INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Talking to a System and Talking to a Human: A study from a Speech-to-Speech, Machine Translation mediated Map Task

Hayakawa Akira[†], *Saturnino Luz*[‡], *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, nick}@tcd.ie`[†], `S.Luz@ed.ac.uk`[‡]

Abstract

This study focuses on the properties of Human-to-Human (H2H) communication in spontaneous dialogues in two different settings. Direct H2H dialogues are compared to the ones that are mediated by a Speech-to-Speech machine translation system. For the analysis, dialogues from the HCRC Map Task Corpus, for direct H2H conversations, and dialogues from the ILMT-s2s Corpus, for computer mediated conversations, were used. In the conversations speakers take the roles of information giver and follower and all the utterances are labelled as instructions, questions or statement, etc. While direct H2H communication enables speakers also to benefit from non-verbal acts, gestures and facial expressions, machine mediated conversation is more complex for the interlocutors. Due to errors made by speech recognition system, speakers adapt their speaking style and also apply repair strategies in order to accomplish the tasks successfully. Comparing the two corpora showed that in the case of computer mediated communication the utterances of the speakers contained less words than in the case of direct H2H interaction where utterances were longer. Also, different word count was found depending on the role of the speaker as well as on the type of the utterance.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Speech-to-Speech Machine Translation (S2S-MT) systems are becoming a reality as a way of communication. Microsoft has already released the Skype Translator and the Japanese Ministry of Internal Affairs and Communication has announced that the Tokyo Olympics in 2020 is to use information systems that use multilingual machine mediated communication for 17 languages in Speech-to-Speech (S2S) form, and to achieve this they will first be tested in hospitals, tourist cites and shopping centres soon.Though there is an increase in the usage of such systems, there is still little research on how dialogues change between direct Human-to-Human (H2H) communication and computer mediated multi-lingual communication. We as humans are already experienced with direct H2H communication and even communication into a foreign language. However, few of us will be familiar with communication via an interpreter, let alone an interpreter that cannot be interrupted, which S2S-MT systems are offering. In this study we present preliminary results from the comparison of these two types of communication which result in a less efficient method of communication, due to the difficulties that the S2S-MT system causes.

2. Material

This study investigates the communication pattern in H2H and S2S-MT communication of task oriented conversation using the Map Task technique. For the H2H communication, sixteen dialogues that used maps 01 and 07 (Figure 1) from the HCRC Map Task corpus [1] were used (§ 2.1), and for the S2S-MT communication data, the fifteen dialogues of the English subjects from the ILMT-s2s corpus [2] were used (§ 2.2).

2.0.1. The Map Task Technique

In both corpora, maps from the HCRC Map Task corpus[1] were used to elicit the task oriented conversation between the subjects. The subjects in each recording were given a role of either Information Giver (IG) or Information Follower (IF), where the IG has a map with a route drawn on it. From this map, the IG is to instruct the IF to draw a copy of the route on his/her unmarked copy of the map. Each map contains a number of landmarks (e.g., "white mountain", "baboons", "crest falls") which may or may not be common to both maps (Figure 1). This difference between the IG's and IF's map, combined with the fact that neither subject can see the other's map adds to the complexity of the task.

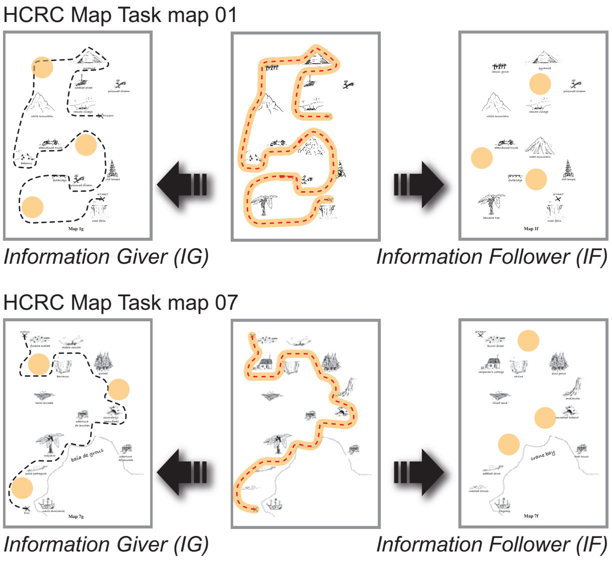


Figure 1: *Map used, with differences highlighted*

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

http://dx.doi.org/10.21437/Interspeech.2016-1623

2.1. Data from the HCRC Map Task Corpus

The HCRC Map Task corpus contains 128 dialogues of subjects using the map task technique. Of the 128 dialogues, 16 dialogues that use the same map as what was used in the ILMT-s2s corpus were used in this study.

