The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus

Hayakawa Akira[†], Saturnino Luz[‡], Loredana Cerrato[†], Nick Campbell[†]
[†]ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[‡]Usher Institute of Population Health Sciences & Informatics, University of Edinburgh, UK
[†]{campbeak, cerratol, nick}@tcd.ie, [‡]S.Luz@ed.ac.uk

Abstract

This paper presents the multimodal Interlingual Map Task Corpus (ILMT-s2s corpus) collected at Trinity College Dublin, and discuss some of the issues related to the collection and analysis of the data. The corpus design is inspired by the HCRC Map Task Corpus which was initially designed to support the investigation of linguistic phenomena, and has been the focus of a variety of studies of communicative behaviour. The simplicity of the task, and the complexity of phenomena it can elicit, make the map task an ideal object of study. Although there are studies that used replications of the map task to investigate communication in computer mediated tasks, this ILMT-s2s corpus 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, where the instruction giver and the instruction follower speak different languages. This paper details the data collection setup and completed annotation of the ILMT-s2s corpus, and outlines preliminary results obtained from the data.

Keywords: interlingual map task, computer mediated communication, natural task oriented dialogue

1. Introduction

The aim of this paper is to present the multimodal Interlingual Map Task Corpus (ILMT-s2s corpus) (Hayakawa et al., 2015) collected at Trinity College Dublin, and discuss some of the issues related to the collection and analysis of the data.

Fifteen dialogues were collected with the map task elicitation technique in a setting where the interaction is mediated by a speech-to-speech (S2S) translation system. The system was implemented using off-the-shelf technology to enable speakers of different languages to communicate with each other (remotely, over the network) in their native languages. This prototype speech-to-speech translation system (ILMT-s2s system) adds three elements: Automatic Speech Recognition (ASR), Machine Translation (MT) and Text To Speech (TTS) synthesis into the communication — we call these "filters". Our main interest is to understand how such filters affect the subjects in terms of cognitive load, adaptation of communicative acts to the technology, prosodic alignment and repair strategies, and other factors that might have implications for the design of dialogue systems and, in particular, S2S translation systems.

To this end we have collected a variety of synchronised and finely time-stamped data streams, including: high quality audio of the subjects' utterances, video and ASR, MT, TTS events. In addition, heart rate, skin conductance, blood volume pressure and electroencephelography (EEG) signals of one subject in each dialogue has also been recorded during the whole interaction with biosignal monitoring devices.

So far the corpus has been used to investigate two speech related interaction phenomena observed in a map navigation setting where the collaborating subjects speak different languages and have their utterances translated by the aforementioned S2S translation system. These studies comprise an investigation of how speakers adjust their speaking style in relation to errors from Automatic Speech Recognition (ASR), and an analysis of possible associations between speech recognition performance and three cognitive states that arise in dialogues.

For the future, besides the traditional analysis of speech, gestures, and facial expression, we plan to investigate possible correlations between the subjects' brains' electrical activity (through EEG), their blood volume pulse and their skin conductance with difficult situations during the interactions.[1] The collected corpus not only represents a new source of data for analysis of different repairs strategies and cooperative behaviour but also several aspects of human adaptation to the technology. It also allows to directly compare the dialogues with the original HCRC Edinburgh Map Task (Anderson et al., 1991), as well as its replications (Newlands et al., 2003; Louwerse et al., 2006).

The knowledge acquired by analysing the data can be used to provide baseline material for component development and testing and will also enable testing of methods for "affect sensing" from acoustic, video and biometric data recorded during the interaction.

2. Data Collection

The data for the ILMT-s2s corpus was collected at Trinity College Dublin between August 2014 and December 2014, using an S2S MT system designed specifically for this purpose.

2.1. ILMT-s2s System

The Interlingual Map Task Speech-to-Speech translation (ILMT-s2s) system is a 700 line Python script that combines three off-the-shelf technologies in a clean simple User Interface (UI) (Figure 1). The ILMT-s2s System runs on both Mac OS X and Linux[2] with the three main components being;

[1] I.e., those in which the communication technology creates problems, communication issues arise, but not obviously due to the intervening technology or the task itself creating difficulty.

[2] For the data collection, only Mac OS X computers were used.

1. The Google Speech API for ASR.
2. The Microsoft Bing translation system for MT.
3. For TTS:
   The Apple TTS system voices provided with Mac OS X computers (English: Kate, Portuguese: Joana).
   or
   The eSpeak speech synthesiser for Linux computers.

