**General comments for the editor and reviewers:**

After giving careful consideration to the comments made by reviewer 3, we have decided to focus our analysis on the 50 ms bins. As (s)he points out, bin size is an arbitrary decision when using eye-tracking data and we feel that we can present a more parsimonious analysis by electing one bin size and using it to respond to all of our research questions. As a result, we have been able to streamline the results section and present a clearer analysis that does not entirely depend on arbitrary decisions, or researchers degrees of freedom. We now utilize the growth curve analysis and resulting model estimates to test all of the hypotheses put forth in the manuscript. In conjunction with the larger sample size, we now present a more robust analysis.

**Reviewer: 1**

**Comments to the Author**

**The authors have satisfactorily addressed my comments. However, I found some inconsistencies in citations.**

**“Anticipation in monolinguals” section, first para:**

**One of the references to tone studies is not about tone but rather about the predictive strength of segmental phonemes. Please rather move this reference to places where segmental cues to prediction are mentioned. The same confusion is repeated at several points in the manuscript.**

Our response: Thank you for noticing this confusion. We have fixed the references throughout the manuscript.

p. 6 . “Relevant to our study, native speakers make use of suprasegmental and segmental information to predict morphology within a word. For suprasegmentals, the Swedish use tone to predict number (singular/plural) (Roll, Horne, & Lindgren, 2010; Söderström, Horne, & Roll, 2015; Roll, Söderström, & Horne, 2013) and tense (present/past) (Söderström, Roll, & Horne, 2012; Roll, 2015)”

p. 10/11 “such that initial segments with fewer possible and more frequent endings trigger stronger preactivation (Roll et al., 2017). In other words, the syllable structure of a lexical item might aid anticipatory processes before morphological information becomes available. ”

**Page 5, first para: Please check which references are concerned with number and which with tense contrasts. Perhaps some other study was meant for the tense contrast?**

Our response: We appreciate this comment, we changed it as shown in the previous response.

**Reviewer: 2**

**Comments to the Author**

**The authors revised the manuscript according to the reviewers’ comments and suggestions, and the manuscript has been improved substantially. Thank you you’re your revisions. After reading the manuscript a few of times, I still have the below concerns.**

**1. What is language experience in the manuscript? The first time it was introduced was in page. 4 about old monolinguals versus younger monolinguals. In this sense, I may guess the definition for language experience may be related to the length of time a speaker being exposed to his native language; But in the later part, the authors also paid much attention to cross-linguistic differences and interference effects. In this sense, I may feel puzzled if the language experience refers to a speaker’s language background concerning the inclusion of a monolingual control group and a language background questionnaire. Only when I read until the end of the discussion part can I know the authors may mean an L2 speaker’s exposure to the second language. I suggest that the definition should be given in a straight-forward way, and use it consistently in the manuscript.**

Our response: we agree the term was vaguely defined. Since our initial intention was to match both L2 groups in terms of L2 proficiency, we will only keep that variable to ensure comparability of both groups.

**2. The IN and NIN L2 groups have not matched on measures indexing language experience. Thus, it seems implausible to state the research purpose without strict control for the participants, instead in a post-hoc manner.**

Our response: we matched learner groups in terms of L2 proficiency and we omitted comparaisons referring to language experience.

**3. Based on previous studies, authors distinguished English-Spanish bilinguals with and without interpreting experience. This is a crucial manipulation for the present study. However, how do you know the previous findings were reliable and can be applied in the present study? I suggest the authors use a measure to examine the anticipatory ability between IN and NIN groups.**

Our response: We are a bit confused about what “previous findings” in this comment exactly refers to. Belonging to either bilingual group (IN, NIN) was established using the background questionnaire. Interpreters were found in official professional registers or through other interpreters. No previous study (to our knowledge) has investigated predictive language processing with interpreters (although we are aware that other research groups are currently working on it). Hence, there is no previous literature on predictive processing with interpreters.

The eye-tracking experiment is our measure of anticipatory ability in this study (we are measuring linguistic anticipatory ability). Possible connections between different types of anticipation (linguistic and non-linguistic) is an interesting question that we are working on for our new project. However, it is beyond the scope of the manuscript we’ve submitted.

**4. The number of participants were 25 monolinguals and 26 non-interpreters, but only 12 interpreters. The number of participants in IN condition were too small. I suggest the authors recruit more participants to complete the experiment.**

Our response: we have added data from 10 more participants (number suggested by the editor).

