**Note:**

Comments made by reviewers are bolded. Following each comment, you will find our answers (not bolded). Thank you for all your comments and suggestions, we believe that they have significantly improved the manustript.

**Reviewer 1**  
**The manuscript reports on an eye-tracking study investigating the role of experience with anticipation vs. proficiency in predictive processing of second language learners. Specifically, it investigates use of prosodic cues (stress) for morphological anticipation in Spanish. A group of interpreters are contrasted with a group of non-interpreter learners with the same level of proficiency and a group of native speakers. Interpreters are assumed to have more experience with anticipation than other non-native speakers. Interpreters and native speakers, but not**

**non-interpreter learners are found to predict suffixes already at the offset of the preceding syllable. Interpreters are somewhat slower to start predicting than native speakers when the first syllable has CV structure, but virtually identical when structure is CVC. There seems to be more variation for the results within the interpreter group. Thus, it seems that increased experience with anticipation does improve predictive processing, but might give a result that is slightly different from native processing.**

**This is an extremely valuable study for better understanding the mechanisms behind predictive processing. I find it a particularly clever move to use interpreters to get at the factor of experience in anticipatory processing. It is also impressive to have such a homogenous control group. The manuscript is very well written, and clearly organized, and the data analysis is sound. However, I have some general remarks and some more specific comments.**

**General remarks**

**In the Discussion/Conclusions, it would be good to get back more to and discuss the idea of increased experience with anticipatory processing influencing predictive behavior, which seemed to be the major research question.**

Our reply: Thank you for this comment, we added the following information in the conclusion “practice with anticipation during interpreting situations improves anticipatory abilities in the L2 during non-interpreting situations.” We also reiterated the idea throughout the discussion.

**It would be good to have summaries of the results in the Results section, more suitable for readers who are less familiar with the statistical models.**

Our reply: We added a new section, *Statistical Analyses*, in which we provide an overview of the three analyses conducted. Furthermore, we have attempted to make the prose for each statistical test more clear by explicitly stating the purpose of the analysis before reporting the results, followed by a short summary afterwards that we believe to be more accessible to readers that are less familiar with the statistical models. The revised prose is included in the new *Statistical Analyses* section as well as the subsequent *Results* section.

**In some tests, interpreters do not differ significantly from either of the other groups. Do they show more variability? This could be rather natural, since they have learned their skills as adults.**

Our reply: Interestingly the variability in interpreter responses does not appear to be greater than that of the other groups. Qualitatively, one can observe this by checking the confidence intervals/standard errors around the point estimates. Furthermore, the GLMM and growth curve analyses provide an assessment of by-subject variability (and controls for it) in the random effects structure. We do not observe any variability above and beyond that of the other groups.

**Expect for possibly increased variability, the interpreter group is very similar to the native speaker group, except in CV syllables, where they start predicting somewhat slower than the native speakers, whereas in CVC syllables they behave basically the same. Could this be an effect of how the interpreters have learned to use anticipation? Maybe in their experience, they need to have a slightly higher level of certainty before deciding what to predict, because if they do not, the error-rate might become unacceptably high for a smooth on-line translation. I am trying to say that the penalty for misprediction might be higher for interpreters than for native speakers. This might be interesting to discuss, but might also be a bit far-fetched.**

Our reply: We agree with the reviewer, this is a plausible hypothesis that goes in line with previous theories of prediction. We have added the following paragraph to the discussion

“This strategy could be influenced by the use of prediction during interpreting, when prediction error becomes more costly. IN might have adopted a different approach to prediction, in which they gathered as many cues as possible before commiting to a prediction. This hypothesis is consistent with how Kuperberg & Jaeger (2016) view prediction, as a graded phenomenon that depends on the comprehender’s goals, prior knowledge and the expected utility of making a prediction.”

**I am not sure how much figure 2 contributes to understanding the manuscript. Figure 3 is very nice. Generally, figures seem to need higher resolution. Even better if the journal accepts, use original vector graphic images (eps, pdf, but not raster graphic saved in eps format).**

Our reply: Thank you for this observation. We have included pdf (vectorized) versions of the figures so that the resolution is appropriate for publication. We have removed the original Figure 2 (the plot of GLMM parameter estimates) and made the appropriate adjustments to subsequent figures/labels.

**Please write out the exact p values instead of ranges (also in tables), i.e. p = 0.005, rather than p < 0.006, and p = 0.08, rather than p > 0.05. It is interesting to know the p values although they are not significant and p < 0.006 and then later p < 0.004 gives the impression you have different alphas. You can decide on a number of decimals, e.g. 3, and write p < 0.001 if lower than that, and if not give the exact p value. When you use Bonferroni-correction, please**

**correct the p values rather than the alpha. You still have an alpha of 0.05, but Bonferroni- corrected for six comparisons. Thus, e.g. uncorrected p = 0.001 would be reported as p = 0.006.**

Our reply: We have included exact p-values for all statistical tests except for those that are less than 0.001, for which we have maintained the “p < 0.001” standard. With regard to the Bonferroni correction, we respectfully disagree with the reviewer. The reason why we do not change the p-values is because the Bonferroni correction is, indeed, an alpha correction. Furthermore, p-values are directly related to t/z values and standard error, which we report for all tests. For this reason, changing the p-values, rather than alpha, would potentially be misleading for any readers interested in double checking our reported results. We respectfully refer the reviewer to the following resources in the case they may be of interest.