The dialogues were between native English speakers, mostly from Scotland, all undergraduate students from the University of Glasgow, half male, half female. The familiarity aspect was controlled by pairing participants once with a familiar subject and once with an unfamiliar one. Participants were on a face-to-face setting following a Latin Squares design, with small barriers between them to prevent them to see each other's maps.

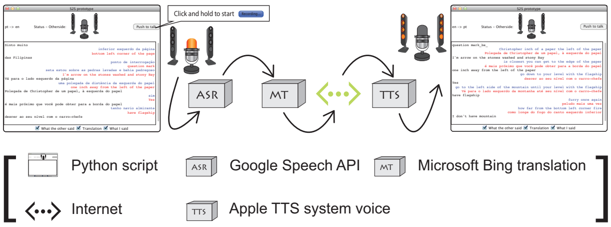
The resulting corpus of audio and video files has been transcribed and annotated intensively, for disfluencies, gaze, moves, prosody, syntax, etc., the last version using XML Annotations on the NXT-format. The dialogue act annotation segments and text were extracted from the release version 2.1 and the segmentation was verified using the dedicated annotation tool ELAN [3] after converting the format.

2.2. Data from the ILMT-s2s Corpus

The ILMT-s2s corpus contains fifteen dialogues between native English and Portuguese subjects speaking to each other in their native language via a Speech-to-Speech (S2S) translation system (ILMT-s2s system). Since this study compares the dialogues of this corpus with the HCRC Map Task corpus, which is only in English, only the dialogue act data (§ 2.2.5) from the English subjects were analysed.

2.2.1. *The ILMT-s2s System*

Two subjects, seated in two different rooms, used the ILMT-s2s system (Figure 2) to communicate with each other. The ILMT-s2s system, is a system that uses off-the-shelf components — Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech synthesis (TTS) — to perform Speech-To-Speech Machine Translation. It 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. Neither subject can hear the other's voice since the output of the ASR and MT is provided by a synthetic voice.



Sinto muito
inferior esquerdo da página
bottom left corner of the page
das Filipinas
ponto de interrogação
question mark
seta estou sobre as pedras lavadas e bahia pedregoso
I'm arrow on the stones washed and stony Bay
Vá para o lado esquerdo da página
uma polegada de distância da esquerda do papel
one inch away from the left of the paper
Polegada de Christopher de um papel, à esquerda do papel
sim
Yes
é mais próximo que você pode obter para a borda do papel
tenho navio almirante
have flagship
descer ao seu nível com o carro-chefe

question mark_bs_
Christopher inch of a paper the left of the paper
Polegada de Christopher de um papel, à esquerda do papel
I'm arrow on the stones washed and stony Bay
is closest you can get to the edge of the paper
é mais próximo que você pode obter para a borda do papel
one inch away from the left of the paper
go down to your level with the flagship
descer ao seu nível com o carro-chefe
Yes
go to the left side of the mountain until your level with the flagship
Vá para o lado esquerdo da montanha até seu nível com o carro-chefe
have flagship
furry once again
peludo mais uma vez
how far from the bottom left corner fire
como longe do fogo do canto esquerdo inferior
I don't have mountain

Figure 2: *ILMT-s2s system used to collect the data*

2.2.2. *The Subjects and Recording Environment*

The subjects were recruited from the Trinity College Dublin digital noticeboard or 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. Each recording session was conducted in a working office and lasted between 20 and 74 minutes, containing between 33 and 201 On-Talk[2] utterances and between 44 and 212 On-Talk dialogue acts. One subject during each recording session was fitted with biosignals recording device, while the other subject was not (Figure 3).[3]

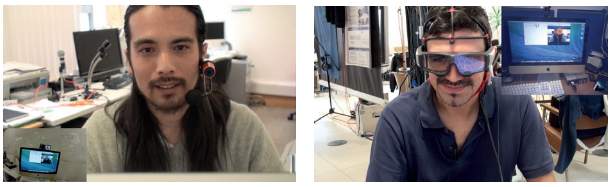


Figure 3: *Subjects during recordings*

2.2.3. *Recorded Media Data*

Audio and video recordings were recorded and are included in the ILMT-s2s corpus. Of these, the transcription and annotation from the audio recorded from the two video cameras that captured the images in Figure 3 were used for this study.

2.2.4. *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 [4]. 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. 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.

2.2.5. *Annotation*

The two students who transcribed the data also annotated the data using the dedicated annotation tool ELAN. Video and audio files were used for the annotation with the following freely definable multi-layered annotation scheme tier:

- 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 [5], 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 [6]. Also, a "Solo" variant of the dialogue acts, and "Interjection" were added due to the "Off-Talk" characteristics [7, 8, 9] of the collected data.

However, to standardise the scheme with the HCRC Map Task scheme, the differences of "*Acknowledgement CP/CPU*" have been ignored and combined as simply "*Acknowledge*" and "*Interjection*" has been ignored.

[2] When the subject is talking to the ILMT-s2s system.

[3] Data from the biosignal recordings were not used in this study.

2.3. Initial Summary of the Two Corpora

The data used from the HCRC Map Task corpus consists of sixteen dialogues using maps 01 and 07 with a total of 32 subjects — two subjects per dialogue. The data from the ILMT-s2s corpus consists of fifteen dialogues with a total of 30 subjects. However, since we are looking into the English dialogues, the dialogue from the Portuguese counterpart was removed leaving only 15 subjects (IG : IF = 7 : 8). Therefore with all things equal there should be approximately 2.1 times more dialogue acts and words in the HCRC Map Task corpus data than the ILMT-s2s corpus data. However as can be seen from Table 1 there are approximately 2.8 times more — HCRC word count per dialogue act: $\mu = 5.953, \sigma = 6.685$, ILMT-s2s word count per dialogue act: $\mu = 5.241, \sigma = 6.420$.

Table 1: *Summary of ILMT-s2s and HCRC Map Task corpora*

Corpus	Total	On-Talk	
	Duration	Dialogue Acts	Word Count
HCRC	02:02:22	3,790	21,119
ILMT-s2s	09:39:27	1,407	7,293

An initial plot of the word count per dialogue act of the two corpora shows that there is a clear difference in the way the subjects of the two roles (IG and IF) correspond to their interlocutor — HCRC IG word count: $\mu = 7.415, \sigma = 7.727$, HCRC IF word count: $\mu = 4.114, \sigma = 4.446$, ILMT-s2s IG word count: $\mu = 5.558, \sigma = 7.849$, ILMT-s2s IF word count: $\mu = 4.793, \sigma = 3.468$. In other words, the subject with the role of IF in the ILMT-s2s corpus uses more words per dialogue act when compared to the HCRC Map Task corpus, but the IG uses less words per dialogue act (Figure 4).

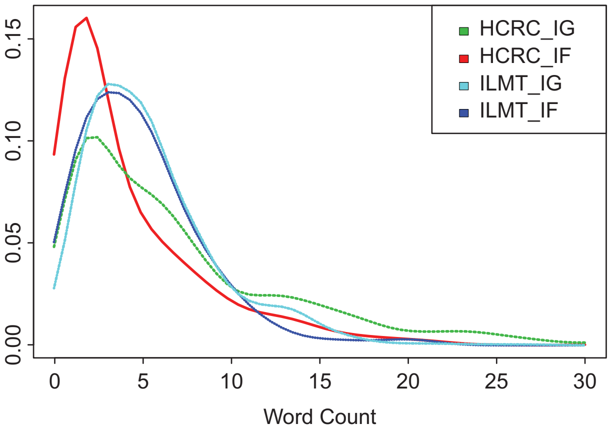


Figure 4: *Kernel density plot — Comparison of Subject Roles*

3. Method

Following this initial finding, further analysis were made to see where this difference comes from and what could be the probable cause.

3.0.1. Word Count Comparison

From the initial summary of the data (§ 2.3), the data indicates that there are differences in the communication method. To verify that there is an actual difference, the following null hypothesis is tested.

H_0**:** The means of dialogue act word count differences are the same for each subject role (IG and IF) in each corpus (HCRC and ILMT-s2s)

To determine which of the eleven dialogue acts have a different word count, an ANOVA test followed by a Post-hoc comparison (Tukey HSD test) was performed to single out the changing acts by adding the dialogue acts to the H_0 hypothesis.

3.0.2. Dialogue Act Comparison

The dialogue acts of the HCRC Map Task corpus and the ILMT-s2s corpus were compared to see if and how the information the subject convey, changes when the dialogue is mediated by a S2S-MT system. To do this the number of occurrences of each acts were counted and the lists of each corpus was compared — further divided by the role of the subject.

4. Results

4.0.1. Word Count Comparison

As already indicated in Figure 4, for H_0, the ANOVA test show that there is significant difference between the word count of the subject role of each corpus ($F_{3,4840} = 79.94; p < 0.0001$) with the effect size, calculated using Cohen's d, reflecting the curves of Figure 4 in Figure 5.

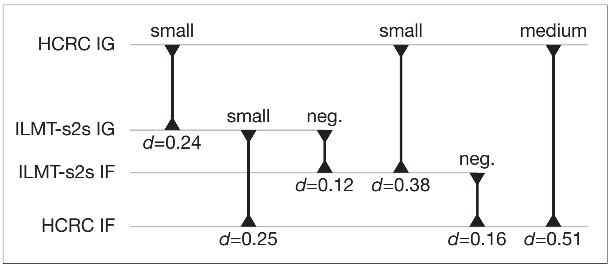


Figure 5: *Cohen's d Effect Sizes of Corpus Role Comparison*

For the results that include the dialogue acts, the Post-hoc comparison results indicated that of the eleven dialogue acts only "*Clarify*", "*Explain*" and "*Instruct*" had a significant difference (Figure 6) for the IG and only "*Clarify*" and "*Explain*" for the IF (Figure 7).

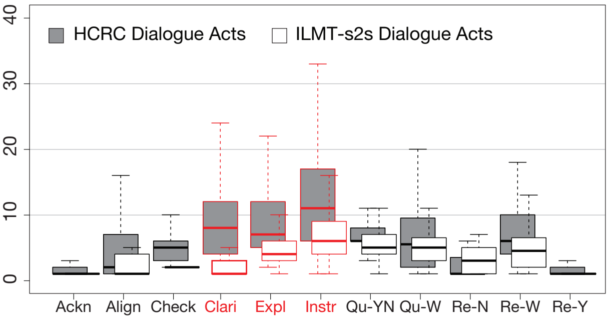


Figure 6: *Word Count Boxplot of IG Dialogue Acts of Each Corpus — Significantly Different Dialogue Acts Coloured Red*

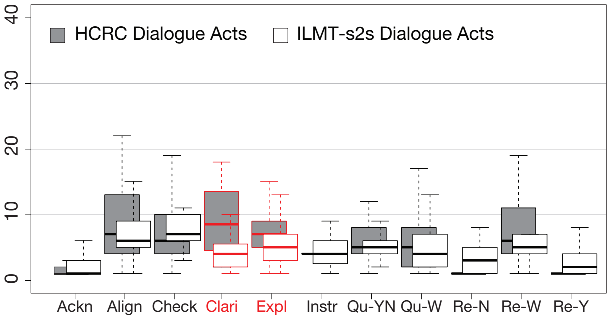


Figure 7: *Word Count Boxplot of IF Dialogue Acts of Each Corpus — Significantly Different Dialogue Acts Coloured Red*

4.0.2. Dialogue Act Comparison

The frequency of dialogue acts changed dramatically. The top three dialogue acts for the role of IF that make a combined 73% in the HCRC Map Task corpus have only a share of 22% in the ILMT-s2s corpus with now four acts required to obtain the same 73% share. A similar phenomenon also happens for the role of IG. Apart from the apparent top item of "*Instruct*" the next four items which make a combined 48% share are all moved lower than fifth position in the ILMT-s2s corpus (Figure 8).

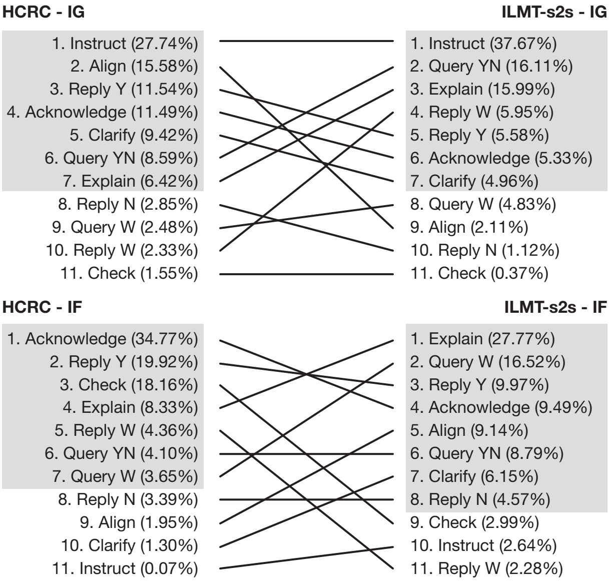


Figure 8: *Change in Dialogue Act Frequency between Corpora*

This change of order has also indicates an increase in the number of high word count dialogue acts being more frequently spoken (Table 2).

5. Discussion and Future Work

Before going into details about the results of this study, it must first be noted that the annotation of the HCRC Map Task corpus and the ILMT-s2s corpus were not done by the same people and unfortunately an inter-coder agreement can not be verified since the annotators of the ILMT-s2s corpus did not annotate any of the HCRC Map Task corpus files to make an inter-coder comparison possible. Therefore there remains the question about the agreement level of the dialogue act annotations between the two corpora. However we are optimistic that the results presented in this paper show an accurate illustration of computer mediated multi-language communication.

Table 2: *Top Acts of each Corpus*

Role	Acts	HCRC	Acts	ILMT-s2s
IG	Instruct	$\mu = 13.26$	Instruct	$\mu = 7.76$
	Align	$\mu = 4.62$	Query YN	$\mu = 5.78$
	Reply Y	$\mu = 1.87$	Explain	$\mu = 4.78$
	Acknow.	$\mu = 1.75$	Reply W	$\mu = 5.06$
IF	Acknow.	$\mu = 1.72$	Explain	$\mu = 5.45$
	Reply Y.	$\mu = 1.72$	Query W	$\mu = 4.86$
	Check	$\mu = 7.27$	Reply Y	$\mu = 2.73$
	Explain	$\mu = 8.08$	Acknow.	$\mu = 1.93$

Due to the characteristics of the ILMT-s2s system where the subject speaks to the ILMT-s2s system, the system converts the speech to text, translates the text to the target language, and then the system sends the translation to the interlocutor's computer for synthesised output, inter-cooperation between the subject and the interlocutor — that results in spreading evenly the cognitive effort required in communication — is no longer as efficient as would be the case in face-to-face communication. Acknowledgement related feedback has little meaning since the interlocutor already has to listen to the whole utterance of the subject. This phenomenon can be seen by the more Gricean interaction that is forming with the higher usage of dialogue acts that use a higher word count from the IG and IF. In this case it might be presumed that visual contact seems more important as a replacement of the acknowledgement related feedback, but previous studies [10] have presented that subjects of the ILMT-s2s corpus had a better perspective of the communication method without visual contact.

From this starting point with this data, it will be interesting to investigate further how the subject repair errors within their communication — the self correction of utterances that resulted in ASR errors, the cooperation from the interlocutor following repeated ASR error correction attempts, etc. — and how this compares/differs from face-to-face task oriented communication.

6. 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 Trinity College Dublin.

7. References

[1] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[2] A. Hayakawa, S. Luz, L. Cerrato, and N. Campbell, "The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA), May 2016, pp. 605–612.

[3] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “Elan: a professional framework for multimodality research,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*. Genoa, Italy: European Language Resources Association (ELRA), 2006, pp. 1556–1559. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf

[4] K. Sjölander and J. Beskow, “Wavesurfer - an open source speech tool,” in *Proceedings of the 6th International Conference on Spoken Language Processing*. Beijing, China: ISCA, 2000, pp. 464–467. [Online]. Available: http://www.isca-speech.org/archive/icslp_2000/i00_4464.html

[5] J. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J. C. Kowtko, and A. H. Anderson, “The reliability of a dialogue structure coding scheme,” *Computational linguistics*, vol. 23, no. 1, pp. 13–31, 1997. [Online]. Available: http://dl.acm.org/citation.cfm?id=972684.972686

[6] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, “The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena,” *Language Resources and Evaluation*, vol. 41, no. 3, pp. 273–287, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10579-007-9061-5

[7] D. Oppermann, F. Schiel, S. Steininger, and N. Beringer, “Off-talk-a problem for human-machine-interaction?” in *Proceedings of INTERSPEECH’01: the 2nd Annual Conference of the International Speech Communication Association*. Aalborg, Denmark: ISCA, 2001, pp. 2197–2200. [Online]. Available: http://www.isca-speech.org/archive/eurospeech_2001/e01_2197.html

[8] A. Hayakawa, F. Haider, S. Luz, L. Cerrato, and N. Campbell, “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, 2016, pp. 776–780. [Online]. Available: http://www.isca-speech.org/archive/sp2016/pdfs_stamped/136.pdf

[9] A. Batliner, C. Hacker, and E. Nöth, “To talk or not to talk with a computer: On-Talk vs. Off-Talk,” in *Proceedings of the Workshop on How People Talk to Computers, Robots, and Other Artificial Communication Partners*, K. Fischer, Ed., Hansewissenschaftskolleg, Delmenhorst, Germany, 2006, pp. 79–100.

[10] L. Cerrato, A. Hayakawa, N. Campbell, and S. Luz, *Future and Emergent Trends in Language Technology: First International Workshop, FETLT 2015, Seville, Spain, November 19-20, 2015, Revised Selected Papers*. Cham: Springer International Publishing, 2016, ch. A Speech-to-Speech, Machine Translation Mediated Map Task: An Exploratory Study, pp. 53–64. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-33500-1_5