OMB Control No. – 0693-0043 – NIST Generic Clearance for Usability Data Collections

DARPA BOLT Phase 3 Speech-to-Speech Machine Translation Data Collection – System User Questionnaire

FOUR STANDARD SURVEY QUESTIONS

1. Explain who will be surveyed and why the group is appropriate to survey.

Background information:

NIST carries out testing of speech-to-speech machine translation systems developed by three different performers for the Defense Advanced Research Projects Agency's (DARPA's) Broad Operational Language Translation (BOLT) program. Data for assessment is collected by several pairs of users speaking different languages (one side English, the other Iraqi Arabic), separated by a soundproof glass barrier, carrying out short dialogs guided by a scenario background. For each dialog, one of the users is the designated initiator with a predefined goal; the other is the designated respondent. The machine translation system enables the communication between them by automatically translating from English to Iraqi and from Iraqi to English. All of the audio and text of each dialog, including both user and system input and output, is logged electronically by the system.

There are two separate component evaluations.

1. In the **Main Evaluation**, users communicate via each system using only speech input and audio output. The systems run on laptop computers, but the users do not see the screen or operate the keyboard. The systems in this evaluation operate under two different conditions:

    a. With system-solicited clarification option turned on: The system may ask the user for clarification of parts of the user's speech that the system is not confident about, before issuing a translation to the other side.

    b. With system-solicited clarification option turned off: The system issues a translation immediately after each utterance, without engaging in any clarification with the user first.

2. In the **Utility Evaluation**, the systems may employ other input and output modalities in addition to speech, by providing mobile devices (smart phone or tablet) to the users for interacting with the system. These modalities may include viewing and reading the mobile device's screen, using the mobile device's touch screen functionality, and typing on the mobile device's virtual keyboard. System-solicited clarification is allowed for all dialogs in this evaluation.

Each performer provides separate systems to be tested for the two Main Evaluation tests (1a and 1b above) and the Utility evaluation test (2 above), resulting in three systems per performer. The two systems for the Main Evaluation will be tested together in single testing sessions for each performer and user pair. The systems for the Utility Evaluation will be tested in separate single testing sessions for each performer and user pair after testing of the Main Evaluation systems has concluded.

Who will be surveyed and why this group is appropriate to survey:

Two groups of persons will participate in different roles for the BOLT speech-to-speech machine translation test, and will complete different questionnaires to assess the data collected as described above for achievement of dialog goals, success of interactions, translation adequacy, and usability aspects of the machine translation systems:

1. The users of the machine translation systems, as described above. These users are appropriate for the questionnaire because they are the group using the machine translation systems and thus have the first-hand user experience necessary to answer the questions.
2. The bilingual experts will not use the translation systems, but will review a log of each dialog in its entirety provided by the translation system. These experts are appropriate for the questionnaire because, due to their command of both English and Iraqi, they are able to review both sides of each interaction in their entirety and assess the log from the translation system accordingly.

This document addresses the questionnaire for Group 1 above; Group 2 is addressed in a separate clearance request.

2. Explain how the survey was developed including consultation with interested parties, pre-testing, and responses to suggestions for improvement.

The questionnaires were developed by NIST researchers with input from DARPA regarding the kind of information DARPA needs to gather on the performance and usability of the machine translation systems.

3. Explain how the survey will be conducted, how customers will be sampled if fewer than all customers will be surveyed, expected response rate, and actions your agency plans to take to improve the response rate.

A user questionnaire will consist of one to two questions after each completed dialog and one to two questions after each system session, depending on whether the system-solicited clarification option was off or on. The users will complete their questionnaires in the testing room, either using pen and paper or an electronic device with a dedicated Graphical User Interface (GUI).

Each user will complete a questionnaire during each of his/her six system sessions. Completing the questionnaire for one system session is expected to take a user 50 minutes. All system users are anticipated to complete their respective questionnaires; there will be no sampling. The expected response rate is 100%.

4. Describe how the results of the survey will be analyzed and used to generalize the results to the entire customer population.

The results will be tallied and then analyzed comprehensively in several ways:

- For the Main Evaluation, the three perfomers' systems will be compared against each other on assessed achievement of goals, success of interactions, translation adequacy, and usability.

- For the Main Evaluation, each perfomer's system with the system-solicited clarification option turned on will be compared to that same perfomer's system with the system-solicited clarification option turned off counterpart on assessed achievement of goals, success of interactions, translation adequacy, and usability.

- Each perfomer's systems from the Main Evaluation will be compared to that same performer's Utility Evaluation counterpart system on the assessed achievement of goals, success of interactions, translation adequacy, and usability.

As there will be no sampling, generalization to a larger population, to other languages, or to other systems not tested does not apply.