Replication Package

This package is intended to offer transparency into the design, analysis, and results for the paper:

Thinking Aloud about Confusing Code: A Qualitative Investigation of Program Comprehension

The information contained in this package falls broadly in to three categories, each of which have received their own directory: preparation, interview_instructions, data, and analysis.

preparation/

These files were all used in the preparation of the study. They helped define and select which snippets of code would be evaluated, as well as create the instruments used in each interview.

- **instrument-0000.pdf**: An example instrument that could have been issued to a subject.
- **questions_snippets.csv**: A CSV listing every snippet chosen for inclusion in this study. Includes headers:
  - atom: Which atom of confusion is contained or removed from the snippets.
  - type: Whether the snippet contains an atom (C), was hypothesized but not confirmed to contain an atom (HC), has an atom removed (NC), or was hypothesized but not confirmed to have an atom removed (HNC).
  - qid: The "question identifier", or the number used to represent snippet sample. Also called *snippet_id* in other files.
  - source: The C source code of the snippet.
- **question_orders.csv**: A CSV listing the orders of snippets shown to subjects during each interview, containing two rows:
  - subject_id: Which subject does this line pertain to
  - question_order: A space-separated list of snippet id's. These correspond to the *qid* column in the above *questions_snippets.csv* file.
- **questions_orders.rb**: A ruby script for semi-randomly generating instrument orders. The number of each type of snippet is dictated by variables in the script, and are defaulted to containing 5 atom-containing snippets, and 1 each of all other types of atoms. The output of this file is a csv, of the same structure as *questions_orders.csv*.
- **build_instruments/**: This directory contains scripts and templates necessary for creating instrument pdfs.
  - **instrument-*.mustache**: Mustache templates of Latex files used to create instruments. The template is required to substitute in subject-specific information such as their subject_id and the specific set of snippets assigned to them.

- **build_instrument.rb**: The ruby script that generates instrument pdfs from the templates described above. The Mustache ruby gem must be installed. In order to run, build_instrument.rb requires a subject_id and a list of snippet_id's (sometimes called qid's) to create the pdf. The easiest way to do this is to supply as an argument a *question_orders.csv* from the parent directory.
- **out/**: A blank directory that will be filled with pdf files after running build_instrument.rb.

interview_instructions/

These documents were provided to the study leader before every interview to serve as a refresher on how to conduct the study.

- **Preflight-Checklist.pdf**: A checklist of items to bring, and procedures to perform before each interview to condition the study leader to behave as uniformly as possible during the study.
- **Meta-protocol.pdf**: A specific set of instructions about how to behave during the interviews, with high-level recommendations as well as scripts for specific interactions that commonly arise.
- **universal_answer_key.csv**: The correct outputs for each program sample shown to subjects. It can be useful to review these during an interview to make sure the study leader doesn't accidentally miss an error from the subject.

data/

The raw output from the interviews, both transcripts of the audio and the written artifacts generated by subjects.

transcripts/

The audio from each interview was transcribed into .txt format. Personally identifying information was stripped. The text is annotated with comments in square brackets []'s with comments including when new snippets were shown to the subject, e.g. [Snippet 1] would indicate that the subject was just shown the snippet with ID 1. Other comments include things like [laughter], or [company] to indicate non-verbal interactions or personally identifying information, respectively.

instruments/

The raw instruments were originally generated with several pages that were removed from the scans included in this package. The pages were removed for privacy concerns or to limit redundancy, as listed below:

- Page 1: Introductory page describing the nature of the research and signature line for informed consent [removed for privacy concerns]

- Page 2: Directions page that explains how to complete the tasks in the study [removed due to redundancy]
- Pages 11-12: Demographic survey [removed for privacy concerns]

An unused example copy of the instrument, as well as the latex templates that created it are available in the *preparation* directory of this package.

Regarding the color of writing on each page, the pen given to subjects was changed during different phases of the study. In the beginning when subjects were working on their own without researcher interaction, they used a black pen. After they completed the first phase of the study and the study leader engaged them in a dialog, they were given a red pen. When viewing the written instruments then, black ink indicates thoughts formulated by the participant alone, while writing in red ink may be affected by interactions with the study leader.

analysis/

The codes (labels) assigned to the text are found in this directory.

- **code_descriptions.csv**: lists each of the 153 types of codes assigned and a brief description of when they are applicable.
- **codes.csv**: lists the 1808 applications of the codes, columns are:
  - subject: The ID of the subject, and consequently which transcript, the code is applied to.
  - start_offset: Where in the document the labeled text begins, counted in number of characters.
  - end_offset: Where in the document the labeled text ends, counted in number of characters.
  - code: The name of the label applied.
  - text: The segment of the interview to which the code is applied.