Reviewer 1

With some revisions, I think this will be a strong contribution to the conference proceedings.

Please omit the results section from the previously published original study. Readers can simply refer to the original publication for the results. Publishing them a second time risks overrepresentation in a meta-analysis (though the authors do make it clear the results have been published before). The list of differences between the present and the prior study is helpful.

* I'm not sure what to do about this. You’d have to be pretty dumb to double-include the first study in a meta-analysis... I could just point to the original paper, and then summarize what was and was not significant?? Thoughts?
  + E.g. “A Wilcoxon signed-rank test determined that participants answered significantly more questions correctly in the Plain condition (median = 5) than in the Markup condition (median = 4.5, p=0.04). Among the 77 participants that showed an asymmetry in their accuracy counts across conditions, 46 (60%) scored higher in the Plain condition.” 🡪 “Participants answered significantly more questions correctly in the Plain condition than in the Markup condition.”

MV: I think that this is a good solution.

✓

From my perspective, ratings of workload and preference are just that. They are not subjective measures of comprehension. Measuring them is certainly helpful; I would just refer to them as separate constructs (p.1).

* It looks like I messed this up. I should have referred to workload and preference as subjective measures of "performance", not "comprehension". Does this seem ok? Do they want workload and preference to be treated as separate constructs from each other, or together as separate from the objective measures?

MV: I think she means to refer to them as separate from the objective measures, but taken together, rather than separate from each other.

MV: Your solution, “separate measures of performance” may be criticized with arguments that preference can’t be assumed to be related to performance. **I would play it safe and just call them subjective measures, unless there is indication in literature that preference and performance are linked.**

* Could word as measuring output usability objectively/subjectively
  + I’m thinking I’ll include workload and preference as subjective performance and add a note about how preference is definitely not the same? Maybe? (This is probably the easiest thing to do, if not the clearest)
* Preference and performance certainly don’t always correlate well/positively – Andre & Wickens 1995 (typically when users don’t actually do the task or don’t learn their performance), does Sue have anything from Hololens she might like to cite?
* Sauro & Lewis (2009) – task-level satisfaction ~ completion rate = .51 (average of all aggregation levels, Table 15 shows correlation matrix (test satisfaction ~ completions = .35, test satisfaction ~ time = -.23, test satisfaction ~ errors = -.23)

A few more details in method and results would be helpful: How was the two-minute cutoff selected? Are all response times given in seconds? Is 6 correct answered questions a good score (in rereading, I'm having trouble knowing the total number of questions the score is based on)? Why bin the NASA-TLX scores? More variance would be included by using it quantitatively. The resulting loss of power may explain the lack of significant differences between plain and markup conditions.

* If we're cutting out some of the original study, we should have room to clarify (I think I did remove some reminders in the interest of space, and given that things are first introduced with an abbreviated description of the first experiment, I’m not shocked that there was confusion).
  + Does anyone know where the second confusion came from? As far as I can tell, response times are always given explicitly in minutes.
  + I actually need to think a bit about the TLX binning. It took a while to convince myself of the chi-square plan in the first place (a reviewer on the first experiment asked for it), but I’m having second thoughts.

MV: I don’t know where the second confusion came from. It is clear to me. However, I would still respond to her comment and, if possible, somehow add something about the times being in minutes.

With a non-significant Chi-square test, I would suggest not interpreting the results further (fourth page, right column), because the change in the median cannot be distinguished from sampling error. A lack of significance in a quantitative study does not make the finding qualitative. Please remove all conclusions suggesting differences in trust in automation across conditions. If the conclusions about strategies (following the trust section) came from observations or participant comments, then those are appropriate.

* This seems reasonable, buys more space.

MV: OK with me

The authors state, "In this paper, a preliminary look at recently collected data was presented, and there is much further analysis that can be done." If this paper is presented, I would like this analysis to be completed. This is because these data will already be published. To avoid double-publication, and to ensure CogSIMA readers get the full benefit of this research, please complete all analysis for the final presentation.

* We'll have to think about how to balance this with AHFE.

MV: Right, when do we have to submit final copy to CogSIMA?

Minor comment: Situation awareness does not need to be capitalized.

* I like to capitalize it before providing the abbreviation SA. Thoughts?

MV: It might be an Army thing to capitalize the expression. I have seen other lower case expressions which have capitalized acronyms, such as machine learning (ML) and electroencephalogram (EEG), among others. So, I agree with the reviewer.

* Sure, I’m noticing that I didn’t capitalize “information extraction” in the abstract, so at the very least I should be consistent.

-----

Reviewer 2

The paper describes the result of an experiment aimed on comparing the comprehension of text documents with and without markup from information extraction pipeline.

The paper is well-written and easy to follow. Research methodology fits well with the established standards in HCI area. The authors measure both subjective and objective effects of markup on human text processing and discuss the (somewhat surprising) results.

