Department of Psychology
Stanford University
Building 420 (Jordan Hall)
450 Serra Mall
Stanford, CA 94305

650-724-4003
longouyang@post.harvard.edu

03 March 2015

Editorial Board
Cognitive Science

Dear Dr. Louwerse,

Please accept our resubmission of the manuscript “Semantic coherence facilitates distributional learning” to be considered for publication. Thank you very much for your thoughtful comments and your attention to the details of this work. We very much appreciate your and the reviewers’ feedback. We apologize for the delay, which was due to our collection of additional empirical data. Point-by-point responses to the reviews are attached below.

Please do not hesitate to contact us as well with any other questions or concerns. Thank you again for your consideration.

Sincerely,

Long Ouyang
Stanford University

Action editor

On page 8 you write "However, while suggestive, this [distributional semantics] work is only indirect evidence of distributional learning. Researchers studying acquisition of grammatical categories have employed methods that can in principle provide stronger evidence. These studies expose learners to artificial languages with certain co-occurrence regularities and measure whether learners form categories on the basis of these regularities." In general, I agree, except that the paper now assumes that the artificial language participants are tested on is similar to a regular language. Clearly, this is not the case (I assume you would agree with this). The richness of natural language can simply not be copied in an artificial language (Christiansen & Chater, 2008). So at best this is a first step towards measuring whether learners form categories on the basis of co-occurrence structures. Also, finding no results in an artificial language environment does not mean that learners do not form categories on the basis of co-occurrence structures. It would be useful to emphasize the relationship (or rather differences) between a natural language and an artificial language environment.

> We agree with this interpretation of the limitations on our data. In our revision, we have added some caveats regarding the naturalism of our stimuli (first paragraph of p.11).

A second issue concerns the experimental design. You presented participants with binomials (e.g., "car and chuv"). We know from existing research on binomials and language statistics that various factors have an effect on the ordering of the two nouns in binomials, including perceptual features, phonological ordering aspects, as well as frequencies in ordering (e.g., markedness) (Benor & Levy, 2006). The question is to what extent these factors might have affected your results. Similarly, the meanings of some of the words (the referent assignment task) might be influenced by the salience of the picture (color and animacy might have an effect on memory), the naturalness of the non-word or the naturalness of the speech synthesis of the non-word. So question is to what extent might your findings be driven by factors that are far more trivial than the ones you find in the results?

> Thank you for this challenging point, which on reflection we agree is an important one to address. To test this hypothesis directly, we collected new data at the 3/3 coherence level where we replaced “and” with a nonce connective “tezz” (Experiment 1b). We still observe an effect of coherence; in fact, task scores for “tezz” are almost identical to those with “and.” These new data strongly suggest that our results are not driven by semantic constraints of "and"; we feel these new data help strengthen and clarify the findings and we are grateful for your suggestion and the opportunity to address it.

Finally, I have read your conference proceedings article this paper is based on and wondered about the different participant numbers that are mentioned in the two papers. These are small differences (654 vs. 678, 151 vs. 162) but they are different. Are the groups of participants in this study different than those in the proceedings article? If that is the case, the results - which are very similar between the two papers - are even stronger, and you want to mention this in the paper. If the participants are not the same, what explains the difference in counts. Why were a handful participants added?

> Thank you for your attention to detail here. We investigated this issue and found that we had made an error in the analyses for the conference paper. We did not collect a handful of participants after the conference paper – rather, by the time we had written the conference paper, we had already collected all participants for Experiments 1 (N = 678) and 3 (N = 162) but an error in the conference paper analysis scripts meant that a handful of participants were excluded from analysis.
>
> While investigating the discrepancy in sample sizes, we also discovered that there was a small error in our current analysis code which meant that we failed to exclude some participants who had not passed all catch trials. We have addressed this issue and found that our results are unchanged.
>
> As is our practice more generally, we will publish our (anonymized) data and analysis scripts in an online repository so that other interested researchers can repeat our analyses or make other uses of the data.