* Bonferroni, C. E. "Teoria statistica delle classi e calcolo delle probabilità." *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3-62, 1936.
* Dewey, M. "Carlo Emilio Bonferroni: Life and Works." <http://www.aghmed.fsnet.co.uk/bonf/bonf.html>.
* Miller, R. G. Jr. [*Simultaneous Statistical Inference.*](http://www.amazon.com/exec/obidos/ASIN/0387905480/ref=nosim/ericstreasuretro) New York: Springer-Verlag, 1991.
* Shaffer, J. P. "Multiple Hypothesis Testing." *Ann. Rev. Psych.* **46**, 561-584, 1995.

**Specific comments**

**(see Kamide/Roll/Kaan/Authors, 20XX, for a review) I think it is enough and less repetitive to just write (Kamide, 2008) etc. It is clear from the context that not only a single result is referred to.**

Our reply: change made

**Page 4:3: shed more light into this issue --> shed more light on this issue**

Our reply: change made

**4:10-14: “This has led some scholars to propose…” Not clear what “this” refers to. Perhaps this sentence could just be removed. Continuation between the previous and following sentences would actually be clearer.**

Our reply: sentence removed

**Page 5:11-12: don’t --> do not**

Our reply: change made

**While this conclusion sounds logical --> “logical” sounds a bit odd here, perhaps “reasonable”?**

Our reply: we have modified this sentence.

**Page 6:44-45: Corpora data --> Corpus data (corpora is plural, like saying “cars tires” instead of “car tires”)**

Our reply: we have modified this sentence.

**Page 6, last para: “they anticipate every 85 seconds” I am not sure what this means. Aren’t they anticipating all the time?**

Our reply: We added a clarification at the beginning of this paragraph anticipation during simultaneous interpreting refers to the production (not only mental preactivation) of a speech segment before it is uttered by the speaker: “Anticipation plays a central role in interpreting. It allows interpreters to pre-active and produce pre-activated information before hearing it, and is commonly taught in simultaneous interpreting courses (Li, 2015) to decrease cognitive load and to interpret efficiently (Seeber & Kerzel, 2011)”

**“and make a prediction 60% of the time when they encounter an L2 syntactic structure that differs from their L1” And if the L2 structure does not differ, don’t they predict? Again, not sure what was intended here.**

Our reply: Again, this definition of prediction also entails production. We have reformulated this part: “To predict, interpreters employ discourse redundancy (Chernov, 2004), contextual and syntactic knowledge (Moser-Mercer, 1978). This allows interpreters to anticipate often (a sentence every 85 seconds, Van Besien, 1999), and effectively (they predict accurately 95% of the time, Liontou, 2012, and increased levels of prediction correlate with fewer errors and with a more complete interpretation with fewer omissions from the source speech, Kurz & Färber, 2003).”

**“interpreters’ predictions are mostly accurate” Compared to what?**

Our reply: see comment above with reformulation of this sentence.

**“and correlate negatively with errors, but positively with interpretation completeness” Please explain.**

Our reply: see comment above

**Page 7:3 interpreters --> interpreters’**

Our reply: change made

**Page7 “Results showed that they produced the expected referent when interpreting into their L2 (therefore making a prediction error), but were able to render the unexpected item when interpreting into their L1.” Maybe “expected (incorrect) referent” would make it clearer, and I assume it should perhaps be something like “were better able to render”**

Our reply: we have modified this section. “Chernov (2004) investigated interpreters’ anticipation of highly constraining sentences with unexpected endings while performing simultaneous interpreting. Results showed that they generated more accurate predictions, measured by production of expected referents, when interpreting from their L2 to their L1 than from their L1 to their L2. However, the participants’ L1s were mixed, the variables were unclear, and statistical analyses were absent. ”

**Page 8, first para. I understand what the authors mean, but this paragraph can be somewhat confusing, since prosodic “stress” is contrasted with vowel reduction, which can be seen as a correlate of stress in English. Please consider reformulating.**

Our reply: We added an explanation “For instance, in Spanish and English monolinguals, a prosodically matched prime facilitates perception, but a mismatched prime inhibits (slower RTs) perception in Spanish monolinguals (Soto-faraco, Sebastián-gallés, & Cutler, 2001), but not in English monolinguals (Cooper, Cutler, & Wales, 2002). These differences suggest that lexical stress in Spanish is used to reduce the number competitors for lexical access, but this does not seem to be the case in English, likely due to the fact that vowel reduction can efficiently fill this role.”

**Page 8:16 “reduce the number competitors” --> reduce the number of competitors**

Our reply: change made

**Page 8:44-45 “when more time and acoustic information are present” This might partly be based on the same mechanism that produces longer simple reaction times for longer stimuli (Raab, 1962, The American Journal of Psychology). Mention if you find it relevant.** Our reply: This is an interesting possibility, thank you for the reference. We have added this possibility to the discussion: "An anonymous reviewer suggested that this might be based on the same mechanism that yields shorter reaction times for longer stimuli (Raab, 1962). While this is a plausible explanation, it must be taken with caution because this study is based on noise stimuli (not language). “

**9:16 rooted on --> rooted in?**

Our reply: change made

**10:18: at same time the speaker is talking --> at the same time the speaker is talking**

Our reply: change made

**11:40 they contain --> the syllable contains… and has longer duration (or the syllables contain)**

Our reply: we have reformulated this part.

