**Does the paper make a novel contribution to the understanding of the topic under investigation?  [max 250 words]**

This paper makes a novel contribution to the empirical literature on the L2 acquisition of null subjects through (i) examining L1 German speakers from a large corpus, and (ii) examining a broader range of types of subjects than just null and overt in a wider range of contexts than are typically examined. The novelty of (i) lies principally in the use of corpus data, as (to the best of my knowledge) English and German are actually very similar with respect to the distribution of pronouns and full DP subjects, making the linguistic basis (interface or feature based) for the acquisition challenge similar for both groups; nonetheless, it is good to confirm this cross-linguistically. For (ii), the inclusion of full DPs does seem to be highly relevant to the data given the importance of information structure in the findings, although it is not clear how well this is integrated with the two theoretical approaches discussed. Contra many previous studies, especially those by Sorace and colleagues through which the Interface Hypothesis was formulated, this study suggests that contexts involving a topic shift are not uniformly challenging for L2 learners, whereas other contexts that do not involve an interface phenomenon are also challenging. This latter finding is not especially interesting in itself, as the hypothesis is simply that interface phenomena will be challenging even for advanced speakers, not that other aspects of the grammar will not be challenging.

**Is the empirical content of the paper sound (i.e. the data are collected and presented properly, the experiments are well designed, the statistics is well done, the examples contain no spelling mistakes, etc)? [max 400 words]**

The data reported in this study are extracted from an existing publicly available corpus (CEDEL2). The video description task employed in the study is appropriate, although the description of the task in Section 4.1 is a little under-informative. One question I have in connection to the written nature of the task is whether this means that the interpretation of the results with respect to the Interface Hypothesis is appropriate. In the 2011 article, Sorace seems to move to claiming that interface difficulties arise due to difficulty with online processing of interfaces, rather than the interfaces themselves. Consequently, the task design may reduce the likelihood of seeing residual difficulty in the use of null subjects, due to the reduced processing difficulty due to (presumably) less time pressure completing the task, which is not adequately addressed. Nonetheless, the size of the L1 and L2 datasets are appropriate, allowing the complex statistical analysis to be completed in a way that gives confidence in the conclusions the author draws. The statistical analysis technique is mostly well-described in Section 4.3, with two small exceptions. First, more information is necessary to clarify the validity of basing the L2 predictions on an L1 model, and especially if there are criteria around what models are good candidates for this in the MuPDARF procedure. I say this because the preparatory steps in Section 4.4.1 indicates that the L1 model has moderate accuracy of 64%, meaning that it incorrectly predicts 36% of the data. While I recognise higher accuracy is not always achievable, it seems odd to conclude that the L2 learners are noticeably

less accurate in their use of pronouns based on the models predictions, when we (i) know that the predictions do not fully align with the L1 data and (ii) do not know whether the lower predictive accuracy in pronoun contexts for the L2 speakers mirrors similar results in the model for the L1 speakers (as the observed/predicted percentages across subject types are not reported for this group). Second, the brief description of surrogate modelling does not make it clear how this approach achieves higher interpretability of the final predictor model. As a final note, example (4d) on p. 3 has a minor spelling mistake in the English translation, with German was translated 'was' rather than 'what', but otherwise the English and Spanish examples are fine (I cannot speak to the accuracy of German examples).

**Is the argument coherent and sound, with no major flaws and/or shortcomings, within the context of the theoretical assumptions made by the author? [max 500 words]**

