Dear Prof. Groner, dear Rudolf,

Many thanks to you and the two reviewers for your evaluation of our manuscript, “Reading English-Language Haiku: Processes of Meaning Construction Revealed by Eye Movements” (H. J. Müller, T. Geyer, F. Günther, J. Kacian, & S. Pierides”, and the very generous and insightful comments.
We have tried to act on (virtually) all of these – and doing so, we believe, helped us greatly improve the quality of the manuscript (regarding both the presentation of the results/statistical analyses as well as the points for further discussion).
Please see below our detailed responses to the individual points raised by the reviewers and for how we addressed these points in the revised manuscript. Please also note that, for ease of navigation, all rewritten and newly added sections are highlighted in blue in the revised manuscript (in particular, see the revised “Limitations and Outlook” section of the Discussion, on pp. 39-45, which has been substantially added to and completely restructured).

We very much hope that you’ll find the revised manuscript acceptable for publication in JEMR.

Best regards,

Hermann

(On behalf the authors)

Response to Reviewers

Reviewer A (A. Jacobs)

General comments and summary of recommendation
This is a pioneering interdisciplinary study using a method adequate for capturing some of the complexities of more natural literary reading and I congratulate the authors for this daring approach which hopefully finds many follow-up studies. This very original, mainly exploratory experiment offers rich and extensive eye movement analyses allowing to generate a number of interesting heuristic hypotheses for future interdisciplinary studies. Both theory and methods are well elaborated and the excellently-written paper is pretty close to a potentially high-impact publication: I have nothing to add to the innovative expert data analyses, but only a few more general recommendations the authors may wish to consider in their revision. If I recommend extension of the referenced literature (see list below) this is to make the present ms even more attractive to the rapidly growing Neurocognitive Poetics (NCP) community: a problem in this field was (and to some extent still is) that experimental reading/eye movement/neuroscience researchers and NCP scholars did/do not read the same journals/papers: I have been trying to change this for years and continue these efforts here and elsewhere (cf. Willems & Jacobs16TICS). Of course, it is up to the authors to follow this or not.

Response: We thank reviewer 1 – Arthur Jacobs – for his generous remarks and insightful recommendations. In what follows, we provide a (brief) response to the various points he raised, and we describe how we have addressed these points in the revised manuscript.

Point 1: "**NCPM.** The authors discuss the Neurocognitive Poetics Model, which at present only allows coarse hypotheses regarding eye movements. In their discussion section they do not come back to the NCPM but could perhaps add a few words on where the model is underspecified and how it could be shaped-up in this respect in the light of their results."

Response: To address this, we have added the following paragraph to the revised "Limitations and Outlook" section: "Implications for NCPM. The present findings have implications for the neuro-cognitive poetics model (NCPM; e.g., Jacobs, 2015). Using an eye-movement measure deriving from NCPM, we could show that 'reading fluency' – assessed in terms of the ratio of forward- to backward-directed oculomotor activity within lines – exhibits patterns characteristic of particular classes of haiku, including systematic changes in the speed and, thus, pacing in which particular lines are scanned for the first time and then (selectively) re-sampled – to construct and check global meaning. In this sense, the NCPM provided us with both a framework to explore the reading of haiku and 'tools' that permitted us to depict some (across haiku types and cut positions) relatively stable patterns of the reading dynamics, in terms of shifts between (predominantly) 'background (BG)' /fluent and (predominantly) 'foreground' (FG)'/disfluent processing modes – central concepts within the NCPM. Given this, our findings can be taken to reinforce some of the basic distinctions fundamental to the NCPM. Arguably, though, to formulate more precise, 'local' hypotheses as to oculomotor activity, one would have to 'zoom in' at the level of individual poems and poem lines, rather than more 'global' classes of poems and structural characteristics of such classes – which was the level at which the present study was pitched. This remains a task for future research.
As for the present findings: Given that ELH is still a relatively uncommon literary genre (despite its rapidly growing popularity), the responses to haiku type and structure that we observed in

'naïve' readers cannot be attributed to overlearned ways of approaching this kind of poetry. This marks our research as different from many other studies that build on conventional/highly familiar types of text. Taking this into account, it would appear non-trivial that our findings are interpretable in line with FG-BG theory, which is at the core of the NCPM: they may be taken to suggest that this reader behavior characterizes 'literary reading/processing' in general, that is, even if participants have no awareness of genre-specific formal features and/or of text-type-associated (pragmatic) knowledge structures acquired from previous (reading) experiences. (See also section Developing a 'sense' for haiku below)" (see pp. 39-40).