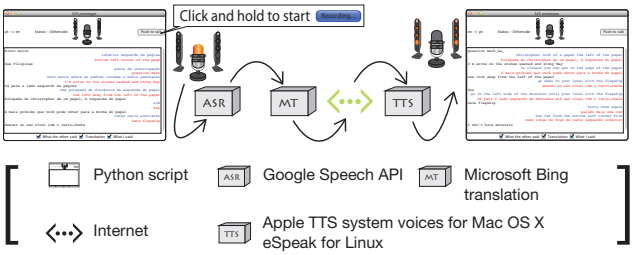

Figure 1: ILMT-s2s System used to collect the data

The ILMT-s2s System takes the speech uttered to the system, by the subject, and provides a Speech-to-Text (S2T) and Speech-to-Speech (S2S) translation of the subject's utterance to the interlocutor situated in another room.[3]

2.1.1. Basic Structure of the System

The ILMT-s2s System is activated by a "Push to talk" button that the subject will click-and-hold for the duration of the utterance, and release once the subject has finished the utterance. For each utterance the ILMT-s2s System executes the following processes and logs the five actions Click-and-hold, ASR, MT, TTS send, and TTS received, in a XML file (refer to § 2.2.6. for details):

1. Record and save an audio file sampled at 96 kHz, 24 bit PCM format for the duration of the click-and-hold action using the *sox* command.
2. Downsample the audio file recorded in step 1 to 16 kHz, 8 bit FLAC format using the *sox* command.
3. Send the FLAC format audio file created in step 2 to the ASR via the internet with the *wget* command.
4. Send the text result of the ASR to the MT service for translation via the internet with the *wget* command.
5. Send the translated MT text result to the interlocutor's client computer via the internet.
6. The TTS component on the interlocutor's client computer converts the text to speech with the TTS command and outputs the translated text as synthesised speech in the target language.

2.1.2. User Interface

The principle of the UI design is to be obvious as "*[well]-designed objects are easy to interpret and understand.*" (Norman, 2002), and also to inform the subjeect what the system is doing by providing feedback of the current status, "*[feedback] and communication encompass far more than merely displaying alerts when something goes wrong. Instead, it involves keeping subjects informed about what's happening by providing appropriate feedback and enabling communication with your app.*" (Apple, 2013).
With this in mind, the UI was specifically designed to look minimalistic (Figure 2) so that emphasis could be placed on the displayed text but feedback indicators were also added to provide the subject information of the actions of the system and their interlocutor.

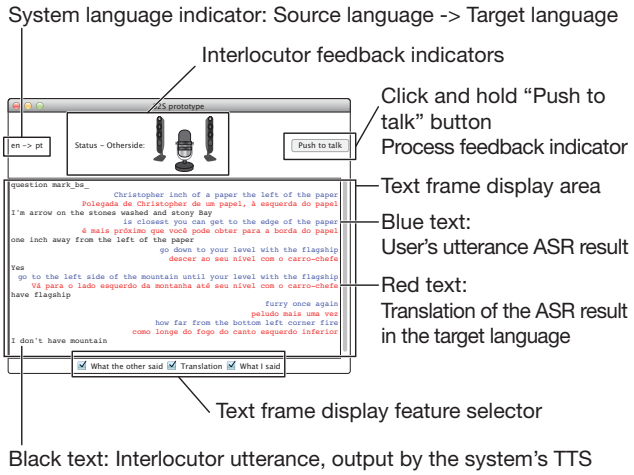

Figure 2: User Interface of the ILMT-s2s System

The design of the textual display is taken from the basic structure of popular texting software. This is based on the assumption that due to the popularity of texting applications on smartphones and computers, the subjects would be familiar with the right-side displaying the subject's text in colour and the left-side displaying the interlocutor's text with no colour. The text was timed to be displayed after all the system's internal processes were completed. For the subject's display this means that the ASR and MT (the blue and red text of Figure 2) is displayed immediately after the MT translation is sent to the interlocutor's computer. For the interlocutor, the S2T (the black text of Figure 2) is displayed on their UI as soon as it is received.
As for the feedback indicators, two systems were included in the UI. One was an indicator of the process being performed by the ILMT-s2s System in the click-and-hold "Push to talk" button (Figure 3), the other was an indicator that displays the *talking* or *listening* action of the interlocutor at the top-centre of the UI (Figure 4).

ILMT-s2s process indicator: Text within the click and hold "Push to talk" button was used to display the current process that was being performed by the ILMT-s2s System (Figure 3). The start time and end time of the five processes, "Recording…" to "Sending text", where finely time-stamped and saved to the XML log file (§ 2.2.6.). Though the terms "Con. to flac", "Con. to text" and "Sending text" are not easily understandable by the subject, they still indicate that the system is doing something and provides the subject with feedback that the system is active.

[3] For the data collection, S2T translation was not displayed to the speaker, but only to the interlocutor.

ILMT-s2s interlocutor action indicator: The microphone and loud-speakers placed at the top-centre of the UI were used to illustrate the action of the interlocutor (Figure 4). While the interlocutor holds the "Push to talk" button, the microphone mesh lights orange to indicate that the interlocutor is speaking to the ILMT-s2s System and then turns back to the original grey colour once the button is released. When the ILMT-s2s System outputs the TTS on the interlocutor's computer, the loud-speaker cones light orange on the subject's computer to indicate that the utterance has been sent and output to the interlocutor and turns back to the original grey once the output is finished. This feature could be described in the same line of thought with popular texting applications where a bubble with dots indicate a interlocutor writing text or the description of "Delivered" and "Read" that appears beside messages that are sent.

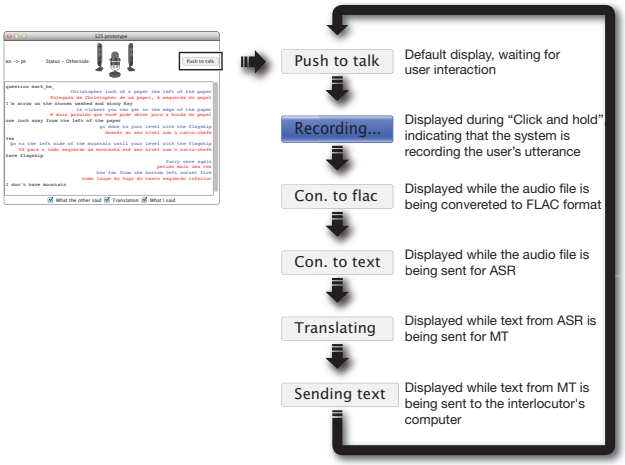

Figure 3: System action indicator

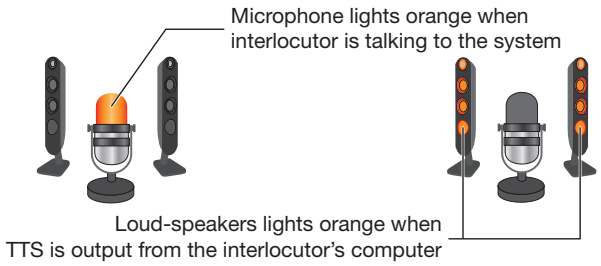

Figure 4: Interlocutor action indicators

2.2. ILMT-s2s Data Collection

The ILMT-s2s System (§ 2.1.) was used to collect fifteen dialogues between native English and Portuguese speakers. Apart from the audio utterance recordings and XML log that the ILMT-s2s System saves, HD video, eye-tracking video and biological signals (heart rate, skin conductance, blood volume pressure and EEG) were recorded using the equipment indicated in Table 1, and worn as displayed in Figure 5 — the eye-tracking video and biological signals were only recorded from one subject per recording.

w/o bio signal monitor	w/ bio signal monitor
Audio-Technica HYP-190H for utterances to the ILMT-s2s System recorded at 96 kHz 24 bit PCM	Sennheiser MKE104 for utterances to the ILMT-s2s System recorded at 96 kHz 24 bit PCM
Sony HDR-XR500 for subject front view, recording at 1080i, 29.97 fps	Sony HDR-XR500 for subject front view, recording at 1080i, 29.97 fps
Panasonic HX-A100 for subject view range, recording at 1080p, 29.97 fps	SMI Eye Tracking Glasses 1.1 recording at 960p, 30 fps (eye tracking data recorded at 24 sps)
–	Mind Media Nexus-4, for bio sensor recording (upto 1024 sps)
Sony HDR-CX370 for subject back view, recording at 1080i, 29.97 fps	–

Table 1: Recording devices per subject


w/o bio signal monitor (during recording)


w/ bio signal monitor (during recording)


Detailed

Figure 5: Subjects during recordings

In total, fifteen dialogues totalling approximately 9.5 hours of edited data have been collected and named as the ILMT-s2s Corpus (Table 3 of Appendix 1 for details).

2.2.1. The Map Task

The recorded dialogues do not have a script, so they are elicited according to the map task scheme, in which the two dialogue subjects have a specific role designated before the recording starts: an instruction giver (IG) and an instruction follower (IF). The IG has a map with a route drawn on it and has to guide the IF so they can draw the same route on his/her unmarked copy of the map — the subjects cannot see each other's map. The subjects were shown a map similar to what was used for the recordings and given an explanation of their role as IG and IF only once they arrived for the recording.