**12:40 pseudo-randomize --> pseudo-randomized**

Our reply: change made

**14:25 bonferroni corrected --> Bonferroni-corrected**

Our reply: change made

**The first paragraph of the Results section would be better in a subsection of Methods called “Eye-tracking data analysis” or similar.**

Our reply: since we only report data on one eye-tracking experiment, we have named this new section “Statistical Analyses”

**14:29 random random --> random**

Our reply: change made

**15:42-44: “there was a group by coda interaction on linear slope for the IN group (Estimate = 3.30, SE = 0.37, p < 0.001), indicating faster fixations to the target.” what is an interaction with group for a certain group?**

Our reply: This sentence was poorly phrased and has been rewritten. It was in reference to the slope adjustment of the linear term for the interpreter group when there is a coda. In other words, the interaction of the linear term and the coda is different for the interpreter group with regard to the baseline (monolinguals). The revised prose is as follows:

“Finally, there was a group by coda simple interaction on the linear slope for IN (Estimate = 3.30, SE = 0.37, p < 0.001), which indicated a faster rate of fixation on the target words for IN with respect to M.”

**16:14 even though --> although**

Our reply: change made

**16:29-40: You could be a little bit more specific about the results here. 16:38 bilinguals --> learners?**

Our reply: change made. We also corrected other cases where bilinguals was used instead of learners.

**17, first para: Perhaps the general tendency that longer time allowed for preparation gives shorter reaction times (e.g. Raab, 1962) might be worth mentioning. Again, only if you find it relevant.**

Our reply: See previous comment

**18:42-44 “faster although later in the time-course” Faster than whom? Later than whom?**

Our reply: “faster although slightly later in the time-course than monolinguals in CV syllables”

**Reviewer 2**  
**This study examined how anticipatory experience of L2 affects suprasegmental cues processing in Spanish monolinguals and English-Spanish bilinguals (with and without interpreting experience). Authors thought that L2 proficiency levels were matched but anticipatory experience were distinct between interpreters and non-interpreters. Eye-tracking data revealed all groups can use stress information to anticipate morphological information for CVC target structure, while only monolinguals and interpreters can anticipate morphological information for CV target structure. However, I have several concerns about this study regarding background, experimental design, data analysis, and whether the observed results indeed warranted the conclusions.**  
  
**1. Previous studies suggest that L2 proficiency is an important factor in determining the anticipation abilities. The focus is whether increased anticipation performance is caused by experience with the language (proficiency), experience with general anticipation, or a combination of the two. Whether participants with different proficiencies in the previous studies also differ in their anticipation experience? According to the experimental design, participants in the current experiment can only explore how anticipation experience moderates without the interference from language proficiency. The author should focus more on the role of anticipation experience instead of L2 proficiency in the literature review.**

Our reply: We agree with the reviewer. We made substantial changes to the section on anticipation in L2 learners. We deleted any interpretation of previous L2 studies as evidence of proficiency effects on anticipation, and we added a sentence emphasizing that these studies mix proficiency with anticipation experience. Thank you.

**2. Two key independent variables in this study were stress position and syllabic structure, but their influence on L2 anticipation abilities were not clearly illustrated in the introduction, especially stress positions. Besides, the authors neither put forward a hypothesis of stress, nor an interpretation to the null effect of stress in the discussion. Taken together, why stress stands as an independent variable here?**

Our reply: we added information about why we used these two variables in the Introduction, and we added hypotheses regarding the two variables in The Present Study.  
  
**3. This paper has reviewed in the introduction that most experiments have found anticipation effects in native and non-native speakers. So, why eye-tracking technique in current experiment? What are the advantages of eye-tracking measure when dealing with this issue?**  
  
4. **Authors just mentioned visual world paradigm briefly in the introduction. This paradigm is important for understanding the experimental methods and logic, and it should be described in detail.**

Our reply: Thank you for this comment. Due to space limitations, we have combined the information requested by the reviewer in comments 3 and 4. We added the following information in introduction: “Finally, the visual world paradigm was employed because it measures attention to upcoming linguistic information before it has been disclosed, by time-locking listeners’ eye-movements to a visual stimulus (e.g., a written word) in response to an oral stimulus (e.g., a sentence) (see Huettig, Rommers, & Meyer, 2011, for a review).”

**5. Authors varied anticipation experience by interpreting and non-interpreting experience between interpreter and non-interpreter in their study. What is anticipation ability? Are there objective measures in the literature? How to know interpreter and non-interpreter groups have significant different anticipation experiences or anticipation abilities?**

Our reply: This clarification was also requested by reviewer 1. We added a clarification at the beginning of this section explaining that previous studies on the use of anticipation during simultaneous interpreting refer to anticipation as the production (not only mental preactivation) of a speech segment before it is uttered by the speaker: “Anticipation plays a central role in interpreting. It allows interpreters to pre-active and produce pre-activated information before hearing it, and is commonly taught in simultaneous interpreting courses (Li, 2015) to decrease cognitive load and to interpret efficiently (Seeber & Kerzel, 2011)”

**6. Studies demonstrated that the interpreting experience enhances general cognitive abilities. What exactly does general cognitive abilities refer to in the current study? In the methods, besides language background questionnaire and proficiency test, participants have also taken some other tests. What are the aims and significances of these tests? In the results, the authors only reported that participants were matched in terms of L2 proficiency and WM, however, they didn’t conclusively report that participants were comparable in general cognitive abilities. If mismatch in general cognitive abilities, how can the authors prove that differences in anticipation abilities are resulted from rich anticipation experience in interpreters, rather than the increased cognitive abilities.**