Reviewer 1

What is surprising to me is the selection of a familiar and highly restricting English construction as the linguistic context. The conjunctive predicate "and" imposes the restriction that its arguments belong to the same syntactic and mostly semantic category. This means that a speaker of English who hears a phrase "dog and dax" can immediately make an informed guess about the probable semantic properties of "dax". Since similar semantic properties will be guessed for "dog and ziv" (or even "cat and ziv"), it is not surprising that s/he judges dax and ziv to be similar in meaning, or picks the correct referent for them. It has been shown in previous studies that humans use the selectional restrictions of familiar predicates to narrow down the semantic characteristics/category of their novel arguments (see, for example, Altman & Kamide, 1999; Koehne & Crocker, 2010). ... This is not the same as the case of postman and mailman, in fact it is not at all clear whether any distributional learning is happening in Experiment 1 here, since all the observed effects can be explained in terms of the successful application of the selectional restrictions of the main predicate ("and"). It is also not surprising that a similar effect is not *observed in Experiment 2 (since the predicate does not impose any phonological restrictions on its arguments), or Experiment 3 (since contradictory semantic predictions are provided by the same predicate for the same target word in different usages).*

> We agree that this is an important point. As described above, we have added Experiment 1b specifically to address this comment. We hope you will agree that this control experiment rules out the (plausible) confound of selectional restrictions for “and.”

Ideally, I would have expected the stimuli to consist of multi-word sentences, where all words are novel but some are paired with familiar referents in isolation prior to the participants' exposure to sentences. In such a hypothetical case, the participants would have been exposed to an unfamiliar language, and their acquired knowledge about target words could have been attributed to distributional learning. I do think that investigating semantic coherence as a facilitator of distributional learning is a promising idea, and I would be very much interested in the outcome of its proper examination.

> We thank the reviewer for this suggestion. The proposed experiment might be an extension for future work, but we do not believe that it is necessary for the claims we are making. Our goal here was to investigate how semantic coherence facilitated distributional learning, with the explicit motivating idea that learners may already know some words and these words can help the process of distributional learning. So the use of real familiar words in our paradigm leverages meanings that are already known outside of the experimental context, just as language learners can during acquisition.
>
> In addition, our design was created to optimize the chance of seeing a semantic coherence effect in what has previously been a very difficult learning problem. If we were to substitute the word “dax” (which the participant has learned means “dog”) for the word “dog,” we would be adding another layer of memory demand to our task (which is already fairly long and complex; the 392 exposure level could take upwards of 45 minutes). In addition, because the exposure would be to distributional regularities surrounding the novel word (“dax”=DOG), the effects we saw would likely be decreased by the weaker association between form and meaning for the novel word (as well as the weaker phonological memory for the novel word).
>
> In sum, while we understand the desire for a fully-artificial language, we believe that our mixed known/novel technique adds value to a literature that has already been focused exclusively on such languages.

Mechanical Turk is not familiar to all readers, so technical terms such as "HIT approval rate" need to be defined. There is no indication of the demographics of the participants or their indicated first language.

We replaced "HIT approval rate" with a less technical term. We did not provide demographic information because we didn't collect it, however. We also did not ask for the first language, although we did restrict participation to US workers.

List of words: what was the criteria for selecting the set of familiar words? They do not seem to be quite comparable in terms of their frequency ranking or phonological properties. Also, was there any pretest for potential semantic associations of the novel words?

Following Lany and Saffran (2010), we selected words for Experiment 1 from the set of animals and vehicles. The two categories of words are roughly comparable in frequency ranking:

Car - 1
Dog - 2
Bus - 3
Cat - 4
Truck - 5
Hamster - 6

The mean ranking for vehicles is (1 + 3 + 5)/ 3 = 3 and the mean ranking for animals is (2 + 4 + 6)/3 = 4.

There was no pretest for potential semantic associations of the novel words (it seems likely, for example, that "moke" had associations with "smoke"). Nevertheless, the randomization scheme ruled out any potential confounds due to semantic associations – such associations would only serve to add variance.

For the original scores (e.g. mean familiarity ratings), it would be informative to give the range of possible values. I could not understand why the difference scores range between [-4,4].

We clarified this in the first paragraph of p.18. Difference scores range between -4 and 4 because the scores themselves range from 1 to 5.

Some conditions (or is it some subset of data points?) contain catch trials and some don't, and it is stated that the inclusion or exclusion of failed catch trials does not affect the results. Yet, the *portion of data with failed catch trials is excluded from the analysis, which in my opinion provides some inconsistency in the final dataset (since the similar subset of undetected bogus datapoint is included).*

> We agree that this feature of the dataset is not ideal. We could correct it by rerunning those particular participants, but we believe that their effect is minimal and conservative with respect to our hypotheses. As we noted in footnote 6, it is mostly participants in the 3/3 condition who did not receive catch trials. If anything, this would weaken any effect of semantic coherence, as it introduces noise at the highest level of semantic coherence but not at any other level.