The maps used to elicit the conversation in this ILMT-s2s data collection were taken from the HCRC Map Task corpus (Anderson et al., 1991). The postscript files of the maps were downloaded[4] and modified using Adobe Illustrator and the original landmark illustrations, placement and names were used but the landmark names where re-written with the "Bradley Hand" font so that all translations

[4]Link to download HCRC Map Task maps: `http://groups.inf.ed.ac.uk/maptask/maptasknxt.html`

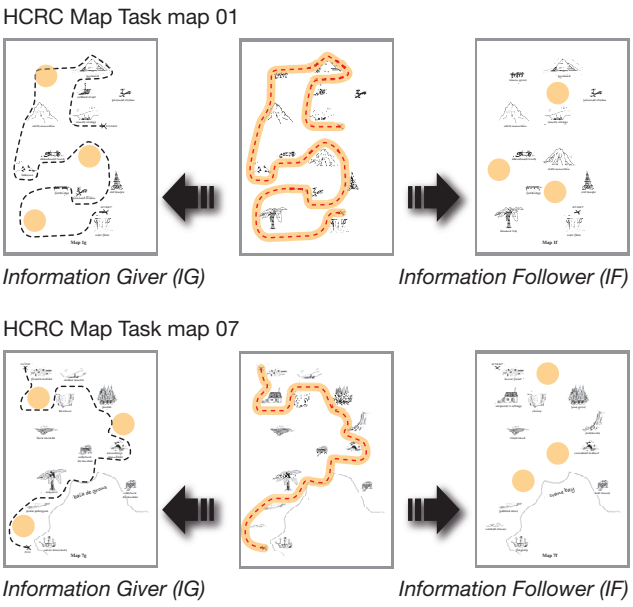

Figure 6: Maps used in the ILMT-s2s Data Collection with differences highlighted

could be displayed similarly.
Of the sixteen maps, only two maps, map 01 and map 07, with different route patterns[5] were selected to be used for the ILMT-s2s data collection (Figure 6).

2.2.2. The Subjects

The subjects of the ILMT-s2s data collection were recruited from the Trinity College Dublin digital noticeboard and also via personal connections. Fifteen recordings of fifteen native English speakers (♀5, ♂10), and fifteen native Portuguese speakers (♀11, ♂4), between the ages of 18 and 45 were collected and balanced as illustrated in Figure 7 and indicated in Table 4 of Appendix 2.

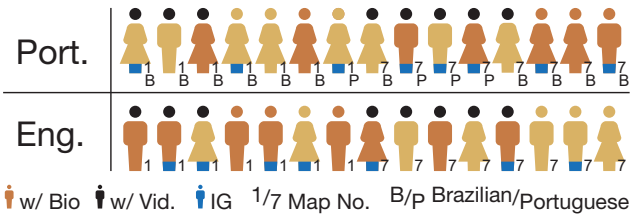

Figure 7: Image of subject pairs and recording situation

2.2.3. Recording Environment

The subjects of the ILMT-s2s data collection were informed about the experiment via an information sheet which was sent to them at least 24 hours before the expected time of data collection. At the time of the data collection, the two subjects were together given a short verbal instruction on how to use the ILMT-s2s System and how the ILMT-s2s System is expected to operate and also a simple explanation of their roles as IG and IF. Each recording session lasted between 20 and 74 minutes and contains between 33 and 199 utterances to the ILMT-s2s system, and this changes to between 43 and 219 if audible utterances that were not directed to the ILMT-s2s System are included. The recording of the data was conducted in a working office of Trinity College Dublin, so environmental noise is included.
The two main setup differential of the data collection were "w/o video" – "w/ video" (below) and "Map 01" – "Map 07" (Figure 6).

w/ video: A constant live video stream of the other subject is displayed on the computer screen, but audio is not transmitted with the video stream. TTS output of the interlocutor's utterance and S2T translation is displayed on the ILMT-s2s System when the system's processes have been completed. Hence the interlocutors ***can*** see each other but can only hear and read the ILMT-s2s System output.

w/o video: The same setup as "w/ video" but without the live video stream. Hence the interlocutors ***cannot*** see each other and can only hear and read the ILMT-s2s System output.