Point 2. "… the authors discuss the aesthetic trajectory but do not address a key issue of poetry reception, i.e. aesthetic liking (cf. Jacobs et al.16 FrontiersPsy). Even if they did not collect any liking ratings for the Haiku, they may discuss (in Intro or Disc/Limitations sections) if/how eye movement parameters could be meaningfully related to this important aspect or whether other (indirect) methods must be used (cf. debate about methods in SSOL: Dixon & Bortolussi15SSOL, Kuiken15SSOL, Jacobs15/16SSOL), e.g., pupillometry can reasonably be related to liking and memory processes under certain conditions (Kuchinke et al.09PACA, Vo et al.08Psychophys; but see Kuchinke et al.07IJP)."

Response: "Aesthetic appreciation – one of the key issues in poetry reception (including the NCPM; Jacobs, 2015) as well as haiku theory (see, e.g., Kacian 2015; Kendal, 2016 [Footnote 8]) – was not directly explored in the present study. Given our (secondary) focus on memory performance as an indicator of the depth of processing and construction of meaning ('understanding achieved'), we had decided to limit our study to these more 'cognitive' factors rather than including further subjective ratings of aesthetic appreciation and the feelings associated with it (which is, in itself, a complex issue; see, e.g., Lüdtke, Meyer-Sickendiek, & Jacobs, 2014, and Jacobs, Lüdtke, Aryani, Meyer-Sickendiek, & Conrad, 2016). In contrast to (direct) oculomotor measures, (indirect) measures such as changes in pupil diameter might be more readily related to aesthetic liking and memory processes, at least under certain conditions (e.g., Kuchinke, Trapp, Jacobs, & Leder, 2009; Võ, Jacobs, Kuchinke, Hofmann, Conrad, Schacht, & Hutzler, 2008; though see Kuchinke, Võ, Hofmann, & Jacobs, 2007). However, since we only collected (subjective) memory responses, but no ratings of aesthetic experience(s), we would be able to relate pupillometric measures only to the former, but not to the latter, and so be unable to discern influences of the two on changes of pupil diameter. Reasonably, however, one might assume that (our measure of) 'understanding, or closure, achieved' (rather than perceived 'haiku difficulty') might well correlate with aesthetic appreciation – implying that aesthetic experience correlates with, or contributes to, explicit ('recollective') recognition of the respective haiku. These are issues that await further, dedicated research [Footnote 9: We thank both reviewers for this valuable suggestion for future research.]. Note though that this research might profitably be guided by our findings of the differential 'pacing' with which haiku are read in the various passes. We predict that pupillometric changes would more be informative about both memory performance and aesthetic experience during the second or third reading passes, where, based on the present data, much of the work is done to construct global meaning." – We've added this paragraph, headlined "Aesthetic trajectory and aesthetic liking", to the Limitations and Outlook section of the Discussion (see p. 43).

Point 3. "**Haiku as paradigmatic study material.** The role of micropoetry not only as a paradigm but also for the development of the poetic sense has recently been discussed in several papers (Jacobs & Kinder15; Jacobs15SSOL) and could be elaborated a bit in this section.

However, reference to Hemingway may not be as ideal as Kliegl thought, since Hemingway's sentence achieves its effects mainly by what is not said (cf. Jacobs et al.16 FrontiersPsy): to what extent eye tracking is a method of choice for measuring what is not seen on a page may deserve some discussion (cf. also Jacobs15FNHUM for discussing the role of ellipse in poetry). I'd rather recommend the authors sell their haiku study as a blueprint for coming line- by-line poetry eye tracking experiments."

Response: Here, the reviewer raises a number of important questions on which we should have been clearer in the (original version of our) manuscript.

