Interlingual Map Task Corpus Collection

Hayakawa Akira¹, Nick Campbell², Saturnino Luz¹

¹School of Computer Science and Statistics, Trinity College Dublin, Ireland
²School of Linguistics, Speech and Communication Sciences, Trinity College Dublin, Ireland

campbeak@scss.tcd.ie, nick@tcd.ie, luzs@cs.tcd.ie

Abstract
We present a prototype interlingual communication system that is being used to collect a corpus of task based dialogues between speakers of different languages. This corpus will be used to assess human reactions to an automated speech-to-speech translation system. In this demonstration we show how the HCRC Map Task can be adapted to support data collection in this interlingual environment, and how we used easily accessible speech and language technology for the rapid prototyping of the system used for data collection. An explanation of the nature and purpose of the data we are collecting is also presented.

Index Terms: repair strategy, speaker alignment, adaptation, cognitive load, language expectation

1. Introduction
The HCRC Map Task Corpus was initially designed to support the investigation of linguistic phenomena. However it has been the focus of a variety of studies of communicative behaviour [1, 2, 3]. The simplicity of the task, and the complexity of phenomena it can elicit, make the map task an ideal object of study.

Here we investigate a map task based on speech-to-speech machine translated interaction where the instruction giver and the instruction follower speak different languages. Although there are studies that used replications of the map task to look into communication in computer mediated tasks [2, 3], this Interlingual Map Task is, to the best of our knowledge, the first investigation of communicative behaviour in the presence of three additional “filters”: Automatic Speech Recognition (ASR), Machine Translation (MT) and Text To Speech (TTS) synthesis. We are interested in understanding how such filters affect the participants in terms of cognitive load, adaptation of communicative acts to the technology, repair strategies, play, and other factors that might have implications for the design of dialogue systems and, in particular, speech-to-speech translation systems. To this end we are collecting a variety of synchronised and finely time-stamped data streams, including: high quality audio of the participants’ utterances, video, eye tracking, physiological signals (EEG, BVP, SC), and ASR, MT, TTS events.

2. An Interlingual Map Task
The Map Task is an instructional task that elicits clarification dialogues. An instruction giver and a follower work on maps that were initially identical, but had some items deleted or modified to create subtle differences. The instruction giver, who has the route from start to finish, does not know what items are missing or modified in the follower’s map, or how the follower is understanding and interpreting each instruction, hence miscommunication is likely to occur. Once either of the subjects feels the need to question the current understanding, the subjects will try to reestablish a common understanding of the situation and continue to finish the task. However, since the reestablishment of the common understanding is only verbal, it is only presumptive.

Changes have been made from the original HCRC Map Task to create the Interlingual Map Task. We have translated all 32 maps into German, Japanese, Portuguese, and Spanish (with our immediate interest being English and Portuguese) and created a speech-to-speech translation prototype (ILMT-S2S prototype) to enable two speakers of different languages to communicate with each other (remotely, over the network) in their native languages.

With the modifications, we have added new obstacles. The ASR, MT, and TTS are “Black Box” applications, so performance can not be tuned to this task. The ASR may find the uncommon object names difficult to recognise and the MT may not translate them as indicated in the other language map causing misalignments that is difficult to correct [4]. The second is the adaptation to the speech-to-speech translation system where identifying the system’s level of competence should cause difficulties.

3. Technical Setup
Our corpus collection work is supported by the ILMT-S2S prototype (Fig. 1). Written in Python, it enables us to accurately log and timestamp all relevant events. It implements a “push to talk” interface that requires the subject to press and hold a button with the mouse while they speak, and release it once they finish the sentence or instruction. Feedback is provided to the user in the form of a microphone that lights while the other subject is speaking, and a pair of loudspeakers that light when TTS is being output from the other subject’s speakers. In addition, the experimenter’s can choose to show (or omit) the text from the ASR and the MT modules.

