LREC 2018 Workshop

LB-ILR2018 and MMC2018 Joint Workshop:

Language and Body in Real Life
Multimodal Corpora 2018: Multimodal Data in the Online World

PROCEEDINGS

Edited by

Hanae Koiso, Patrizia Paggio

ISBN: 979-10-95546-16-0
EAN: 9791095546160

7 May 2018

Proceedings of the LREC 2018 Workshop “LB-ILR2018 and MMC2018 Joint Workshop”
7 May 2018 – Miyazaki, Japan

http://real-life-interaction.jdri.org/jointws2018/

Organising Committee

Language and Body in Real Life

Yasuharu Den	Chiba University, Japan
Tomoko Endo	Seikei University, Japan
Kristiina Jokinen	National Institute of Advanced Industrial Science and Technology, Japan
Hanae Koiso	National Institute for Japanese Language and Linguistics, Japan
Ryoko Suzuki	Keio University, Japan
Sandra Thompson	University of California, Santa Barbara, USA
Anna Vatanen	University of Oulu, Finland

Multimodal Corpora

Patrizia Paggio	University of Copenhagen, Denmark / University of Malta, Malta
Kirsten Bergmann	University Bielefeld / University Osnabrück, Germany
Jens Edlund	KTH Royal Institute of Technology, Sweden
Dirk Heylen	University Twente, The Netherlands

Programme Committee

Language and Body in Real Life

Yasuharu Den	Chiba University, Japan
Tomoko Endo	Seikei University, Japan
Kotaro Funakoshi	Kyoto University and HRI-JP, Japan
Kristiina Jokinen	AIST, Japan
Mary Kim	University of Hawaii, USA
Hanae Koiso	NINJAL, Japan
Jarkko Niemi	University of Helsinki, Finland
Tsuyoshi Ono	University of Alberta, Canada
Mirka Rauniomaa	University of Oulu, Finland
Ryoko Suzuki	Keio University, Japan
Sandra Thompson	University of California, Santa Barbara, USA
Khiet Truong	University of Twente, The Netherlands
I-Ni Tsai	National Taiwan University, Taiwan
Anna Vatanen	University of Oulu, Finland

Multimodal Corpora

Jens Allwood	University of Göteborg, Sweden
Jan Alexandersson	DFKI Saarbrücken, Germany
Philippe Blache	LPL - CNRS and Université d'Aix-Marseille, France
Susanne Burger	Carnegie Mellon University, USA
Kristiina Jokinen	AIST, Japan
Bart Jongejan	Copenhagen University, Denmark
Maria Koutsombogera	Trinity College Dublin, Ireland
Sebastian Loth	Bielefeld University, Germany
Costanza Navarretta	Copenhagen University, Denmark
Catherine Pelachaud	CNRS at ISIR and UPMC, Framc
Ronald Poppe	Utrecht University, The Netherlands
Albert Ali Salah	Bogazicy University, Turkey
David Traum	University of Southern California, USA

Preface

This proceedings collects the papers summaries of the oral and poster presentations from two workshops that joined hands at LREC 2018 to provide the respective research communities with a larger audience. The two original workshops were Language and Body in Real Life and Multimodal Corpora: Multimodal Data in the Online World. While the two workshops focused on different aspects of research in multimodal communication, they also shared the same fundamental interest in how speech and body are used in human communication. We believe that the programme of the joint event combines the final contributions into an interesting and varied whole.

Patrizia Paggio and Hanae Koiso

Programme

14:00 – 14:05 Opening

14:00 – 15:25 Language and Body in Real Life I

Hanae Koiso, Yasuyuki Usuda, Haruka Amatani, Yoshiko Kawabata, and Yasuharu Den
Design and Preliminary Analysis of the Corpus of Everyday Japanese Conversation

Hiroko Tokunaga, Masaki Shuzo, and Naoki Mukawa
Preliminary Analyses of Spatial Positions of Poster Session Audience and Their Joining in/Leaving Behaviors

Mizuki Koda
Language and Body as Resources for Distributing Orientation: The Organization of Participation in Leaving the Ongoing Conversation

Rui Sakaida and Yasuharu Den
Sitting Down and Standing Up as Resources for Reorganization of Participation Framework: Analysis of Preparatory Meeting for Nozawa Onsen Fire Festival

15:30 – 16:30 Poster Session (including coffee break)

Emanuela Cresti, Lorenzo Gregori, Massimo Moneglia, and Alessandro Panunzi
The Language into Act Theory: A Pragmatic Approach to Speech in Real-Life

Hajime Murai
Factor Analysis of Japanese Daily Utterance Styles

Yasuyuki Yoshida, Takuichi Nishimura, and Kristiina Jokinen
Biomechanics for Understanding Movements in Daily Activities

Yasuharu Den
F-formation and Social Context: How Spatial Orientation of Participants' Bodies Is Organized in the Vast Field

Patrizia Paggio and Costanza Navarretta
Temporal Coordination of Facial Expressions and Head Movements in First Encounter Dialogues

Soumia Dermouche and Catherine Pelachaud
Expert-Novice Interaction: Annotation and Analysis

16:30 – 16:50 Language and Body in Real Life II

Saori Daiju and Ono Yoshi
Studying Japanese Distal Demonstrative 'are' Using Video Corpus

16:50 – 17:50 Multimodal Corpora

Ingo Siegert, Julia Krüger, Olga Egorow, Jannik Nietzold, Ralph Heinemann, and Alicia Lotz
Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon ALEXA

Justine Reverdy, Akira Hayakawa, and Carl Vogel
Alignment in a Multimodal Interlingual Computer-Mediated Map Task Corpus

Kristiina Jokinen
Conversational Gaze Modelling in First Encounter Robot Dialogues

17:50 – 18:00 Closing

Table of Contents

Design and Preliminary Analysis of the *Corpus of Everyday Japanese Conversation*

Hanae Koiso[†], Yasuyuki Usuda[†], Haruka Amatani[†], Yoshiko Kawabata[†], Yasuharu Den[‡,†]

[†] National Institute for Japanese Language and Linguistics
10-2 Midoricho, Tachikawa, Tokyo 190-8561, Japan
{koiso, usuda, h-amatani}@ninjal.ac.jp

[‡] Graduate School of Humanities, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, Japan
den@chiba-u.jp

Abstract

Conversations emerge in various ways in everyday life. To capture the diversity of real-life conversations, we started the compilation of a large-scale corpus of everyday Japanese conversation, the *Corpus of Everyday Japanese Conversation*, CEJC. The CEJC is designed to contain various kinds of everyday conversations in a balanced manner so as to capture the diversity of everyday conversations and to observe natural conversational behavior. The CEJC targets conversations embedded in naturally occurring activities in daily life, without the exogenous intervention by researchers imposing topics or displacing the context of action. Since the start of the project in 2016, we have compiled 94 hours of conversations in the CEJC, corresponding to about a half of the target size of the entire corpus, and have morphologically annotated 38 hours of data. In this paper, we first outline the design of the CEJC including corpus size, recording methods, and annotations to be included in the corpus. Then, we conduct a preliminary analysis on some linguistic aspects of the corpus, based on the morphologically annotated data, showing that the CEJC captures the diversity of real-life conversations.

Keywords: Corpus of everyday Japanese conversation, corpus design, corpus analysis, linguistic aspects, morphological annotation

1. Introduction

Conversations emerge in various ways in everyday life. To capture the diversity of real-life conversations, we have to collect a variety of conversations occurring in natural settings in our daily life. Yet, most of the corpora constructed so far have targeted conversations in artificially created settings in terms of topics and recording situations, such as task-oriented, experimental dialogs or chats among participants recruited for recording purpose.

To overcome these undesirable circumstances, we started the compilation of a large-scale corpus of everyday Japanese conversation, the *Corpus of Everyday Japanese Conversation*, CEJC (Koiso et al., 2018). The CEJC is designed to contain various kinds of everyday conversations in a balanced manner so as to capture the diversity of everyday conversations and to observe natural conversational behavior. The CEJC targets conversations embedded in naturally occurring activities in daily life, without the exogenous intervention by researchers imposing topics or displacing the context of action (Mondada, 2012). Since the start of the project in 2016, we have collected more than 400 hours of conversations and compiled 94 hours out of them in the CEJC, corresponding to about a half of the target size of the entire corpus. Among them, 38 hours of data have also been morphologically analyzed.

In this paper, we first outline the design of the CEJC including corpus size, recording methods, and annotations to be included in the corpus. Then, we conduct a preliminary analysis on some linguistic aspects of the corpus, based on the morphologically annotated data, showing that the CEJC captures the diversity of real-life conversations.

2. Design of the CEJC

Corpus size We plan to publish more than 200 hours of conversations. Based on the data we have recorded, transcribed, and morphologically annotated so far, the total number of words, conversations, and conversants are estimated at 2.1 million words, 400 conversations, and a total of 1200 conversants, including 600 different participants.

Recording In order to record convesations embedded in naturally occurring activities in daily situations, we mainly adopt a recording method called the *individual-based* method. In this method, we recruit 40 informants balanced in terms of sex and age (man/woman $\times$ 20s/30s/40s/50s/over 60 $\times$ 4 informants), provide them with portable recording devices for approximately two to three months, and have them record about 15 hours of conversations in their daily activities. The informant carries portable recording devices and record his/her everyday activities in a variety of situations such as at home, at a restaurant, and outdoors. The project members do not mediate their field recordings. About four to five hours of conversations, among 15 hours, per informant, i.e., total of about 180 hours, are selected for the CEJC by taking into account the balance of conversation variations, quality of recorded data, and legal and ethical issues.

In order to precisely understand the mechanism of our real-life social conduct, not only audio but also video data are collected and published. Two types of compact action cameras, Kodak PIXPRO SP360 4K and GoPro Hero 3+, are used when recording indoors (Koiso et al., 2016a). Figure 1 shows video images of a conversation between a customer and a barber at a barbershop.

Figure 1: Video images of a conversation between a customer and a barber at a barbershop.

Annotation The speech is manually transcribed, and two types of POS information, short–unit word and long–unit word, are automatically annotated. The *Core* data set, consisting of 20 hours of conversations, is designed to contain manually corrected POS information and manually annotated dependency structures, dialog acts, and intonation labels.

See Koiso et al. (2018) for the details of the design of the CEJC.

3. Preliminary Analysis

So far, we have compiled 94 hours of conversations in the CEJC. Based on this data, Koiso et al. (2018) provided a preliminary evaluation on the issue of balancedness by reference to the survey results of everyday conversational behavior described in Koiso et al. (2016b). In this paper, on the other hand, we conduct a preliminarily analysis on some linguistic aspects of the CEJC based on the morphologically annotated data. As a result, we show that the CEJC captures the diversity of real-life conversations.

The data analyzed here amounts to about 38 hours, which consists of 81 conversations, and contains a total of 159 conversants, including 275 different participants.

3.1. Variance in the number of words

In this section, we focus on the number of words in a unit of time. Since the CEJC is designed to target conversations in various activities that occur naturally in daily life, participants may, on one occasion, be involved in an activity in which they do not have to speak all the time such as cooking, and, on another occasion, engage in an activity that is mainly conducted verbally such as business meetings. Therefore, the diversity of activities involved in the CEJC may result in a wider range of numbers of words compared to corpora recorded in restricted situations.

We compared the distribution of the words per minute (WPMs) in the CEJC with those in the *Chiba Three-party Conversation Corpus* (Chiba3Party) (Den and Enomoto, 2007) and in casual conversations included in the *Corpus of Spontaneous Japanese* (CSJ) (Maekawa, 2004). Conversational data in the Chiba3Party and the CSJ were recorded in rather artificial settings. The Chiba3Party consists of 12 conversations among university friends recruited for recording purpose, and amounts to 2 hours. Casual conversations in the CSJ includes 16 dyadic conversations between recruited interviewers and informants, and amounts to 1 hour. The WPM for each conversation in each corpus was calculated as the number of words in the conversation divided by the duration, in minute, of the conversation.

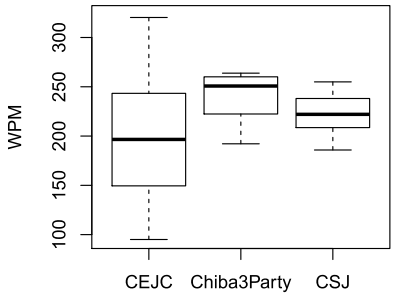


Figure 2: Distributions of the words per minute in the CEJC, the Chiba3Party, and the CSJ

Figure 2 shows the distributions of the WPMs in the three corpora. It is found that the mean of the WPM is smaller and the variance of the WPM is larger in the CEJC than in the Chiba3Party and the CSJ. In the Chiba3Party and the CSJ, the participants were recruited for recording purpose, and, thus, might feel pressure to continue speaking, resulting in greater and condensed WPMs. In the CEJC, by contrast, participants often engage in other activities while having a conversation, such as eating, cleaning, cooking, and doing homework. In these situations, they sometimes concentrate on non-conversational activities without talking. The larger variance of the WPM in the result might suggest that the CEJC captures the diversity of real-life conversation embedded in naturally occurring activities.

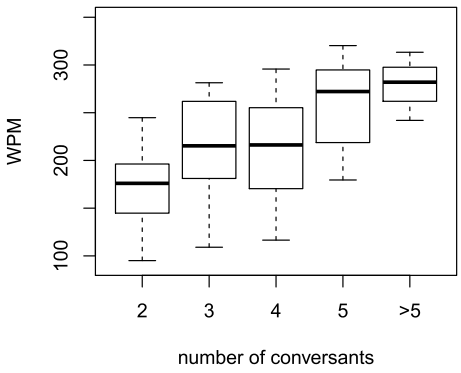


Figure 3: Distributions of the words per minute according to the number of conversants

There may be, however, another factor that could account for the larger variance in the CEJC. The numbers of conversants in the CSJ and the Chiba3Party are fixed at two and three, respectively, while the CEJC has much variety in the number of conversants. Since the number of conversants may have an influence on the variance of the WPM, we next investigate the distributions of the WPMs in the CEJC according to the number of conversants.

Figure 3 shows the result. It is found that the WPM increases when the number of conversants increases. This tendency would predict that the WPM in the CEJC, on average, is larger than those of the CSJ and the Chiba3Party, because the CEJC contains conversations with more than three conversants, which are never included in the CSJ or the Chiba3Party. What we saw in Figure2, however, is opposite to this prediction; the median of the WPM is smaller in the CEJC than in the CSJ and the Chiba3Party. This result reinforces the argument that conversants in the CEJC sometimes concentrate on non-conversational activities without talking, which is one of the key features of real-life conversations.

Increase in the WPM with increasing number of conversants in Figure 3 might be caused by response tokens, such as "hai" and "un," produced by listeners during a speaker's utterance, since there would be more response tokens in a conversation with more conversants. The tendency, however, is retained even after response tokens, as well as other interjections, have been removed from the data. One reason for this might be that when there are more than three conversants, they can be divided into two or more groups to have a separate talk (Sacks et al., 1974). This result suggests that the number of conversants could affect conversational structures.

Conversations with four or more conversants account for about 35% of the data collected so far, and the whole corpus maintains the balancedness of the numbers of conversants by reference to the survey results of everyday conversational behavior (Koiso et al., 2018), i.e., at least 24% for more-than-three-party conversations.

3.2. Proportion of polite forms

In Japanese, polite and non-polite forms are expected to be distinguished in a proper way, according to the conversants' relationships and speaking situations. If the CEJC captures a broad range of conversational situations in real life, various patterns of the use of polite and non-polite forms between conversants would be observed. In this section, we investigate how conversants use polite and non-polite forms differently depending on their relationships.

In this analysis, we used 16867 utterance units (Den et al., 2010) containing verbs or adjectives as their main predicates. Utterance units with auxiliary verbs "desu," "masu," and "gozai-masu" in their predicates were defined as polite forms. We also distinguished utterance units with and without final particles "ne," "yo," "na," "sa," "wa," and "no," because some studies pointed out that polite forms with final particles show lower degree of formality (Ijuin, 2004; Ogi, 2014; Satake, 2016). Utterance units for which the addressee's relation to the speaker cannot be uniquely determined were excluded from the analysis; for instance, in a conversation with a teacher and two or more students, the teacher's utterances were included in the analysis, but the students' utterances were excluded because they may be addressed to either the teacher or other students, i.e., friends.

Figure 4 shows the proportions of polite and non-polite forms according to the conversants' relationship, i.e., the addressee's relation to the speaker. First, we take a look at the general tendency of the use of polite and non-polite forms. It is found that the proportion of the use of polite forms is the highest when talking with customers, and the second highest when talking with business connections. This suggests that people tend to speak more formally when they are talking in business situations. In other cases, familiarity and power relation among participants may influence on how politely they speak. For instance, people use more polite forms to their acquaintances than to their family members and friends. Similarly, senior and similar-age colleagues at an office are more likely to be spoken in a polite way than junior colleagues, and teachers are more likely to be addressed politely than students, presumably because of asymmetric power relations.[1]

Let us, next, glance at characteristics of utterance units with and without final particles. Overall, final particles are employed more often with non-polite forms, with a notable exception of conversation between teachers and students, which shows relatively high likelihood of the use of polite forms with final particles. In the data observed in this analysis, there were only four teacher–student conversations, and they were all informal chats outside the classroom. This may lead more use of polite forms with final particles, which is less formal.

Now, we look at our data from another perspective. Focusing on a particular conversant, Informant A, we explored the relation between how often polite forms and final particles are used and with whom the informant is talking (Figure 5). In the case of Informant A, a male professional officer in his 30s, his "family" category was

[1] In the cultural context of Japan, there is a rather clearly asymmetrical power relation between seniors and juniors. This causes juniors usually use polite forms to their seniors, which may be unusual in Western cultures. Note that demands on the use of polite forms also depend on the company's/organization's custom.

