

Epistemic Discourse Annotation for Arabic: Building a Gold-Standard Corpus from AL-Risāla

Eid Mohamed¹ and Tariq Yousef²

¹Department of English Literature and Linguistics, Qatar University

²Department of Mathematics and Computer Science, University of Southern Denmark

We present ongoing work on the construction of an annotated corpus for rhetorical and epistemological analysis of Arabic discourse, using the journal AL-RISĀLA as a case study. While Arabic NLP has made significant progress in tasks such as morphology and semantic processing (Yousef et al., 2023), higher-level discourse phenomena, particularly those grounded in Arabic rhetorical traditions, remain largely underexplored. Existing computational approaches often rely on frameworks developed for Western languages, which fail to capture the layered aesthetic, cultural, and interpretive dimensions of Arabic texts. This work is part of a broader effort to develop culturally grounded computational methods for Arabic Digital Humanities (Mohamed, 2024). Building on prior large-scale processing and topic modeling of Arabic periodicals (Mohamed and Mohamed, 2024)(Mohamed and Hassan, 2026), we move toward modeling rhetorical structures rooted in *balāgha*. Instead of adapting existing annotation schemes, we design a framework that directly reflects Arabic rhetorical theory and its role in meaning construction.

The annotation guidelines were defined by a senior domain expert in Arabic rhetoric and focus on a set of 16 rhetorical devices, including figurative language such as *jinās*, *saj*, and repetition; and structural and discourse-level phenomena such as *taqdīm wa-takhīr*, deletion, emphasis, contrast, and parallelism. These labels capture core mechanisms through which meaning, persuasion, and aesthetic effect are constructed in Arabic texts.

The full AL-RISĀLA corpus consists of more than 13,000 articles covering diverse genres and published across the period 1933–1953. The annotation corpus was selected to represent this diversity, both in terms of genre and historical coverage. The full article is provided to annotators in order to preserve context, while the annotated data can later be split into sentences for sequence-labeling experiments. The annotation process is implemented

using Doccano (Nakayama et al., 2018) in a controlled collaborative setup. The team consists of three trained annotators, each assigned 110 articles, resulting in a total of 330 annotated articles. A subset of 15 shared articles is annotated by all annotators and used to compute inter-annotator agreement and construct a gold-standard dataset, while the remaining 285 articles serve as training data for downstream modeling. Annotation is currently ongoing. To ensure consistency and reliability, all annotations are reviewed by two domain experts, and weekly meetings are organized to resolve disagreements and refine the guidelines. A central challenge in this work is the inherent interpretability of rhetorical phenomena. Many devices, such as metaphor, implication, or structural emphasis, depend on contextual and cultural knowledge, making them difficult to annotate consistently. Our approach addresses this through iterative guideline refinement and continuous feedback between annotators and experts. Preliminary observations suggest that disagreements are primarily driven by interpretation of implicit rhetorical meaning rather than by label ambiguity, highlighting the need for context-aware modeling. The project will move toward model-driven annotation and large-scale analysis. The annotated corpus will be used to train and fine-tune transformer models for automatic sequence labeling of Arabic rhetorical features. Further, we plan to integrate large language models into the annotation workflow in a human-in-the-loop setup, where model predictions assist annotators and are iteratively refined. This approach enables scalable annotation while maintaining quality and aligns with recent work on complex annotation tasks at the boundary of human agreement (Thalke et al., 2023). In parallel, the corpus will be expanded beyond AL-RISĀLA to include AL-MANĀR and other Nahḍah periodicals. This will enable comparative and diachronic analysis of rhetorical strategies and intellectual discourse across sources.

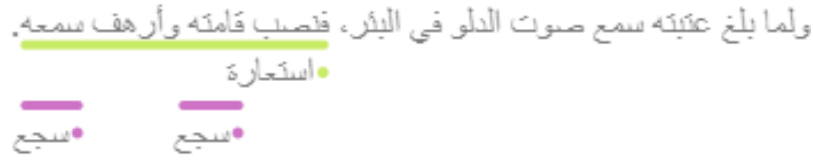


Figure 1: Example of sentence-level annotation in Doccano, illustrating the identification of a rhetorical device (istiāra, metaphor) within an Al-Risāla article, along with token-level span selection and label assignment.

References

- Eid Mohamed. 2024. [The potential and limits of arabic digital humanities](#). *Journal of Cultural Analytics*, 9.
- Eid Mohamed and Said Hassan. 2026. [Cultural analytics and the politics of representation: Mapping the jewish presence in egypt’s al-risālah \(1933–1953\)](#). *Digital Scholarship in the Humanities*, 41(1):302–319.
- Eid Mohamed and Talaat F Mohamed. 2024. [Rational imaginary and discursive formation of arabo-islamic identity in al-manār and al-risālah: A topic modeling study](#). *Journal of Cultural Analytics*, 9.
- Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. *Software available from <https://github.com/doccano/doccano>*, 34.
- Rosamond Thalken, Edward Stiglitz, David Mimno, and Matthew Wilkens. 2023. Modeling legal reasoning: Lm annotation at the edge of human agreement. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9252–9265.
- Tariq Yousef, Lisa Mischer, Hamid Reza Hakimi, and Maxim Romanov. 2023. [Enhancing state-of-the-art NLP models for classical Arabic](#). In *Proceedings of the Ancient Language Processing Workshop*, pages 160–169, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.