001

Replciation:

		β	S.E.	Z	p-value
	Prime + (1 + Prime | Subject)				
some $\rightarrow$ *number4*	Prime	−0.080	0.162	−0.492	0.623
number4 $\rightarrow$ *some*	Prime	0.354	0.238	1.488	0.137

Observations

So, the motivation was to see what kind of priming processes there were, and whether these were distinct or shared. And, from the data gathered, Bott and Chemla argue that enrichment can be primed.
They note that the between-category priming effect illustrates that there are shared mechanisms across the EVA categories. However, the greater within-category priming results demonstrate that there are also some additional effect of EVA specific mechanisms.
I'm no sure that this was replicated.

Observations

We observed that there was a quite different effect of priming for within category trials for *some* and *number*4. This is not in line with what Bott and Chemla observed.
For between category trials we didn't have a sufficient number of categories to do the same analysis as Bott and Chemla did. However, we did observe that there was quite an effect on which categories the prime and target came from.
As Bott and Chemla argue that there's a shared priming process between *some* and *number*4, it doesn't seem that our data supports this. For, we should expect that effects would be similar in both directions. We have close to significant results for the within category case, but not for the betweeen category case.

Summary

The aim of Bott and Chemla (2016) was to better understand how people use alternatives to enrich the basic meaning of a sentence.
Bott and Chemla observed significant effects of priming both within and between categories, and in particular between *some* and *number*4 categories.
In our replication we found significant effects of priming within categories, but no significant effects of priming between categories, even when the prime and target categories were controlled for.
Unclear to me exactly what this means.

Don't Go Past This Slide ...

Additional details follow ...

Combined Results from Bott and Chemla

		β	S.E.	Z	p-value
Overview	Prime $*$ WithBet $+$ (1 $+$ Prime $*$ WithBet $\mid$ subject)				
	(Intercept)	−0.594	0.198	−2.991	.003
	Prime	0.563	0.034	16.342	<.001
	WithBet	0.126	0.029	4.284	<.001
	Prime:WithBet	−0.430	0.033	−13.177	<.001
Within simple	Prime	0.993	0.059	16.950	<.001
Between Simple	Prime	0.133	0.033	4.082	<.001
Within detail	Prime $*$ WithCat $+$ (1 $+$ Prime $*$ WithCat $\mid$ subject)				
	(Intercept)	−2.088	0.255	−8.185	<.001
	Prime	1.239	0.109	11.374	<.001
	WithCatNUM4	2.068	0.195	10.588	<.001
	WithCatSOME	1.823	0.157	11.598	<.001
	Prime:WithCatNUM4	0.174	0.166	1.046	.269
	Prime:WithCatSOME	−0.138	0.137	−1.007	.314
Between detail	Prime $*$ BetCat $+$ (1 $+$ Prime $*$ BetCat $\mid$ subject)				
	(Intercept)	−0.691	0.204	−3.384	<.001
	Prime	0.145	0.058	0.058	.012
	BetCatSOMEADH	−0.054	0.089	−0.611	.540
	BetCatSOMENUM4	0.889	0.112	7.915	<.001
	Prime:BetCatSOMEADH	−0.069	0.079	−0.873	.383
	Prime:BetCatSOMENUM4	0.078	0.088	0.888	.374

Combined Results from the Replication

		β	S.E.	Z	p-value
Overview	Prime $*$ WithBet $+$ $(1 +$ Prime $*$ WithBet $\mid$ subject)				
	(Intercept)	0.962	0.346	2.778	<.010
	Prime	0.310	0.074	4.196	<.001
	WithBet	-0.006	0.067	-0.089	.929
	Prime:WithBet	0.294	0.071	4.135	<.001
Between Simple	Prime	0.016	0.089	0.181	.857
Within Simple	Prime	0.603	0.114	5.277	<.001
Within Detail	Prime $*$ WithCat $+$ $(1 +$ Prime $*$ WithCat $\mid$ subject)				
	(Intercept)	1.361	0.460	2.960	<.010
	Prime	0.759	0.206	3.678	<.001
	WithCat	-0.784	0.432	-1.816	.069
	Prime:WithCat	-0.164	0.265	-0.618	.536
Between detail	Prime $*$ BetCat $+$ $(1 +$ Prime $*$ BetCat $\mid$ subject)				
	(Intercept)	0.899	0.506	1.777	.076
	Prime	-0.086	0.160	-0.541	.589
	BetCat	0.861	0.451	1.910	.056
	Prime:BetCat	0.362	0.282	1.281	.200