(i) First of all: "… by suggesting 'haiku as paradigmatic study material' (see Introduction), we do not wish to imply that the study of haiku reception should replace studies of longer literary/poetic texts – not least for reasons of ecological validity, as the genre of haiku occupies but a small niche within the realm of literary/poetic forms (though see our remarks on potential advantage of studying the reader response to such a relatively unfamiliar genre of poetry in section Implications for NCPM above). Arguably, however, the study of short poems, like haiku, allows for more systematic variation of experimental conditions (in our study: haiku types and cut positions) than would be possible with longer texts – and in this sense, it provides an interesting paradigm for exploration in future studies. With longer texts, of course, innovative quantitative narrative analysis (QNA) tools (e.g., Franzosi, 2010; Jacobs, Lüdtke, Aryani, Meyer-Sickendiek, & Conrad, 2016) provide an apt basis for relating 'processing' (reflected in eye-movement measures, pupillometric and peripheral-physiological measures, BOLD measures, etc.) to the variables measured by these tools. However, from an experimental point of view, systematic variation of certain variables and observing the effects of these variables can provide additional information beyond the 'relationships' revealed by alternative approaches. Ultimately, though, it would be ideal to combine both approaches – such as applying QNA tools to haiku – in future research (see also Willems & Jacobs, 2016)." – We've added this paragraph, headlined Haiku as paradigmatic study material" to the Limitations and Outlook section of the Discussion (see pp. 44-45).

(ii) Concerning the reference to Hemingway's "For sale: baby shoes, never worn", we believe it is not inappropriate. Haiku – especially those of the 'juxtaposition' type studied in the present study – exploit what is referred to as the 'gap' between the juxtaposed images, which is to be filled in by the reader. For instance, the reader might have to come up with a situation model that permits the (only) implied relationship between the two images to be worked out – that is, to create a coherent meaning Gestalt. According to haiku theorists, this process is driven (or 'semantically constrained') by the energy contained in the images themselves: "*We might consider the images to be the two poles of an electrical element, like a Tesla coil, and the relationship between them to be the spark which shoots the gap. The more powerful, clear and certain the choice of images, the brighter and surer the spark, ... And the stronger the spark, the more likely we will find secondary sparks as well, which in haiku we term resonance. Our goal in haiku is to find the correct images to serve as poles, and to allow the energy in the things themselves, the images and the language, to provide the spark inherent in them*" (Kacian, 2006, p. 56). In a sense, Hemingway's six-word 'short story' fulfills this description quite well (without being a haiku): what is not said – the gap between 'For sale: baby shoes' and 'never worn' – is what sparks the mind into filling in the gap.

(iii) Thus, "…haiku, by their very nature, are brief, highly condensed poems that necessarily

leave many things unsaid: "... the true subject of a haiku is never mentioned in the haiku. It is what a haiku implies that makes it a great or worthless haiku" (R. H. Blyth, quoted in Kacian, 2006, p. 39). Much of haiku's impact derives from the 'gap' between the juxtaposed images, which the reader has to fill in to achieve closure of the meaning Gestalt – a process that, according to haiku theorists, is driven by the energy contained in the images themselves [Footnote 7]. Like in vision, where the perceptual Gestalt formed includes elements not actually present in the distal stimulus, the unsaid is 'elaborated' in the meaning Gestalt constructed from the images presented. This raises an interesting question (for neuro-/cognitive poetics), namely: "to what extent is eye tracking a suitable method for gauging what is not seen, or absent, on a page?" (Jacobs, personal communication, December 2016). As for the present study, of course, our eye-movement analyses – examining oculomotor activity collapsed across haiku, or haiku classes, and readers – cannot tell how a specific reader filled in the gap in a specific poem, but we gleaned information about some of the major sampling strategies (of intial reading and selective re-rereading) that readers adopt for arriving at an ultimate solution. In this sense, our analyses go beyond the reading of the words that make up a poem. However, to paraphrase Jacobs (2015), this may not be far enough: "neurocognitive poetics research needs testable hypotheses about what those things 'absent' from a text elicit in a reader's mindbrain" (p. 6). Accordingly, appropriate analyses would need to be conducted at the level of individual readers, taking into account their "'apperceptive mass' (Kintsch, 1980), i.e., their knowledge (e.g., semantic and autobiographical memory), motivations, expectations, preferences" (Jacobs, 2015, p. 6); and at the level of individual poems, taking into account their larger meaning potential (i.e., multiple meanings). On the reader side, such analyses would conceivably involve direct poem-specific analyses of understanding achieved or (reproduction) memory of the meaning constructed (cf. Yaron, 2002). And modeling of how readers arrive at something new from initially (seemingly) incompatible information might profitably draw on established cognitive-linguistics approaches such as 'conceptual blending theory' (e.g., Fauconnier, 1997; Fauconnier & Turner, 1998, 2008, Turner, 2014). Arguably, though, haiku might provide an apt material for such analyses and modeling attempts in future work." – We've added this paragraph, headlined "Examining what is 'unsaid' in poetry", to the Limitations and Outlook section of the Discussion (see pp. 40-41).

Point 4. "While I agree with the authors on the importance of studies on micropoetry, I know from experience that the NCP community – in which the present ms hopefully will achieve a broad acceptance – may be more sceptical. Longer (poetic) texts are their main object of interest and, of course, eye tracking can very well be used for such materials. Although prose was used, a case in point is van den Hoven et al.16COLLABRA who tested some NCPM predictions using eye tracking and LMMs (the ms in press could be made available by the authors). While the potential drawbacks of the complexities of longer texts can be addressed by using a combination of quant. narrative analysis (QNA tools, Jacobs et al.16SSOL) and LMM (van den Hoven16), the potential benefits, i.e. higher ecological validity, weigh a lot (cf. Willems & Jacobs16TICS)."

Response: We do acknowledge the value of studying the reading (eye movements) of longer (poetic) texts, and the power that the use of innovative quantitative narrative analysis tools and linear mixed-model analyses can yield in this regard (see also Introduction). And we do not argue that that study of (structured) short forms of poetry, such as haiku, should replace the study of reading longer (poetic) texts (especially given that haiku is just one, relatively 'rare' form of poetry). Instead, we believe the former could usefully complement the latter, for the reasons discussed above (see also our response (i) to the reviewer's point 3 above).

Point 5. "Even though Haiku uses "ordinary language" without Rhyme, they have Rhythm (meter, stress). R&R is a key feature of poetry and even in silent reading it has measurable effects (Chen et al.16COG; Menninghaus et al.15PACA). Another aspect not to be completely left out is sound! Haiku also must affect reader responses through phonological iconicity and "mental" sound. Again, even in silent reading this has been shown to play a role (Aryani et al.16PACA). Perhaps the authors would like to comment a bit on the possible roles of rhythm and sound for Haiku reception and to what extent eye movement parameters could (or not) be meaningfully related to these important aspects. In general, the NCP community would benefit from an expert and honest discussion of the strengths and weaknesses of eye tracking for poetry reception studies along the lines the authors already try on lines 15ff p. 39, but this could be usefully extended."

Response: This is, of course, entirely correct! "While haiku tend not to use rhyme (because "rhyme remains such a compelling device that its presence in this fragile form is often overpowering"; Kacian, 2006, p. 84), elements of musicality, specifically stress and rhythm as well as sound and timbre, are carefully crafted by the poet to 'energize' the images in a haiku and create the 'spark' between them. To quote Kacian (2016): "... often we are attempting to give voice to the wordless, and it is only through mastery of the musical [i.e., rhythm and timbre] elements of a poem that we can approximate the effect of the experience upon us" (p. 89). In ELH, "it is unusual to have fewer than one or more than three stresses per line", with "stresses [occupying] the center of attention in each line, and … the unstressed syllables [serving] to bridge the time between these stressed moments, creating a rhythm specific to the poem" (p. 87). "… in such a brief [form], what matters … is that the rhythm be suggestive of the experience, that it contain the energy of the moment and attract the reader to it" (p. 87). As to tonal quality, "some syllables are susurrant, some percussive, some nasal. [Their] combination … across the duration of the haiku account[s] for its timbre. ... In each case, we are choosing words not just for meaning, but for tonal quality" (p. 87). This combination of rhythm and timbre elements is what makes the sound of a haiku. Given the complexities involved, again, future, dedicated research would be required to examine the effect of these musical elements (including aspects of 'phonological iconicity' and 'mental sound'), on haiku reception and aesthetic liking (for effects of these elements even in silent poetry reading, see Aryani, Kraxenberger, Ullrich, Jacobs, & Conrad, 2016; Chen, Zhang, Xu, Scheepers, Yang, & Tanenhaus, 2016; Menninghaus, Bohrn, Altmann, Lubrich, & Jacobs, 2014)." – We've added this paragraph, headlined "'Musicality' in haiku and aesthetic liking", to the Limitations and Outlook section of the Discussion (see p. 43-44).

Minor point: l. 24, p. 7: the reference should be: Hsu, C. T., Jacobs, A. M., Citron, F., and Conrad, M. (2015). The emotion potential of words and passages in reading Harry Potter – An fMRI study. *Brain Lang*. 142, 96–114. doi: 10.1016/j.bandl.2015.01.011

Response: replaced!

Reviewer 2

Müller and colleagues study an eye movement pattern generated during reading English language haiku (ELH) and its relationship with different types of ELH and behavioural measures. It is a very interesting attempt by interdisciplinary-minded researchers, which challenges unravelling a mechanism underlying the meaning reconstruction using poetic material (haiku). The study is well-designed (e.g., controlling stimulus features visually, rhetorically, and lexically) in general. However, some points (mostly about the analysis procedure) were not very compelling for me, which I hope could be revised. I discuss the comments and questions below, in the order they arise in the text. There are two issues that I indicated with asterisks. These issues would be critical to the interpretation and conclusions.

Response: We thank the reviewer for his/her detailed assessment of our manuscript and valuable recommendations. Please see below for how we have addressed these points in the revised manuscript.

[Procedures]

(1). P.17 L.7

Please briefly recap the classification criteria ('context-action' vs. 'juxtaposition' types) proposed by Kacian (2006), since it is critical when the results are interpreted.

Response: Done! Please see p. 17.

(2) P.19 L.8

Would be good if you can motivate your choice for the two scales ('yes'/ 'no' binary choice and following 1-4 confidence-level choice) in the memory test; it could be instead, for example, simply more continuous like scale [e.g., -3 (definitely not seen) to +3 (definitely seen)] with which you can study the parametric modulation of oculomotor parameters.

Response: Having participants make a binary, 'yes'/'no' recognition choice and then, in case of a positive response, following this up by a question about the explicitness of recollection ('remember'/'know' response) is not unusual (if not a standard procedure) in 'recognition memory' research (see, e.g., Gardiner, Ramponi, & Richardson-Klavehn, 1998, 2002). Also, given that some have argued that the 'remember'/'know' distinction is just an expression of the confidence in the recognition response (rather than being indicative of a qualitative difference between the underlying processes; see, e.g., Dunn, 2004), we introduced a (confidence-) rating procedure in part two of the recognition test.

The reviewer's query made us aware that we had actually not properly explained the rating required. To clarify this, we added the following sentence (p. 16): "– the specific question being: "How certain are you that you have seen this haiku earlier on? (1 = 'I definitely recollect having seen the haiku', and 2–4 = 'I feel I have seen the haiku'", with various (degrees of) strengths associated with this 'feeling of familiarity')."

Also note that the current study was meant to be 'exploratory', our main interest being to collect as many as possible, qualitatively different data (reading behavior, recognition performance, debriefing ratings). Of course, we were aware that this might come at the expense of not obtaining fully conclusive data in some respects, which would make it necessary to conduct further, more quantitative investigations (as suggested by the reviewer). Concerning the current data, we found that (i) almost all haiku were correctly recognized in the yes-no part of the test (accuracy: 86%) and (ii) most of the (correctly recognized) haiku were associated with re-collective experience (70%) – so we ended up with insufficient data for systematically looking at the gradations of the confidence-rating responses.

As we discuss on page 29, this extraordinarily good memory performance is likely owing to the fact that our current recognition test was too easy. Given this, follow-on investigations would need to introduce a set of better controlled 'foil' poems (e.g., purpose-created variations of poems presented for reading), which would make the old-new discrimination more difficult. In this way, a parametric assessment of haiku memory memory might indeed offer new insights into the relationship between memory and oculomotor scanning behavior. (Alternatively, one might also consider using 'recall tests'; see, e.g., Yaron, 2002, and our discussion of this on p. 37).

In any case, we thank the reviewer for his/her suggestions of using a more continuous scale, which we may be able to follow up in future investigations.

[Analysis and Results]

(3) P.20 L.6

The name of the test used in the 'frequentist' statistics must be given clearly, although it is supposed to be rmANOVA in the most cases. The name of the test at [P.21 L.24] was clearly given as ANOVA, but those for all the others are missing. Additionally, most of the results from ANOVA were reported about the interactions only – in that case, the main-effects were not significant? If so, it should be stated clearly (e.g., 'Only statistically significant results were reported.').

Response: We clarified these issues in the revision. First, we added information about the specific tests used in the individual analyses. Second, on p. 23 (Footnote 3), we now explicitly state that we limit the result presentation to the 'highest' effects possible, that is, we do not report main effects and lower-order interactions in case they were qualified by a higher-order interaction.

(4) P.20 L.31
(4a) Please give a definition of 'fixations per line' – clarify the parameters used when you detect it. Did you use an average x-y coordinate of the fixation period? How did you treat the fixations, for example, spotted at a few pixels away from the letter in the line? How about the ones at the middle of two lines?

Response: We added the following information about the definition of interest areas (lines) as well as fixations to the revised manuscript (see p. 20): "The eye-movement record was stored and later on analyzed off-line with purpose-written C++ software. For this, we defined three different rectangular 'regions-of-analysis' (ROA) areas (size: 10.63° x 1.66°) positioned on top of the

three poem lines, with identical, display-centered coordinates for each observer: ROA 1 was positioned at x-y coordinates 12.35°-10.21°, ROA 2 at 12.35°-11.90°, and ROA 3 at 12.35°-13.59°, with a vertical separation of 0.03° between adjacent ROAs. Only fixations that fell in any of the three ROAs were considered for further analysis. This led to the loss of some 4.5% of all fixations. Saccades were separated from fixations based on standard velocity and acceleration criteria (SR Research default settings: velocity exceeding 35°/sec and acceleration exceeding 9500°/sec^2). The x-y coordinates of a given fixation were determined by averaging the x-y coordinates across all 4-ms (i.e., 250-Hz sampling frequency) time bins during the duration of a given fixation."

(4b) And this is just a recommendation: The definitions of the 'first', 'second', and 'third pass' are now given at [P.23 L.2], and those of 'progressive' and 'regressive' saccade appear at [P.22 L.25], but I would recommend to bring them in Method (maybe Analysis?) section, as those definitions should be considered during the analysis procedures (when you actually detect the 'pass' or classify the 'saccade').

Response: Many thanks for this recommendation. We acted on it, moving these 'definitions' from the Results section up to the Method/Analysis section (see pp. 20-21). However, we did not go for a 'clean' separation of definitions from analyses/results (i.e., the definitions were briefly reiterated, without technical details, in the Results section), as we felt that such a separation would have been too disruptive to the flow (and therefore understandability) of the results presentation.

(5) ** P.21 L.24
(5a) I cannot see why the argument 'dwell time per word is longer in the fragment line compared to the other lines' could be supported solely by showing the significant interaction of 2 x 2 x 3 [haiku type x cut type x line] ANOVA. Why not try 2 x 2 x 2 [haiku type x cut type x line type (frag / phrase)] ANOVA, and check if the main-effect of line type and any other interactions were significant? The same thing can be said for the other ANOVAs - why not use 2 x 2 x 2 rather than 2 x 2 x 3 style? In most comparisons you are not interested in the difference between 1st and 2nd phrase lines, thus the 'line' condition (now 3 levels: line1, 2, 3) could be sorted into 2 levels (fragment, phrase).

Response: We thank the reviewer for this suggestion. While a 2 (line: combined phrase vs. fragment lines) x 2 (poem type) x 2 (cut position) analysis may indeed simplify the presentation of the results, the fact that there were some, theoretically likely important differences in the (1st-, 2nd-, and 3rd-pass) reading of the two phrase lines (dependent on poem type and cut position) made us decide to continue with the 3 x 2 x 2 analytical design throughout (rather than opting for a nested design, e.g., comparing the combined phrase lines to the fragment line in step 1 and then looking for differences between the phrase lines in step 2) – we state this aspect explicitly on p. 20, ll.5-7. Although this may at first appear somewhat more demanding for the reader, using this design cuts down on the number of ANOVAs and, thus, switching between what the 2 levels of the line factor refer to in each particular ANOVA.

Further, although again somewhat more demanding for the reader, the Fisher Least Square Difference (FLSD) values given in Tables 2–4 (for the three-way interaction) already allow for a comparison of the oculomotor measures across the three poem lines (without the need for 'nested analyses'). In essence, the FLSD scores provide a measure of whether a statistical effect (main

effect or interaction) was due to differences in the processing of the fragment line compared to the phrase lines, or whether the effect resulted from differences between the two phrase lines.

Thus, overall, although 'using' the FLSD score requires a little more work on the part of the reader, we belief that this way of carrying out and presenting the analyses 'unifies' scrutiny of the various result patterns and thus, ultimately, allows for a 'simplified' understanding of the re-/ reading dynamics (compared to a breakdown of each ANOVA in terms of main effects and lower- and higher-order interactions). – We hope that the reviewer finds this analysis approach acceptable.