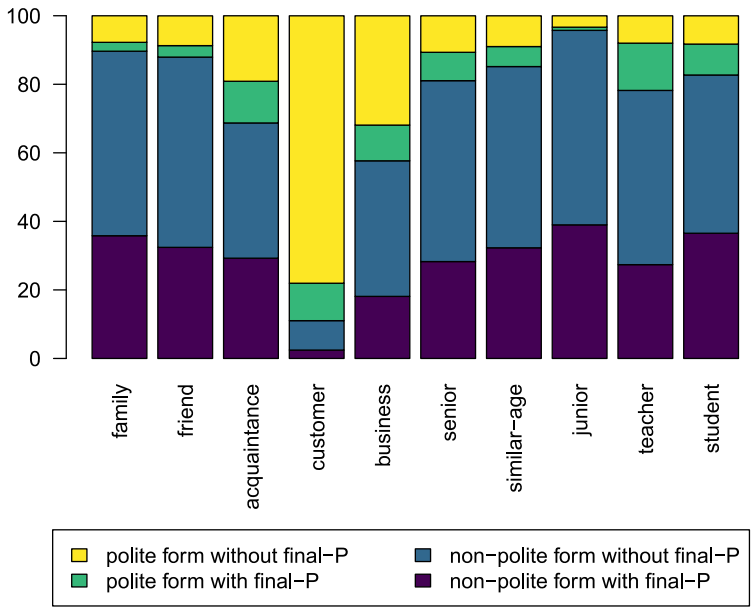


Figure 4: Usage of polite and non-polite forms according to the conversants' relationship

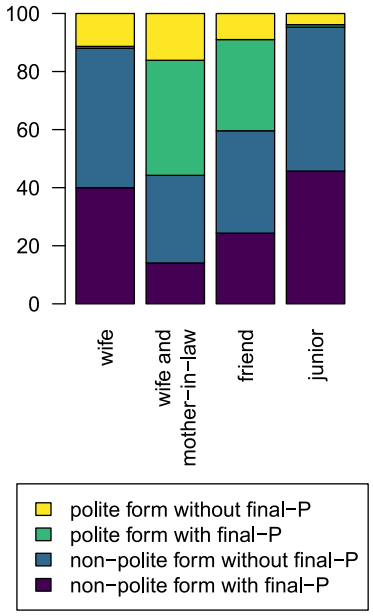


Figure 5: Informant A's usage of polite and non-polite forms according to the relation to the addressees

divided into "wife only" and "wife and mother-in-law (his wife's mother)." It is found that Informant A tends to use far more non-polite forms when talking with his wife and junior colleagues than other people. Polite forms are less likely to be employed in very close relationship or when he is in power. On the other hand, he speaks using polite forms when talking with his mother-in-law. Interestingly, in quite a few cases when talking with his mother-in-law, he uses more polite forms with final particles than those without final particles. This suggests that he downgrades formality so that he avoids a sense of unfamiliarity caused by being too polite.

In this way, the characteristics of the use of polite and non-polite forms reveal the diversity of the CEJC in terms of relationships among participants, which is another key feature of real-life conversations.

4. Conclusions

In this paper, we outlined the design of the *Corpus of Everyday Japanese Conversation*, which we have been constructing since 2016. We also presented some results on a preliminary analysis based on the morphologically annotated data from linguistic aspects, such as the variance in the number of words according to the number of conversants and the proportion of polite forms according to conversants' relationships. We pointed out that the CEJC covers conversations in various situations in real life. We plan to publish a part of the CEJC, about 50 hours, on a trial basis in 2018, and the entirety in 2022.

5. Acknowledgments

The work reported in this article is supported by the NINJAL collaborative research project "A Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation."

6. Bibliographical References

Den, Y. and Enomoto, M. (2007). A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In T. Nishida, editor, *Conversational informatics: An engineering approach*, pages 307–330. John Wiley & Sons, Hoboken, NJ.

Den, Y., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M., and Yoshida, N. (2010). Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of LREC 2010*, pages 2103–2110, Valletta, Malta.

Ijuin, I. (2004). A comparison of speech style adaptation between native situations and contact situations (in Japanese). *The Japanese Journal of Language in Society*, 6(2):12–26.

Koiso, H., Tanaka, Y., Watanabe, R., and Den, Y. (2016a). A large-scale corpus of everyday Japanese conversation: On methodology for recording naturally occurring conversations. In *Proceedings of LREC 2016 Workshop: Just talking — Casual talk among humans and machines*, pages 9–12, Portoroz, Slovenia.

Koiso, H., Tsuchiya, T., Watanabe, R., Yokomori, D., Aizawa, M., and Den, Y. (2016b). Survey of conversational behavior: Towards the design of a balanced corpus of everyday Japanese conversation. In *Proceedings of LREC 2016*, pages 4434–4439, Portoroz, Slovenia.

Koiso, H., Den, Y., Iseki, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y., and Usuda, Y. (2018). Construction of the Corpus of Everyday Japanese Conversation: An interim report. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Maekawa, K. (2004). Design, compilation, and some preliminary analyses of the *Corpus of Spontaneous Japanese*. In K. Yoneyama and K. Maekawa, editors, *Spontaneous speech: Data and analysis*, pages 87–108. The National Institute for Japanese Language and Linguistics, Tokyo.

Mondada, L. (2012). The conversation analytic approach to data collection. In J. Sidnell and T. Stivers, editors, *The handbook of conversation analysis*, pages 32–56. Wiley-Blackwell, Hoboken, NJ.

Ogi, N. (2014). Language and an expression of identities: Japanese sentence-final particles *ne* and *na*. *Journal of Pragmatics*, 64:72–84.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Satake, K. (2016). Nichijo-danwa-ni mirareru keigo shiyono jittai (An empirical study on honorifics in everyday discourse) (in Japanese). In Gendai Nihongo Kenkyukai, editor, *Danwa-shiryo: Nichijo-seikatsu-no kotoba (Discourse material: Language in everyday life)*, pages 191–212. Hitsuji Shobo, Tokyo.

Preliminary Analyses of Spatial Positions of Poster Session Audience and Their Joining in/Leaving Behaviors

Hiroko Tokunaga, Masaki Shuzo, Naoki Mukawa*
School of System Design and Technology, Tokyo Denki University
5 Senju Asahi-cho, Adachi-ku, Tokyo 120-8551, Japan
*mukawa@mail.dendai.ac.jp

Abstract

The aim of our research is to construct criteria for evaluating participant behaviors that will help ensure fruitful communication experiences in poster sessions. We recorded on video the proceedings of a simulated poster session. Then, we analyzed the audience' spatial position at each poster presentation and their joining in/leaving behaviors. The results highlighted the key behaviors of an audience in the front row of a poster presentation that wanted to leave from the discussion, and those of an audience in the rear row that wanted to join in the discussion. These findings suggest that it would be effective to provide an encouragement that helps audiences to join in/leave from presentations suited to the situations.

Keywords: interactive communication, communication skill

1. Introduction

One format for presenting research at academic forums and conferences is the poster session. Presenters and attendees stand close to each other in which the latter can interact with the former during the presentation. Attendees can also observe tentatively the poster from a short distance, or walk away if they are uninterested. During a poster session, attendees choose those posters that arouse their interest or concern; as such, a poster session serves as a discussion forum.

However, when poster presentations attract many attendees, it can be difficult for attendees to join in a discussion once the presentation has started. Conversely, if a discussion drags on, those wishing to move to another poster presentation may find it difficult to leave from the scene.

Thus, poster sessions feature a mixture of attendees: those who want to listen attentively, those who want to observe tentatively, those who want to join in the discussion, and those who want to leave from the discussion. Therefore, to ensure fruitful discussions at poster sessions, presenters must be competent in both managing the floor and presenting. For their part, attendees must possess communications skills that will allow them to join in and leave from presentations suited to the occasion. Few studies have explored the specific behaviors of presenters and attendees in the dynamically changing environment of a poster session; these studies have derived criteria for evaluating the abovementioned skills from such data.

Therefore, we attempted to derive criteria for evaluating the above behavioral skills. This study aims to clarify the attendees' spatial positions and behaviors during poster presentations, with a focus on audience drop in/out process. We set up a laboratory to simulate poster presentations, and recorded on video the proceedings of the presentations. We used the video footage to analyze the attendees' behaviors during the presentation. We also examined specific examples that illustrated the behaviors of attendees exhibiting when they want to "join in" or "leave."

2. Previous Work

When people of a specific group converse with each other, the participants therein are aware of their own participation status and that of the others. Goffman (1981) proposed a "participation framework" for analyzing the interactional roles in conversations involving three or more people. This participation framework classifies participants into "speakers," "addressees," and "side participants," according to the centrality of their role in the conversation. Clark (1996) elaborated on this framework, adding the roles of "bystander" and "eavesdropper." Bono (2004) then applied Clark's Goffmanian model to the context of a dynamic poster presentation. In a poster presentation, the way the audience participates changes over time. Accordingly, when applying the model in this context, describing the model diachronically is important. Bono (2004) described participation diachronically as follows: "Nonparticipants" become "bystanders" when they approach intentionally the conversational space and are recognized by the existing participants. Once they are recognized by all existing participants, they become "side participants." Then, when a "speaker" addresses them, they become "addressees," and when they address an existing participant, they become "speakers."

In a conversational scene involving many people, it is not only the structure of participation framework that comes into play; another important element is spatial organization, which refers to the relative spatial positions and orientations of the participants. Kendon (1990) proposed the "F-formation" as a concept for describing spatial organization in conversational scenes involving three or more participants. The F-formation describes three kinds of functional spaces that extend outward from the participants. The first is the orientation space (o-space), which is the central space formed in front of the individuals who are engaging each other in a conversation. The second is the participants' space (p-space), which is a ring-shaped space surrounding the o-space. Then, there is the region space (r-space), which lies beyond the p-space.

McNeill (2006) used Kendon's F-formation system and broke the concept down into social and instrumental F-formation. The latter refers to conversational space in

which communications are mediated through an object. According to Bono (2009), poster presentations typically have an instrumental F-formation, in that participants gaze at a poster. Bono also argued that the spatial organization typically consists of a semicircular alignment, and that the o-space in this configuration becomes smaller as people draw closer to the poster.

Previous research on poster sessions have focused on the ways audiences change over time and the structure of the interactional relations embedded in the conversation. Our current goal is to use these insights to construct criteria for evaluating the behavioral skills of poster session participants. As a first step toward this goal, we analyzed participants' joining in/leaving behaviors in the video-recorded poster sessions. We believe that the findings derived from this analysis will help enhance poster presentation skills in research conferences, business meetings, and educational settings.

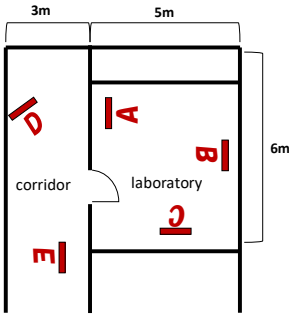


Figure 1: Arrangement of five posters in the experimental area.

Figure 2: A scene of Poster B presentation. Movies for analysis are edited from four angle cameras (overhead, backward, left, and right).

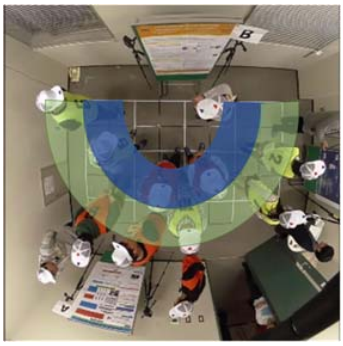


Figure 3: A position of each audience is labeled as p-space (blue) and r-space (green).

3. Experimental Setup of Poster Presentation

An experimental poster session was set up at our laboratory of Tokyo Denki University. The poster session comprised five presenters, 19 attendees.

Presenters were one assistant professor, three graduate students, and one undergraduate of 4th grade. Poster A, which was presented by a graduate student, and Poster B, by an assistant professor, were analyzed. These two presenters were selected because of their abundant presentation experiences and high presentation skills. Before the session, they were instructed that the presentations would be simulated academic conferences or symposiums, must be completed in about 10 minutes, including discussions, and communicate proactively to attendees who were interested.

All attendees were university students aged between 20 and 24 years. They obtained informed consent. The poster session lasted for 40 minutes. We instructed attendees to attend each of the five presentations within the 40-minute period.

We also had one facilitator who was responsible for prompting the attendees to join in and leave for temporary period of experimental session. We, however, excluded the facilitator from our analysis in this study.

Experimental layout is shown in Figure 1. There were five poster presentations. Three of these (Posters A, B, and C) were in the room area, and two (Posters D and E) were in the corridor area (see Figure 1).

Twelve video cameras were set up for recording. For Posters A, B, and C in the laboratory area, we set up cameras to the left and right sides of each of three poster panels and on the ceiling right above the panel and rear upper wall. We then edited the video footage to prepare it for analysis (see Figure 2). As for Posters D and E, no video was recorded.

4. Annotation Method for Video Data

The behaviors of attendees who visited posters were annotated from the edited video data. Two posters of A and B were discussed for a preliminary study. Other posters will be discussed in the future. We annotated the spatial positions of attendees, their movements, and postural configurations, among others, by using the free software ELAN.[1]

When an attendee was gazing at a poster in a stationary position, he/she was assigned as the audience of that poster. As shown in Figure 3, two labels were indicated for the spatial alignment of the attendees: the front row where the p-space is formed (shown in blue in the figure), and the rear row where the r-space is formed (shown in green).[2]

Subsequently, the movements and postural configurations of attendees for Posters A and B were annotated. We indicated several labels for head tilting, head turning to side, gaze, body inclination, body twisting, arm and hand posture, and leg motion. The annotated labels were provided by one of the authors.

[1] ELAN : https://tla.mpi.nl/tools/tla-tools/elan/

[2] F-formation is a concept of interpreting the spatial configuration of conversational scenes from the aspect of interaction. In this study, to simplify interpretations, we regarded the front row of participants as p-space and the rear row as r-space.

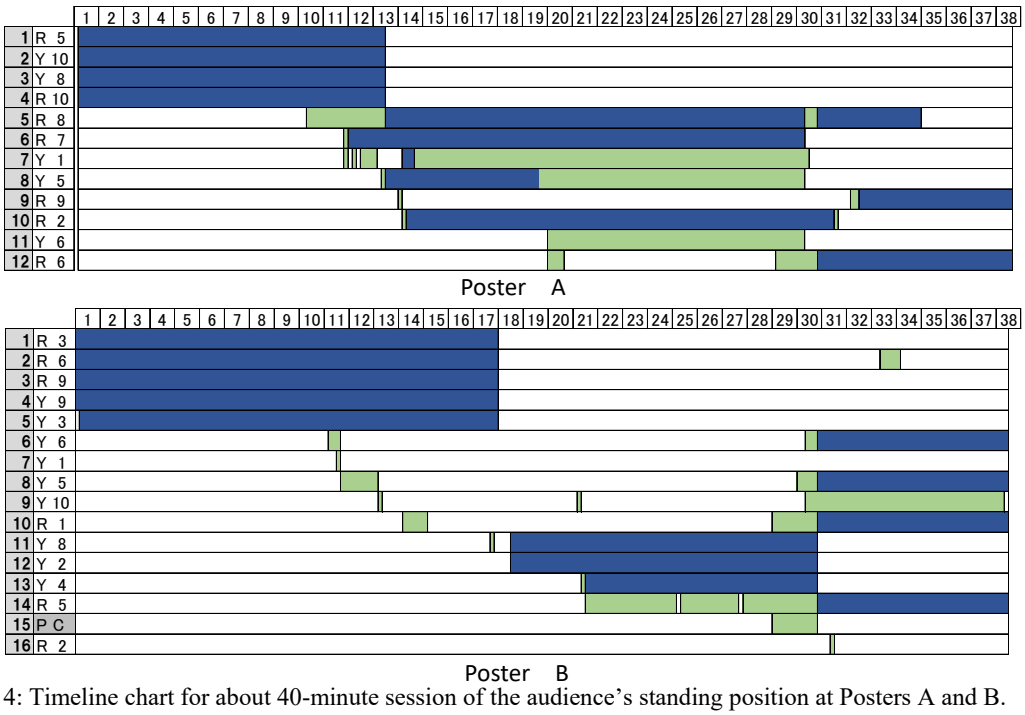


Figure 4: Timeline chart for about 40-minute session of the audience's standing position at Posters A and B.

5. Audience's Spatial Positions

Figure 4 shows a timeline of the audience's positions in Posters A and B, in which the horizontal and vertical axes indicated elapsed time (minutes) and attendee ID (bib color and number), respectively. In the table, a blue and a green bar indicated the time that the attendee stood in the front row (p-space) and the rear row (r-space), respectively.

The presenters of Posters A and B delivered their presentation for about three times. Each presentation had different audiences. In the first presentations of both Posters A and B, the attendees standing in the front row remained in this position for the duration of this presentation. In the second and last presentations, other attendees joined in or out during the presentations.

6. Behavior Analysis

In this chapter, we analyzed the attendees' behaviors in each row separately, to determine whether the front-row attendees engaged in the discussion throughout the presentation, and the circumstances in which the rear row attendees joined in/ out.

6.1 Typical Attendees in P-space

When we analyzed the behaviors of the attendees, we found that many in the front row were nodding, leaning forward, and looking at the spots on the poster to which the presenter was pointing (see Figure 5(a)). We assume that the action of standing in the front row is an indicative of an attendee's desire to hear the presentation or his/her signal of curiosity in the presentation. We also assume that actions of leaning forward or nodding at appropriate moments are expressions of concern and curiosity.

Such behaviors probably help audiences attract the gaze of the presenter and obtain opportunities to ask questions

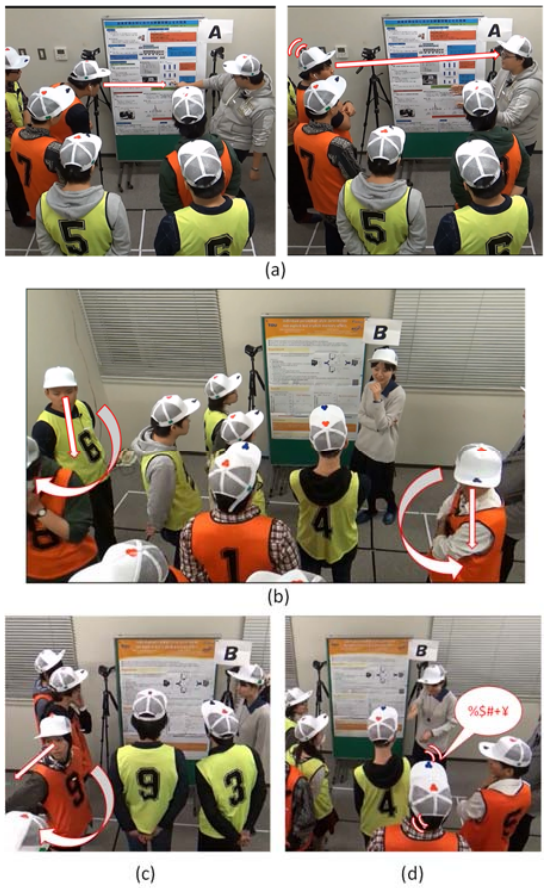


Figure 5: Example behaviors are shown at around 28(a), 28(b), 16(c), and 30(d) minutes of presentations.

or raise comments. Thus, the characteristics of an audience that is involved actively in the presentation are exhibited in the behaviors of the front-row audiences.

6.2 Typical Attendees in R-space

As for the behaviors among rear-row audiences, at times, they watched the presenter, and at other times, they exhibited a postural configuration called "body torque" (Schegloff, 1998), in which they twisted their upper body and gazed around (see Figure 5(b)). Attendees who stood in the rear row were farther away from the presenter; hence, they were less likely to attract the gaze of the presenter. Additionally, with front-row audiences obstructing their view, they might have found it hard to view the poster. Thus, it is difficult for rear-row audiences to engage actively in the discussion, making moving to another poster a preferable action for them.

Takanashi (2016) reported that the more peripheral a participant's spatial position, the more likely he or she is to shift attention to a different activity. Likewise, in our examples, the behaviors of the rear-row audiences, in which they exhibited body torque and turned their faces toward other posters and presenters, are presumably the typical behaviors that audiences exhibit when they are wondering whether to continue listening to the presentation or move to another presentation.

6.3 Transitional Audiences Between P- and R-space

In certain cases, a front-row audience that was engaging actively in the discussion started to exhibit behaviors similar to those of a rear-row audience; conversely, there were cases in which a rear-row audience started exhibiting behaviors similar to those of a front-row audience. Figure 5(c) shows an example of the former: a front-row audience of Poster B, R9. At around the 16-minute mark, R9 exhibited body torque and cast his gaze at other posters.

Gazing at the presenter is an appropriate action for an audience; conversely, averting one's gaze from the presenter signals "disengagement from the conversation" (Goodwin, 1981). Sakaida (2017) observed interactions while standing between organizers from an excerpt of the recorded video of the preparatory work for traditional fire festival in Japan. He described a stepwise process of leaving from a conversational scene, in which participants first avert their gaze and then start walking away from the scene. In the above example, R9 first averted his gaze from the presenter and then looked toward another poster. This behavior presumably denoted that R9 wanted to leave from the front row and move to another poster.

Figure 5(d) shows an example of the latter: a rear-row attendee of poster B, R1. At around the 29-minute mark, R1 was facing the presenter, nodding frequently and venturing a few comments. In exhibiting behaviors similar to those of front-row audiences, R1 was presumably trying to get the surrounding audiences to approve of his own participation in the discussion.

7. Conclusion

Poster sessions feature a mixture of audiences, each with their own purposes for listening to the presentations. To help ensure fruitful discussions, audiences must join in and leave from discussion circles in such a way that each audience can participate in discussions with multiple poster presenters. With this in mind, we discuss the behaviors of the two cases we highlighted above: R9 and R1.

While standing in the front row, R9 twisted his upper body and gazed round to other areas, signaling that he was leaving from the discussion. However, until the discussion came to an end, R9 never actually moved away from the front row to another poster. Despite signaling his intention to do so, if R9 was unable to leave from the Poster B discussion circle and move to another poster, this situation would have been disadvantageous for him. Therefore, it might need someone to assist R9 to leave from the discussion and engage in another poster discussion. Additionally, it might be necessary to provide a few floor management tips to the presenter of Poster 9, such as how she could have given a nod or similar acknowledging gesture to R9, which would have conveyed her approval of his leaving.

While standing in the back row, R1 gazed at the presenter, nodded frequently, and commented. However, the presenter did not look at him, and the front-row audiences did not look around to acknowledge his presence. If we analyze this case based on Goffman's participation framework, we could say that participation of R1 in the discussion circle was not approved by the other participants. To ensure that someone like R1 can join as a member of the discussion, it is necessary to have the presenter a skill to recognize R1 as an audience and to give all of addressee her presentation. It might also be effective to instruct the front-row audiences (such Y4 and R5) on behaviors, such as making space in the front row. Thus, audiences must be more skilled at joining in and leaving from presentations. At the same time, however, all participants must be skilled at assessing accurately when an audience wishes to join in or leave, and behaving to assist her/him to join or leave.

The behavioral skills referred to above apply to research presentations, but they can also be applied to a wide range of interactive communication scenes, including group discussions.

This study was unable to analyze the participants' actual utterances. The audience's behaviors may be affected by verbal presentation skills of the presenter. Thus, we must analyze relationship between a presenter's speech skills and the audience's joining in and leaving behaviors. We intend to accumulate more case studies, chronologically analyze the audience's joining in/leaving behaviors and their inter-poster movements, and analyze quantitatively the behavior data.

Our next goal is to build skill evaluation criteria that are applicable to specific behaviors of presenter and audience. In the future, we will present insights that can help people improve the way they communicate in business or education settings.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 17K00282. and a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Bibliographical References

Clark, H. H. (1996). *Using Language*, Cambridge University Press.

Goffman, E. (1981). *Forms of Talk*, University of Pennsylvania Press.

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearer*s, Academic Press.

Katsuya, T. (2016). *Conversation and Communication Analysis* (in Japanese), Nakanishiya Publication.

Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*, Cambridge University Press.

Bono, M., Suzuki, N. and Katagiri, S. (2004). Conversation: Do Interaction Behaviors Give Clues to Know Your Interest? (in Japanese), *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society,* 11(3): 214–227.

Bono, M. and Takanashi, K. (eds.) (2009). *Analysis Method of Multi-party Interaction* (in Japanese), Ohmsha.

McNeill, D. (2006). Gesture, Gaze, and Ground. In: Renals, S., Bengio, S. (eds.) *Proceedings of Machine Learning for Multimodal Interaction: Second International Workshop, 2005*. Berlin/Heidelberg: Springer Verlag, pp. 1–14.

Sakaida, R. and Den, Y. (2017). Reorganizing Interactional Bround by Virtue of Spatial Configuration, Allocation of Involvement, and Membership Categorization (in Japanese), *SIG Report of the Japanese Society for Artificial Intelligence*, *SIG-SLUD-80,* pp. 1–6.

Schegloff, E. A. (1998). Body Torque. *Social Research,* 65(3): 535–596.

Language and Body as Resources for Distributing Orientation: The Organization of Participation in Leaving the Ongoing Conversation

Mizuki Koda
Graduate School of Humanities and Studies on Public Affairs, Chiba University
mizu.07181008@gmail.com

Abstract

The purpose of this paper is to describe one aspect of the organization of participation in interaction. Especially, I investigate how participants reorganize their participation by focusing on a transitional phase in interaction. The "transitional phase" in interaction is part of interaction where some changes happen to participants' distribution of orientation. To investigate what happens during the transitional phase, this study focuses on the procedure for leaving the current conversation. Based on detailed analyses of the practice in leaving the current activity, this paper argues that displaying double orientation to the current and next activity is one characteristic feature for accomplishing the smooth transition in an interaction, and that the departure from the ongoing interaction is finely coordinated with talk-in-interaction. Moreover, there is a possibility that the utterance is oriented to the ongoing activity and the bodily movement is oriented to another activity. Finally, I discuss the contribution of distributing orientation to a successful interaction.

Keywords: distribution of orientation, multiparty conversations, multimodality

1. Introduction

This paper aims to illustrate one aspect of the organization of participation in interaction. Especially, I investigate how participants reorganize their participation by focusing on a transitional phase in interaction. By saying "transitional phase," I mean the part of interaction where some changes happen to participants' distribution of orientation. As their orientation becomes stable, a new phase in interaction begins. In this sense, "activity" is defined as a course of action with a certain state of participants' orientation. Therefore, the transitional phase in interaction can be seen as a transition between activities. However, note that the next activity is not planned in advance, and created by the way in which participants change the distribution of their orientation. In what follows, I consider the very moment when participants are reorganizing their participation by distributing orientation differently compared to the previous as a transitional phase in interaction. That is, this paper investigates how conversationalist manage their behavior in the transition between activities.

In order to see the phenomenon, this paper focuses on the procedure for leaving the ongoing interaction. Since the pioneering paper by Schegloff and Sacks (1973), how to accomplish and coordinate closing conversation is a recurrent analytical topic in conversation analytic studies. On this background, Broth and Mondada (2013) has provided insightful observation focusing on walking away activity in interaction between guides and guided persons. They demonstrate that walking away as a coordinated and negotiated practice raises normative expectations among the participants. Paying more attention to the transition between phases of activities, Deppermann, Schmit, and Mondada (2010) have examined how participants collaboratively accomplish a written agenda of a meeting in local interactional work. They argue that the fine-grained multimodal coordination of bodily and verbal resources provides for opportunities of sequentially motivated relevant next actions. The difference between these previous studies and this study is that there are not officially planned activities shared among participants in this data. Of course, the setting of dinner party provides a rough outline of activities, such as having a meal after all the guests arrive there. However, how and when the interaction moves on to the next phase is more depending on the local environment compared to a meeting or a guided walk.

To see what happens in the procedure for leaving the current interaction, this paper deals with the following excerpts observed in face-to-face multiparty conversation among friends: 1) a participant leaves the ongoing conversation to join another one; 2) a participant physically leaves the current interaction to go to another place, and 3) a participant leaves the ongoing interaction by standing up. All the excerpts are part of the video-recording of face-to-face multiparty interactions among seven people having a dinner party at the host's house. For analyzing this data, this study adopts the approach of Conversation Analysis, that is analyzing interaction focusing on the sequential organization of it. The transcript was written by the author following the conventions originally developed by Gail Jefferson (2004). The multimodal description was inspired by Lorenza Mondada (Mondada, 2011; 2014). The symbols used in the transcript are explained in the list below.

The list of symbols

[	a starting point of overlapping talk
(0.0)	silence represented in seconds
(.)	a micro pause.
::	the prolongation or stretching of the sound
hh	audible exhalation
.hh	audible inhalation
( )	inaudible word(s)
(words)	likely possibilities of what was said
* *	delimit descriptions of Ivy's gaze and actions
+ +	delimit descriptions of Doris's gaze and actions
∫∫	delimit descriptions of Thea's gaze and actions
¥ ¥	delimit descriptions of Lucy's gaze and actions
Δ Δ	delimit descriptions of Asa's gaze and actions
*-->	gaze or action described continues across subsequent lines
---*	gaze or action described continues until the same symbol
#im.	the exact point where screen shot has been taken
H	hand
R	right
L	left
UP	upper
P:	actions conducted in a preparation phase of a gesture
S:	actions conducted in a stroke phase of a gesture
R:	actions conducted in a retract phase of a gesture

2. Leaving the Ongoing Interaction

I start the analysis with observing the organization of smooth departure from the ongoing interaction. It is observable that the behavior of participants who leave the current conversation displays double orientation to the current conversation and something else. Moreover, there is a possibility that the utterance is oriented to the ongoing conversation and the bodily movement is oriented to another activity.

2.1 Smooth Departure from the Ongoing Interaction

The focus of excerpt (1) is on how Lucy leaves the current conversation between John. Before this transcript, Lucy starts to talk about her experience of being asked if she is married or not by one of the teachers in the high school before she began to work there. However, she is interrupted by another participant (Thea), gives up her storytelling, and becomes a recipient of Thea's talk.

(1) *Application*

```
18 JOH:  #For your [school?
19 DOR:            [Oh [re-?
20 IVY:                [Okay.
    im   #im.1-1
21 IVY:  [.h h h ¥h[hhhh   ¥#[Yeah, I wa-
22 DOR:  [ahahaha
23 THE:            [I'm sor¥#[ry.
24 LCY:                      ¥#[Yeah,(when)
   lcy         ->¥gaze IVY-¥JOH-->
    im                       #im.1-2
25 LCY:  mine [like when we [were
26 IVY:       [actually     [uh:hm
27 LCY:  crossing [( )they asked for married
28 JOH:           [No way.
29 LCY:  [female. That's why ¥#a lot
30 IVY:  [Nakagawa       sen¥#sei
   lcy                    ->¥gaze IVY-->
    im                       #im.1-3
31 LCY:  [of   [em
32 IVY:  [says.
33 MAR:        [( ) half of them
34       (0.2)
35 JOH:  O::h, oh, oh
```

im.1-1

im.1-2

im.1-3

Triggered by John's question about previous Lucy's story in line 18, Lucy resumes her storytelling in line 24. When talking about her experience, Lucy's verbal and non-verbal behaviors show her double orientation. Her utterance and face directions are oriented to the conversation with John. On the other hand, the lower part of her body keeps facing to the coffee table, the center of all the participants. Her body is torqued. Schegloff (1998) argues that body torque displays involvement in more than one activity or a course-of-action. Also, lower segments of the body are oriented to prior activities, he says. From this point of view, Lucy's behavior is displaying her main involvement in the co-occurring conversation, and the side involvement in the conversation with John (Goffman, 1963).

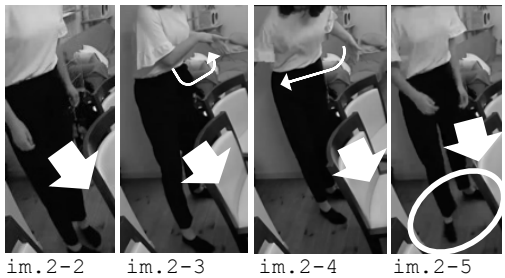
When leaving the conversation, Lucy's behavior also displays her double orientation. In line 29, even though Lucy keeps talking to John, she shifts her gaze direction from John to Ivy (im.1-3), the recipient in the other conversation. Ending her turn in line 31 at a syntactically incomplete point, Lucy completely leaves the conversation with John and starts another activity (listening to Thea's talk) at the same time. To sum up, when leaving the ongoing interaction, Lucy's behavior displays her double orientation both to the current activity and something else: her utterance is oriented to the current conversation, and her body and gaze direction are positioned for participating in another conversation.

The next fragment (2) also describes that the person who leaves the ongoing interaction displays double orientation. In this scene, Ivy, Doris, and Thea are looking for the host's cat, while the other participants are washing hands in the bathroom for having dinner. Ivy joined the conversation on the way to the bathroom.

(2) *Cat 1*

```
40 DOR:  *I think it ( )+*#like run away.
   ivy   *walks to the stairs----- *gaze down stairs->
    im                          #im.2-1
41       (1.2)*(0.2)
   ivy         *steps to L-->
42 THE:  Like a blu*#r(0.5)[hhhhhhhh
43 IVY:                    [Actually
   ivy              *body orients to the table>
    im               #im.2-2
44 IVY:  I foun*#d like(0.1)*went*#(0.4)
   ivy          *steps to L--*body orients to the bathroom>
                *P:lifts arms-----*S:shows the path->
    im           #im.2-3            #im.2-4
45 IVY:  *so *#quickly.*
   ivy   *R:lowers arms*
              *steps fwd/walks away-->
    im         #im.2-5
46       (0.3)
47 THE:  #Aha
    im   #im.2-6
```

im.2-1



im.2-2 im.2-3 im.2-4 im.2-5

im.2-6

Ivy starts to walk away before the end of the sequence. In lines 43 to 45, Ivy reports what she saw about the cat describing the path which he took (see im.1-3 and im.1-4). Ivy's turn is responded to by Thea's utterance in line 47. "Aha," which is equivalent to the "change-of-state" token *oh* (Heritage, 1984), proposes that Ivy's talk is informative to Thea. Also, "Aha" does not require any responses, the sequence is closing on line 47.

Even though Ivy's departure is ignoring Thea's response in line 47, her leaving procedure is accomplished gradually by displaying her double orientation. As Ivy is ending her turn, she performs a gradual turning away from the conversation with her body movement. After looking at downstairs, she turns around (im.1-2) and shows the path with her right hand (im.1-3 and im.1-4). As the images captured, the direction of her lower body is gradually shifting to the bathroom, which is on the right side of her. As was mentioned above, lower segments of the body are oriented to prior activities. Therefore, Ivy's main involvement is shifting from searching the cat to leaving there for the bathroom. That is, her utterance is displaying her orientation to the current activity, while her body movement is showing the orientation to another activity.

2.2 Leaving the Interaction Intermittently

The previous examination illustrates that the smooth departure is supported by the behavior displaying double orientation to the current activity and another activity, which is usually the next thing done by participants. The last excerpt shows that the leaving procedure is coordinated with talk in interaction. In the fragment (3), Ivy, who leaves the ongoing conversation, fails to leave, shows the full involvement in the conversation, and finally leaves there.

In this scene, all the participants are sitting at the dining table to start the meal. Asa, the host suggests washing hands in lines 3 and 4 (also see im.3-1). In response to her behavior, other participants stand up to go to the bathroom.

(3-1) *Cat 2*

```
1  THE:  *( )[( )you're
2  DOR:       [( )hhhhhhhhh
3  ASA:       [Sorry, should ∆we-
   ivy   *gaze THE-->
   asa                        ∆raises hands-->
4  ASA:  should we*::
5  THE:  right in *front of
   ivy           ->*gaze down-->
6  THE:  the came∆[ra(0.6) ha
7  DOR:           ∆[Actually I *#was
8  ASA:           ∆[like       *#ah:
   asa          ->∆holds hands-->
   ivy                         *ASA-->
    im                          #im.3-1
```

```
9  DOR:  there to[o and *I *#was like
10 MAR:          [Y e   *a *#h
   ivy       -gaze ASA *R-->
                        *Touches the seat->
    im                   #im.3-2
11 DOR:  [u:m (0.2)[Think I kind of
12 ASA:  [Sorry ( )[but
13 THE:            [.hhahhuhu
   ivy   -touches the seat-->
14 DOR:  wanna avoid it.(0.3)[.h.h
15 THE:                      [hu
16       (0.2)+*#(0.2)∫(0.2)
   dor        +points IVY w/ RH-->
   ivy         *gaze fwd-->
   the                ∫gaze IVY-->
    im          #im.3-3
17 DOR:  I+vy was *like +*#I think I don't
   dor    +gaze THE-----+table-->
   ivy   -gaze fwd*DOR---*places hands on lap
    im                   #im.3-4
```

im.3-1

im.3-2 im.3-3 im.3-4

Ivy, who is gazing at Asa (im.3-1) also tries to leave there. In the leaving procedure, her behavior displays double orientation. Ivy seems to be seeking the good timing for standing up by looking at the right side of her (im.3-2). However, her orientation is distributed into two in lines 16 and 17 because Doris points at Ivy and starts to talk about her. As the image 3-3 shows, Ivy's gaze direction is oriented to the conversation between Doris and Thea, while her arm position is showing her orientation to standing up. After that, Ivy shows her full involvement in the conversation by resting her hands on the lap, gazing at Doris and facing the lower part of her body to the table.

(3-2) *Cat 2*

```
18 DOR:  wanna *sit *here
   ivy            *gaze R-->
                  *touches the seat->
19 DOR:  +huhah[h     +*#.h( )yeah
20 THE:        [Oh my +*#gosh,
   dor   +gaze IVY----+THE-->
   ivy          -gaze R-*faces fwd-->
    im                  #im.3-5
21 THE:  *really?
   ivy   *gaze THE-->
22 DOR:  *#she was like *(oo*ps)
   ivy                 ->*gaze DOR-->
         *grabs the back/leans fwd------*leans back-->
    im    #im.3-6
23 DOR:  sorry *to*(o)(soon)hhhhh
   ivy          *gaze table-->
                  *leans fwd-->
24 THE:  hhhhhha
25       (0.1)
```

However, Ivy resumes the leaving procedure soon. Her behavior again displays her double orientation to the current conversation and the departure for the bathroom. She touches the seat again looking at the right side of her in line 18. She does not stand up quickly and gazes at Doris and Thea grabbing the back of her seat (im.3-5 and im.3-6). Finally, Ivy stands up and leaves the conversation between Thea and Doris in line 26.

The timing of her standing up links with the sequence organization of talk. Ivy leans forward during Doris's turn line 23 to stand up. However, she does not stand up until after line 25. In line 24, Thea is laughing, and no one starts a new turn. That means that the sequence is closing in lines 24 and 25. Therefore, Ivy's departure for the bathroom is initiated at transition-relevance place, in which the transition to a next speaker becomes possibly relevant (Schegloff, 2007).

3. Conclusion

The episodes presented in the excerpts (1) to (3) offer various instances of the same phenomenon: one participant leaves the ongoing interaction. It has been shown that leaving procedure is accomplished by distributing orientation to the current activity and another activity, which is the next thing done by participants. According to the observation, the departure from the ongoing interaction is finely tuned to the organization of talk. Also, there is a possibility that the current activity is oriented to by utterances, and the next activity is oriented to by bodily behaviors.

How participants use their bodies in interaction for coordinating with others has been discussed in various ways since many scholars conducted an intensive analysis of video recordings in the 1970s and 1980s (e.g., Goodwin, 1981; Heath, 1984). In the stream of interaction, different parts of utterances and one body display distinct orientations that are differently distributed toward the environment and other participants. Such an integrated contexture of orientations is constituted in response to, and constitutive of, the current progress of the ongoing activity (Nishizaka, 2017). By revealing the complex practice of distributing orientation, this study strengthens the concept of language and body as resources for locally succeeded interaction.

Finally, this paper considers why people distribute their orientation in a transitional phase in interaction. One of the possible answers is that projecting the next movement supports a smooth, successful interaction. "Projection" is regarded as an essential element of the turn-taking system. In the sequence of talk, the possible completion, where the transition of speakers can happen, is projected by the design of each utterance. Because of this mechanism, interaction is conveyed without any significant delays or gaps. From this point of view, it can be said that distributing orientation to the current and next activities supports the smooth transition between phases of interaction in interaction in the same way the projection does.

4. References

Broth, M. and Mondada, L. (2013). Walking away: the embodied achievement of activity closings in mobile interaction. *Journal of Pragmatics*. 47: 41-58.

Deppermann, A., Schmitt, R., and Mondada, L., (2010). Agenda and emergence: contingent and planned activities in a meeting. *Journal of Pragmatics*. 42 (6): 1700-1718.

Goffman, E. (1963). Behavior in public places: Notes on the organization of gatherings. New York: Free Press.

Goodwin, C. (1981). Conversational Organization: Interaction Between Speakers and Hearers. New York: Academic Press.

Heath, C. (1984). Talk and recipiency: Sequential organization in speech and body movement. In M. Atkinson & J. Heritage (Eds.), Structures of social action: Studies in conversation analysis, 247-265. Cambridge: Cambridge University Press.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In J. M. Atkinson and J. Heritage (Eds.). Structures of social action: Studies in conversation analysis, 299-345. Cambridge: Cambridge University Press.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), Conversation analysis: Studies from the First Generation, 13–31. Amsterdam: John Benjamins Publishing Company.

Mondada, L. (2011). The organization of concurrent courses of action in surgical demonstrations. In J. Streeck, C. Goodwin & C. LeBaron (Eds.), Embodied interaction: Language and body in the material world, 207–226. Cambridge: Cambridge University Press.

Mondada, L. (2014). The Conversation Analytic Approach to Data Collection. In J. Sidnell & T. Stivers (Eds.), The Handbook of Conversation Analysis, 32-56. NJ: Wiley-Blackwell.

Nishizaka, A. (2017). The Perceived Body and Embodied Vision in Interaction. *Mind, Culture, and Activity*. 24 (2): 110-128.

Schegloff, E. (1998). Body torque. *Social Research* 65: 335-596.

Schegloff, E. (2007). Sequence organization in interaction: A primer in conversation analysis. Cambridge: Cambridge University Press.

Schegloff, E. and Sacks, H. (1973). Opening up closings. *Semiotica*. 8: 289-327.

Sitting Down and Standing Up as Resources for Reorganization of Participation Framework: Analysis of Preparatory Meeting for Nozawa Onsen Fire Festival

Rui Sakaida[1], **Yasuharu Den**[2]
[1] National Institute of Informatics, [2] Chiba University
[1] 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
[2] 1-33 Yayoicho, Inage-ku, Chiba, 263-8522 Japan
[1] lui@nii.ac.jp, [2] den@chiba-u.jp

Abstract

In real-life social settings, especially in open fields or public spaces, people often talk while standing, fluidly changing their posture, position, or spatial formation. Therefore, the participation framework is continually reorganized. This paper suggests that body movements for changing posture, especially sitting down and standing up, can serve as resources for reorganization of participation framework. Sitting down and standing up result in the change of the level of eyesight of participants, which dynamically contributes to reorganization of their interactional space, F-formation.

Keywords: participation framework, posture, body movement

1. Introduction

During conversations conducted in experimental settings or closed rooms, participants usually remain seated. Such stable posture of participants lets them continue to talk at the same place and/or with the same interlocutors. On the other hand, in real-life social settings, especially in open fields or public spaces, people often talk while standing, fluidly changing their posture, body position, or spatial formation. Therefore, the participation framework is continually reorganized.

This paper demonstrates the way in which people reorganize participation framework while chatting in an open space, focusing on the change of the participants' posture, especially sitting down and standing up. When we talk in public spaces, we usually keep standing with our body faced to each other, organizing *F-formation* in circular arrangement (Kendon, 1990). However, for some reason, we sometimes sit down and stand up again while talking, and such body movements can influence the participation framework, i.e., who attends the conversation as ratified participants, who is supposed to speak, and who is addressed by the speaker (Goffman, 1981). Sitting down and standing up can serve as resources for reorganization of participation framework, since their movements change the level of participants' eyesight. This paper, by observing excerpts of video-recorded conversation, explicates how the movements of sitting down and standing up make the participation framework reorganized.

2. Data

In this paper we analyze some excerpts from the video recordings of the preparatory work for the fire festival called *Dosojin Matsuri* in Nozawa Onsen, Nagano, Japan. In Nozawa Onsen village, a group of people, called *San'yako*, work for the preparation and operation of the fire festival (Figure 1)[1]. *San'yako* consists of almost all the men around 42 years old in the village. The principal members of *San'yako* are 42 year-old men, who play main roles of the preparation and operation of the festival. Every year the principal members of *San'yako* change, and every three years all the members of *San'yako* change to the next generation. After completing their roles in *San'yako*, the executives of the preceding *San'yako* support the current *San'yako* as advisers or supervisors.

In this paper, we analyze conversation in an informal preparatory meeting from the data recorded in January 2017 (fiscal 2016) (Figure 2). The conversation was conducted by one adviser (AD) from the second preceding *San'yako* (*Kyooai*), two supervisors (SV1, 2) from the preceding *San'yako* (SV1 from *Seishoo* and SV2 from *Kooshin*), and two executives of the current principal members of *San'yako* (*Reishoo*), i.e., the chairman (C) and the vice-chairman (VC).

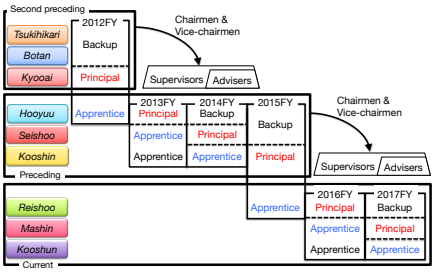


Figure 1: Organization chart of *San'yako*.

Figure 2: Participants of the preparatory meeting.

[1] For more detailed information about *Dosojin Matsuri* and *San'yako*, see Den (2018).

```
01  SV1  ichiban:, roku ji kara ku ji no aida ni:, (0.4) soo shuuchuushite, soo kite hoshii ttsutte sa, (0.4)
          the_most  six o'clock from nine o'clock GEN between at        yes  densely  yes come want  say  FP
         "Between six o'clock to nine, ((we)) want the most ((people)) to come, ((we)) would say."
02       hi#ru*ma ko(h)re(h)ru(h) yatsu +w*[a kite mo ii n[da #kedo *ttsutte. (0.4)+(0.9)
           daytime  can_come  guys  TOP come may  JD though say
         "Those who can come in the daytime may come, though, ((we)) would say."
03  SV2                                    [(arigatoo)    [(ssu::.)
                                            "Thank you."
    sv2      *turns his face to C      *receives the coffee     *turns his face to SV1
    c                                 +gives the coffee to SV2              +walks to the back of SV2->
    fig   #1                                                #2
04  AD   ma shigoto dekiru yatsu wa na. (0.4)
         INJ  work  capable  guys  TOP FP
         "Well, ((only)) those who are capable of the work ((may come))."
05  SV2  hehehehe *(1.8)+(0.6)
    c              ->+stands behind SV2->
    sv2          *opens coffee->
06  AD   osaki(ni)# (itadaite masu). (0.2)*(0.5)
          already_PL  have_PL  PL
         "((Sorry but we are)) drinking ((coffee)) already."
    sv2                              ->*
    fig          #3
07  SV2  (konomae) jimukyoku kuru tsuttetta nda kedo, ita tte shibare(h) ne(h)e(h) shi [hehehe *.h hehe .h +.h
          the_other_day  secretariat  come  said  JD though exist though  can_tie  NEG  and
         "The other day ((our)) secretariat said that ((he)) would come, but even if ((he)) comes, ((he)) cannot tie ((the ropes))."
08  VC                                                                                [h
    sv2                                                                                    *brings coffee to mouth->
    c                                                                                           +looks to the snow
09  SV1  (   *       ) ya+ttero ya. (1.1)
                        do_IMP  FP
         "((We could say to him,)) 'Do (   )."
    sv2    ->*drinks coffee
    c                 +sits on the snow
10  SV2  iya(h), sore fujin no ie itte nawa nattero ya(h)* 'ttsu(h)'tte(h) (0.#6)%(1.5)*(0.6)√(0.4)*(1.5)
          no  it  women GEN house go  ropes twine_IMP  FP  say
         "'No, go to the Women's House and twine the ropes', ((we)) would say ((to him))."
    sv2                                          *brings coffee to mouth and drinks it
    sv2                                                                *looks down *puts coffee on the ground->
    ad                                                          %brings coffee to mouth and drinks it
    sv1                                                                 √gazes at C->>
    fig                                                      #4
11  AD   ((a *long si[gh))
12  SV1             [Koo]da*i ga ite ichiban shita tte kuro da kke.
                     Kodai  NOM exist  the_most  young  QT  black  be  Q
                    "Kodai ((is the next grade)), and as for the youngest, is it black?"
    sv2    ->*looking down-->*looking up->>
13  C    (0.5) soo (de)su.
             yes  PL
            "Yes."
14  SV1  (0.1) herumetto. [a, nara ii na, dokomo kabutte [nee kara na:.
            hard_hat  oh  then  good FP  anywhere  overlap  NEG because FP
           "Hard hats. Oh, then you have no problem, because ((the colors of hard hats of)) all ((the grades)) are different."
15  C                     [hai.
                           "Yes."
16  SV2                                          [ehahahahaha.
```

Figure 3: Transcript of excerpt 1.

3. Analysis

3.1 Excerpt 1: Sitting Down and Rejoining Conversation

In conversation during preparation of the festival, the participants often go out of and rejoin the conversation for various reasons, e.g. in order to get some stuff relevant to the conversation. Slightly before excerpt 1 (Figure 3, 4), the current chairman (C) had gone out of the conversation, and after a while he came back there. Here we observe how the change of his posture helped him to rejoin the conversation. In the first part of this excerpt (Figure 3, lines 01-03; Figure 4, #1), AD, SV1, SV2 and VC are talking together. In line 01, SV1 suggests when they want their colleagues to come to help them, saying *ichiban:, roku ji kara ku ji no aida ni:, soo shuuchuushite, soo kite hoshii ttsutte sa* ("Between six o'clock to nine, we want the most people to come, we would say."). And SV1 adds a conditional suggestion to his own previous statement, *hiruma koreru yatsu wa kite mo ii nda kedo ttsutte* ("Those who can come in the daytime may come, though, we would say.") (line 02). Before this scene all the four participants were standing and facing each other, but now AD, SV1 and SV2 are squatting. Up to this moment, AD first squatted down and started to drink his coffee, and the SVs followed him, squatting down too.

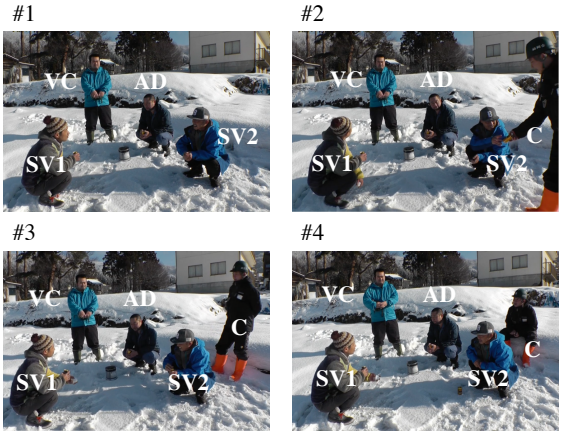


Figure 4: Transition of participants' posture in excerpt 1.

While SV1 is claiming the conditional suggestion (line 02), C, who had left the conversation in order to get another coffee, comes back and give it to SV2, who had not got coffee before (lines 02-03; Figure 4, #2). First SV2, noticing C approaching him, turns his face to C, and then C gives the coffee to SV2. Receiving the coffee, SV2 says *arigatoo* ("Thank you.") (line 03). Even while SV2 is getting the coffee from and thanking C, SV2's lower body

continues to be mainly involved (Goffman, 1963) in the current conversation, in a body-torqued position (Schegloff, 1998). As soon as he receives the coffee, SV2 turns his face to SV1 again, strongly reorienting to the current interactional space.

Whereas SV2 keeps participating in the conversation with AD, SV1 and VC, C walks to the back of SV2 (lines 03-05), not spatially joining the conversation, that is, continuing to be out of F-formation (Figure 4, #3). In line 04 AD responds to SV1's suggestion (line 02), *ma shigoto dekiru yatsu wa na* ("Well, only those who are capable of the work may come."). Immediately SV2 laughs (line 05), displaying alignment with AD's strict but laughable comment, and opens his coffee (line 05). After a long gap (line 05), AD says to SV2, who is preparing to drink coffee, *osakini itadaite masu* ("Sorry but we are drinking coffee already.")[2], and the topic of the conversation is tentatively suspended.

Afterward SV2 laughingly says *konomae jimukyoku kuru tsuttetta nda kedo, ita tte shibare nee shi* ("The other day our secretariat said that he would come, but even if he comes, he cannot tie the ropes.") (line 07), thereby retopicalizing the concern which had been introduced by AD in line 04, about who should come to help. At the end of the utterance SV2 brings his coffee to his mouth, when C looks to the snow on the ground. And SV2 drinks his coffee while SV1 is responding to him (line 09) (although the content of his response cannot be clearly transcribed), during which C sits on the snow (Figure 4, #4).

After C sits on the snow, SV2 says *iya, sore fujin no ie itte nawa nattero ya ttsutte* ("'No, go to the Women's House and twine the ropes', we would say to him.") (line 10), which is a joking statement imaginarily addressed to their *jimukyoku* (secretariat). From the end of the utterance SV2 again starts to drink his coffee, and AD also drinks his. Subsequently SV2 looks down and puts his coffee on the ground, when SV1 begins to gaze at C, who is sitting on the snow (line 10).

After a rather long silence, in which multimodal conducts by several participants are observed (line 10), overlapping with a long sigh by AD, SV1 asks a question *Koodai* ((the name of the next chairman from *Mashin*)) *ga ite ichiban shita tte kuro da kke* ("Kodai is the next grade, and as for the youngest, is it black?") (line 12). He does not use any address terms, but SV1 keeps gazing at C, addressing the question to him. Although what he is asking by the question is not clear, in the next turn C answers *soo (de)su* ("Yes.") (line 13). In line 14, responding to C, SC1 utters *herumetto* ("Hard hats."), which is an increment to his own previous question, and subsequently says *a, nara ii na, dokomo kabutte nee kara na:* ("Oh, then you have no problem, because the colors of hard hats of all the grades are different."). SC1's question is about the colors of hard hats of *San'yako* members, which can be similar and thus confusing between the three grades. Through this sequence C is invited to rejoin the conversation by SC1.

When he came back to the conversation, C was standing outside of the F-formation (Figure 4, #3), and therefore he was not a main participant of the conversation, i.e., neither a speaker nor an addressee, but a side-participant (Clark, 1996). On the other hand, afterward he sat down and his level of eyesight was lowered (Figure 4, #4), which seems to enable him to be directly seen by SV1. The utterance of SV1 was a question which could be answered by either C or VC, as Kodai's superior, in terms of epistemics. However, interestingly, the addressee was not VC, who had been present in the conversation and kept standing all the time, but C, who just arrived back there. C succeeded in rejoining the conversation by sitting down and being gazed at by one of the participants.

3.2 Excerpt 2: Standing up and Integration of Two Conversations

Excerpt 2 (Figure 5, 6) is a conversation a few minutes after excerpt 1. At the beginning of this excerpt, unlike in excerpt 1, the conversation is separated into two groups, that is, they are schisming (Egbert, 1997). AD, SV1 and VC are talking on the left side of the picture, and C and SV2 are talking on the right side (Figure 6, #1). The first half of the transcript indicates that they are respectively conversing, although much of the conversation is unfortunately inaudible because of the recording condition, especially for C and SV2. Now two of the participants of the conversation on the left side are squatting, but afterward they stand up, and the change of their posture contributes to the transformation from two separated conversations into a single conversation.

In lines 01-03 (Figure 5) AD, who works as a woodcutter, talks about his job during the preparation of this festival, although what he says is not clearly understandable. After saying *juuni gatsu ni natte kara yama ( )* ("From the beginning of December, the mountain, ( )") (line 01), AD, who has been squatting till then (Figure 6, #1), stands up and says *ni kai shika dekite nai* ("I have been able to do it only twice.") (line 03). Responding to him, SV1 nods twice (line 04). At the same time AD slightly steps forward, and then looks to the left, where C and SV2 are talking. However, after two seconds AD returns to the front, thereby reorienting to the current interactional space he is involved in (line 04; Figure 6, #2). In line 05 AD continues to talk to SV1 and VC, *kanzenni (gekkyuu nanoni) dare ga (yaru) ka ( )* ("Who does it though monthly paid?"), to which VC responds with laughter (line 06). Subsequently SV1, who works as a woodcutter with AD, also talks about his own job, *kyoo yuki furi soo dakara ika nai kedo ( )* ("Today, it is likely to snow, so I won't go.") (line 07). During the utterance of SV1, VC slightly steps forward as if he responded to AD's stepping. At the end of his turn, SV1 stands up, when all the participants are no longer squatting.

While SV1 is standing up, SV2, who has been talking with C, also changes his posture. SV2 turns around and looks back, saying something unclear (*(Rei) ( ) (suk ka)*) (line 09; Figure 6, #3). Rei is C's first name, so SV2 seems to be responding to C's inaudible utterance in lines 04-07. In the direction to which SV2 attends there is a vast field for the fire festival, where other members of *San'yako* are working. After SV2 started to turn around, AD and subsequently SV1 look to the field as well. Through this sequence, the participants, who has been talking separately, come to orient themselves to the same object, that is, the festival field.

After the participants looked to the festival field together, SV2 and SV1 returns to the front again (line 10), when all

[2] This is a Japanese idiomatic statement to display politeness toward a senior when a junior does something selfish before the senior do. In this sequence, AD, who is the eldest, is using the idiom as a kind of joke.

```
01  AD   juuni {1} gatsu ni natte kara yama (       ).
         December in become from mountain
         "From the beginning of December, the mountain, (     )."
02       (0.4)%(0.1)#(0.2){2}(0.2)
    ad        %stands up->
    fig            #1
03       ni kai shika dekite nai. (0.2)
          twice only can_do NEG
         "((I)) have been able to do ((it)) only twice."
04  SV1  ((nod %{3} nod)) (0.2)%(2.0)%(0.4)
    ad       ->%slightly steps forward
    ad                    %looks to the left
    ad                              %turns to the front
05  AD   kanzenni (gekkyuu# nanoni) dare ga (yaru) ka [(  ).
          perfectly monthly_pay though who NOM do Q
         "Who does ((it)) though monthly paid?"
06  VC                                              [hahahaha
    fig                 #2
07  SV1  kyoo$ yuki [furi soo dakara ika{4} [na√i kedo (    ).
         today snow fall likely because go NEG though
         "Today, it is likely to snow, so ((I)) won't go."
08  VC             [.h .h                [(    )i.
    vc      $slightly steps forward
    sv1                                  √stands up->
09       {5}(0.5)%(0.6)#(0.1){6}√
    ad          %looks to the field->
    sv1                  √looks to the field->
    sv1                ->√slightly steps forward
    fig              #3
10       (0.4)*(1.3)√(0.7)
    sv2    ->*returns to the front
    sv1           √returns to the front->
11  AD   ((a long√ sigh))
    sv1       ->√looks to the field
12       (3.9)
13  AD   a, %ashi√ta::, (1.1)%(0.7) +ashi%ta (0.5)
         oh tomorrow tomorrow
         "Oh, tomorrow, tomorrow,"
    ad    ->%looks at SV1    %turns to the left
    ad                              %looks down and looks at SV1->>
    sv1         √looks at AD
    c                          +looks at AD
14  SV1  ashita made matte mite sono ato:, dooro u[mere#ru yooni natte.
         tomorrow until wait try that after road can_cover come_to
         "((We would)) wait until tomorrow, and after that, ((we)) come to be able to cover the road."
15  SV2                                             [((nod nod nod nod nod))
    fig                                                  #4

C    {1}(          [          ).{2}

SV2               [((nod))

     (2.0)

C    {3}(                     )

     (                        )

     (          *             ).{4}
sv2             *turns around

     (0.4)

SV2  {5}(Rei)(   *    )(suk ka).{6}
          Rei              do  Q
sv2            ->*looks back->
```

Figure 5: Transcript of excerpt 2.

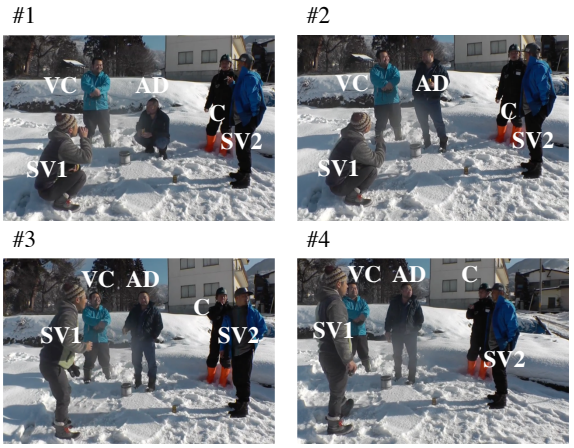


Figure 6: Transition of participants' posture in excerpt 2.

the participants are faced to each other. However, no one starts speaking, but instead AD produces a long sigh and SV1 reorients to the field (line 11). Here the participants' mutual orientation looks weakened, but after a long silence (line 12) AD starts to speak, *a, ashita::* ("Oh, tomorrow,"), looking at SV1 (line 13). Responding to the AD's action, SV1 looks at AD, displaying of recipiency (Goodwin, 1981; Heath, 1986), but AD suspends his turn and turns to the left, where SV2 and C are standing. C responds to AD's movement and gazes at him. At this moment, all the participants being faced to each other again, their interactional space for conversation, that is, a common F-formation is set up. Subsequently AD says *ashita* ("tomorrow") again, looking down and then at SV1. Gazed at by AD twice, SV1 starts to say *ashita made matte mite sono ato:, dooro umereru yooni natte* ("We would wait until tomorrow, and after that, we come to be able to cover the road."), as if he took over AD's utterance. Then SV2 nods several times as a response to the SV1's utterance. Here both SV2 and C, who were talking separately from AD, SV1 and VC, are invited to join the conversation with them, and the conversation between all the participants is restarted (Figure 6, #4).

In this way, the change of posture of AD and SV1 successfully triggered their conversation, which had been separated into two parts, to be integrated into a single conversation again.

4. Concluding Remarks

This paper suggested that body movements for changing posture, especially sitting down and standing up, can serve as resources for reorganization of participation framework. Sitting down and standing up result in the change of the level of eyesight of participants, which dynamically contributes to reorganization of their interactional space, F-formation.

5. Acknowledgments

The work was supported by JSPS KAKENHI Grant Number 15H02715, led by Mika Enomoto.

6. References

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Den, Y. (2018). F-formation and social context: How spatial orientation of participants' bodies is organized in the vast field. In *Proceedings of the LREC Joint Workshop on Language and Body in Real-Life and Multimodal Corpora 2018*.

Egbert, M. M. (1997). Schisming: The collaborative transformation from a single conversation to multiple conversations. *Research on Language and Social Interaction*, 30 (1): 1–51.

Goffman, E. (1963). *Behavior in public places: Notes on the Social Organization of Gatherings*. New York: Free Press.

Goffman, E. (1981). *Forms of Talk*. Pennsylvania: University of Pennsylvania Press.

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York: Academic Press.

Heath, C. (1986). *Body Movement and Speech in Medical Interaction*. Cambridge: Cambridge University Press.

Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge: Cambridge University Press.

Schegloff, E. A. (1998). Body torque. *Social Research*. 65 (5): 536–596.

The Language into Act Theory: A Pragmatic Approach to Speech in Real-Life

Emanuela Cresti, Lorenzo Gregori, Massimo Moneglia, Alessandro Panunzi
LABLITA – University of Florence
{elicresti, lorenzo.gregori, moneglia, alessandro.panunzi}@unifi.it

Abstract

This paper briefly introduces the Language into Act Theory (L-AcT), that proposes a pragmatic framework for the corpus-based collection and analysis of spontaneous speech. The L-AcT methodology takes the utterance (i.e. the counterpart of a speech act) as the reference unit for analysis. A set of large-scale Romance corpora has been collected in accordance with the L-AcT methodology (LABLITA Corpus, C-ORAL-ROM, C-ORAL-BRASIL, Cor-DiAL). Data for each corpus can be compared across languages, since they are built using the same corpus design, which entails a set of variation parameters relevant for representing spontaneous speech and, specifically, its pragmatic variation. LABLITA-C-ORAL corpora are text/sound aligned at the utterance level. Empirical research carried out by LABLITA has verified a systematic correspondence between stretches of speech ending with a terminal prosodic break and the accomplishment of an illocutionary force, thus identifying utterances. Within the latter, a correspondence between chunks separated by non-terminal breaks and information functions has been identified. The IPIC database was created for the cross-linguistic comparison of information structure in Romance languages. With regard to the pragmatic classification of utterances, a working repertory of illocutionary types has been established, induced empirically from pragmatic and prosodic features shared in Romance corpora.

Keywords: Pragmatics, Prosody, Spoken romance corpora

1. Introduction

1.1 The L-AcT Framework

The Language into Act Theory has been in development in Italy since the nineteen-eighties and aims at providing a pragmatic framework for the corpus-based collection and study of spontaneous speech (Cresti 2000). L-AcT focuses on four crucial aspects: a) a corpus building strategy for both the representation of the speech universe and for comparative studies; b) the exploitation of prosody for the identification of the linguistic reference units in the flow of speech; c) the information structure of the utterance; d) illocutionary types in spontaneous speech.

Within the tradition stemming from Austin (1962), L-AcT assumes that the utterance is the counterpart to a speech act and constitutes the primary reference unit for the analysis of speech. Its main innovation is to consider spoken activity as manifested through prosodic devices, specifically with regard to the core aspects of illocutionary force and information structure (IS). Therefore, the processing of prosody is taken as a mandatory step for the identification of both utterances and their information structure, and is achieved through the perceptual evaluation of prosodic breaks.

2. Corpus building

2.1 Collection criteria

The corpus design of the LABLITA resources entails a set of variation parameters that are considered relevant for representing natural interactions in spontaneous speech (Biber, 1988; Mello 2014) and, specifically, its dia-phasic variation (Berruto, 2000), selected to ensure probability of occurrence to the maximum number and variety of speech act types. The recording parameters are: a) informal, non-regulated and formal, regulated turn-taking; b) public, private, family context; c) dialogue, multi-dialogue, monologue exchange; d) public domain (law, religion, business); e) media and telephone production (Table 1). The recording strategy focuses on the acoustic data only, which given the relatively unobtrusive technology used in its recording allows the collection of a broad set of situations and domains, difficult to achieve with more invasive equipment such as for video.

2.2 Resources

Using the aforementioned corpus design framework, LABLITA has archived a resource with high dia-phasic (approx. 950 recording sessions) and dia-stratic (more than 2000 speakers) variation. From this huge collection, an Italian corpus has been derived whose recordings contain approx. 988,000 transcribed words and 107,000 reference units (Cresti et al. forthcoming). The recordings were transcribed in the CHAT-LABLITA format (Moneglia Cresti 1997; McWhinney 2000) and session metadata are in both the CHAT and IMDI format. The orthographic transcriptions (in txt files) are enriched by the tagging of terminal and non-terminal prosodic breaks. Each utterance has been aligned to its acoustic source in XML files, following L-AcT protocol. The text-to-speech synchronization was achieved through WinPitch, which allows real time F0 displacement of large speech excerpts.

Beyond the Italian corpus, the L-AcT framework has been deployed and tested in the collection and annotation of comparable Romance corpora: C-ORAL-ROM (Cresti & Moneglia 2005), C-ORAL-BRAZIL (Raso & Mello 2012), Cor-DiAL (Nicolas Martinez 2013). The C-ORAL-ROM resource is a multilingual corpus of the main Romance languages (Italian, French, Spanish, European Portuguese), containing 1,200,000 words, 1,426 speakers, 772 spoken texts, and 123:27:35 hours of speech. The four corpora were collected using the same corpus design for reasons of later comparability.

The C-ORAL-BRASIL resource (2006-2010) was collected by Raso & Mello (2012) in the Minas Gerais metropolitan district using the C-ORAL-ROM sampling and annotation criteria. It presents 362 recorded speakers, 139 spoken texts, 21:08: 52 hours of speech, and 209,000 words, and focuses on informal dia-phasic variation.

2.3 Corpus Design and speech variability

The corpus design parameters of the LABLITA resource capture basic generalizations of the variability of spoken language. We are able to focus on the spoken performance, considering, for instance, basic phenomena such as the middle length of utterances and information units, the

noun-verb ratio, and the percentage of verbal and verbless utterances. Such properties are at the core of the linguistic constructions characterizing speech.

CORPUS VARIATION PARAMETERS			S.	W.	UTT.
TURN TAKING	CONTEXT	STRUCTURE OF EVENT			
Free Informal	Family Private	Monologue	26	48,606	4,866
		Dialogue MultiDial	141	242,896	46,133
	Public	Monologue	3	3,112	227
		Dialogue MultiDial	41	59,756	11,569
	Telephone	Dialogue	74	23,004	4,445
	Talking Children	Dialogue MultiDial	276	260,595	N.C.
Sub-total			**561**	**637,969**	**67,240**
Regulated Formal	Family Private	Monologue	1	3,139	193
		Dialogue MultiDial	28	53,126	8,582
	Public	Monologue	39	77,442	5,082
		Dialogue MultiDial	53	107,666	14,820
	Broadcast		69	108,553	11,031
Sub-total			**190**	**349,926**	**39,708**
Total			**751**	**987,895**	**106,948**

Table 1: Design of the LABLITA Corpus

The quantitative measures of each of the above phenomena show a systematic variation across textual diaphasic typologies, demonstrating the appropriateness of the corpus design. The Graph in Figure 1 analyses one of the main lexical aspects of speech: that it supposedly records a higher number of Verbs with respect to the written variety (Halliday 1976; Biber 1999). The figure shows however that the Verb vs Noun Ratio follows this prediction only in informal dialogues, and that it actually favors nouns in Formal - Monologic contexts.

From a syntax point of view, the presence of verbless utterances has been considered a very particular feature in speech performances (Blanche-Benveniste 1997); again, however, this feature strongly characterizes informal dialogues, where the ratio of Verbal to Verbless utterances is almost 50/50. Conversely, the number of verbless utterances decreases significantly in Formal contexts and is markedly reduced in Monologues. In summary, one of the relevant parameters turns out to be different to its predicted value for "formal / monologic" and "informal / dialogic" cases, both at the lexical and syntactical levels.

Given that the C-ORAL corpora have been collected and built using the same corpus design, it is worth noting that the quantitative variation of the above phenomena repeats with the textual variation of the four Romance languages and Brazilian Portuguese (Cresti & Moneglia 2005; Panunzi & Mittman-Malvessi 2014; Moneglia & Cresti 2015). This cross-linguistic trend is proof of the consistency of the correlation between the parameters and the core linguistic phenomena considered.

However, it must also be noted that in our interpretation the variation of the linguistic properties is grounded in pragmatics (illocutionary activation), which distinguishes the speech performance achieved in informal interactive Dialogic Contexts from that in Formal Monologues. It is worth exploring, in the context of this workshop, that the high-level distinction of "Formal" vs "Informal" which characterizes the L-AcT corpus design is not compliant with the model proposed in the most relevant corpus building strategy proposed nowadays i.e. the Balanced Corpus of Everyday Japanese Conversation by NINJAL (Koiso et al. 2016).

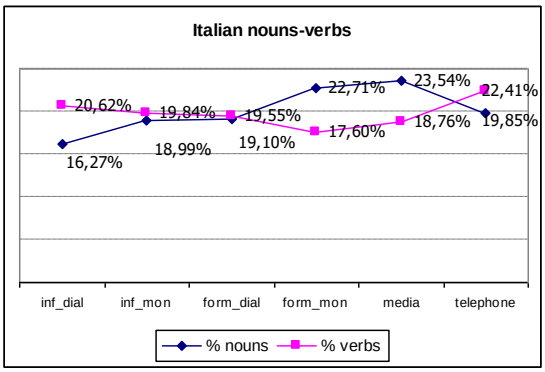


Figure 1: The Variation of Verb / Nouns Ratio

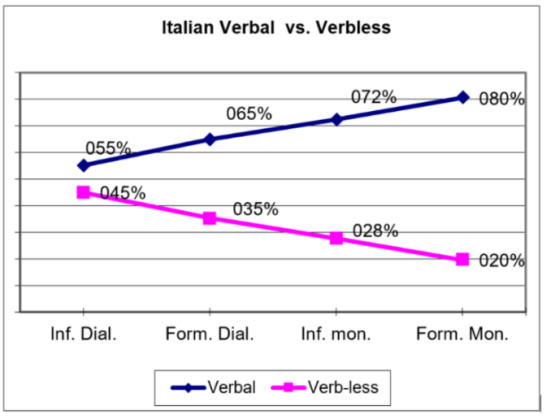


Figure 2: The Variation of Verbal vs. Verbless utterances

The pragmatic viewpoint of L-AcT focuses on the representation of speech act typologies, and their occurrence is not a function of the behavior accompanying the speech (eating, leisure, work, transfer, rest), as suggested by the NINJAL survey. Each speech act is accomplished as a function of the subjective initiative of the speaker toward the addressee. The L-AcT corpus design strategy is aimed at ensuring coverage of the maximum number of speech act types.

3. Exploitation of prosody

The L-AcT methodology assumes a systematic correspondence between stretches of speech ending with a terminal prosodic break and the accomplishment of an illocutionary force, and, within the utterance, between chunks segmented by non-terminal breaks and information functions (Cresti & Moneglia, 2005). The idea of the perceptual relevance of prosodic breaks traces back to the IPO tradition, which stresses the relevance of intentionally performed prosodic cues ('t Hart & al., 1990). Their correlation with acoustic features in speech has been

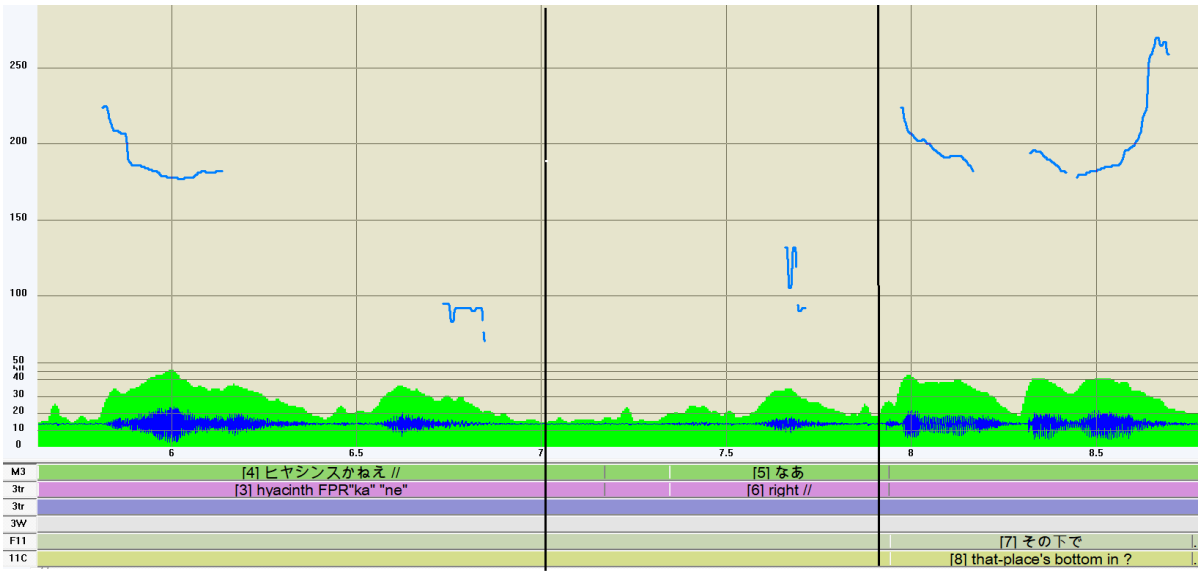


Figure 3: Text-to-speech alignement per utterance of two turns (WinPitch software)

extensively debated (Swerts & Geluiken 1993; Swerts, 1997; Firenzuoli, 2003; Martin, 2015). For all LABLITA and C-ORAL corpora, text to speech alignment at the utterance level according to prosodic cues (terminal breaks), and the scanning of the utterance into prosodic units (non-terminal breaks), has been implemented using WinPitch. This methodology ensures significant segmentation of speech into reference units, forming counterparts to speech acts as pragmatically defined. The annotation of prosodic breaks has been validated (Danieli et al 2004; Raso & Mittmann 2009; Moneglia et al., 2010; Mello et al., 2012).

Beyond the Romance languages, the methodology has been extended to the English language and is in progress for Japanese (Cresti & Fujimura forthcoming). The example in Figure 3 shows of how a dialogic turn by a Japanese speaker appears when segmented into independent utterances.

4. Information Structure

Within L-Act, the scanning of the utterance into prosodic units using non-terminal breaks reveals the prosodic interface for the Information Sstructure (IS). IS has its center in the pragmatic accomplishment of the illocution, which is developed by a necessary information unit i.e. the Comment. The Comment may be accompanied by optional components, forming the information pattern, which may be composed of many information units each developing different functions: textual (Topic, Parenthesis, Appendix, Locutive Introducer) and dialogical (Discourse markers) (Moneglia & Raso 2014). Each information unit is performed by a dedicated prosodic unit type.

This conception is retraceable to Chafe (1970; 1994) and moves away from one of the most popular nowadays that of Krifka (Krifka 2007; Krifka & Musan 2012). The latter is grounded in natural logic and finds the conditioning origin of information structure, and finally of speech, in the context (i.e. Common Ground (Stalnaker 1999)). In contrast, at the core of its conception L-AcT focuses on the subjective initiative of the speaker toward the addressee, who reacts to the context but does not depend on it.

L-AcT was also used to ground the cross-linguistic comparison of Information Structure in spontaneous speech. For this, the IPIC database was created by LABLITA (Panunzi & Gregori 2012) and applied to comparable Italian and Brazilian-Portuguese mini-corpora, that were tagged according to L-AcT criteria (Mittmann-Malvessi & Raso 2012; Panunzi & Mittmann-Malvessi 2014). Quantitative data for the comparison between Italian and Brazilian Portuguese can be found in Panunzi & Mittmann- Malvessi (2014) and in Moneglia & Cresti (2015). The database was also extended to compare information structure for an American English selection taken from the S. Barbara corpus (Du Bois et al., 2000) by the LEEL laboratory in Belo Horizonte (Cavalcante & Ramos 2016). A Spanish selection from Cor-DiAL (Nicolas 2013) is forthcoming.

5. Repertory of illocutionary activities in spontaneous speech

Within L-AcT, the pragmatic analysis of speech is grounded in illocution, defined briefly as a "mental/affective reaction to an external input which is transformed into a conventional linguistic action towards the addressee" (Cresti 2018). Realistically, the classification of an illocution has always been a challenge (Kempson, 1977; Sbisà, 1989; Sbisà & Turner, 2013; Leech 2014). Beyond the well-known illocutionary types such as assertion, order, question - reducing the illocutionary variety to the syntactic typologies of the sentence: declarative, jussive, interrogative (Fava, 1995) - many other new illocutionary types may be envisaged. Over the past twenty years the LABLITA team has carried out empirical research on corpora to identify illocutionary types and their prosodic profiles, following a corpus-based

methodology (Cresti & Firenzuoli 1999; Firenzuoli 2003; Cresti et al. 2003; Cresti 2005, forthcoming ; Rocha 2016). The systematic analysis of entire spoken texts allowed the recognition of several illocutionary types that were not considered in the standard taxonomy (Searle 1969), but which recur within dia-phasic and dia-stratic variations of Romance corpora. Correlations between specific illocutionary types and sets of communicative, pragmatic, cognitive features have been discovered and hypotheses on models of prosodic units conveying illocution are in development. The value for an utterance depends on the speaker's affective activation toward the addressee.

LABLITA's corpus-based research has led to an initial repertory of almost 90 illocutionary types which are grouped into 5 illocutionary classes; i.e *representation*, *direction*, *expression*, *ritual*, which record a variation among types, and *refusal*, which does not record a variation among types. In turn, the illocutionary classes can be divided into 14 sub-classes which present intermediate pragmatic levels within each class. This repertory is a working set of concepts which have been induced from corpus based analysis, although at present no corresponding operational criteria for speech acts annotation has been defined into L-AcT.

Table 2 shows that for instance the assertive class, which is the most common in speech, presents speech act types that have not been dealt with in the literature before, since they could only be observed in corpora. Assertion foresee an intermediate level of categorization composed of two sub-classes: weak assertion and strong assertion. Sub-classes can be distinguished for the degree of relevance of the semantic content in the utterance, the (speaker's) commitment to the content's truth, and the degree of the speaker's involvement with respect to the addressee. So far, within the weak sub-class, self-conclusion and assertion taken for granted types are high frequency in corpora. When the speaker accomplishes a self-conclusion, he seems to suddenly become distant from the flow of the exchange and rather unconcerned with the addressee's involvement, so without looking at the latter, he performs the utterance with a low or even whispered voice, executing it through a prosodic unit with a falling f0 movement. Conversely, assertion taken for granted type is fully integrated in the speaker / addressee exchange. The speaker reports information already known or expected, presupposing the agreement of the addressee. In this case, he performs the utterance with a long ascending f0 movement ending at top values (Cresti forthcoming).

The L-AcT repertory of illocutionary types has been compared with other systems, among which we would like to cite that proposed by Yuki, Abe & Lin (2005) for Usage Based Linguistic Informatics, which is one of the few based on different language corpora. The UBLI taxonomy is composed of 50 substantive functions in the conversation which are strictly dependent on the most frequent content of the linguistic action performed (asking price, time, number, existence, place, …). Beyond the differing theoretical assumptions, it is interesting to observe how a corpus-based approach brings to light some interesting points of agreement (Cresti 2006; Moneglia 2011).

Assertion	**Direction**		**Expression**	**Rituals**	**Refusal**
WEAK Self-conclusion On-going comment Confirmation Explanation Assertion taken for granted Literal citation	**COMMUNICATIVE INVOLVEMENT** Distal recall (visible / non-visible addressee) Proximal recall Functional recall	**LINGUISTIC BEHAVIOUR** Partial question Polar question Alternative question Confirmation request	**BELIEF** Contrast Softening Obviousness Irony Doubt Admission Waiver Rhetorical question	**COURTESY** Thanks Greetings Welcome Excuses Wishes Congratulations Condolences Compliments	
STRONG Answer Ascertainment Assertion of evidence Hypothesis	**CHANGE OF THE ATTENTION** Distal deixis (still / moving object) Proximal deixis Prompt Event presentation	**NON LINGUISTIC BEHAVIOUR** Order Interdiction Prohibition Invite Offer Agreement	**FEELINGS AND MOODS** Protest Complain Grumbling Imprecation Surprise Wish Easement	**SOCIAL** Legal declarations Convictions Judgments Penalties Examination Diagnoses Dedications Religious rites	
	MENTAL TRANSFORMATION Instruction Person introducing Agreement request Self-correction Reported speech Warning	**ENDORSEMENT** Committeemen (bet, promise) Proposal Authorization	**SPEAKER ADDRESSEE RELATION** Approval Disapproval Derision Challenge Reproach Hint Concession	**DIALOGIC MOVES** Assent Repetition request Request of stop Request of waiting	

Table 2: Repertory of illocutionary types

6. Bibliographical References

Austin, J. L. (1962). How to Do Things with Words. Oxford: Oxford University Press.

Biber, D. (1988). Variation across Speech and Writing. Cambridge: Cambridge University Press

Biber, D., Johansson, S. and Leech, G. (1999): The Longman Grammar of Spoken and Written English. London: Longman.

Blanche-Benveniste, C. (1997). Approches de la Langue Parlée en Français. Paris: Ophrys.

Berruto, G. (2003). Fondamenti di sociolinguistica. Roma-Bari, Laterza,

Cavalcante, F. and Ramos, A. (2016). The American English spontaneous speech minicorpus. Architecture and comparability. *CHIMERA*, Special Issue.

Chafe, W. (1970). Meaning and the structure of language. Chicago: University of Chicago Press.

Chafe, W. (1994). Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing. Chicago: University of Chicago Press

Cresti, E. (2000). Corpus di italiano parlato. Firenze: Accademia della Crusca.

Cresti, E. (2005). Per una nuova classificazione dell'illocuzione a partire da un corpus di parlato (LABLITA). In E. Burr, editor, Tradizione e innovazione: il parlato. Atti del VI Convegno internazionale SILFI (giugno 2000, Duisburg), pages 233-246. Pisa : Cesati.

Cresti, E. (2006). Some comparisons between UBLI and C-ORAL-ROM, In Y. Kawaguchi, S. Zaima and T. Takagaki, editors, Spoken Language Corpus and Linguistic Informatics, pages 125-152. Amsterdam: Benjamins.

Cresti, E. (2018). The illocution-prosody relation and the information pattern in the spontaneous speech according to the Language into Act Theory (L-AcT). In M. Moroni and M. Heinz, editors, Prosody: Grammar, information structure, interaction. Special issue Linguistik Online, 88/1.
https://bop.unibe.ch/index.php/linguistik-online/index

Cresti, E. (forthcoming). The pragmatic analysis of speech and its illocutionary classification according to Language into Act Theory. In S. Izre'el, H. Mello, A. Panunzi and T. Raso, editors, In Search for the Reference Unit of Spoken Language: A Corpus Driven Approach. Amsterdam: Benjamins

Cresti, E. and Firenzuoli, V. (1999). Illocution et profils intonatifs de l'italien. *Revue française de linguistique appliquèe* IV/2: 77-98.

Cresti, E., Moneglia, M. and Martin, P. (2003). L'intonation des illocutions naturelles répresentatives: analyse et validation perceptive. In A. Scarano, editor, Macrosyntaxe et pragmatique: l'analyse linguistique del'oral, pages 243-264. Roma: Bulzoni.

Cresti, E. and Moneglia, M., editors (2005). C-ORAL-ROM. Integrated reference corpora for spoken romance languages. DVD + vol. Amsterdam: Benjamins.

Cresti, E., Moneglia, M. and Panunzi, A. (forthcoming). The LABLITA Corpus & the Language into Act Theory: analysis of Viterbo excerpts. In: A. De Dominicis, editor, Atti del Convegno Internazionale "Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis". Roma: Editrice Accademia dei Lincei.

Cresti, E. and Fujimura, I. (forthcoming). The information structure of spontaneous spoken Japanese and Italian in comparison: a pilot study. in A. De Meo, and F. Dovetto, editors, Atti del LI Congresso Internazionale SLI Le lingue extraeuropee e l'italiano. Problemi didattici, sociolinguistici, culturali. Napoli: Liguori.

Danieli, M., Garrido, J. M., Moneglia, M., Panizza, A., Quazza, S. and Swerts, M. (2004). Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech C-ORAL-ROM. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa and R. Silva, editors, Proceedings of the 4th LREC Conference, pages 1513-1516. Paris: ELRA.

Du Bois, J. W., Chafe, W., Meyer, C., Thompson, S. and Meyer, Ch. (2000). Santa Barbara Corpus of Spoken American English, Part 1. Philadelphia: Linguistic Data Consortium.

Fava, E. (1995). Tipi di atti e tipi di frase. In L. Renzi, Lorenzo, G. Salvi, and A. Cardinaletti, editors, Grande Grammatica Italiana di Consultazione, pages 19-48. Bologna: Il Mulino.

Firenzuoli, V. (2003). Le Forme Intonative di Valore Illocutivo dell'Italiano Parlato: Analisi Sperimentale di un Corpus di Parlato Spontaneo (LABLITA). PhD Thesis. Università di Firenze.

Halliday, M. (1987). Sistema e funzione nel linguaggio. Bologna: Il Mulino.

't Hart, J., Collier, R. and Cohen, A. (1990). A Perceptual Study on Intonation. An Experimental Approach to Speech Melody. Cambridge: Cambridge University Press.

Yuki, K., Abe, K. and Lin, C. (2005). Development and assesment of TUFS Dialogue Module-Multilingual and Functional Syllabus. In Y. Kawaguchi, S. Zaima, T. Takagaki and M. Usami, editors, Linguistics Informatics. State of the Art and Future, pages 313-333. Amsterdam: Benjamins.

Kempson, R. M. (1977). Semantic Theory. Cambridge: Cambridge University Press.

Koiso, H., Tomoyuki, T., Ryoko, W., Daisuke, Y., Masao, A. and Yasuharu, D. (2016) . Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation. In N. Calzolari et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Paris: ELRA.

Krifka, M. (2007). Basic notions of information structure. In C. Féry, G. Fanselow and M. Krifka, editors, Interdisciplinary Studies of Information Structure 6, pages 13-55. Potsdam: Universitätsverlag.

Krifka, M. and Musan, R., editors (2012). The Expression of Information Structure. Berlin/Boston: De Gruyter Mouton.

Leech, G. (2014). The Pragmatics of Politeness. Oxford: Oxford University Press.

Martin, Ph. (2015). The structure of spoken language. Intonation in romance. Cambridge: Cambridge University Press.

McWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. Third Edition. Mahwah: Lawrence Erlbaum Associates.

Mittmann-Malvessi, M. and Raso, T. (2012). The C-ORAL-BRASIL Informationally Tagged Mini-Corpus.

In H. Mello, A. Panunzi and T. Raso, editors, Illocution, modality, attitude, information patterning and speech annotation, pages 151-183. Firenze: Firenze University Press.

Mello, H. (2014) Methodological issues for spontaneous speech corpora compilation: The case of C-ORAL-BRASIL. In T. Raso and H. Mello, editors, Spoken Corpora and Linguistic Studies, pages 27-68. Amsterdam/Philadelphia: John Benjamins.

Mello, H., Raso, T., Malvessi-Mittmann, M., Vale, H. P. and Côrtes, P. O. (2012). Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In T. Raso and H. Mello, editors, C-ORAL-BRASIL I: Corpus de referência de português brasileiro falado informal, pages 125-176. Belo Horizonte: Editora UFMA.

Moneglia, M. (2006). Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective. In Y. Kawaguchi, S. Zaima and T. Takagaki, editors, Spoken Language Corpus and Linguistics Informatics, pages 153-179. Amsterdam, Benjamins.

Moneglia, M. (2011). Spoken Corpora and Pragmatics. *Revista Brasileira de Linguìstica Aplìcada* 11/2: 479-519.

Moneglia, M. and Cresti, E. (1997). L'intonazione e i criteri di trascrizione del parlato adulto e infantile. In U. Bortolini and E. Pizzuto, editors, Il Progetto CHILDES Italia, pages 57-90. Pisa: Del Cerro.

Moneglia, M. and Cresti, E. (2006). C-ORAL-ROM Prosodic Boundaries for Spontaneous Speech Analysis. In Y. Kawaguchi, S. Zaima and T. Takagaki, editors, Spoken Language Corpus and Linguistics Informatics, pages 89-114. Amsterdam: Benjamins.

Moneglia, M. and Cresti, E. (2015). The Cross-linguistic Comparison of Information Patterning in Spontaneous Speech Corpora: Data from C-ORAL-ROM ITALIAN and C-ORAL-BRASIL. In S. Klaeger and B. Thörle, editors, Interactional Linguistics: Grammar and Interaction in Romance Languages from a Contrasting Point of View, pages 107-128. Tübingen: Stauffenburg.

Moneglia, M. and Raso, T. (2014). Notes on the Language into Act Theory. In T. Raso and H. Mello, editors, Spoken Corpora and Linguistics Studies, pages 468-494. Amsterdam: Benjamins.

Moneglia, M., Raso, T., Mittmann-Malvessi, M. and Mello, H. (2010). Challenging the Perceptual Prominence of Prosodic Breaks in Multilingual Spontaneous Speech Corpora: C-ORAL-ROM/C-ORAL-BRASIL. In Speech Prosody 2010. Chicago.

Nicolas Martinez, C. (2012). Cor-DiAL (Corpus oral didáctico anotado lingüísticamente). Madrid: Liceus.

Panunzi, A. and Gregori, L. (2012). DB-IPIC. An XML Database for the Representation of Information Structure in Spoken Language. In H. Mello, A. Panunzi and T. Raso, editors, Pragmatics and Prosody. Illocution, Modality, Attitude, Information Patterning and Speech Annotation, pages 133-150. Firenze: Firenze University Press.

Panunzi, A. and Mittmann-Malvessi, M. (2014). The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. In T. Raso and H. Mello, editors, Spoken corpora and Linguistic Studies, pages 129-151. Amsterdam: Benjamins.

Raso, T. (2014). Prosodic Constraints for Discourse Markers. In T. Raso and H. Mello, editors, Spoken Corpora and Linguistic Studies, pages 411-467. Amsterdam/Philadelphia: John Benjamins.

Raso, T. and Mittmann-Malvessi, M. (2009). Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. *Revista de Estudos da Linguagem* 17(2): 73-91.

Raso, T. and Mello, H., editors (2012). C-ORAL-BRASIL I: Corpus de referência de português brasileiro falado informal. Belo Horizonte: Editora UFMA.

Rocha, B. (2016). Uma metodologia empírica para a identificação e descrição de ilocuções e a sua aplicação para o estudo da Ordem em PB e Italiano, PhD. Dissertation, Belo Horizonte UFGM.

Sbisà, M. (1989). Linguaggio, ragione, interazione: per una teoria pragmatica degli atti linguistici. Bologna: Il Mulino.

Sbisà, M. and Turner, K., editors (2013). Pragmatics of speech actions. Berlin: Mouton de Gruyter.

Searle, J. (1969). Speech acts: an essay in the philosophy of language. Cambridge: Cambridge University Press.

Stalnaker, R. (1999). Context and Content. Oxford: Oxford University Press.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America* 101: 514-521.

Swerts, M. and Geluykens, R. (1993). The prosody of information units in spontaneous monologues. *Phonetica* 50: 189:196.

Factor Analysis of Japanese Daily Utterance Styles

Hajime Murai
Future University Hakodate
116-2 Kamedanakano-cho, Hakodate, Hokkaido, Japan
h_murai@fun.ac.jp

Abstract

It may be more difficult to extract fundamental utterance styles in real-life daily conversation than those in fictional utterances because the characteristics of utterance styles are exaggerated in fictional utterances. However, by utilizing a large-scale corpus of daily conversation, it is possible to identify the fundamental patterns of Japanese utterance styles. In this study, the NUCC was targeted and extraction of the characteristics of utterance styles was carried out using the statistical method of factor analysis. As a result, five factors ("Average style in NUCC," "Avoid affirmation style," "Frank teenager style," "Dialect style," and "Polite style") were extracted quantitatively. Compared to fictional utterance styles, "Avoid affirmation style" is unique in real daily conversation. On the other hand, "Crude style" and "Hearsay style" do not appear. Although the similarities between the fictional corpus and the NUCC support the validity of the result, the factors were impacted by bias in the corpus. It would be desirable to utilize a speaker-balanced daily conversation corpus for a more precise analysis.

Keywords: Utterance, Style, Japanese

1. Introduction

Utterance styles are affected by various attributes, such as gender, age, situation, cultural settings, social backgrounds, personalities of the characters, and the mood of a scene. In the case of Japanese fictional utterances (in novels or general story texts), each character is differentiated based on their utterance styles; it is a popular technique used to help readers understand each character's personality (Kinsui, 2003). These utterance styles can be detected by comparing frequencies of function words in utterances (Murai, 2017A). Moreover, fundamental patterns of utterance styles can be extracted by a factor analysis of a fictional corpus (Murai, 2017B).

In the field of Japanese real conversation in daily life, the main research topics have been general grammatical characteristics, pragmatic semantics (Seto, 2015), and the relationships between specific single attributes (such as politeness and gender) and utterance styles (Kurosawa, 2010). A fundamental total pattern of Japanese utterance styles has not been examined quantitatively based on a real corpus. It may be more difficult to extract fundamental utterance styles in real-life daily conversation because the distinct utterance styles may tend to be exaggerated in conversations between fictional characters (particularly in entertainment content). Therefore, case study approaches and psychological experimental approaches have been used in the field of Japanese utterance styles of daily conversation (Miyazaki, 2014; Shen, 2012).

However, by utilizing a large-scale corpus of daily conversation, it is possible to identify the fundamental patterns of Japanese utterance styles. In this study, the Nagoya University Conversation Corpus (NUCC) was targeted and extraction of the characteristics of utterance styles was carried out using the statistical method of factor analysis.

2. Corpus for Utterance Analysis

The NUCC is composed of transcriptions of 129 uncontrolled, natural conversations between or among friends, family members, or colleagues. Each conversation has two to four participants and lasts 30 to 60 minutes. The participants are 198 native Japanese speakers of various ages and from diverse academic backgrounds (Fujimura, 2012). For the factor analysis, utterances were grouped by each speaker in 129 conversation scenes. In total, 296 utterance sets were obtained from the NUCC (excluding one very reticent speaker for statistical reasons). The attributes of the speakers of the 296 utterance sets are given in Table 1.

	Female	Male	Total
10s	15	2	17
20s	116	26	142
30s	43	1	44
40s	21	8	29
50s	22	4	26
60s	26	4	30
Over 70s	7	0	7
Unknown	1	0	1
Total	251	45	296

Table 1: Speaker details for utterance sets in the NUCC

It clear that the gender and age balance of the NUCC is biased. However, it is the only large-scale everyday conversation Japanese corpus available. From this table, it is expected that the characteristics of the utterances of young women will be prominently featured.

3. Characteristics in Utterance Styles

In this study, the frequencies of function words in utterances were adopted as characteristics of text style because in many Japanese novels, different usage patterns of function words are used to exhibit characters' personalities (Kinsui, 2003). In the Japanese language, function words mainly correspond to particles and auxiliary verbs. Therefore, the statistical significances of the frequencies of particles and auxiliary verbs were analyzed using factor analysis (Murai, 2017B). The NUCC provides morphologically analyzed data sets for the included conversation texts. Therefore, particles and auxiliary verbs in utterances were extracted and counted from the 296 data set units.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Case particle "Ga"	**0.86**	0.16	-0.06	0.02	0.03
Case particle "No"	**0.84**	0.18	-0.07	0.02	0.00
Auxiliary particle "Nante"	**0.81**	-0.30	-0.04	-0.08	-0.02
Connective particle "Kara"	**0.79**	0.07	0.07	-0.21	-0.01
Incidental particle "Ha"	**0.78**	0.19	-0.14	0.09	0.12
Case particle "Kara"	**0.74**	0.07	-0.07	0.04	0.08
Case particle "Ni"	**0.71**	0.32	0.01	0.03	0.00
Auxiliary verb "Ta"	**0.70**	0.23	0.10	0.08	-0.08
Connective particle "Te"	**0.69**	0.37	-0.05	0.09	-0.05
Case particle "Wo"	**0.67**	0.27	-0.25	0.05	0.09
Final particle "Wa"	**0.66**	-0.31	0.00	0.29	-0.16
Connective particle "To"	**0.63**	0.10	-0.09	-0.08	0.15
Final particle "Ne"	**0.60**	-0.06	0.16	-0.03	0.10
Case particle "De"	**0.60**	0.38	0.07	0.03	0.00
Auxiliary verb "Chau"	**0.56**	0.11	0.08	-0.17	-0.14
Final particle "No"	**0.56**	-0.34	**0.53**	-0.02	-0.25
Auxiliary particle "Nanka"	**0.55**	-0.04	-0.01	0.06	-0.02
Auxiliary particle "Tte"	**0.51**	0.39	0.12	-0.03	-0.09
Auxiliary particle "Dake"	**0.51**	0.07	0.07	0.17	0.02
Connective particle "Ba"	**0.51**	-0.08	0.23	0.05	0.06
Auxiliary verb "Nai"	**0.51**	0.17	0.37	-0.26	0.02
Auxiliary particle "Made"	**0.50**	0.22	-0.07	0.02	0.06
Auxiliary verb "Teru"	**0.50**	0.34	0.24	-0.23	-0.03
Quasi-particle "No"	**0.47**	0.29	0.16	-0.10	0.29
Auxiliary verb "Tuu"	**0.47**	0.01	0.25	0.14	-0.06
Auxiliary verb "Rareru"	0.33	0.26	0.00	-0.02	0.01
Auxiliary verb "Reru"	0.28	0.23	0.15	0.05	0.20
Auxiliary particle "Ka"	0.02	**0.90**	0.11	-0.03	-0.13
Case particle "To"	0.31	**0.71**	0.00	0.05	-0.02
Final particle "Ka"	-0.16	**0.61**	0.19	0.07	**0.41**
Connective particle "Shi"	0.00	**0.60**	0.21	0.08	-0.12
Incidental particle "Mo"	**0.42**	**0.56**	0.02	0.03	0.03
Connective particle "Keredo"	**0.47**	**0.56**	0.00	-0.05	-0.02
Auxiliary particle "Tari"	0.24	**0.53**	-0.29	0.00	-0.04
Final particle "Na"	-0.05	**0.51**	0.28	0.32	-0.04
Auxiliary verb "Rashii"	-0.06	0.34	0.30	-0.03	-0.12
Final particle "Yo"	0.10	-0.11	**0.77**	0.04	0.28
Final particle "Jan"	-0.08	0.00	**0.67**	0.05	-0.13
Auxiliary verb "Da"	0.32	0.31	**0.56**	-0.07	-0.07
Final particle "Sa"	-0.15	0.36	**0.52**	0.04	-0.24
Final particle "Mono"	0.14	-0.16	**0.51**	**0.46**	0.05
Auxiliary verb "Tai"	-0.21	**0.41**	**0.47**	0.09	0.03
Final particle "Ke"	0.04	0.08	**0.42**	-0.08	-0.10
Auxiliary particle "Shika"	0.13	0.08	0.39	0.00	0.12
Auxiliary particle "Kurai"	0.27	0.25	0.29	-0.09	0.07
Auxiliary verb "Zu"	-0.07	0.08	0.18	**0.88**	0.15
Auxiliary verb "Toru"	-0.10	-0.03	0.18	**0.83**	0.09
Auxiliary verb "Ya"	0.04	0.24	-0.33	**0.58**	-0.15
Auxiliary verb "Desu"	0.13	-0.17	-0.17	-0.01	**0.98**
Auxiliary verb "Masu"	0.11	-0.04	-0.21	0.07	**0.85**

Table 2: Results of factor analysis of frequently appearing function words in the NUCC

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Female 10s	-0.04	0.32	1.06	0.35	-0.14
Female 20s	-0.16	0.19	0.18	-0.09	-0.12
Female 30s	-0.03	0.36	-0.16	-0.19	-0.03
Female 40s	0.05	-0.05	-0.50	-0.05	0.17
Female 50s	-0.34	-0.83	-0.80	0.07	-0.31
Female 60s	0.97	-0.34	0.04	-0.29	0.46
Female over 70s	1.27	-0.59	0.01	0.20	-0.53
Male 10s	0.22	0.76	1.39	2.82	0.14
Male 20s	-0.13	-0.06	0.30	0.53	0.17
Male 30s	-1.56	-1.29	-1.33	-0.31	-0.52
Male 40s	-0.02	-0.17	-0.71	0.24	0.73
Male 50s	-0.64	-0.86	-0.54	-0.40	0.16
Male 60s	0.26	-0.64	-0.58	0.12	0.87

Table 3: Average factor scores for each gender / age category in the NUCC

4. Factor Analysis for Utterance Styles

To extract the typical utterance styles of Japanese daily conversation, a factor analysis for frequencies of particles and auxiliary verbs was performed. Because of statistical limitations, the top 50 most frequent function words (particles and auxiliary verbs) were selected and 50 dimensional word frequency vectors were extracted for each speaker in each scene. The rotation method used was Promax and a parallel analysis was performed to determine the number of factors. As a result, five factors were identified. The resultant factor scores are shown in Table 2; the bold font signifies cells whose factor scores exceeded 0.4.

In order to investigate the meanings of each factor, the average factor scores were calculated as Table 3 in each of the categories from Table 1. The factor score shows the relationships between each factor and each data set. If a factor score is regularly high in some data sets, this suggests the correlation of the factor and the data sets.

The five factors corresponded with the frequently appearing utterance patterns in Japanese daily conversation in the NUCC. The characteristics and naming of each factor are as follows:

Factor 1: This factor includes general function words such as the case particles “Ga,” “No,” “Wo,” “Kara,” and “Ni.” Therefore, Factor 1 reflects neutral general usage in Japanese utterances. However, Factor 1 also includes some feminine characteristic words such as the final particles “Wa” and “No” as well as informal words frequently used by young speakers such as the auxiliary verbs “Chau” and “Tuu.” This combination of “neutral,” “feminine,” and “youth” characteristics may be occurring because of the bias of the NUCC. Table 1 clarifies that the NUCC includes more feminine and youth usage of utterances characteristically. Therefore, Factor 1 may represent “Average style in NUCC.”

Table 3 shows that this factor has a strong relationship among older females. This result may suggest that the traditional feminine utterance style is only applicable for older females in real conversation in modern Japan.

Factor 2: This factor includes such auxiliary particles as “Ka” and “Tari,” as well as the case particle “To,” the connective particle “Shi,” and the incidental particle “Mo.” These particles have the common functions of juxtaposing and continuation. This may reflect an utterance style of continued speaking without specifying the end of the sentence. This factor also includes the final particles “Ka” and “Na.” These two particles show some nuances of the interrogative form. These utterance styles may relate to both avoiding assertions and speaking in an ambiguous way. Above “Ka” and “Tari,” also have been utilized in similar way. This may be a result of some pragmatic strategy employed to avoid collisions and to enhance empathy. Therefore, Factor 2 is referred to as “Avoid affirmation style.”

This factor is commonly related to young females (10s, 20s, and 30s) and also to10s males. It may be characteristic of young female utterance styles in modern Japan. However, the 10s-male category includes only two people in the NUCC and therefore it cannot be concluded that Factor 2 is related to the young male demographic.

Factor 3: This factor includes such final particles as “Yo”, “Jan,” “Sa,” “Mono,” and “Ke.” These may characteristically reflect informal, frank communication styles. Moreover, Table 3 shows that this factor strongly related to 10s females and males. Therefore, Factor 3 is referred to as “Frank teenager style.” Although factor 1 also includes frank style, the differences from factor 1 are

gender free, youth only, and separation from general utterance style.

Factor 4: This factor includes the auxiliary verbs "Zu," "Toru," and "Ya." Although the auxiliary verb "Zu" is a somewhat general word, "Toru" and "Ya" are characteristically used in various dialects. In the NUCC, some speakers also have dialect tones, and those may be reflected on this factor. Therefore, it was labeled "Dialect style."

This factor is strongly related to 10s male in Table 3 because one of the two 10s male speakers has strong dialect tone. However it cannot be generalized because of too small sample size.

Factor 5: This factor includes the auxiliary verbs "Desu" and "Masu." These are clearly related to Japanese honorific utterance styles. Therefore, Factor 5 was referred to as "Polite style."

There is no certain tendency in the correlating categories of this factor in Table 3. Honorific utterance styles are dependent on the social relationships between the speaker and the listener. Therefore, the categories of gender and age seem not to be related meaningfully in this factor.

The results from the examination of the NUCC were compared to those observed in the utterance styles in Japanese fictional texts (Murai, 2017A, 2017B), and three of these factors also appeared in the fictional texts: "Frank style," "Kansai dialect style," and "Polite style." In the cases of those three factors, the included words are not exactly the same, but they are very similar between real and fictional utterances. The "Average style in NUCC" seems to be combination of "Neutral style" and "Feminine style" in the factors of fictional utterances.

Though in previous research seven fictional utterance styles were obtained, "Crude style" and "Hearsay style" have not been observed in the real-life daily conversation corpus. "Crude style" often reflects hostile relationships between fictional characters and therefore would not appear in the experimental daily conversation apart from rare situations of quarrel. "Hearsay style" is frequently used in fictional conversation in order to diversify narrative forms. However, such diversification may not be necessary in everyday conversation.

On the other hand, "Avoid affirmation style" has not been observed in the fictional utterance corpus. It may be a new utterance style in modern Japan. Therefore, fictional writers may not recognize this utterance pattern. Instead that, fictional writers adopt traditional feminine speech style for their fictional feminine characters. However, traditional feminine speech style is used mainly in over 60s (factor 1) in real corpus data. This result would be help to understand the time span of utterance style change.

5. Conclusion and Future Work

The characteristics of fundamental utterance styles in Japanese daily conversation were analyzed by a factor analysis method based on the NUCC. As a result, five factors ("Average style in NUCC," "Avoid affirmation style," "Frank teenager style," "Dialect style," and "Polite style") were extracted quantitatively.

Compared to fictional utterance styles, "Avoid affirmation style" is unique for real-life daily conversation. On the other hand, "Crude style" and "Hearsay style" did not appear in that corpus.

Although the similarities between the fictional corpus and the NUCC support the validity of the result, the factors were likely affected by the bias of the corpus. It would be desirable to utilize a speaker-balanced daily conversation corpus for more precise analysis.

Moreover, knowledge of the relationships between the speakers and the listeners would be useful for obtaining detailed characteristics of utterance styles.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 26730168, the NINJAL collaborative research project 'A Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation', and the NINJAL project 'Corpus of Everyday Japanese Conversation'.

7. Bibliographical References

Kinsui, S. (2003). Virtual Japanese: Mystery of Functional Words. Iwanami Shoten, Tokyo. (In Japanese)

Kurosawa, A. (2010). The sentence-final forms used in Meidai Dialogue Corpus: Does the plain style differ from the polite style? Yamagata University Working Papers in International Education, 2, pp. 3–11. (In Japanese)

Miyazaki, C., Hirano, T., Higashinaka, R., Makino, T., Matsuo, Y., & Sato, S. (2014). Fundamental Analysis of Linguistic Expression that Contributes to Characteristics of Speaker. In the Proceedings of the Association for Natural Language Processing, pp. 232−235. (In Japanese)

Murai, H. (2017A). Situational Effects on Functional Word Frequencies within Conversational Sentences in Japanese Novels. Proceedings of JADH Annual Conference 2017, pp. 40—42.

Murai, H. (2017B). Characteristics of Utterances in Japanese Fiction-writing. IJCAI 2017, the 2nd. International Workshop on Language Sense on Computer, pp. 6—10.

Seto, K. & Kishi, Y. (2015). Construction of a Dialogue System Using a Speech Type of Estimation by Adjacency. Proceedings of Information Processing Society of Japan 2015, pp. 131—132. (In Japanese)

Shen, R., Kikuchi, H., Ohta, K., & Mitamura, T. (2012). Towards the text-level characterization based on speech generation. Journal of Information Processing Society of Japan, 53(4), pp. 1269–1276. (In Japanese)

8. Language Resource References

Fujimura, I., Chiba, S., & Ohso, M. (2012). Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a Lexical Profiling Approach to Comparing Spoken and Written Corpora, Proceedings of the VIIth GSCP International Conference. Speech and Corpora, pp. 393—398.

Biomechanics for understanding movements in daily activities

Yasuyuki Yoshida, Takuichi Nishimura, Kristiina Jokinen
AIRC AIST Waterfront
Aomi, Koto-ku, Tokyo JAPAN
{ firstname.lastname}@aist.go.jp}

Abstract

We discuss biomechanics and its use in studying human movements especially in sports and exercise events, and how sensor information from the devices such as acceleration sensor, gyroscope, force plate, and motion capture system can be effectively used to gain a greater understanding of human movements in every-day activities and communicative situations as well. Using AI and IoT technology, we propose to apply the approach to collect, analyse, and annotate motion data in common activities.

Keywords: biomechanics, movement, gesturing, everyday activity analysis

1. Introduction

In language communication, interlocutors effectively accompany their speech with gestures and body movements. These movements range from unconscious moving to intentional gesturing, and they have various functions such as giving rhythm to one's speech, indicating engagement in conversation, pointing, coordinating interaction, and of course, performing certain physical actions. Until recently, studies have been based on video analysis and manual annotations (cf. Allwood et al. 2007, Jokinen 2011). Several annotation tools such as Praat (Boersma and Weenink 2009) for speech and Anvil (Kipp, 2001) and Elan (ELAN) for video data can be used for detailed analyses. However, it is time-consuming and often difficult to manually annotate timing and amplitude of the various actions and activities accurately, and so advance video analysis has been used to extract movements of conversational participants using OpenCV toolkit (Bradski and Koehler, 2008), see e.g. Vels and Jokinen (2015) who experiment with bounding boxes, and Jongejan (2016) who provides a plugin to include velocity and acceleration of head movements from video analysis to Anvil-annotations. Sensor and tracking technology has been developed especially in medical domain, and used to analyse e.g. non-verbal behavior (Philippot et al. 2003) and measure movement in Parkinson disease (Galna et al 2014). The Human Communication Dynamics framework (Stratou and Morency, 2017) aims at a unified approach to address challenges in multimodal behavior analysis, and to jointly analyse the participants' language, gestures and social signals for efficient computational perception algorithms in behavioral sciences and real world applications.

In this paper we present a new methodological approach to study movement in human conversation and daily activities, based on Biomechanics. We follow the approach of Human Communication Dynamics, but differ from this in that we especially aim to study human movements and motor learning in everyday activities where the movement analysis is not necessarily used to infer communicative intentions of the participants, but to perform certain actions better, as when instructing learners how to move their body in a correct way in DanceSport, care-taking, etc.

We explore biomechanics in automatic detection and analysis of human motion, and the results of our experiments show that the joint use of various sensor data enables us to achieve accurate perception of human motion. It is thus possible to achieve a better understanding of the different aspects of human motion and to study how they function in everyday communication and signal the participants' engagement in interactive situations. The data can be used in various practical applications developed for the health and well-being of the people.

Another important contribution of the paper is the new methodology that can be used in human-human and human-robot interaction studies. Sensor information allows us to observe human motion and gesturing in everyday activities, and we can then analyse it automatically using machine-learning techniques. Using IoT possibilities to share the sensor information with a communicating robot, the data can be directly used in the control and coordination of the interaction between the human and the robot. If the robot is equipped with the knowledge of the motions in general, e.g. annotations and ontologies of the motion data, it is possible to explore how a robot agent can learn common activities by imitation and explicit instruction.

The paper is structured as follows. We will first briefly introduce biomechanics and the sensors used in the motion and gesture analysis in Section 2. We will then describe the experiments in motion data collection in Section 3. Finally, we discuss methodological issues concerning the application of biomechanics and sensory data for the understanding of the human every day activities in real life.

2. Biomechanics

The new technology on sensors has been significantly advanced in the recent years. Various robust high speed and sensitive devices, such as the acceleration sensor, gyroscope, electromyography, force plate, and motion capture system, have been developed to measure motion and body posture with high accuracy and precision. Information from the sensors can then be effectively used to collect and analyse data on human movements.