Details for the bar plots

Prime		Response		From the replication			From Bott and Chelma	
Type	Category	Category	mean %	Raw mean	Raw S.D.	Raw S.E.	mean %	Raw S.E.
Strong	Num4	Num4	0.6767956	2.634409	1.653619	0.1714723	0.615	0.018
Weak	Num4	Num4	0.5675553	2.184783	1.683334	0.1745536	0.339	0.018
Strong	Num4	Some	0.5762712	2.193548	1.702032	0.1764925	0.553	0.019
Weak	Num4	Some	0.5833029	2.239130	1.750162	0.1814834	0.484	0.019
Strong	Some	Num4	0.7414502	2.511364	1.597371	0.1656396	0.544	0.020
Weak	Some	Num4	0.6498584	2.466667	1.643510	0.1704240	0.474	0.019
Strong	Some	Some	0.6966165	2.329545	1.713514	0.1776831	0.604	0.019
Weak	Some	Some	0.4703510	1.978261	1.728737	0.1792617	0.340	0.018

Relevant cell mean and S.E. from Bott and Chemla included.

Analysis of the experiment by halves

		β	S.E.	Z	p-value
Within by half	Prime $*$ WithCat $*$ Half $+$ (1 $+$ Prime $*$ WithCat $*$ Half $\mid$ Subject)				
	(Intercept)	−0.221	0.378	−0.584	.559
	Prime	0.141	0.320	0.442	.658
	WithCat	1.068	0.481	2.222	<.050
	Half	0.850	0.371	2.294	<.050
	Prime:WithCat	0.770	0.542	1.423	.155
	Prime:Half	0.172	0.243	0.706	.480
	WithCat:Half	−0.671	0.334	−2.011	<.050
	Prime:Withcat:Half	−0.534	0.379	−1.407	.159
Half 1 only	(Intercept)	0.675	0.3157	2.139	<.050
	Prime	0.314	0.1154	2.719	<.010
	WithCat	0.377	0.1801	2.092	<.050
	Prime:WithCat	0.217	0.1975	1.098	.272
Half 2 only	(Intercept)	1.518	0.6091	2.493	<.050
	Prime	0.519	0.2572	2.016	<.050
	WithCat	−0.203	0.299	−0.680	.497
	Prime:WithCat	−0.392	0.361	−1.086	.277

Possible Explanations

One observation may be the lack of *ad hoc* categories. This is important as Bott and Chemla included extra priming to get these going. But then it's kind of odd that the effect of these would only show up elsewhere. If this was borne out, then it would support some kind of shared mechanism, but in a way that is very unclear. And, it wouldn't seem to fit with the kind of theory that Bott and Chemla propose, for the relevant effect would be the consideration of alternatives in general, rather than specific alternatives.

Alternatively, something that may affect things is how easy the pictures were to process. We get more of an effect with some, and these certainly were easier to look at, especially as sometimes the relevant stuff could be placed at the bottom of the picture, and therefore be harder to see. Again, this is perhaps something that we should have kept information about, but did not.

References

- Baayen, R Harald (2011). "languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"". In: *R package version* 1.1.
- Bates, Douglas et al. (2014). "Fitting linear mixed-effects models using lme4". In: *arXiv preprint arXiv:1406.5823*.
- Bott, Lewis and Emmanuel Chemla (2016). "Shared and distinct mechanisms in deriving linguistic enrichment". In: *Journal of Memory and Language* 91, pp. 117–140.
- Elff, Martin (2012). "memisc: Tools for management of survey data, graphics, programming, statistics, and simulation". In: *R package version 0.95-38, URL http://CRAN. R-project. org/package= memisc*.
- Team, R Core et al. (2013). "R: A language and environment for statistical computing". In: