Max Planck Institute
for Psycholinguistics

Wundtlaan 1
6525 XD Nijmegen, NL
Phone: +31 (024) 3521266
E-mail: marisa.casillas@mpi.nl

January 26, 2016

Dear Profs. Dr. Gerrig and Dr. Pickering,

I hereby resubmit a manuscript on "The development of children's ability to track and predict turn structure in conversation" to the *Journal of Memory and Language*. I apologize for the delay in getting the manuscript resubmitted. It took quite a long time to implement one of the suggestions from Reviewer 3, mostly because I had to acquire some new technical skills in order to carry out his/her suggested analysis on our computation-heavy data set. Nevertheless, I believe that the added analysis has substantially strengthened the paper as a whole, so I appreciate your patience very much.

Please note that, because of a miscommunication, I prematurely resubmitted the manuscript a few weeks ago (without my co-author's final sign-off). I immediately withdrew the manuscript when our miscommunication came to light and we have since then made minor improvements to the paper. With my co-author's final approval, I am ready to return the paper to JML. However, now my only option is to resubmit the manuscript as "new". If it is at all possible, we would request that the paper go back to the original reviewers, if only because we did quite a lot of work to address their specific (and substantial) requests. In case this is possible, I have included our responses to the reviewers below.

For example, in addition to the new random permutation analysis suggested by Reviewer 3, we have also included two further follow-up analyses in the Appendix, each inspired by the other reviewers' comments. We have also tried to make the text much more succinct, with distinct and straightforward results and discussion sections, following requests from all three reviewers.

We hope that, with these changes, we have greatly improved the paper's quality. Please find a description of specific comments addressed below. Thank you very much for your reconsideration. Please let us know if there is any further information that you require about the submission.

Sincerely,

Marisa Casillas

Manuscript identifying details:

ID:	JML-14-235 (from the previous submission)
Title:	The development of children's ability to track and predict turn structure in conversation
Authors:	Marisa Casillas[a] and Michael C. Frank[b]
Affiliations:	a) Max Planck Institute for Psycholinguistics and b) Stanford University

We thank all three reviewers for their thorough and helpful comments on the previously submitted manuscript. We have tried to address all of their comments, in the manuscript and here in the response letter.

Reviewer 1

1. 333ms may be too long for older children's saccadic planning time and may therefore inflate the anticipatory looking rates for older children.

Thank you for this comment; we were concerned about this issue as well, but we unfortunately did not collect independent saccadic planning norms for children ages 1–6 (and we do not know of any authoritative source for these norms). In the newly analyzed data, we take the conservative approach of using adult-like planning times (200ms) for children at all ages in both experiments.

2. The description of the random baseline permutations was not clear.

In our revision, we have completely redone all random permutation analyses and have attempted to clarify their description (Section 2.2.2).

3. Are anticipatory gaze shifts driven by boredom (and not by anticipation)? Do children generally just look away as the speaker goes on?

Thanks for pointing out an important alternative hypothesis. To address this comment, we modeled the proportion of participants looking at the current speaker for our real data and for (hypothetical) boredom-driven lookers (who look away from the current speaker at a constant rate, starting 1 second after the onset of speech) across turns of different length. Even in short turns, where there is very little temporal room for children to behave differently from the boredom-driven lookers, we still see a clear difference between the boredom-driven simulations and the real data (Figure D.2). We take this result to be evidence against the alternative hypothesis; we report the details of the analysis in the Appendix (Section C).

4. The reported results from the models models are hard to follow.

We apologize—the data from this experiment are very complex. To aid comprehension, we have added more information about how each variable was coded and have put summary tables of every model's output into the text. In addition, we have reorganized our results sections in hopes of making them more straightforward.

5. "No speech" should be the reference level against which the linguistic conditions are compared in Experiment 2.

This is another very useful comment, thank you! We agree and have changed our analyses to set "no speech" as the reference level. We had started with the "normal speech" condition as the reference level because this choice mirrors the coding in related work on adult turn-end prediction (de Ruiter et al., 2006). But the additive model implied by the "no speech" baseline is easier to interpret.

6. Looking at the data, it seems possible that the children are not using any linguistic information.

Thanks for raising this challenging and important point. To summarize: We think that the robust question (speech act) effects we observed in both experiments provides strong evidence for the use of linguistic information.

In Experiment 2, children's data showed a 2-way interaction of age and transition type (question vs. non-question), but only in the "normal" speech condition, suggesting that it was the only condition in which they could extract the linguistic cues needed to accurately identify questions. We speculate that this is because prosody and lexical cues to questionhood often work together in forming questions during everyday speech, but it could also be that the manipulated speech was simply too unfamiliar for them to predict effectively. Either way, the existence of the question effect in the normal condition indicates that children do, in fact, use linguistic information (to identify questions) and the interaction with age suggests that they get better at doing so as they get older. These effects cannot be explained by duration alone (and duration is controlled in our analysis), and there is no visual cue advantage difference between the normal condition and the others. We are then left to conclude that they use linguistic information to support the question effect.

A similar story can be told for the data in Experiment 1: non-verbal cues to questions vs. non-questions are available in both the English and non-English stimuli, but participants (adults and children) made more anticipatory switches when they had access to the lexical information. Again, this cannot be explained by gap duration or unequal visual advantages across the stimuli. Thus, we believe that both experiments provide support for the use of linguistic cues, albeit for a subset of these cues.

7. It is too hard to link the conclusions in the discussion back to the statistical results and the effects themselves are not talked about in a clear enough way.

We apologize. In our revision, we have tried our best to make these links clearer, first by clarifying the statistical analyses and outputs, and second by trying to explicitly link each claim to an individual result (through the use of figure and table references). We hope that this will help readers better connect the results to our interpretations.

8. There are too many references to unpublished work.

The single upside to our delayed resubmission is that all of the unpublished work referred to previously is now published or in press!

(Reviewer 1's minor points/corrections have all been addressed directly in the text.)

Reviewer 2

1. It is not clear which claims are backed up by which analyses. Analyses should be carried out for each group separately.

As discussed above, we have tried to improve the linkage between our claims and analyses and now have presented the statistical results in tables, as requested. Please see responses 4 and 7 under Reviewer 1 for a little more detail.

The main focus of our analyses was the effect that age (as a continuous variable) has on anticipatory gaze. However there are two cases in which our main analyses are not satisfying without further, age-group analyses: (a) age sometimes showed itself to interact with other factors (e.g., age and language condition), and (b) without further tests we can not say whether particular age groups themselves statistically differ from chance, across or within conditions.

Following your suggestion, we have thus made two additional changes. First, we added follow-up two-tailed t-tests, making pairwise comparisons between age groups for significant interactions so that we could find out how the interacting factor (e.g., language condition) significantly changes across age groups (Sections B.1–3). Second, we have added individual models of the youngest age groups (ages 1, 2, and 3), comparing their real looking behavior to our randomly permuted data across conditions (at the ends of Sections 2.2.2 and 3.3.2). These follow-up analyses, which are summarized in the text, are primarily described in the Appendix (Section B).

2. The results and discussion for each study would be better if they were separated into different sections.

Our original intention in combining results and discussion was to make the findings easier to understand. Since we were not able to achieve this in the prior submission, we have followed Reviewer 2's advice and now use a more traditional "results-first" structure. We hope this change has improved the clarity of the paper.

3. The methods sections need more elaboration in a few places, and the baseline analysis section could be much clearer.

We thank the reviewer for careful attention to our methods section (and its missing details). We have added in the experimental methods information requested and have re-checked each section for missing information that might be useful in replicating our methods. We've also worked to clarify the description of the baseline analysis (now reported in depth in the Appendix), including a new figure (Section 2.2.2; Figure 4).

4. The adult data deserve further discussion because they raise the question of whether this measure taps into linguistic processing at all.

We actually think the adults' results are well in-line with prior work showing that (a) adults' predictions primarily rely on lexical cues (E1 & E2) and (b) adults are better anticipators than young children (E1 & E2). The complicating factor is that we report a new and important interacting factor with their linguistic processing: speech act. Like the children in our data set, adults made more anticipatory switches following questions, which means that most of our evidence about their use of linguistic cues comes from question transitions. This enriches the picture of linguistic processing for turn prediction, as painted by prior work: linguistic processing is brought into play more often in some contexts than others. From these points we conclude that our experiments do indeed tap into participants' use of linguistic cues to predict upcoming turn structure.

5. There is a lot of unpublished work referred to in the manuscript, with errors in the reference list.

These papers are now all published or in press. We apologize for errors in the reference list; we have checked them again and hope that there are no remaining errors.

Reviewer 3

1. The manuscript could be shortened substantially without loss of content.

We have significantly shortened the introduction and tried to remove redundancies where possible in the rest of the text. We have also moved some of the secondary analyses to the Appendix so that the main text can be read without getting into the nitty gritty of every single analysis. We hope that this has helped make the main text more straightforward and approachable length-wise.

2. There is a confound between linguistic condition and puppet pair in Experiment 2.

We thank Reviewer 3 for this point. We repeat here some of the information now available in the Appendix (Section D) for easy summary: Our design does not fully cross puppet pair (e.g., robots, blue puppets) with linguistic condition (e.g., "words only" and "no speech"). Even though each puppet pair is associated with different conversation clips across children (e.g. robots talking about kitties, birthday parties, or pancakes), robots were only associated with "words only" speech, merpeople were only associated with "prosody only" speech, and the puppets with fancy clothes were only associated with the "no speech" condition. We made this choice to increase the pragmatic felicity of the experiments for the older children (i.e., robots make robot sounds, merpeople's voices are muffled under water, the fancy-clothed puppets are in a room with main other voices). It is therefore fair to point out a possible confound between linguistic condition and puppet pair.