2.2.4. Biosignal Recording

As mentioned in Table 1, to record the biosignals, a Mind Media B.V., NeXus-4 was used to collect Skin Conductance (SC), Heart Rate (HR) using the Blood-Volume Pulse (BVP) readings, and the brains electrical activity through Electroencephelography (EGG). The SC sensor was placed on the middle and ring finger and the BVP sensor was placed on the index finder. EEG sensors were placed in the F4, C4, P4 position with a ground channel placed at A1 of the 10 – 20 location system (Figure 8). The sampling frequency for the SC, HR and EEG were 32 kHz, 32, kHz and 1,024 kHz respectively.

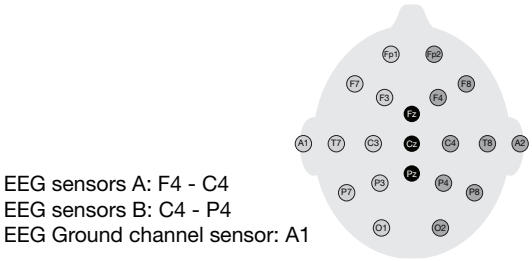

Figure 8: 10 – 20 system layout map

2.2.5. ILMT-s2s System User Survey

After each recording the subjects completed a survey to express their disagreement or agreement on fifteen statements designed to gather their perception of the system's usability — *Ease of use*, *Effectiveness* and *Satisfaction*. Responses were given on a 7 point Likert scale (evenly spaced from 1 = Strongly disagree to 7 = Strongly agree), presented together with an open text field for possible further comments. Five extra open questions, presented in an open text field format, were added to ask about the difficulties the subjects experienced during the interactions (Table 2).

[5]There are sixteen HCRC maps which are made of four basic patterns, where the same route is used but the landmarks are changed to make four different maps for each basic pattern.

7 point Likert scale statements
1. Overall, I am satisfied with how easy it was to use this system.
2. It was simple to use this system.
3. I could effectively complete the tasks using this system.
4. I was able to complete the task quickly using this system.
5. I was able to efficiently complete the task using this system.
6. I felt comfortable using this system.
7. It was easy to learn to use this system.
8. I believe I could become productive quickly using this system.
9. Whenever I made a mistake using the system, I could recover easily and quickly.
10. The interface of the system was pleasant.
11. I liked using the interface of this system.
12. This system has all the functions and capabilities I expected it to have.
13. I was satisfied with the voice of this system.
14. I was satisfied with the output of this system.
15. Overall, I am satisfied with this system.
Open text field questions
1. Please indicate why you changed the style of communication.
2. Please indicate what made you give up clarifying the intention of the other participant.
3. Please indicate all the things that irritated you.
4. Please indicate all the things that pleased you.
5. Please indicate what you felt was most difficult.

Table 2: List of survey statements and questions

2.2.6. ILMT-s2s System XML Log

Each time the subject clicks and holds the "Push to talk" button, the ILMT-s2s System will record a log of the following four actions: Recorded utterance, ASR, MT, TTS send. Also when the ILMT-s2s System receives an utterance from the interlocutor's computer, the subject's ILMT-s2s System will record a log of the received TTS. This means that of the six basic steps of the ILMT-s2s System explained in § 2.1.1., all but the conversion of the .wav audio file to a .flac audio file are recorded in the log.

Apart from this, the system also records the source and target language, date and time of recording, and subject details such as ID, gender, age and origin, that are manually input into the system before the recording starts.

3. Data to Corpus

3.1. Synchronisation

Since all the devices used during the dialogue sessions recorded video or audio, Final Cut Pro X of Apple Inc. was used for the synchronisation. The ILMT-s2s System's audio utterance recordings were first synchronised with the video recordings from the SONY HDR-XR500 camcorder. The audio from the camcorder was converted into monaural audio so the resulting video had one channel with the audio from the camcorder and the other channel with the audio from the ILMT-s2s System. This made it possible to combine and see the audio that was spoken to the other subject and the audio of the whole session in the same audio file, but still keep them separate. This new video file was then synchronised with the video and audio files from the other devices. Once all files from each subject were synchronised, the files from each subject pair were synchronised with each other so that files from both sides of the conversation started as they had done in reality (Figure 9).