Our reply: We agree that further explanation was needed, thank you. There is still debate around the notion of an interpreter advantage (similar to the debate around the bilingual advantage). We have incorporated references that discuss this debate, as well as explained which brain areas seem to be impacted by interpreting training: “this explains why interpreters are better at: (1) detecting written errors than interpreter students, non-interpreter bilinguals, and monolinguals (Yudes, Macizo, Morales, & Bajo, 2013), (2) adapting their strategies to tasks (e.g., repeating information vs. interpreting into their L2) (Togato, Paredes, Macizo & Bajo, 2015), (3) reading speed, comprehension, WM (Bajo, Padilla & Padilla, 2000) (but see Dong & Cai, 2015, for a review of studies also against this WM-interpreter advantage). Furthermore, interpreters exhibit increased cortical thickness in brain areas related to phonetic processing, higher-level formulation of propositional speech, conversion of items from WM into a sequence, and domain-general executive control and attention (Hervais-Adelman et al., 2017). We examine whether this “interpreter advantage” extends to non-interpreting situations, specifically, L2 anticipation. ”

Due to space limitations, we cannot discuss in detail all measures that have been used to test the interpreter advantage, neither was it feasible to collect a battery of tests tapping into other executive function components. In order to make our groups comparable, we chose one of the executive function components that has shown differences between interpreters and non-interpreters, namely working memory.

We also made cleared that we used WM as a measure of cognitive abilities in the participant’s section: “To rule out the possibility of IN performing better than NIN due to higher WM or L2 proficiency, we tested for homogeneity of variance for WM (all groups) and L2 proficiency (L2 groups), and then conducted TOST (two one-sided tests) of equivalence for all pairwise comparisons (Lakens, 2017).”  
**7. In the section of Lexical stress in Spanish and English, the authors didn’t clearly explain why they chose lexical stress to explore the anticipation abilities in English-Spanish bilinguals after a simple introduction to the definition of lexical stress. Apart from that, the important contents in this section are not realized in latter discussions on the results. I suggest authors should establish closer relationships between the listed contents and the reflections on results, and enrich the discussion.**

Our reply: We agree that the inclusion of lexical stress was not sufficiently motivated. We have rewritten the section *Lexical Stress in Spanish and English* section, which is now titled *Lexical Stress and Syllabic Structure in Spanish and English* (starting on page 8 of the revised manuscript), and we have added a close connection to this issue throughout the *Discussion*.

**8. Authors proposed two research questions: 1) Does monolinguals, interpreters, and non-interpreters use lexical stress to anticipate morphological information? If yes, does syllabic structure (CV/CVC) affect anticipation? 2) Does practice with interpreting facilitate anticipation in non-interpreting contexts? My question is about the necessity of independent existence of research question two. Obviously, the present study didn’t explore the question in both interpreting and non-interpreting contexts. The answer to question one seems enough to answer question two. Meanwhile, non-interpreting contexts have been frequently mentioned, but never been specified in the current study.**

Our reply: we agree. We have merged the two questions.  
  
**9. In this experiment, there were 18 practice sentences, 32 fillers, but only 16 experimental sentences. The average number of sentences in each condition is relatively few, so how to ensure the reliability of results.**

Our reply: We agree with the reviewer that having a larger number of experimental items would be ideal, as is typically the case. We have taken care to use robust statistical modeling in order to aid our inferences. This includes reporting parameter estimates and measurements of spread as an indication of effect size. We believe that with the information provided we make a strong case in favor of the reliability of the results, though we can only truly be sure of this through replication.  
  
**10. Reaction time can also reflect the anticipation abilities. My question is whether the RTs of button press after the onset of target word were recorded? If yes, why they were not reported in the results?**

Our reply: Our experiment did indeed record reaction time of button presses after the onset of target items. While including RT data would certainly complement the present manuscript, we are unable to report them at this time due to word limitations. Moreover, we are measuring anticipation at the word level, which happens within milliseconds. Participants might have waited until the end of the sentence in order to press the button, which is not informative about whether they anticipate or not. Fortunately, we feel that we make a strong case for the role of prosodic information and syllable structure with the analyses that are reported.   
  
**11. Eye-tracking data were analyzed in three methods, one-sided t-tests, generalized linear mixed effects model, and growth curve analysis. The authors should make it clear the different aspects of data each method focuses on, and specify the advantages growth curve analysis has in comparison with the former two. Besides, why one-sided t-tests focuses only on paroxytone and oxytone, while ignoring CV and CVC structures?**

Our reply: We agree that it was necessary to clearly state the purpose of each analysis and how they were used to answer our research questions. We have included a new section, *Statistical analyses*, and added more information before each of the three analyses in order to make this more clear. Specifically, we believe the reader can now see more explicitly how the information obtained from each analysis distinctly contributes to our understanding about the role of prosodic information and syllabic structure with regard to anticipation.

The reviewer also brings up an important question with regard to the one-sided t-tests, which focus on lexical stress. A priori lexical stress was the primary variable for our study. As such, it was given causal priority in all models. For this reason it was also the object of focus in the t-tests. We considered the possibility of subsetting the data further in order to test syllabic structure as well, however, we ultimately decided the benefits of doing so did not outweigh the costs, particularly to our alpha, which was already Bonferroni corrected for 6 comparisons. We believe that we were able to fully explore the role of syllabic structure with the data we currently have via the GLMM and GCA. In sum, in order to maintain statistical power and protect alpha, we made the decision not to conduct more t-tests, but syllable structure was included in subsequent models. We currently have other studies underway that will focus on syllable structure.  
  
**12. Overall speaking, the structure of the paper is unbalanced, specifically, some redundancies and restatements in the introduction, while insufficiency in the discussion. Meanwhile, partial discussions only rest on the enumeration of the phenomena, with fewer in-depth explorations.**

Our reply: We agree. We reduced the Introduction, we expanded the discussion, and we offered a more critical review of the literature, as long a more in-depth explanation of the results in the discussion section.  
  