(5b) And from Table 1: the difference between fragment and phrase line for L1-CA (668ms vs ~566ms = ~102ms) and L2-CA (704ms vs ~660ms = ~44ms) conditions failed to meet the Fisher's LSD (210ms), which seems against the argument above. Thus, I could interpret the results as follows; the significant difference of the dwell time between the fragment and phrase lines was shown for juxtaposition haiku, but not for context-action ones. As with this, for the other ANOVA tests, please consider the results of pairwise comparisons in the manuscript, since the LSD always appears in the Table.

Response: The reviewer is right! With a FLSD difference of 210 ms, the finding of increased processing times for the fragment line applies – in a strict, statistical sense – only to L.1- and L.2-cut juxtaposition haiku. For context-action haiku, by contrast, there is only numerical difference in the processing times between the fragment and phrase lines. However, in order to get the 'gist' of the data across, in the revision we still use relatively general, 'haiku-independent' terms when describing reading differences between the fragment and phrase lines. However, we have qualified (clarified) the critical statements, making the presentation in the text consistent with Table 1. Examples:

p. 23, l. ll. 6-7: "…is particularly marked for juxtaposition haiku…" -> "…(and, given the Fisher Least Square Difference provided in Table 1, statistically significant only) for juxtaposition haiku…"

p. 23, l. 14: "…with this pattern being more marked in…" -> "… with this pattern being more marked (i.e., statistically significant only) in…"

(6) ** More generally, please check again that all of the arguments are supported directly by the results of statistical tests. For example, in P.23 L.26, 'overall dwell time per word decreases with the number of reading passes', although this is obvious from the average values, no tests were performed actually. Dealing such (eye tracking) dataset, one should be aware that the data variances are expected to be large - the results must be always confirmed statistically.

Response: We apologize for the omission of this test, which we now incorporated in the (revised) manuscript (p. 25): "A 2 x 3 repeated-measures ANOVA comparing the dwell times between pro- and re-fixations across the three reading passes revealed the interaction to be significant (besides a significant main effect of pass): $F(2,20)=16.17$, $p<.01$, BF BF=5.3+e9, FLSD: 28 ms. Pro-fixations showed a marked decrease in dwell times from the first to the second pass and then remained stable (200 ms vs. 76 and 75 ms). Re-fixations, by contrast, exhibited a dwell-time decrease only from the second to the third pass (73 vs. 46 ms), while being statistically comparable between the first and the second pass (92 vs. 73 ms)."

(7) P.23 L.12

Statistics on the rank order variables - are the calculation of average values and following ANOVA applicable to the rank order variables? As my experience they are tested by nonparametric methods normally.

Response: We thank the reviewer for pointing this out. In the revised manuscript, we now report results from (nonparametric) Friedman (and Wilcoxon) tests. Since these tests (in essence, one-way ANOVAs for nonparametric data) allow assessment of the effects of only a single factor, we conducted separate tests for each haiku x cut position condition (see p. 24). The changed section now reads as follows (with changes marked in blue): "second-pass re-reading was equally likely to start with line 1 or line 2 (average ranks of 1.52 and 1.48, respectively), and line 3 was re-entered following lines 1 or 2 or both lines 1 and 2 (average rank of 2.46) – a pattern that was seen in all haiku type x cut position conditions; Friedman (one-way ANOVA) tests on the rank-order data confirmed this pattern to be significant for all four conditions (see Supplement Table S2 for the full data set). This priority for lines 1 and 2 relative to line 3 is also evident in the third pass (average ranks of 1.49, 1.48, and 2.02 for lines 1, 2, and 3 respectively), though it is less marked, owing to the greater variability in when the third pass occurred for a particular line (see above). Again, Friedman tests revealed this pattern to be significant for all (haiku type x cut position) conditions, except for L.2-cut context-action haiku for which the ranks were statistically indistinguishable among the three lines (1.81, 1.64, and 2.02 for lines 1, 2, and 3, respectively; Friedman $\chi^2 = 2.95$, $p=.23$). [In addition, for the third pass, line 2 had (some modest) priority over line 1 for all (haiku type x cut positions) conditions except for L.2-cut juxtaposition haiku, where re-reading was highly likely to start with line 1 (Friedman $\chi^2 = 13.91$, $p<.001$; ranks: 1.18 vs. 1.61 and 1.95 for lines 1, 2, and 3, respectively; the priority of line 1 over line 2 (and line 3) in L.2-cut juxtaposition haiku was substantiated by a Wilcoxon signed-rank test, which revealed a significant result: $p<.05$).Note that the (significant) Friedman χ^2 values (range: 8.95 to 16.54) were significant even when using relatively conservative alpha-levels (.01 or .001), to prevent an inflation of type-I errors."