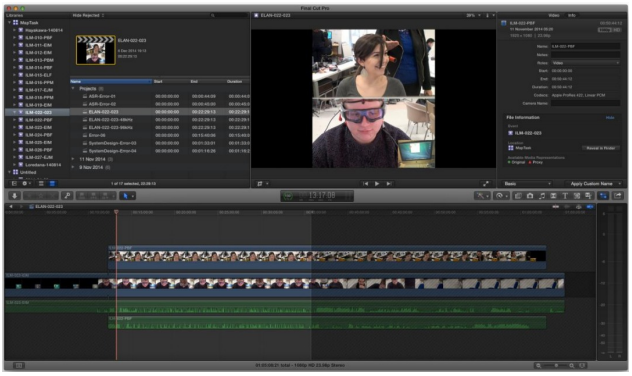

Figure 9: Representation of synchronised files

The synchronised dialogue is then cut between the points where the subjects start and end the map task to removing unwanted data, as displayed in the lighter grey section of Figure 9.

3.2. Transcription

Two students (one native speaker of English and one native speaker of Portuguese) where recruited to orthographically transcribe the edited audio files using the open source software Wavesurfer (Sjölander and Beskow, 2000). The English transcription text was verified by the author to double check the utterances and correct any misinterpretation of the speech – Since the author had spoken to all the subjects, the author had a better understanding of the subjects interests and background that it was possible to correct utterances that were difficult to hear and understand. One example of this would be "thanks for not responding" being changed to "python not responding". As for the Portuguese transcription text, it has also been checked by a second native Brazilian Portuguese speaker with the request to verify the transcription, spelling and usage of accents. Once completed, the transcribed files were again checked to verify that start and end point of the transcription segmentation have been correctly implemented and that no utterances audible have been missed out.

3.3. Annotation

The two students who transcribed the data also annotated the data using the dedicated annotation tool ELAN (Wittenburg et al., 2006). Video and audio files were used for the annotation with the following freely definable multi-layered annotation scheme tiers — For the annotation of the cognitive states we calculated the inter-coder agreement on one of the dialogues and the results are well above 60%[6]:

- Dialogue acts [25]: *Acknowledgement CP/CPU*, *Align*, *Check*, *Clarify*, *Explain*, *Instruct*, *Interjection*, *Query y-n/w*, *Reply y/n/w*, with also a *Solo* variant.

  These labels were based on the Dialogue Structure Coding scheme (Carletta et al., 1997), but with modifications to "Acknowledgement" to differentiate the simple acknowledgement (CP) of the utterance from the acknowledgement with the actual understanding of the utterance (CPU) which is defined under the MUMIN coding scheme (Allwood et al., 2007). Also, a "Solo" variant of the dialogue acts, and "Interjection" were added due to the "Off-Talk" characteristics (Oppermann et al., 2001) of the collected data.

- Cognitive states [3]: *Surprised*, *Amused*, *Frustrated*.

  Understanding that there are numerous cognitive states proposed by various annotation schemes (McCowan et al., 2005; Popescu-Belis, 2005; Juel Henrichsen and Allwood, 2013), we chose to initially annotate only *Surprised*, *Amused* and *Frustrated*.

  Since the data collection is task based, we thought the subjects would be focused on providing and receiving clear instructions. However, due to "*the vocabulary problem*" (Furnas et al., 1987), clear communication is difficult even in human-to-human situations. By adding the *filters* of the ILMT-s2s System, further complication will arise. We assumed this would bring out the selected cognitive characteristic in the interaction with the system.

- Facial expressions [3] and head movements [2]: *Smile*, *Laughter*, *Surprise*, *Feedback-nod*, *Feedback-shake*.

  These labels were based upon the "General face" and "Head movement" attributes of the MUMIN coding scheme (Allwood et al., 2007). Of the original "General face" attributes, "Scowl" and "Other" were removed and "Surprised" was added since the ASR and MT results may not provide what the subject was expecting and result in a surprised rather than a scowled expression. The "Head movement" attributes were summarised into "Feedback-nod" for vertical, and "Feedback-shake" for horizontal head movements, following the expected ease of visual communication during the "w/ video" recording setup mentioned in § 2.2.2..

Additionally to the above tiers, the transcription, log data from the ILMT-s2s System XML files and results from further analysis were also added to the ELAN annotation files as tiers. Transcriptions, start and end times of utterances to the system, ASR, MT, TTS delivery and TTS output, On-Talk[7], Off-Talk Self[8], Off-Talk Other[9], word count etc. were added for each subject as indicated in Figure 10.

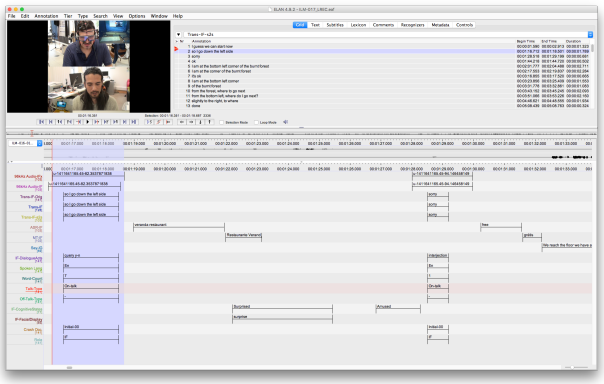

Figure 10: Representation of ELAN data files

4. Corpus Analyses and Future Work

To date, the data from the ILMT-s2s corpus has been used in four preliminary studies.