**Minor comments  
The first part of results section is to test the differences among groups on WM and L2 proficiency. This should be reported in the methods section.**  
Our reply: change made. We included this information in Results, because, in the past, whenever we included it in Methods, a reviewer would ask us to move it to Results. We agree with this reviewer in that the appropriate section to report homogeneity results is Methods. Thank you.

**Reviewer: 3**  
**Summary:  
The study tests two groups of English-Spanish late bilinguals, interpreters and non-interpreters, matched on Spanish proficiency against a monolingual control group in a visual world task using written stimuli that vary in lexical stress and phonological weight of the first syllable. The investigators hypothesize that the interpreter group will show greater predictive ability compared to the non-interpreter group due to their overall experience in anticipation derived from interpretation. The results indicate that the interpreter group shows a similar ability to predict target words using stress, whereas the non-interpreter group shows less prediction.   
  
Evaluation:  
I commend the authors for including multiple analyses for testing their research question. The experimental design seems simple and clear. The topic itself is really innovative; however, with the descriptions provided thus far, I am not as certain as the authors that they have dissociated language experience (which they term proficiency) from training in anticipation derived from interpretation. I expand on these comments below and ask for additional details as well. I believe that the study will be interesting to the journal audience, but I am just not certain that the strong claims made by the authors are warranted from the design or the results.   
  
Comments:**  
**1. Definition of anticipation.** **Authors define anticipation as “preactivation of a linguistic item that has not been mentioned yet.” This definition sounds like what others call prediction (Federmeier, 2007; Kaan, 2014; Kuperberg & Jaeger, 2016). One important current debate is whether prediction involves actual pre-activation or is more probabilistic. I would like to see the authors touch on this debate. (p. 3, 1st paragraph)**

Our reply: we have added the following information in the manuscript to clarify current debates around the definitions of prediction: “There is still a wealth of unsolved problems and unanswered questions regarding how humans anticipate information. Does prediction involve pre-activation or just a state of preparedness (Ferreira & Chantavarin, 2018)? Is pre-activation probabilistic or all-or-nothing (DeLong, Urbach, & Kutas, 2017)? What exactly do people predict (Huettig, 2015)? How pervasive is prediction (Huettig & Mani, 2016; Nieuwland et al., 2018)? Future research investigating these issues must take place before a comprehensive understanding of the cognitive mechanisms underlying prediction is possible. ”

**2. Clarification on contrast being studied here. Authors make strong claim that anticipation in L2 may be a specialized skill that needs to be developed (as in interpreters) and may not just come with increased proficiency. I believe this is a fair hypothesis; however, I don’t think the terminology used specifies this well. Specifically, the authors describe proficiency as “increased experience with language” (p. 3, 2nd paragraph). I would argue instead that experience (with language) is separable from proficiency and that the results from the interpreters (demonstrating L1-like processing) may be due to experience rather than specifically anticipation—this is related to the point above on what is meant by prediction—whether it is actual pre-activation or more probabilistic. In order to convince me that the L1-like prediction seen in interpreters is due to training in pre-activation and not experience (i.e., sheer number of hours exposed to the L2), the authors would need a separate measure on frequency of L2 use and have two groups matched on frequency of L2 use rather than proficiency alone. With the measures provided by the authors, I do not see evidence for specifics of interpreting providing the benefit beyond greater frequency of L2 use. This same framing (experience with language v experience with “general” anticipation) is presented again on p. 6, first full paragraph and in the goals of the study, outlined in pp. 8-9. I’d like to see this potential confound addressed in the discussion. It does appear that the authors have collected percent of time using L1 and L2 for both interpreter and non-interpreter groups. At the very least, the authors could report this information and confirm or disconfirm whether there is any difference in amount of time spent in the L2 between the two groups.**

Our reply: we agree. We deleted all references associating L2 proficiency with language experience, we added a *t-*test comparing the frequency of L2 use in the interpreter and the non-interpreter group, and we discussed this issue at length in the discussion.

**3. Literature review on predictive processing in L2. On p. 5, the authors summarize that L1 speakers consistently use a myriad of cues to generate predictions but that L2 learners are not as effective. The mixed findings in L2 processing are discussed as unrelated to age of acquisition, L1 characteristics or L2 proficiency. I believe that the authors’ reading of the literature warrants greater depth. Overall, the authors are treating prediction as a categorical ability, either one shows it or doesn’t, yet even within the L1 literature, prediction is being viewed in a more nuanced way. For example, the authors cite Martin et al. (2013) as an illustration of failure to show prediction in L2 processing, yet it is based on a finding from the L1 literature (DeLong et al., 2005) that has recently failed to replicate in a multi-site attempt using the same materials and methods (Nieuwland et al., 2018). Moreover, individual differences in predictive processing, again in L1 populations, indicate that even amongst supposedly homogenous L1 participants predictive processing will differ (Tanner, in press; Tanner & Van Hell, 2014) and can change as L1 speakers age (work from Federmeier group: e.g., Wlotko et al., 2010; Federmeier et al., 2010; Leckey & Federmeier, 2017).**