(8) P.28 L.21

Please clarify the reason why difficulty and understanding rating scores were sorted into two levels ('easy' / 'difficult') from five (1 to 5)? Are there any technical reasons? It seems advantageous to keep the original five-levels as it offers more fine-grained statistical results.

Response: We have added a new footnote (Footnote 4, see p. 48) to explain the reasons: "Note 4. The reason for collapsing the 5-point rating scales into binary scales was that, averaged across the two ('difficulty' and 'understanding achieved') ratings, most haiku received ratings of "1" (42.2%), whereas ratings of "2", "3", "4", and "5" were relatively, and increasingly, rare (21.5%,14.3%, 13.1%, and 8.9%, respectively). The latter means that, for instance, there would have been only some 4–5 ('unrepresentative') poems that were assigned a rating of "5" – which would have rendered a more fine-grained analysis questionable. Thus, especially also with regard to linking the rating to eye-movement data, we decided to convert the 5-point scales into dichotomous scales".

[Discussion]

(9) Could you clarify the relationship between the 'aha' experience and 'emotional-related' processing of language (e.g., liking, surprising, interesting, romantic, dramatic…)? For me, it seems that 'aha' experience you discussed in the manuscript potentially be replaced by 'emotional' experience. Are there any functional differences between the two? What is the direct evidence for your argument that 'aha' experience in particular, not the (other) emotional experiences, was evoked when readers achieve understanding and enhanced the memory test performance?

Response: See response to point (10) below.

[Further comment]

(10) Maybe in the later (next) study: it would be interesting to study the pupillary dataset which you must recorded with the eye movements. As the visual features of stimuli were controlled very carefully, the contamination by perceptual factors (e.g., light reflex) would be minimised. Although the distortion caused by gaze potions should be corrected, the pupillary dilation would be still informative when you study cognitive and affective processing dynamics in haiku comprehension.

Response: These are two highly important points, which were also raised by reviewer 1 – see his point 1 above and our response to it, which we repeat here: "Aesthetic appreciation – one of the key issues in poetry reception (including the NCPM; Jacobs, 2015) as well as haiku theory (see, e.g., Kacian 2015; Kendal, 2016 [Footnote 8]) – was not directly explored in the present study. Given our (secondary) focus on memory performance as an indicator of the depth of processing and construction of meaning ('understanding achieved'), we had decided to limit our study to these more 'cognitive' factors rather than including further subjective ratings of aesthetic appreciation and the feelings associated with it (which is, in itself, a complex issue; see, e.g., Lüdtke, Meyer-Sickendiek, & Jacobs, 2014, and Jacobs, Lüdtke, Aryani, Meyer-Sickendiek, & Conrad, 2016). In contrast to (direct) oculomotor measures, (indirect) measures such as changes in pupil diameter might be more readily related to aesthetic liking and memory processes, at least under certain conditions (e.g., Kuchinke, Trapp, Jacobs, & Leder, 2009; Võ, Jacobs, Kuchinke, Hofmann, Conrad, Schacht, & Hutzler, 2008; though see Kuchinke, Võ, Hofmann, & Jacobs, 2007). However, since we only collected (subjective) memory responses, but no ratings of aesthetic experience(s), we would be able to relate pupillometric measures only to the former, but not to the latter, and so be unable to discern influences of the two on changes of pupil diameter. Reasonably, however, one might assume that (our measure of) 'understanding, or closure, achieved' (rather than perceived 'haiku difficulty') might well correlate with aesthetic appreciation – implying that aesthetic experience correlates with, or contributes to, explicit ('recollective') recognition of the respective haiku. These are issues that await further, dedicated research [Footnote 9: We thank both reviewers for this valuable suggestion for future research.]. Note though that this research might profitably be guided by our findings of the differential 'pacing' with which haiku are read in the various passes. We predict that pupillometric changes would more be informative about both memory performance and aesthetic experience during the second or third reading passes, where, based on the present data, much of the work is done to construct global meaning" (see paragraph "Aesthetic trajectory and aesthetic liking" on p. 43).