For each utterance (by either participant), the ILMT-S2S prototype executes the following processes: (1) records an audio file sampled at 96 kHz, in 24 bit format, (2) downsamples...
the audio file to 16 kHz, saves it in 8 bit FLAC format and passes it on to the Google ASR interface, (3) passes the result of ASR to the MT service (Microsoft Bing), (4) sends the MT results to the remote subject’s client, where (5) the TTS component converts the text to speech.

3.1. The data collection

3.1.1. XML record

The processes that the system executes are recorded in an XML file. One such file is generated per speaker. It contains accurate time stamps, and descriptions of user and system events, including: MT and TTS process activation times and end times, and the outputs (parsed and raw) of these demonstrations. This XML file also records the participant’s details, such as demographics, source and target languages, etc.

3.1.2. Audio

Three separate audio sources are recorded. The first is a Hi-Resolution Audio (96 kHz, 24 bit) recording of the utterances passed on to the ASR process. The microphone is connected by an XLR connector to a Edirol UA-25EX that feeds the digitised signal via a USB connection to the computer and saved. A Sennheiser lavalier microphone (MKE104P set) is used so the eye tracking and the biofeedback sensors equipment can be fitted with ease. An Audio-Technica headset microphone (HYP-190H) is used for the other subject to keep the microphone at a consistent distance from the mouth. The second is a standard CD quality audio recording using the camcorder’s internal microphone. This is used to record the utterances that where not spoken to the system, i.e. muttering, sighs, and other reactions uttered by the subject. For the third type see “Data alignment”.

3.1.3. Video

A frontal view of the subject is recorded in high-definition video from a Sony NEX-FS700 for the subject that is not using the eye tracking and biofeedback sensor equipment, and standard-definition from a Canon MX2 for the other subject. These camcorders record the whole experiment and the video is used to analyse the status of the subject during the experiment (annotate non verbal information that cannot be retrieved from the audio), and further used as data to test user attention recognition models.

3.1.4. Eye tracking

The SensoMotoric Instruments GmbH (SMI) Eye Tracking Glasses 1.1 (ETG) is used with the iViewETG software to record the subject’s gaze. This ETG has 2 cameras that face the subject’s eyes and records the pupil position, pupil diameter, eye position, gaze vector and status (e.g., blink, saccade, fixation) at 24 fps. From the output ASCII file, the levels of cognitive load can be assessed from the pupil diameter and eye status. A third camera records what the subject is seeing. By using the supplied BeGaze2 software, this video is edited to display the gaze location of the subject.

3.1.5. Biofeedback sensors

To record the bio signals, a Mind Media B.V., NeXus-4 is being used to collect the blood volume pulse, skin conductance and electrical brain activity. The correlations of these signals to affective and attentional [5] states in different communicative situations, and their effect on cognitive load are of interest.

3.2. Data alignment

This is an important aspect of managing this corpus and professional video editing software is used here. By including an audio recording to the biofeedback sensor and eye tracking recordings, we use the audio from the main video recording as the reference that the clip synchronisation function uses to align the other audio files. With the other audio files aligned to the main video recording, the time difference for the XML, eye tracking, bio sensor data time log can be determined. With all the files related to one subject during one experiment aligned, the annotation data can be used with various data sources (e.g., annotation of the audio file can be aligned with the bio sensor output).

4. Closing Remarks

Although this research is at the initial stages, we have been able to observe certain patterns in interlingual task based dialogues that could have implications for system design. Speakers attempted to adapt to ASR problems by paraphrasing (e.g. a participant replaced “youth hostel” with “YMCA”), and to perceived MT problems by being more explicit and formal in their utterances. Such strategies prove less effective for dialogue misalignment caused by ASR-MT interactions [4]. In addition, the lack of backchannels and affirmative cues, which play an important role in spontaneous dialogues [6] caused communication problems even when ASR and MT performed well. Recent work suggest that attentional selection in conversation can be mapped to EEG signals [5] by fairly straightforward methods. We plan to use these methods to analyse listening behaviour.

5. The Demo

A demonstration of the prototype will be presented at the conference. This presentation will illustrate the data collection procedure. Conference participants will be invited to participate in a sample data collection session.

6. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL (www.cngl.ie) at Trinity College, Dublin.

7. References