However, in my point of view, the paper needs more precise positioning and limiting the scope. The title seem a) very broad (it is about automation in general, which includes variety of operations beside the markup), b) a bit misleading (as the markup in the experiment wasn't actually produced by some automated algorithm). Besides, the difference between "plain" text and "processed" text probably depends significantly on type of the text. The authors experiment with a very specific kind of text documents (intelligence messages) which are very concise and are mostly "raw facts". For the "normal" text documents with other information density the effect of markup might be quite different. However, the focus on intelligence messages isn't stated explicitly neither in the title, nor in the abstract.

* I'm not sure what they want us to do here. I can add to the abstract that we used a simulated intelligence task.

MV: Good idea. Or, maybe we should change the title to incorporate the expression, “sense-making” as in sense-making for situational awareness or making sense of intelligence reports to support situational awareness with annotation/mark-up or keep it the same. It is broad but I have seen similarly broad titles elsewhere. Also, I’d be interested in the perspectives of others.

* I don’t think we can change the title, at least not through EDAS

Typos/language issues: - p.2 missing period after the last sentence of the Section II.

* Okay.

Overall, the paper discusses a quality experiment that may be of interest for the scientific community, so I'd suggest to accept it.

Reviewer 3

The author describes a paucity of results surrounding marked up text but the author must not have read up on the plethora of research on hypertext comprehension. The paper has some major limitations including mistaking conditions during explaining results (confusing the Markup and Plain conditions) and does not ever bring in the degree of markup as a variable. Too much markup would inevitably slow-down reading which was not discussed. Highlighting vs changing the color of text is a confound not addressed in the paper (fig 1 vs fig 3). The results of the current experiment show a clear lack of statistical power.

* Huge reviewer pet peeve here. If there's a plethora of research that we haven't found, how can you not scrounge up at least one reference?!? My understanding is that hypertext is a non-linear, semantic network text representation, which has some relevance for the markup in the first experiment (though the only linking we had was at the sentence level), but not much for the second. I'm fine throwing a reference in, but I'm not sure what the big deal is. Of course, if the reviewer had pointed directly to whatever lit they had in mind, it might be clear... which makes me worry that it doesn’t actually exists…
  + I skimmed [this overview](http://www-psych.nmsu.edu/~pfoltz/reprints/Ht-Cognition.html) – the hypertext and reading people seem to be looking at is not super relevant. The hypertext is at a much higher level (e.g., linking topics), and they’re looking at things like the coherence of the text (e.g., hypertext might be more useful with less coherent text), navigation (the more you have to jump around to follow the hypertext, the less helpful it may be), etc. Not completely irrelevant, probably worth citing, but doesn’t seem to touch much on what we’re doing.

MV: Agree! Maybe you could add a footnote saying that findings from the hypertext comprehension research community, while important and [[generally]] relevant [[but not to us]], are beyond the scope of the present work. Then you could a paraphrase of what you found above.

* They seem to be saying that we said Markup when we meant Plain or vice versa somewhere, but I can’t find it. Let me know if you see anything.

MV: Right. Can’t see it now but I’ll look again.

* It's possible that quantity of markup didn't make an appearance in this paper (it did in others, referencing cases where lower quanitify/higher quality forms of markup beat out other forms of markup), maybe I could add a sentence about the 2nd markup arguably being not only higher quality, but lower quantity than the 1st, and it’s possible this was an important factor in no longer seeing a plain preference.

MV: Good idea.

* I feel like we totally talk about the highlighting vs. font color change, and its implications are discussed in the discussion. Is there something else we can do to appease this person?

MV: You/We did talk about it. She missed it.

* I'm getting worked up again - The clear lack of statistical power obviously wasn't clear to us, so what do you think the issue is?!? If there's room, maybe I could add discussion of how we chose 200 participants (more than enough for a medium effect) and add something about the participant attrition hurting our power. But as far as I can tell I included effect sizes, so I don't know what the big huff is about.

To publish these results the authors require a more precise methodology, an IRB to conduct human studies, more statistical power, and a broader literature review.

* What do you think needs to be more precise in the methodology?

MV: Reviewer has neglected to indicate any weakness/imprecision in the methodology, which requires attention, so I think you are probably fine disregarding this reference in the comment.

* Do they want a statement that this is not human subjects research and was deemed exempt by our IRB??

MV: Maybe add a footnote, but I haven’t seen references to IRBs in conference papers, at least not linguistics or computational linguistics papers, and I have looked for it recently.

* Is the more statistical power request a weird way of asking for parametric statistics??

MV: There are people who dis prefer non-parametric measures. You can’t help it if much of the data was unusable for the experiment. What you learned probably was that you need for a turker to spend a minimal amount of time on task and or to fill a minimal number of slots, before you even consider logging their results, so that you have a higher number of usable data points. If you want to address the stat power issue, you could maybe add something to that effect.