**5. The authors mix-use three terms for one concept, namely anticipatory skills, anticipatory experience and anticipatory exposure. I suggest the authors should be more consistent in expression in case of any misunderstanding.**

Our response: there was indeed some mixing, thank you for noticing. We have reduced it to anticipatory experience. “Anticipatory experience” refers to the type of practice the speaker has making predictions (during language processing or during interpreting). We have adjusted the use of the term throughout the manuscript according to the aforementioned definition.

**6. In the part of anticipation in interpreters, my suggestion is that the authors should be more focused on interpreters’ superiority in anticipatory skills in relation to non-interpreter L2 learners, where I consider that the anticipatory skills should be developed from explicit long-term training. Recall a study mentioned in page 6, their study suggests that short-term tone-suffix game training also improved speakers’ anticipatory performance. I think such short-term and long-term training is of interest in studies on anticipation. The authors may emphasize your inclusion of IN, meanwhile making a comparison between two forms of training.**

Our response: We have modified the last paragraph of that section as follows: “Our study stakes out new territory by investigating whether interpreters’ vast anticipatory experience, developed over a prolonged period of time, extends to non-interpreting situations. Our goal is to isolate the role of anticipatory experience from proficiency, to inform instructional practices. As previously mentioned, short-term training on the association between prosodic cues and morphology strengthens prediction (Schremm et al., 2017). The present study makes a contribution to the prediction models by investigating how experience with interpreting could act as long-term training.”

**7. Sentences like “CVC syllables in English and Spanish can be considered marked, at least with regard to CV syllables”. “Marked” is a vague expression, or do the authors mean that CVC syllables are more easily to be activated in word recognition due to its limited number?**

Our response: This sentence refers to markedness from a phonological theoretical standpoint. Under this framework, markedness generally implies that a segment or feature stands out/differs from some norm (unmarked) and is interpreted as different by the phonological module. Specifically, it is assumed to be a universal of syllable typology that all languages prefer syllables that start with a consonant and end with a vowel. Due to space limitations we do not provide a detailed account of markedness theory in the manuscript, though we attempt to clarify what it means in the context of our study: that some configurations may stand out more to listeners due to structural properties (i.e., CV vs. CVC).

In the revised manuscript we have included a footnote with relevant references regarding markedness in phonology and the role of syllable structure in Spanish.

* Hayes, B., and Steriade, D. (2004). "Introduction: The phonetic basis of phonological markedness," in Hayes, Kirchner, and Steriade (eds.), *Phonetically-Based Phonology*. Cambridge: Cambridge University Press, pp. 1-32.
* de Lacy, P. (2006). *Markedness: Reduction and Preservation in Phonology*. Cambridge Studies in Linguistics 112. Cambridge University Press.
* Colina, S. (2009). *Spanish phonology: A syllabic perspective*. Georgetown University Press.

**8. The authors also raise an important question whether late L2 learners integrate suprasegmental and segmental information at different stages of acquisition. However, they did not discuss this point further.**

Our response: although this is an interesting point, investigating this issue would require adding different L2 proficiency groups. In general, it is recommended to start teaching suprasegmentals to L2 learners. However, some production studies suggest that suprasegmental features may develop later with more L2 experience. Given that both L2 groups (interpreters and non-interpreters) have comparable L2 proficiency, we do not believe that we can make any claims related to stages of acquisition. See related bibliography:

- Prator, C. H. (1971). "Phonetics vs. phonemics in the ESL classroom: when is allophonic accuracy important?" In: Tesol Quarterly, pp. 61-72.

- Saito, K. (in press). Advanced segmental and suprasegmental acquisition. In P. Malovrh & A. Benati (Eds.). The handbook of advanced proficiency in second language acquisition. Wiley Blackwell.

9. **The authors expect monolinguals to anticipate earlier than the learners, considering that disambiguation can depend on lexical stress in Spanish but on reduction in English. I find it confusing because such causal relationship is not well-established. Both Spanish monolinguals or English-Spanish learners did the same Spanish task (steady vowels), thus the authors may mean that learners’ native language interfere with perceiving L2 lexical access.**

Our response: We have included the following explanation in the discussion “Our findings suggest that learners’ native language may have interfered with their L2 perception of lexical stress. However, the lack of a language pair with similar stress and syllabic structure in L1 and L2 prevents us from making strong assertions about this issue. To address this limitation, we plan to collect data with Mandarin Chinese learners of Spanish (Mandarin Chinese and Spanish are both assumed to be syllable-timed languages, but English is stress-timed), keeping syllabic structure constant.”

**But L2 learners could still utilize segmental information for disambiguation. Do the authors mean suprasegmental takes effect prior to segmental?**

Our response: We do not mean that suprasegmental takes effect prior to segmental, although we are confused about what this could mean. Specifically, we are referring to the fact that the participants have the same information (suprasegmental and segmental), but the suprasegmental information is used differently depending on the L1. Spanish speakers are accustomed to using lexical stress contrastively, whereas this is less clear for English speakers because lexical stress provides less information.

**Please elaborate on the cognitive significance of faster anticipation.**

Our response: this sentence was added to the manuscript:

“Faster anticipation is important because it facilitates recognition and interpretation of information by limiting the repertoire of potential candidates, saves resources to allow the listener to prepare for upcoming information, and guides top-down deployment of attention by improving information seeking and decision making (Bubic, Cramon, & Schubotz, 2010).”

**10. In the first point of discussion on monolinguals, the authors reach two conclusions 1) monolinguals integrate suprasegmental and segmental information to predict morphological information within a word; 2) natives use suprasegmental information to anticipate syntactic information. I have questions over their second conclusion, since they introduced a new syntactic level but lacked clear descriptions.**

Our response: we did not claim that our data proved syntactic prediction. The sentence in the manuscript was:

"These findings suggest that humans integrate suprasegmental and segmental information during lexical access to predict morphological information within a word, supporting the notion that structural integration and lexical recognition go hand in hand. In addition, our results are in line with studies showing that natives use suprasegmental information to anticipate syntactic information."

Therefore, the link between suprasegmental information and syntactic anticipation referred to other studies different than ours.

**11. In their second point of discussion on L2 groups and syllabic structures, they stated “That advanced English learners of Spanish predict inflectional morphology based on supersegmental and segmental ……”. How does their associate syllabic structures with L2 inflectional morphology. And they claim that L2 learners can integrate suprasegmental and segmental cues used differently in their L1 for anticipation purposes. The same question also applies here, and the authors do recognize that lexical stress are used differently in English and Spanish but don’t recognize syllabic structures are used differently in the two languages. So I think this paragraph is puzzled and needs to be reorganized.**

Our response: we are confused about the wording in this comment and we are not sure about what the reviewer is asking. In the section *lexical stress and syllabic structure in Spanish and English*, we explain that the two languages differ in the use of lexical stress (syllable-timed in Spanish vs. stress-timed in English) but not syllabic structure (CV is the default structure for Spanish and English).

**12. In the final part, the authors acknowledge that their design cannot explain why IN are better than NIN in anticipation, since they mixed anticipatory experience with language experience. To bolster their belief in anticipatory experience, they add a contrast between IN and monolinguals who differed in exposure to a target language. But the two groups also faced the same problem since they also differed in language proficiency, language background etc.**

Our response: we have avoided references to differences in language experience, and we focused on the comparability between learner groups in terms of L2 proficiency. We have removed this comparaison.

**Reviewer: 3**

**Comments to the Author**

**Evaluation: I believe that this is an improved manuscript. There continues to be some issues that should be addressed. Major issues continue to be about the interpretation of the reported analyses.**

**1. On p. 9 (I’m using the BLC-generated pagination and not user-generated manuscript pages), last sentence of the first paragraph, the authors write: “However, the vowel duration study mixed vowel duration (suprasegmental) with syllabic structure (segmental), and to our knowledge there is no study investigating whether natives use segmental cues to make predictions.” The “However…” here seems to be misplaced. I would suggest an edit along the following lines: “In contrast, no study has investigated whether native speakers make use of segmental cues to make predictions, although Rehrigh (2017) mixes vowel duration (i.e., suprasegmental) with syllabic structure (i.e., segmental).” I am wondering if this statement is specific to only eye-tracking studies making use of the visual world paradigm or if a study like DeLong et al. (2012; a v an) would count. Vaughan-Evans et al. (2014, soft mutation in Welsh) also makes use of phonological predictive processing but both of these studies were tested with ERPs (hence the request for clarification above).**

Our response: we have changed the sentence as follows “and Anglophones use vowel duration to predict voice (active/passive) (Rehrigh, 2017) – but this study mixed suprasegmental (vowel duration) and segmental variables.”

We are not focusing only on eye-tracking studies. We have not included the articles suggested by the reviewer because we are focusing on research on prediction at the word level (as mentioned at the beginning of that paragraph: “native speakers seem to use suprasegmental cues to predict morphology within a word”), as opposed to at the sentence level. Moreover, DeLong et al. 2012 uses the article as a marker for prediction of the following word, which mixes grammar rules (*a* vs. *an*) with segmental changes. As far as we can tell, Vaughn-Evans et al. (2014) does not measure prediction but rather cross-linguistic syntactic transfer (they measured facilitation at target onset as opposed to before it was heard/read).

