QuiNE-GT 1.0 Data Statement

Author(s): Yvette Oortwijn, Thijs Ossenkoppele, Arianna Betti
Version: 1.1
Date of creation: March 26th 2021
Date of last modification: March 30th 2021
Document origin:
https://docs.google.com/document/d/1xbVgLKQFVRsLBGz19678lQL1yPnFBSZH_eHW4Wv9d4Q/edit
Aim: To release a data statement cf. Bender & Friedman (2018) for the QuiNE-GT 1.0 dataset
Released together with: The HumEval2021 paper 'Interrater disagreement resolution; A systematic procedure to reach consensus in annotation tasks' (2021), written by the authors of this statement.
On: Forthcoming

A. Curation Rationale. In order to investigate Quine's thought with respect to the relation between epistemology and science, paragraphs relevant to this relation were retrieved from the Quine in Context corpus using semi-automatic string search. The queries consisted of combinations of search terms that were manually curated by three Quine experts. These same experts have carried out the subsequent scoring of the retrieved paragraphs.

Besides the Quine in Context corpus, QuiNE-GT 1.0 consists of:

- **a.** A conceptual model of the concept of naturalism and multiple research questions concerning the relation between science and epistemology.
- **b.** Lists of terms pertaining to the subconcepts of the model at **a**. The terms on these lists are ranked based on their relevance to the subconcept they are categorized under.
- **c.** Annotated paragraphs from the Quine in Context corpus in which combinations of the terms from **b** occur, all provided with a unique identifier and a score that indicates the degree of evidence they contain with respect to one of the research questions at **a**.

Due to copyright reasons, we are not able to release the Quine in Context corpus. For rough reproducibility, please consult our more detailed log of the construction of QuiNE-GT 1.0, available here.

B. Language Variety. Exact data on the various languages used in QuiNE-GT is not available, but the main language is US English (en-US). Incidentally, short fragments of Latin (la-Latn), Greek (grc-Grek), German (de-DE), and French (fr-FR) may occur.

C. Speaker Demographic. The only speaker in our dataset is Quine, whose native language was US English. The passages are from various stages of his career, thus our speaker's age ranges from ±25 to 92 years old. He likely identified as a white male, and presumably had an upper middle to high socioeconomic status.

D. Annotator Demographic. The three annotators ages range from ±25 to ±50 years old, and identify as either white non-binary, white female, or white male. Their socioeconomic statuses range from high lower class/low middle class to high middle class. All three are trained philosophers, ranging from graduate to professor level.

E. **Speech Situation.** Apart from a few informal book reviews and a poem, all texts in the Quine in Context corpus are meant for an academic audience, mainly philosophers, logicians, and mathematicians. All texts were written between 1933 and 2000.

F. **Text Characteristics.** Most of Quine's work is philosophical, and touches upon ontology, epistemology, philosophy of language, and philosophy of logic and mathematics. A smaller portion of his texts are within the domains of logic and mathematics or linguistics. Quine used a rather wide vocabulary, and uses terms that are typical of analytic philosophy and of academic writing more generally.

Although extensive work has been carried out to produce a high-quality corpus (Betti et al., 6693-6694), there is still a small portion of OCR errors present in the Quine in Context corpus. This means that there is an indeterminable chance that paragraphs that would qualify for annotation have been missed.

G. **Recording Quality:** N/A

H. **Other:** N/A

I. **Provenance Appendix:** N/A

J. **Ethical Approval.** Although no ethical approval for research involving human participants has been requested, all authors declare that the research was conducted in accordance with the ACM code of ethics[1], as well as with the institutional code of conduct of the ILLC[2], where the research has been carried out. Informed consent was obtained from all participants included in the research. The authors declare they have no conflict of interest.

References:

Bender, Emily M., and Batya Friedman. 2018. 'Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science'. *Transactions of the Association for Computational Linguistics* 6: 587–604. https://doi.org/10.1162/tacl_a_00041.

Betti, Arianna, Martin Reynaert, Thijs Ossenkoppele, Yvette Oortwijn, Andrew Salway, and Jelke Bloem. 2020. 'Expert Concept-Modeling Ground Truth Construction for Word Embeddings Evaluation in Concept-Focused Domains'. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6690–6702. Barcelona, Spain (Online): International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.586.

[1] https://www.acm.org/code-of-ethics

[2] https://www.illc.uva.nl/cms/uploaded_files/inlineitem/Academic-Practice-Code-of-Conduct.pdf