Overall, the argumentation is sound, although there are some minor shortcomings in the coherency of the paper that need to be addressed. These shortcomings mostly relate to the presentation of the theoretical background. Although the article is positioned as an exploratory study in many places (especially Previous Research), in others (Abstract, Introduction, Interpretation and Conclusion) it seems to be much more closely concerned with testing the Interface Hypothesis. Overall, different sections of the paper give the impression of being an Interface Hypothesis study and an exploratory study, and this lack of coherence detracts from the quality of the evidence presented. In connection with this, the extensive discussion of an alternative feature reassembly-based conception of the learning task based on Slabakova and colleagues' work in the Interpretation and Conclusion Section, whilst excellent, comes out of the blue and does not have sufficient space to adequately convince a reader of the appropriateness of this analysis. For both hypotheses, one way to strengthen the coherence of the theoretical argumentation is to incorporate a new section at the end of Section 2 setting out what the learning task and expected residual difficulty is for null subjects under each approach. An additional shortcoming in the argumentation relates to the presentation and discussion of previous research in Section 3.2, which is seriously under-developed. Whilst the broad picture of the literature with respect to the methodologies and general findings is ably presented, some of the studies need to be described in substantially more detail to help the reader understand the state of the science on the topic. In particular, summaries of the Liceras and Díaz (1999) data and Schmitz et al. (2013) study, having presumably similar participants, would enrich the presentation of these studies, and allow for clearer discussion of the findings of the present study in Section 4.5 and following. In the same vein, fuller summaries of some of the studies with L1 English speakers would also add substantially to the argumentation. For example, Domínguez and Arche (2022) explicitly discuss the Slabakova-style conception of feature reassembly introduced in the Interpretation and Conclusion section of the study, and consequently a fuller discussion of their paper would strengthen your position in this respect. Finally, the substantial amount of new information presented in Section 5 detracts from developing a clear discussion of how the present study fits into the existing picture of the field to the extent that this is presented in Section 3. For example, the near-zero effect of proficiency is striking given previous studies; although Section 4.5 identifies that the predominance of advanced L2 speakers in the study sample may have caused this, it could also be related to the difference in task type and other factors between the present study and previous research that are currently left unexplored.

**Are there any relevant scholarly works that have been overlooked by the author? If the answer is YES, please provide the full references.**
No


**Have you seen this paper, its content, the proposed analysis, or the conclusions published in other venues? [If your answer is YES, please add the relevant reference.]**
No


**If you accept the paper with minor revisions, please list the revisions you would advice (you are not required to proofread the paper)    [max 500 words]**


In addition to the comments in other places in the review, I would ask the author to consider the following revisions. First, while the title is not inappropriate, it is arguably insufficiently specific. I would suggest incorporating both the specific group of L2ers (i.e., L1 German speakers) and broad method (i.e., corpus study) in a subtitle. At present, approximately half of the abstract rehearses the justification for the study, minimising the space for the originality and significance of the study, as well as the findings, to come through. I would suggest reconsidering the material in the abstract to make sure that the study itself, not just the justification, is well summarised. Although Section 2 is generally very clear, the justification for using the term 'referential continuity' given on p. 5 over 'topic shift' is a little unclear, especially since 'topic shift' continues to be used throughout the paper (primarily in connection to the Interface Hypothesis, which makes sense). It might be clearer to frame it as not using 'subject (dis)continuity' (this may be the intended reading). Overall, Section 3 is very short and does not do justice to either the Interface Hypothesis as a primary object of the study, or the previous empirical research. Section 4 is very well presented, with the minor caveats about Section 4.3 and 4.4 outlined above. Section 5 is simultaneously under-developed, in the sense of not interpreting the findings in light of the wider field as presented in Section 3, and over-reliant on presenting new theoretical ideas out of the blue. As suggested above, one recommendation would be to provide a more complete theoretical background of acquisition earlier in the paper, thus making the extended discussion of the two competing accounts here more clearly grounded in the preceding material. Finally, please note the following non-exhaustive list of minor typographical errors that I observed:

p.1 – 4th line of section I, one word reads 'difficulthard'.
p. 3 – line immediately before example (4) reads 'occurring with with foci'.
p. 3 – English translation (i.e., third line) of example (4d) has 'Was' for 'What'.
p. 6 – end of second line before example (9) reads 'in turn i if'.
p. 17 – the first set of nodes given in the discussion of the pronoun branches of the model are (18, 22, 25, 25) but should be (18, 22, 23).