**On the L2 processing side, the authors’ critical reading distills mixed results as failure to find consensus on potential factors that affect predictive processing. However, of those studies that do find anticipatory processing, it is often associated with higher proficiency and L1 may affect how quickly L2 participants begin to exhibit predictive processing. The authors cite Lew-Williams & Fernald (2010) as evidence that L2 processing lacks the ability use morpho-syntactic cues. The test case was grammatical gender in English learners of Spanish using the visual world paradigm. Yet other studies using the same technique do show anticipatory processing at higher levels of proficiency with the same type of speakers (L1 English-L2 Spanish) and ability to show facilitated processing at lower levels of proficiency but only in congruent cross-linguistic structures by Italian learners of Spanish (Dussias et al., 2013). Other studies have also found an effect of L1 on sensitivity to grammatical gender processing (e.g., Sabourin & Stowe, 2008). And as is the case of L1, recent studies have also taken a more individual-level approach to this question, suggesting that L2 speakers look predictive in an L1-like, when we examine predictive effects based on what the L2 speaker knows (Hopp, 2016). These are just a few examples that are simplified in the review provided by the authors, and I would ask the authors to provide a more nuanced account that accurately reflects the state of L1 and L2 predictive processing.**

Our reply: For L1 studies, we added this information “Such predictions depend on speech rate (slower rates increase prediction), preview time (longer times increase prediction), task instructions (explicitly instructing participants to predict increases prediction) (Huettig & Guerra, in press), and age (younger age increases prediction) (Wlotko, Lee, & Federmeier, 2010). Interestingly, older monolinguals with larger vocabularies and higher verbal fluency are as effective as younger monolinguals making linguistic predictions (Federmeier, Mclennan, De Ochoa, & Kutas, 2002), suggesting that increasing language experience may enhance anticipatory abilities.”

For L2 studies “Thus, Dussias, Valdés Kroff, and Guzzardo Tamargo (2013) found that low-proficiency learners of a gendered L1 (Italian) can partially use gender information to make gender agreement predictions in a gendered L2 (Spanish), whereas low-proficiency learners of a genderless L1 (English) cannot. In addition, Hopp (2016) reported that lacking a mental representation of gender marking hinders L2 prediction of gender agreement (Hopp, 2016).”

We could not add the following references: Sabourin & Stowe, 2008; Tanner, in press; Tanner & Van Hell, 2014. We understand the connection between these articles and our research. However, due to space limitations, we had to focus on articles related to predictive processing (as opposed to sensitivity to morphosyntactic violations or grammatical agreement anomalies). We hope that the addition of the previous paragraphs sufficiently explains variability among research in L1 and L2 predictive processing.

**4. Literature review on predictive processing in interpreters. It may be helpful to further expand this section by including work by Teresa Bajo’s group (e.g., Macizo & Bajo, 2006, Togato et al., 2017, Yudes et al., 2012) on how sentence processing in interpreters can change based on what task they are being asked to do (e.g., reading for comprehension v. reading for repetition). Some of these studies are mentioned in the discussion briefly, but they could be moved to the introduction.**

Our reply: We added the following information: “Simultaneous interpreting is cognitively taxing (Gile, 2015), because it requires interpreters to retain information from the source language in working memory (WM) , access meaning, connect to previous information, translate into the target language, and produce the message in the target language (Bajo, Padilla, & Padilla, 2000). This explains why interpreters are better at: (1) detecting written errors than interpreter students, non-interpreter bilinguals, and monolinguals (Yudes, Macizo, Morales, & Bajo, 2013), (2) adapting their strategies to tasks (e.g., repeating information vs. interpreting into their L2) (Togato, Paredes, Macizo & Bajo, 2015), (3) reading speed, comprehension, WM (Bajo, Padilla & Padilla, 2000) (but see Dong & Cai, 2015, for a review of studies also against this WM-interpreter advantage).” Due to space limitations, we did not include Macizo & Bajo, 2006. This article is about translators (they work with written texts) and not interpreters (they work with oral speech).  
  
 **On p. 7, lines 5-11, authors interpret the finding that interpretation into L2 resulted in greater prediction errors than when interpreting into L1 with unexpected continuations as in line with prediction being stronger for L1 than L2 processing. I don’t follow the logic here. Please explain further.**

Our reply: We reformulated the description of this study as follows “Chernov (2004) investigated interpreters’ anticipation of highly constraining sentences with unexpected endings while performing simultaneous interpreting. Results showed that they generated more accurate predictions (measured by production of expected referents) when interpreting from their L1 to their L2 than from their L2 to their L1. However, the participants’ L1s were mixed (impeding Chernov from forming an association between the results and L1 transfer), the materials are not adequately described (preventing to establish which variables were analyzed), and there is not an statistical analysis (limiting the scope of the conclusions the author draws).”

**5. Results. Why was only a 10 ms bin selected for the t-tests analysis? Is this based on protocol from a prior study? Please cite if so.   
Please provide the full model output as an appendix for the GLM as was done for the GCA model. Also, what does parameter stand for in the Appendix II model output? I would usually consider the parameter to be the estimate, which is already reported. Please clarify.  
In the GLM, were stress and syllable structure also coded with reference levels or was some other coding scheme used? Please specify. Also, if monolinguals were the reference level for the Group factor, how was the non-difference between the two bilingual groups determined? Was a separate model created for this contrast?**

Our reply: As the reviewer has correctly observed, our analyses utilize two different bin sizes: 10ms and 50ms. In eye-tracking research, selecting a bin size for analysis is not trivial and always involves a trade-off between statistical power and temporal resolution (see Mirman 2016 for a complete discussion). The motivation for using the 10ms bin for the t-tests and GLMM follows Llompart & Reinisch (2017), which also utilized the visual world paradigm to analyze segmental and suprasegmental information. In our case, it was critical to use a smaller bin at the offset of the first syllable of the target words in order to respond to the main research question: Do participants anticipate verb tense before hearing the critical suffix? At the segmental level using a bin size that is too large would make this question difficult, if not impossible, to answer, as we would not know if responses were above chance before having heard the critical suffix. We reduce the possibility of this happening by reducing the bin size to 10ms, though the trade-off is having to sacrifice data for the analyses. The 50ms bins are used for Growth Curve Analysis to reduce computational strain and facilitate plotting. This decision had no effect on the results.

