Responsible NLP Checklist

Paper title: *Beyond Word Boundaries: A Hebrew Coreference Benchmark and an Evaluation Protocol for Morphologically Complex Text*
Authors: *Refael Shaked Greenfeld, Reut Tsarfaty*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☒ A2. Did you discuss any potential risks of your work?
No. We do not identify specific additional risks arising from this work beyond those generally associated with language datasets and NLP research. The paper introduces a Hebrew coreference dataset, annotation guidelines, and benchmark experiments, and does not involve deployment in a high-stakes setting.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☒ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
We use a pre-existing Hebrew newswire corpus and add a new annotation layer. Because the underlying text is preserved as-is for the linguistic task, the paper does not include a separate discussion of checking for or anonymizing personally identifying or offensive content.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
3.1 (Scope and Document Selection), Table 1. We report dataset statistics including the number of documents, sentences, tokens, mentions, and the train/dev/test split sizes.

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
es 4 (Experimental Setup), Appendix B, Appendix C, and Appendix D. Section 4 describes the evaluation scenarios, models, and metrics; Appendix B provides prompt templates; Appendix C gives the LLM inference pipeline and deterministic settings; and Appendix D reports the neural baseline training regime and key hyperparameters.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean,

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice. ACL 2026 used a subset of ARR checklist form.

etc. or just a single run?
5 (Results and Discussion), Tables 2 and 4. We make clear that neural results are averaged across 5 seeds and LLM results are averaged across 5 runs, and we report variability information in the table captions.

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
The paper summarizes the annotation scheme and Hebrew-specific decisions in 3.2 and provides an operational summary in Appendix E, but it does not reproduce the full annotator instructions inside the paper itself. Instead, the paper states that the complete annotation manual is available in the project repository. For this reason, the full text of instructions is not fully reported in the manuscript.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
The paper reports compensation in 3.3, stating that annotators were paid 50 NIS per hour, but it does not describe the recruitment procedure or provide further demographic/payment-adequacy discussion. Therefore, this information is only partially reported and does not satisfy the full checklist item. T

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
The paper does not discuss a consent procedure for individuals mentioned in the source data. The corpus is based on publicly available news articles, and our contribution is a new annotation layer over an existing text resource rather than new data collection from participants. However, the manuscript does not explicitly discuss consent, so the correct checklist answer is No.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The paper does not report ethics review board approval or exemption determination for the annotation protocol.

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☒ E1. If you used AI assistants, did you include information about their use?
AI assistants were used only for limited support such as language editing, brainstorming, or drafting assistance. All scientific decisions, annotation design, experiments, analysis, and final verification were performed by the authors.