User evaluation of the ILMT-s2s system: Results show that in general, the subjects seemed to have experienced the system as *easy to learn to use* and *simple to use* — mean 4.9. subjects also found the system *pleasant and likeable* — mean 5.8, and they liked the output — TTS obtained the highest grade in the evaluation with a mean 6.0 (Cerrato et al., 2015).

User adaptation to the ILMT-s2s system: Result shows that speaking rate and clear speech are used as repair strategies when ASR errors occur. This behaviour is not a generalised, stable mode of speaking in the dialogues we analysed, since it seems to be a targeted and flexible adaptation strategy (Hayakawa et al., 2015a).

Detection of Cognitive states: Results show that a combined biosignals yields detection performance well above the baseline (72% accuracy) when the time window is restricted to the perceived duration of the state. Extending the window to the end of the utterance following the cognitive state yields poor detection on biosignals alone, but improves considerably when features of the speech signal are added, thus showing the potential usefulness of speech features as a biosignal (Hayakawa et al., 2015b).

Detection of "Off Talk": In this study we analysed "Off Talk" (Oppermann et al., 2001). The characteristics of the three speech types (*On-Talk*, *Off-Talk Self* and *Off-Talk Other*) show significant differences in terms of speech rate ($F_{2,2719} = 101.7; p < 2e-16$), and for this reason a detection method was implemented to see if they could also be detected with good accuracy

[6] Calculated using the modified kappa feature of ELAN 4.9.0's "Inter-Annotator Reliability…" function.

[7] On-Talk: Talking to the ILMT-s2s System.

[8] Off-Talk Self: Talking to oneself.

[9] Off-Talk Other: Talking to someone else.

based on their acoustic and biological characteristics (Hayakawa et al., 2016).

As already mentioned in § 1., we think that the knowledge acquired by analysing the data can be used to provide baseline material for component development and testing and will also enable testing of methods for "affect sensing" from acoustic, video and biometric data of the interactions.

5. Acknowledgements

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at TCD.

6. Bibliographical References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3):273–287.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Apple. (2013). Mac OS X Human Interface Guidelines. *Apple Inc, Cupertino*.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., and Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

Cerrato, L., Hayakawa, A., Campbell, N., and Luz, S. (2015). A Speech-to-Speech, Machine Translation Mediated Map Task: An Exploratory Study. In *Proceedings of the workshop Future and Emerging Trends in Language Technology (FETLT2015)*, Seville, Spain. Springer, LNAI. in press.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Hayakawa, A., Cerrato, L., Campbell, N., and Luz, S. (2015a). A Study of Prosodic Alignment in Interlingual Map-Task Dialogues. In The Scottish Consortium for ICPhS, editor, *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK. The University of Glasgow. Paper number 0760.1-9.

Hayakawa, A., Haider, F., Cerrato, L., Campbell, N., and Luz, S. (2015b). Detection of Cognitive States and Their Correlation to Speech Recognition Performance in Speech-to-Speech Machine Translation Systems. In *Proceedings of INTERSPEECH'15 Sixteenth Annual Conference of the International Speech Communication Association*, pages 2539–2543, Dresden, Germany. ISCA.

Hayakawa, A., Haider, F., Luz, S., Cerrato, L., and Campbell, N. (2016). Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task. In *Proceedings of Speech Prosody 2016 (SP8)*, Boston, Massachusetts, USA. ISCA. In press.

Juel Henrichsen, P. and Allwood, J. (2013). Predicting the Attitude Flow in Dialogue Based on Multi-Modal Speech Cues. In Jens Allwood, et al., editors, *NEALT2012: Proceedings of the 4th Nordic Symposium on Multimodal Communication*, Linköping Electronic Conference Proceedings, pages 47–53. Göteborg Universitet.

Louwerse, M., Jeuniaux, P., Hoque, M., Wu, J., and Lewis, G. (2006). Multimodal communication in computer-mediated map task scenarios. In *Procs. of the Cognitive Science Society*, pages 1717–1722.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al. (2005). The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88.