* Llompart, M., & Reinisch, E. (2017). Articulatory information helps encode lexical contrasts in a second language. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(5), 1040.
* Mirman, D. (2016). *Growth curve analysis and visualization using R*. CRC Press.

With regard to the second set of questions, we have included the full model output of the GLMM in the appendix. In this case, stress and syllable structure were dummy coded. We used the *multcomp* package (Hothorn et al., 2008) to test the non-difference between the two bilingual groups. We have provided all relevant information in the manuscript in order to assure this information is clear to the reader. Specifically, this information can be found on pages 15-16 of the revised manuscript, which includes the following sentence:

“Where appropriate, pairwise tests were conducted using the R package *multcomp* (Hothorn, Bretz, & Westfall, 2008).”

* Hothorn, T., Bretz F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal.* 50(3), 346–363.

**6. Interpretation of results. The authors triangulate the results of the 3 analyses to infer that non-interpreter bilinguals only predict in CVC words whereas interpreters and monolinguals predict in both syllabic conditions. However, I understand the tests as follows: the t-tests test for immediate prediction after target syllable offset (no evidence of prediction for non-interpreter group), the GLM only shows a main effect for group and syllabic structure (doesn’t answer question of prediction in bilingual groups, just shows overall lower looks to target item as compared to monolinguals and overall greater looks in CVC conditions), and the GCA model showed a non-interpret x quadratic time term interaction (slower than monolinguals) and a group x syllable structure interaction, indicating lower fixations on CV syllables for both bilingual groups compared to monolinguals. The t-test speaks on immediate prediction after syllable offset but neither of the other 2 analyses address prediction in my opinion. Rather, they show overall fewer looks to target items and slower rates for the non-interpreter group. This leaves open the possibility that the non-interpreter group is predicting but at a slower time course than the monolinguals and interpreters. In particular, when looking at Figure 3, it appears that by 50 ms after target syllable offset, the non-interpreter group is looking at target items at greater than chance. This is presumably before the end of the word, even if it is slower than monolinguals and interpreters. I’d like to see this possibility mentioned in the discussion. Alternatively, this could explicitly be tested by increasing the time window used in the t-test analysis.**

Our reply: The reviewer correctly observes that only the t-test analysis directly tests the hypothesis that the three groups of participants are predicting the appropriate suffix before actually hearing it. The GLMM and the GCA compare group responses at the target syllable offset and group fixation behavior over the time course, respectively. As noted, the non-interpreter group does not predict at the target syllable offset, i.e., overall they do not fixate on target items above chance at this critical point. For this reason we do not believe that they are predicting at a slower time course than the monolinguals and interpreters, at least not in the sense that we are testing and interested in for the present work. In other words, were we to use a wider time bin at the syllable offset, it is possible that the non-interpreter fixations would indeed be above chance; however, they would already have heard part of the suffix. That is, they would already have morphological information in the acoustic signal that directly indicates to them the verb tense and number, even if at this point it is before the end of the target item. Essentially, this would introduce a confound and we would not be able to say how they were determining where to look, i.e., via anticipatory mechanisms or general perception. For this reason this type of analysis would not respond to the questions that we are focusing on at this time.

**As for the discussion of question 2, I would just iterate that this study at this point doesn’t in my view isolate the greater prediction in the interpreter group to interpretation per se. Rather, this benefit may be due to greater exposure to the language, i.e., if a non-interpreter group were similarly working in their L2 at a greater percentage, they may also demonstrate greater anticipation.**

Our reply: as mentioned earlier, we deleted all references associating L2 proficiency with language experience, we added a *t-*test comparing the frequency of L2 use in the interpreter and the non-interpreter group, and we discussed this issue at length in the discussion.

**Miscellanea:**  
**a. On p. 4, lines 40-42, authors write that anticipatory studies on word recognition are scant. Is this a fragment? Otherwise it seems contradictory to what was written at the beginning of the paragraph and with the prior literature that has copious studies on visual and spoken word recognition. Perhaps I’m missing something. Do the authors mean something more specific here? Please clarify.**Our reply: We have reformulated the introductory section, we hope it is clearer now.

**b. Authors write: “Rehrig’s (2017) findings that upper intermediate (not advanced) Chinese learners of English fail to use vowel duration cues to differentiate between active and passive verb endings provides further evidence for the proficiency explanation to the discrepant findings of L2 prosodic studies” (p. 5, lines 44-50). This sentence is hard to parse. Please consider breaking up into shorter sentences.**

Our reply: we agree, we have reformulated this sentence and included limitations of this study “Rehrig (2017) reported that Chinese learners of English failed to use vowel duration to predict verb suffixes essential to interpret the sentence as active or passive, but the null results could be due to lack of proficiency (it was assessed subjectively via self-ratings), or the use of a contrast known to be acquired late even in monolinguals (active/passive voice).”

**Also, it’s unclear to me that L2 studies use labels such as beginner, upper intermediate, intermediate, advanced in a standardized way, which may also contribute to the mixed findings.**   
Our reply: we agree, we added this as a possible limitation (see comment above).

**c. On p. 10 when describing the interpreter group, the paragraph ends with “…although two interpreters showed the opposite pattern.” I didn’t understand what this phrase was referring to, please clarify.**

Our reply: We clarified as follows “Most of the IN worked in the simultaneous interpreting mode (the interpreter translates the speech at the same time the speaker is talking) and occasionally in consecutive interpreting (the interpreter renders the translation after the speaker finishes one section of the speech), although two of them interpreted mostly in the consecutive mode.”