**2. On p. 15, first full paragraph, the authors lay out their predictions concerning stress, first claiming that interpreters will show better anticipation with paroxytone stress, then citing this prediction as being borne from Kuperberg and Jaeger (2016) and Chernov (2004), and end by stating that lack of sufficient acoustic information may lead interpreters to wait for more information. The link to the prior studies is not clear to me. I’d like to suggest that the authors invert the second and third parts of their paragraph structure, starting with the prediction, then the interpretation, and ending with the link to prior work, which is clearer with Kuperberg and Jaeger (2016). I don’t understand the link to Chernov’s redundancy claims (or I’m just missing a tighter link).**

Our response: we have reformulated this entire section and we have eliminated references to those particular studies.

**3. It is good that the authors provided more context to justify their 10 ms time bin decision for the first set of analyses, outlined in the author response letter. I believe that this justification needs to be included in the manuscript as well as the citation. I am not convinced that selecting a 50 ms time bin would be a confound with hearing the remaining syllable as the authors claim. To me, this is still a point that needs to be discussed in the Discussion. The authors are staking the rather strong claim that the NIN group is not capable of predicting, based on an arbitrary time window selection. Without further analyses or even further word length details, this claim is too strong in my opinion.**

Our response: We thank the reviewer for this comment. As noted above, we have reconsidered our analysis and taken the advice provided. The revised manuscript reflects these changes.

**4. On p. 21, line 20 the authors write: “In sum, at the offset of the target syllable, M fixated on target words more than NIN regardless of stress type, but there was no difference between M and IN.” The reported pairwise comparison also indicates that there was no difference between IN and NIN, is that correct? This in-between state for the IN group—at least based on the results of the GLMM—should also be discussed in the discussion.**

Our response: In the revised manuscript the pairwise comparisons refer specifically to differences in target fixations over the time course (rather than to a specific time point, as was the case in the GLMM in the previous analysis). As suggested, the resulting differences between interpreters and non-interpreters are described in the results section and again in the discussion.

**5. In the discussion on p. 27, last sentence of first full paragraph, the authors write: “On the contrary, if anticipation is what yields the interpreter advantage, M and IN would be similarly efficient anticipating or IN would be better than M. The lack of significant differences between M and IN in the t-tests, GLMMs, and GCAs would point to anticipation experience as the explanatory factor of the interpreter advantage, if it weren’t for M fixating on target words more often than IN.” Didn’t the GCA show a timecourse difference between IN and M? In any case, why would we expect experience and prediction to be linear? At some point there’s likely to be an asymptote where prediction does not get any faster regardless of experience. I’m not so convinced by this monolingual group comparison.**

Our response: The results of the revised analysis, which includes new interpreter data, are now different. The discussion section reflects these changes.

**Miscellanea**

**a. On p. 11, line 28, change “pre-active” to “pre-activate”.**

Done, thanks.

**b. Fix case on last names in Soto-Faraco et al. (first appears on p. 13)**

Done, thanks.

**c. On p. 13, line 19, change “vocoide” to “vocoid”.**

Done, thanks.

**d. On p. 14, in the sentence that begins “These studies suggest that post-puberty learners…” change “…, and generates the need…” to “…and generate the need…”—assuming that “generate” agrees with “These studies”.**

Done, thanks.

**e. On p. 15, first sentence under Participants, let me suggest changing “…12 interpreter advanced English-Spanish learners…” to “…12 advanced English (L1) – Spanish (L2) interpreters…”**

Done, thanks.

**f. On p. 21 line 3, the random effects structure for the GLMM is described as “…by-subject and by-item random intercepts with random slopes for stress and syllable structure.” I believe that random slopes are only justified for subjects in this design (i.e., by design, a target item with CV syllable structure and penultimate stress cannot appear as CVC and ultimate stress). The final model would need to reflect the correct random effects structure (or please clarify, see Barr et al., 2013).**

This analysis has been taken out of the revised manuscript. In any case, we thank the reviewer for this comment. The description was poorly worded and suggested that the by-item random effects included random slopes when they did not.

**References not cited in manuscript**

**Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255-278.**

**Vaughan-Evans, A., Kuipers, J. R., Thierry, G., & Jones, M. W. (2014). Anomalous transfer of syntax between languages. Journal of Neuroscience, 34, 8333-8335.**