For all experiments, a ranking of familiarity is reported for each condition (e.g., F>W>C>P). Are these significant results? On all exposure levels?

> These orderings indicate significant differences, but were determined by collapsing across exposure levels (see Figures B1 and B2).

In Figure 4, you should add the label "semantic coherence" for the grey headers.
In Figure 5, the caption mentions error bars, but there are no error bars on the graphs.
In Figure 6, the label for the forth graph must be "syllable count (Exp2)".

> Thank you for your close reading! We fixed these issues.

In the first two paragraphs of Experiment 2, there are several mentions of previous work but no references is provided.

> We fixed this by adding a reference to Lany & Saffran, 2010, the work that we were specifically referring to.

Experiment 2 (page 24): "The one exception is that the participants in the onset coherence condition were able to distinguish withheld from co-occurence violation sentences, matching 2/3 participants." Shouldn't this be the position variation instead (according to Figure 6)?

> We are actually referring to the Holm-corrected pairwise t-tests in (what is now) Figure B2, not Figure 6. Notice, however, that even in Figure 6, the withheld trials (blue curve) are rated as more familiar than the co-occurrence violation trials (green curve), though Figure 6 does not facilitate parsing the significance of this comparison collapsing across exposure levels.

Reviewer 2

In the abstract, it's written: "Are people able to use distributional statistics in learning language? Results from prior artificial language learning experiments suggest that the answer may be no." This seems a bit misleading in at least two crucial respects: 1) even among the work the authors cite as negative evidence, there is evidence that humans can use distributional statistics when correlated with phonological statistical cues to learn category-based structure; and 2) learning a language encompasses more than the learning of semantics and categories (e.g., humans can use purely distributional statistics to learn patterns relevant to other aspects of language-learning, such as speech segmentation, phonetic categorization and syntactic form-classes). The quoted passage should be sharpened accordingly. Also, from pg. 25, it's written: "While computational models? experiments with humans show consistent failures." I think this statement could be similarly sharpened in light of the fact that relevant work from Lany and Saffran (2010, 2011) reach more positive conclusions of limited success, drawing upon learners' use of multiple cues in distributional learning (which may arguably be the norm, rather than the exception, for natural language input).

> Thank you for encouraging a higher level of precision. We softened the rhetoric in the abstract and the first paragraph of the General Discussion (p. 28)

For the assessment-related tasks (i.e. similarity, memory, referent assignment), it was unclear whether these were completed by subjects in the same fixed order (and what that would be), or whether the presentation order of the tasks was counterbalanced among participants

> We clarified this point (we used a fixed ordering) and justified our sequencing at bottom of p.13.

"we will refer to learning from such learning as distributional learning" (pg.3) => probably meant to write "? refer to such learning" instead of "? refer to learning from such learning" (as I'm not sure what the latter would mean, unless a metacognitive component is implied?)

> We clarified this issue.

word-word matrix of Figure 1: Shouldn't the co-occurrence of b one word after a (i.e., "ab"; b and a: +1) be "2" -- instead of "1" (as currently indicated)? And similarly, shouldn't the co-occurrence of a one word after b (i.e., "ba"; a and b: +1) be "2" - instead of "0" (as currently indicated)?

Thank you for your attention to detail. We have replaced the matrix diagram with a corrected figure.

pg. 13: in the Methods write-up, it states that 14 unique sentences were used in the exposure phase - but if 6 of the 18 possible sentences were withheld, wouldn't there be 12 unique sentences (as illustrated in Fig. 2)? and if so, does this change the number of reported trials for the exposure conditions (assuming each familiar sentence was repeated an equal number of times)?

Again, we thank you for your thoroughness. There were in fact 14 unique sentences; Figure 2 was wrong – we only withheld m1-p1, m2-p2, n1-q1, n2-q2 (we didn't withhold m3-p3, n3-q3). We have fixed Figure 2.

Figure 6: syllable count condition is part of experiment 2 (rather than expt 3)

We fixed this.

for Figures 5 and 7, are the error bars missing?

Yes - we had omitted them for clarity but now have added them back.