**d. On p. 11, line14: for spatial resolution specification, change .32 to 32.**

Our reply: change made

**The averaged calibration error of .01 degrees strikes me as really low—is this accurate? I think normal values are around .25-.5 degrees.**

Our reply: We agree, this was a mistake, the range provided by the reviewer is correct and reflects our data, we changed this information.

**e. On p. 11, lines 37-40, the authors write: “Previous studies show that the coda could make a difference in the ability to anticipate…” Please include citations to prior work here.**

Our reply: We have added a paragraph explaining why we investigate syllabic structure in the new section “Lexical Stress and Syllabic Structure in Spanish and English”. Moreover, we added this information “The Present Study” section: “Second, we expect advanced L2 learners to be able to anticipate (see Schremm et al., 2016, for evidence that advanced L2 Swedish learners use tones to anticipate suffixes), but only in CVC conditions (see Sagarra & Casillas, 2018, for evidence that advanced L2 Spanish learners anticipate with CVC, but not CV, conditions). “

**f. Please indicate size, font, and placement of visual stimuli. I’m also inferring from the 2 experimental lists that position of target words was not counter-balanced. Were word pairs always paroxytone on left and oxytone on right? Please also indicate what keys were used for left and right. Please specify in Materials section**.

Our reply: we added this information to the materials section “Words were displayed in Arial font and 150pt size, were centered in the left and right halves of the screen, and were counterbalanced (half of present verbs appeared on the left, half as targets and half as distractors, and half of past tense verbs appeared on the right, half as targets and half as distractors).”

We used the shift keys “Then they chose one of the two words as soon as they could by pressing the right or left shift key”

**g. I’m wondering why the authors are using Bartlett’s test to test for mean differences across groups in working memory and proficiency. Doesn’t this just test for equal variance? Shouldn’t the test be an ANOVA for working memory and a 2-sample t-test for proficiency?**

Our reply: Thank you for making this observation. This was the result of a miscommunication. In order to test for equivalence amongst the groups for working memory and DELE scores we first tested for homogeneity of variance (reported) and then we conducted a series of TOST (two one-sided tests) of equivalence for pairwise comparisons (now reported). See Lakens (2017) for information regarding this procedure.

* Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8(4), 355-362. doi:10.1177/1948550617697177

**h. Please provide the exact p-values for t-tests in Table 2 for the non-interpreter group.**

Our reply: Exact p-values for the non-interpreter group are now included in Table 2.

References not cited in manuscript  
  
**DeLong, K. A., Urbach, T. P., Kutas, M. (2017). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience, 8, 1117-1121.**

Our reply: added **Dussias, P. E., Valdes Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. Studies in Second Language Acquisition, 35, 353-387.**Our reply: added

**Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. Psychophysiology, 44, 491-505.**Our reply: This article is included in the review by Wlotko et al., 2010, due to space limitations we have added only one.

**Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. Brain and Language, 115, 149-161.**Our reply: added

**Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. Second Language Research, 32, 277-307.**Our reply: added

**Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? Language, Cognition, and Neuroscience, 31, 32-59.**Our reply: added

**Leckey, M., & Federmeier, K. D. (2017). Age-related shifts in hemispheric dominance for syntactic processing. Psychophysiology, 54, 1929-1939.**Our reply: since our study doesn’t focus on age-related effects and due to space limitations, we have added only one article related to this issue (Wlotko et al., 2010).

**Macizo, P., & Bajo, M. T. (2006). Reading for repetition and reading for translation: Do they**

**involve the same processes? Cognition, 99, 1-34.**Our reply: We included other articles by Bajo’s group, however this one is about translation and due to space limitations we haven’t included it.

**Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., …, Huettig, F. (2018). Large-scale replication study reveals limit on probabilistic prediction in language comprehension. eLife, 7, e33468.**

Our reply: added **Sabourin, L, & Stowe, L. A. (2008). Second language processing: When are first and second languages processed similarly? Second Language Research, 24, 397-430.**

Our reply: We understand how this article is related to our topic, although it is not specifically about prediction. Due to space limitations, we have not included it in our literature review.

**Tanner, D. (in press). Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach. https://na01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fdx.doi.org%2F10.31234%2Fosf.io%2Fnjgt6&amp;data=02%7C01%7Ccl1032%40spanport.rutgers.edu%7C029a2447cdc74a2f4bb708d63b08b0f0%7Cb92d2b234d35447093ff69aca6632ffe%7C1%7C0%7C636761306367352976&amp;sdata=Ag1lNS%2F%2FvdyZBUOjmfaFxIQ0lm61WJoM%2FpnXB5EPUVA%3D&amp;reserved=0**Our reply: We understand how this article is related to our topic, although it is not specifically about prediction. Due to space limitations, we have not included it in our literature review.

**Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. Neuropsychologia, 56, 289-301.**Our reply: We understand how this article is related to our topic, although it is not specifically about prediction. Due to space limitations, we have not included it in our literature review.

**Wlotko, E., Lee, C., & Federmeier, K. D. (2010). Language of the aging brain: Event-related potential studies of comprehension in older adults. Language and Linguistics Compass, 4, 623-638.**

Our reply: added **Yudes, C., Macizo, P., & Bajo, M. T. (2012). Coordinating comprehension and production in simultaneous interpreters: Evidence from the articulatory suppression effect. Bilingualism: Language and Cognition, 15, 329-339.**

Our reply: added