Thankfully, we also ran a short follow-up study at the museum with 3–5-year-olds that addresses this issue. In the follow-up study, each child only saw one video—the normal speech conversation about birthday parties—with a randomly assigned puppet pair performing the conversation. Five children watched each puppet pair, for a total of 30 children across the six pairs (shown in Figure 3). This experiment holds all things constant except for the appearance of the puppets. We then used a mixed effects logistic regression of children's anticipatory switches (yes or no at each transition), with puppet pair (robots/merpeople/fancy dress/normal-speech-puppets) as a fixed effect and participant and turn transition as random effects. In four versions of this model we systematically varied the reference level to check for differences between every puppet pair, finding no significant affects of puppet type on switching rate. We take this as evidence that, although we did not fully cross puppet pairs and linguistic conditions in Experiment 2, it was unlikely to have had strong effects on children's looking rates above and beyond the intended effects of linguistic condition. We have included details and graphs for this norming experiment in the Appendix (Section D).

3. Computer animation would have been a better choice than puppets.

We agree with the reviewer that computer animation is one method for providing greater experimental control. We did not have the ability to create custom animated conversation videos at the time the research was conducted, though we do use it in some ongoing work. Nevertheless, we believe that the control experiment reported above provides some data that speaks against a "puppet saliency" hypothesis accounting for the results of Experiment 2.

4. The model design needs to be further justified and the way the effects are talked about needs clarification.

In our revision, we correct some small issues in our model description, for example, making it clear that we used a logistic regression (for our binary response variable of switch-no switch). We have also added a justification of our use of separate models for adults and children and our decisions regarding random slopes. We have also cleaned up the way we talk about the significant effects in the model output. In particular, we believe that the concern regarding dummy coding was a misunderstanding; dummy coding was in fact used for categorical predictors (as is now clearer with the presentation of results in Tables 2 and 5).

5. An ideal-observer analysis of the usefulness of the linguistic cues would give a better idea about what can affect anticipatory gaze.

We agree—this is a great suggestion, but it may venture too far outside the scope of the current paper, especially because our stimulus set is probably too small to support building a generalizable model of cue use. In addition, the interpretation of our experiments is already complex without the addition of a computational model.

6. The random baseline would be better if it were converted to a full permutation analysis so that the baseline rate of anticipatory gazing is not treated as an additive factor, which it probably is not.

We thank Reviewer 3 very much for the suggestion and clear explanation of how to conduct the analysis. Unfortunately, this was not a trivial analysis to implement, causing some delays in the manuscript resubmission. Running 20,000 random permutations with mixed models required a substantial investment in terms of parallel computing infrastructure.

That being said, we think that the new permutation analysis is indeed simpler and more theoretically intuitive than our prior method (described in 2.2.2 and 3.3.2, and visualized in Section A). Please note that there were two problems with conducting the analyses as originally instructed in the reviewer's comments. 1) We used z values instead of β estimates because the standard error estimates for the randomly permuted data were categorically higher (due to data sparsity issues). The use of z values is closer to the practice in neuroimaging data analysis, where statistical significance is computed based on the empirical distribution of test statistics. 2) Many of the models resulted in convergence warnings. We found that models with convergence warnings were more likely to have extreme variance in their estimates (see Appendix Table A.1). We therefore made the decision to report results across all convergent models, but we are happy to revisit this decision if the Reviewer believes there is an alternative approach that we should follow.

7. The theoretical contribution of this paper is not clear.

We have tried to improve the clarity of our findings' theoretical importance in the revised manuscript. I summarize here the points that we cover.

First, one of the most important theoretical contributions of our work is in establishing and replicating the effect of speech act (questions) in children and adults' spontaneous predictions. While participants may always use linguistic information to predict upcoming speaker changes, our results do not support the idea that they always use linguistic information to predict upcoming *turn ends*, as is assumed in metalinguistic measures of turn-end prediction (e.g., pressing a button while listening to speech). Instead, our results suggest that participants' spontaneous predictions are the result of question-monitoring, presumably achieved by recruiting linguistic cues to questionhood from the unfolding signal. The trends in our data suggest that participants are primarily recruiting lexical cues to do this, though establishing this pattern will require follow up work with more focused stimuli.

Second, lexical information alone is not equivalent to full linguistic information for children, as it has been shown to be with adults (e.g., de Ruiter et al., 2006; replicated in our work). While this result will require replication with different stimulus items, it still points to an important developmental effect.

Third, young children (e.g., at ages one and two) do make anticipatory gaze prediction more often than would be expected by chance alone (albeit at low rates). Since older children and adults' gaze shifts are primarily driven by question turns, this may suggest that, although children *can* make predictions about upcoming turn structure, it doesn't count for much until they have acquired the linguistic skill to pick out question turns. This finding, in turn, has potentially important implications for how participant role (first- instead of third-person) and cultural differences (high vs. low parent-infant interaction styles) might affect children's early predictive skill. It also bridges prior work demonstrating that children begin taking non-linguistic turns in infancy (e.g., Hilbrink et al., 2015) with work that shows that children still have trouble integrating linguistic aspects of responding at age 3–4 (Casillas et al., in press; Garvey & Berninger, 1984).