Newlands, A., Anderson, A. H., and Mullin, J. (2003). Adapting communicative strategies to computer-mediated communication: an analysis of task performance and dialogue structure. *Applied Cognitive Psychology*, 17(3):325–348.

Norman, D. A. (2002). *The design of everyday things*. Basic books.

Oppermann, D., Schiel, F., Steininger, S., and Beringer, N. (2001). Off-talk-a problem for human-machine-interaction? In *Proceedings of INTERSPEECH'01: the 2nd Annual Conference of the International Speech Communication Association*, pages 2197–2200, Aalborg, Denmark. ISCA.

Popescu-Belis, A. (2005). Dialogue acts: One or more dimensions. *ISSCO WorkingPaper 62*.

Sjölander, K. and Beskow, J. (2000). Wavesurfer - an open source speech tool. In *Proceedings of the 6th International Conference on Spoken Language Processing*, pages 464–467, Beijing, China. ISCA.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1556–1559, Genoa, Italy. European Language Resources Association (ELRA).

7. Language Resource References

Hayakawa, Akira and Luz, Saturnino and Cerrato, Loredana and Campbell, Nick. (2015). *The ILMT-s2s Corpus*. CNGL Programme, distributed via ELRA, Trinity College Dublin, 1.0, ISLRN 100-610-774-625-0.

Appendix 1: ILMT-s2s Data Summary

Dialogue N^{o} — Setting	Duration	On-Talk					
		Utt. Count	(IG :	IF [% Diff.])	Word Count	(IG :	IF [% Diff.])
1 — Map 01, w/ Video	00:23:02	88	(55 :	33 [−40%])	717	(450 :	267 [−40%])
2 — Map 01, w/ Video	00:48:57	217	(100 :	117 [+17%])	1,522	(593 :	929 [+57%])
3 — Map 01, w/ Video	01:14:20	209	(106 :	103 [−3%])	818	(422 :	396 [−6%])
4 — Map 01, w/o Video	00:22:31	132	(59 :	73 [+24%])	530	(265 :	265 [±0%])
5 — Map 01, w/o Video	00:40:20	262	(201 :	61 [−70%])	1,349	(915 :	434 [−53%])
6 — Map 01, w/o Video	00:44:07	171	(93 :	78 [−16%])	1,310	(744 :	566 [−24%])
7 — Map 01, w/o Video	00:37:55	183	(94 :	89 [−5%])	1,369	(888 :	481 [−46%])
8 — Map 07, w/ Video	00:59:29	174	(121 :	53 [−56%])	1,437	(1,016 :	421 [−59%])
9 — Map 07, w/ Video	00:29:13	202	(98 :	104 [+6%])	1,011	(522 :	489 [−6%])
10 — Map 07, w/ Video	00:36:07	168	(76 :	92 [+21%])	548	(254 :	294 [+16%])
11 — Map 07, w/ Video	00:43:17	157	(101 :	56 [−45%])	942	(481 :	461 [−4%])
12 — Map 07, w/ Video	00:48:19	176	(100 :	76 [−24%])	1,283	(551 :	732 [+33%])
13 — Map 07, w/o Video	00:26:13	98	(49 :	49 [±0%])	652	(345 :	307 [−11%])
14 — Map 07, w/o Video	00:25:59	126	(69 :	57 [−17%])	473	(294 :	179 [−39%])
15 — Map 07, w/o Video	00:19:38	76	(34 :	42 [+24%])	416	(235 :	181 [−23%])
Total	09:39:27	2,439	(1,356 :	1,083 [−20%])	14,377	(7,975 :	6,402 [−20%])

Table 3: ILMT-s2s Corpus Data Summary

Appendix 2: ILMT-s2s Data Balance

	w/ Video	w/o Video	Map 01	Map 07	♂	♀	w/ Bio	w/o Bio	IG	IF	En.	Pt.
w/ Video	16	–										
w/o Video	–	14										
Map 01	6	8	14	–								
Map 07	10	6	–	16								
♂	8	6	6	8	14	–						
♀	8	8	8	8	–	16						
w/ Bio	8	7	7	8	9	6	15	–				
w/o Bio	8	7	7	8	5	10	–	15				
IG	8	7	7	8	7	8	8	7	15	–		
IF	8	7	7	8	7	8	7	8	–	15		
English	8	7	7	8	10	5	8	7	7	8	15	–
Portuguese	8	7	7	8	4	11	7	8	8	7	–	15
Pt-Portugal	3	1	1	3	2	2	2	2	4	0	–	4
Pt-Brazil	5	6	6	5	2	9	5	6	4	7	–	11

Table 4: Balance of Subjects