Biomechanics is a study of human movement. It applies the laws of mechanics and physics to human performance and aims to explain how and why the human body moves as it does by analysing the forces acting on the body (kinetics) and the movements of the body (kinematics). It is used especially in sport and exercises, with two main purposes: to improve physical performance, and to prevent injuries. Besides human movements in sports and exercises, biomechanics can also be used to study daily activities such as walking, sitting and lifting. Using AI and IoT technology, we propose to apply the biomechanis approach to collect, analyse, and annotate motion data in common daily activities, including language communication.

In biomechanical experiments, sensor technology is widely deployed, and a motion capture system and force plates are frequently used (Figure 1). These instruments can quantify the human movements from dynamics. The motion capture system is used to measure the position data of body segments, while the force plates are used to measure ground reaction forces. The data is interpreted with respect to knowledge about the human anatomy and physiology, and inverse dynamics is used to compute the turning effect of the anatomical structures (muscles, ligaments) in joints, which is necessary to perform the particular motion.

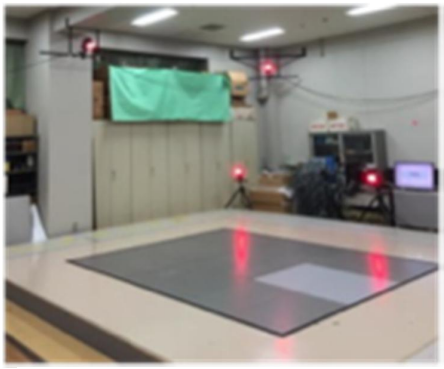


Figure 1 Force Plate and Motion Sensors.

Figure 2 shows a snap-shot of a motion tracker system depicting a person balancing on a force plate. The force plate is a device that measures the three components of a force (along x, y, and z axis) applied to the surface, as well as the vertical moment of force. It is used to measure acceleration, work, and power of locomotion, and can also measure the angle and distance of a move such as a jump. Combined with kinematics of the joint angles, it is possible to determinate torque, work and power for each joint to study movement e.g. for robotics and sports applications.

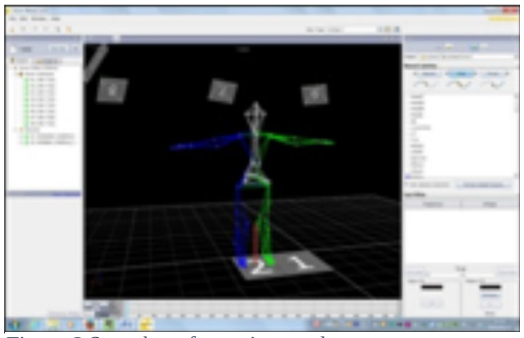


Figure 2 Snapshot of a motion tracker system.

According to Hooke's law, force is directly proportional to extension distance on a linear spring: $F = -kX$, where k is a constant factor and characterizes the stiffness of the spring. Besides the linear force that pushes and pulls an object, movement can also be twisted by a rotational force called torque or moment of force. Torque is defined as the rate of change of angular momentum of an object, and it is directly proportional to angle of rotation on torsion spring (Figure 3). Torque is measured in Newtonmeters (Nm).

Previous studies have shown that muscles have elastic function (Komi 2000). Research about human and animal locomotion have used the spring-mass model to explain the interdependency of the mechanical parameters that characterize the movement, especially running and hopping. The spring-mass model is a simple model that represents the mass of the actor as a single point mass, and the musculoskeletal system as a spring. During running and hopping, lower extremities can be modelled by a linear spring (Farley and Morgenroth 1999), while lower extremity joints can be modelled by torsional spring model (Hobara 2009; Hobara 2010).

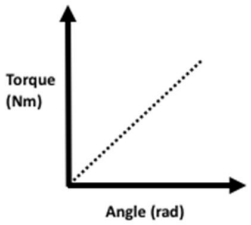


Figure 3 The relationship between torque and angle based on Hooke's law.

Although the actual body is a complicated set of muscles, bones, tendons, and ligaments which act across and upon joints to produce movement, the spring-mass model describes and predicts the mechanics of the movements in an accurate manner. It can be concluded that the individual elements of the musculoskeletal system are integrated in a way that allows the overall system to behave like a simple spring during running and jumping. It is also possible to adopt the spring model to study joints and body parts in various other activities as well, besides running and jumping (see below). Furthermore, it is possible to represent the body's movement ability, or stiffness, by the spring constant k, and much research has focused on determining this constant.

3. Experiments and applications

Development and increased stability of motion trackers as well as sensor technology provide help in quantifying movement. In this section we summarize our research on daily activities, such as walking and dancing, using this information. The purpose of the studies has been to analyze whether the torsion spring model can be applied to the axial twisting movements in ballroom dancing and other activities. We also present Axis Visualizer, a mobile phone application to visualize motions.

3.1 Axis twisting experiment

Axial twisting movements along the longitude axis occur frequently. For instance, during walking the upper body and lower body rotate in opposite direction. We used a motion capture system to measure angle for rib cage, and a force plate to measure torque. The correlation between the angle and torque was calculated and compared with the spring model predictions. The setup of the experiment is as shown in Figure 4. Participants had to do axial twisting

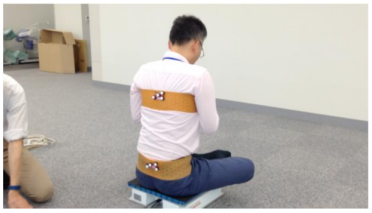


Figure 4 Setup for the axial twisting experiment.

movement sitting on the force plate. After a short practice, the experiment started with a minimum of 10 seconds axis twisting in two conditions: in a slow, relaxed condition, and in a fast, intensive condition. The results are shown in Figures 5 and 6. The smooth harmonic curve and the linear correlation between torque and the angle show that the repetitive axis twisting movement can be modeled using the spring model (more details in Yoshida et al. 2018).

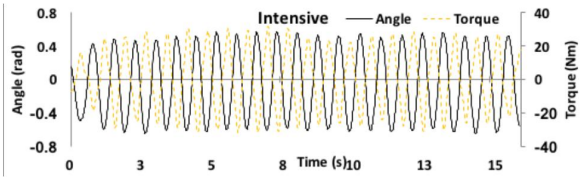


Figure 5 Visualisation of torque and angle in an intensive axis twisting movement. From Yoshida et al. (2018).

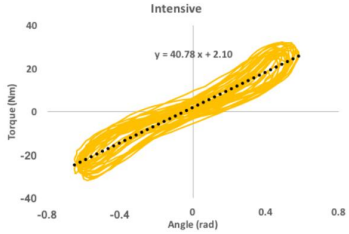


Figure 6 Relation of torque and angle in an intensive axis twisting movement. From Yoshida et al. (2018).

3.2 DanceSport

One of the popular dancing styles in the world is ballroom dancing, nowadays called DanceSport. Dancing can effectively help in fitness and wellbeing, and the exercise effects of DanceSport have already been proved. For instance, Rehfeld et al. (2017) consider the effects of a long (18 months) dancing intervention on elderly people's fitness and well-being, and how it can be efficiently used to enhance motoric capabilities of the elder people thus preventing injuries that stem from inaccurate or fragile motor control.

Dancing is also a good example of a movement which requires balance and smooth locomotion over a large area. Moreover, it requires coordination between two persons. Biomechanical analysis can provide a detailed analysis of the timing, amplitude and speed of the joint movements by the dancer's, allowing accurate quantitative measuring of the coordination in dance configurations. In the preliminary experiments with the Japanese professional dancers, we have noticed e.g., that the amplitude of the joint movements is less compared with the same movements performed by the individual dancer alone (showing the dance movement without the partner), while the rotation speed is slower in individual dancing. We will continue analysing the data from the All Japan Ballroom Dance Competition, to get a clearer understanding of the dance movements. The results can be used for learning and practise purposes, and to train competitors for better individual performance.

Biomechanical data can also be used to investigate human coordination in general, e.g. in joint tasks like cooking, assembling devices, or communication. In particular, since language communication is a cooperative activity whereby interlocutors use gesturing and body posture to coordinate

Figure 7 Axial twisting movement for Axis Visualizer.

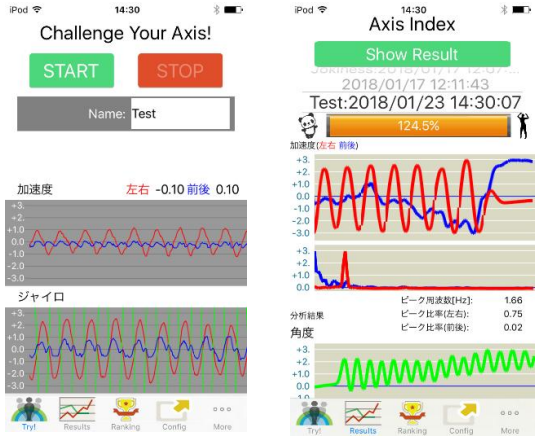


Figure 8 Two screen shots for Axis Visualizer.

the flow of interaction, such accurate measurements of the movements can be used to study engagement in interaction, i.e. to investigate how the interlocutors pursue their communicative goals while simultaneously pay attention to the partner in order to understand the partner's intention. Biomechanical measures allow us to calculate correlations between timing and location of the individual movements, and also include body rotation and speed of the movements as parameters to understand the posture of participants.

3.3 Axis Visualizer

Many people use activity trackers and smartwatches to measure various activities of their daily lives. As a practical application of the biomechanical information for everyday use, we developed an easy-to-use application for mobile terminals which allows the user to assess smoothness of their axial twisting exercise. The application is called Axis Visualizer and it is meant to function as a quick and simple assessment tool. The application deploys iOS Sensors for acceleration, while gyro inside the mobile terminal is used to analyze the spring model (see Section 3.1). The app can be used by simply attaching the mobile terminal to one's chest and doing the axial twisting movement for a short time, as shown in Figure 7. After the exercise, the system analyses the motion, and calculates whether the movement was harmonic. Two screenshots of the app displaying the result of an exercise are shown in Figure 8.

4. Discussion

Accurate biomechanical information has been mainly used for medical testing and rehabilitation tasks as well as for advanced studies on neuro-cognition and biomechanical feedback. We propose to apply the approach to collect, analyse, and annotate motion data in common everyday

activities to increase understanding of human behaviour in real situations and to be able to build models for their computational assessment. We provided two examples of this kind of research and discussed an axis twisting experiment and DanceSupport.

Biomechanics data can also be collected using portable devices. This opportunity provides an interesting option for researchers who aim at studying interaction in real-life situations. So far, the participants' movements have been studied from video recordings or by using specific motion-tracker devices, which require the data collection to take place in laboratories. However, for ecologically valid data, it is important to be able to measure everyday activities in real-life situations, using simple devices and easy-to-use interface. The mobile application, Axis Visualizer, can be considered as the first step in this direction, since it exhibits the possibility to use a mobile phone to record motion and get an overview of the person's real-life activities.

In natural multimodal communicative situations, the connection from the visual scene to cognitive interpretation and appropriate conversational responses is important to understand the relevant mechanisms for human-human communication and for interactions between human and robot agents (cf. Jokinen and Wilcock, 2013). Hand and head movements are effectively used as signs that e.g. point to an object of interest, coordinate turn-taking by mutual gaze, and accompany the speaker's speech with beat movements (Kendon 2004, Paggio et al. 2010, Jongejan 2012, Jokinen 2011).

An interesting area of research is simultaneous timing of hand gestures, eyes, and nodding. The eyes and hands are used together in many everyday tasks, and it has been shown that the eyes generally direct the movement of the hands to targets: the eye-gaze is about one second ahead of the action start (Land 2006). Furthermore, the eyes provide initial information of the object (its size, shape, and possible grasping locations) so that the human can determine the motion of the hand, the hand shape and force to be used in the fingertips in order to exert suitable level of force and coordination to perform a task. The complexity of the coordination of eye and hand to perform everyday tasks is an interesting challenge for studies in cognition and neural control of eye and hand coordination, but it is also important in clinical work concerning disorders and impairments. For instance, in older adults, eye-hand coordination has been shown to decrease especially in tasks involving fast and precise movements, e.g. such everyday tasks as picking up a pen or making tea can become difficult. Having technology which enables training and assistance in such situations is useful for improving independent living and wellbeing. In various sporting performances, computer games, typing, etc. feedback through biosensors and biomechanics can give accurate information about how the task is progressing and what kind of changes in the task procedure are necessary to improve the system design and logistics of the interaction.

Given that the new technology allows several different data flows to be recorded and analysed, a unified approach to data model is necessary, cf. Human Communication Dynamics framework (Stratou and Morency, 2017). Some discussion can also be found in Hall and Llinas (1997), and more recently in Blaauw et al. (2016), from the sensor integration point of view, and we aim at exploring with the Fusion Model to enhance our understanding of gestures and movements in communication, to build models for the conversational rhythm and for the interlocutors' interest and involvement in the interaction and to better estimate human engagement in smooth communication. Moreover, combined with the knowledge of actions and activities, it is possible to experiment with automatic learning, i.e. to learn to recognize gestures and action sequences automatically. In attempts to teach a robot agent to perform certain tasks, e.g. pick up a pen, data about the correct movement patterns is necessary, and the proposed method can be an efficient way to collect accurate data.

Considering the IoT context of intelligent homes and public places, the use of biometrics and sensor data brings in a possibility to record everyday activities in real situations in the ubiquitous environments. The data can be immediately shared with other devices, e.g. with robots, which can thus learn about human motion and be able to provide assistance that is relevant in a given context. For instance, in elder care scenarios, a fall of an elder person onto the floor, irregularities in sleeping patterns or toilet use, wandering around the rooms, or not being able to find keys, can be noticed by a ubiquitous system which can then act in an appropriate manner (call for human help, suggest a keyring location, etc.)

The approach also brings in questions about the reliability of the information which depends on the technology. For instance, force platforms can be inexpensive off-the-shelf consumer products which makes it easy to conduct experiments. However, if used in exercise and health-care applications for measuring a patient's balance and mobility performance, their adoption should be carefully checked, and manufacturing should be in accordance to quality standards as established by ISO.

Like any data collection nowadays, the use of sensors needs to be considered with respect to some ethical aspects. Biomechanics allows people to be accurately identified by their physical features and typical behavior, so it will be possible to uniquely identify people. Data collection thus requires extremely careful consideration and planning and brings in questions about data storage and re-use. Statistical methods allow models for anonymous data source, and the data can be deleted after the analysis, but the issues related to building general models or individual models for certain physical and behavioral characteristics remain. Also, high-quality technology can enable attackers and people may give away information without their consent or knowledge.

5. Conclusions

Biomechanics is an area of research widely used in sport and medical domains for rehabilitation and improving performance. In the context of language communication, we expect that it will be possible to use the same approach to collect data, and through modeling, simulation and measurement gain a greater understanding of performance in everyday tasks and communicative events.

6. Acknowledgements

We would like to thank the participants in the experiments and our colleagues for assistance in carrying out work. We also thank the NEDO project for the support of the work.

7. 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. In Martin et al. (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*.

Blaauw, F.J., Schenk, H.M., Jeronimus, B.F., van der Krieke, L., de Jonge, P., Aiello, M., Emerencia, A.C. (2016). Let's get Physiqual – An intuitive and generic method to combine sensor technology with ecological momentary assessments. *Journal of Biomedical Informatics*, vol. 63, page 141-149.

Boersma, P. and D. Weenink (2009). Praat: doing phonetics by computer (version 5.1.05). Retrieved May 1, 2009, from http://www.praat.org/

Bradski, G. and Koehler, A. (2008). Learning OpenCV: Computer Vision with the OpenCV Linbrary. O'Reilly.

ELAN. http://www.lat-mpi.eu/tools/elan/

Farley, C.T. and Morgenroth, D.C. (1999). Leg stiffness primarily depends on ankle stiffness during human hopping, J. *Biomech*., Vol.32, No.3, pp.267–273.

Galna, G., Barry, G., Jackson, D., Mhiripiri, D., Olivier, P., Rochester, L. (2014). Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease. *Gait and Posture* 39(4): 1062-1068.

Hall, D.L., Llinas, J. (1997). An introduction to multisensor fusion. *IEEE Special Issue on Data Fusion*, 85(1).

Hobara, H., Inoue, K., Gomi, K., Sakamoto, M., Muraoka, T., Iso, S. and Kanosue, K. (2010). Continuous change in spring-mass characteristics during a 400m sprint, *J. Sci. Med. Sport*, Vol.13, No.2, pp.256–261.

Hobara, H., Muraoka, T., Omuro, K., Gomi, K., Sakamoto, M., Inoue, K. and Kanosue, K. (2009). Knee stiff- ness is a major determinant of leg stiffness during maxi- mal hopping, *J. Biomech*., Vol.42, No.11, pp.1768–1771.

Jokinen, K. (2011) Multimodal Information - Collection and Analysis of Interactive Data. HCII, Orlando, U.S.

Jokinen, K., Wilcock, G. (2013). Multimodal Open-domain Conversations with the Nao Robot. In: Mariani, J., Devillers, L., Garnier-Rizet, M. and Rosset, S. (eds.) *Natural Interaction with Robots, Knowbots and Smartphones - Putting Spoken Dialog Systems into Practice*. Springer Science+Business Media

Jongejan, B. (2012). Automatic annotation of head velocity and acceleration in Anvil. European language resources distribution agency, 5. pp. 201–208

Kendon, A. (2004). Gesture: Visible Action as Utterance. Cambridge University Press

Kipp, M. (2001). Anvil – A generic annotation tool for multimodal dialogue. Proceedings of the Seventh European Conference on Speech Communication and Technology, pp. 1367–1370.

Komi, P.V. (2000). Stretch-shortening cycle: A powerful model to study normal and fatigued muscle, *J. Biomech.*, Vol.33, No.10, pp.1197–1206.

Land, M.F. (2006). Eye movements and the control of action in everyday life. *Progress in Retinal and Eye Research* 25, pp. 296-324.

Paggio, P., Allwood, J., Ahlsen, E., Jokinen, K., Navarretta, C. (2010). The NOMCO multimodal Nordic resource - goals and characteristics. Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10).

Philippot, P., Feldman, R. S., Coats, E. J. (2003). *Nonverbal Behavior in Clinical Settings*, New York, NY, USA:Oxford University Press.

Rehfeld K, Müller P, Aye N, Schmicker M, Dordevic M, Kaufmann J, Hökelmann A, Müller NG. (2017). Dancing or Fitness Sport? The Effects of Two Training Programs on Hippocampal Plasticity and Balance Abilities in Healthy Seniors. *Front Hum Neurosci*. 15(11): 305.

Stratou, G. and Morency, LP. (2017). MultiSense—Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case. In: *IEEE Transaction on Affective Computing*, 8(2).

Vels, M. and Jokinen, K. (2015). Detecting Body, Head, and Speech Signals for Conversational Engagement. *The IVA Conference Workshop 2: Engagement in Social Intelligent Virtual Agents*. Delft, The Netherlands.

Yoshida, Y., Liang, Z., Nishimura, S., Konosu, H., Nagao, T., Nishimura, T. (2018). Quality Evaluation For Sports Coaching - Evaluate Trunk Torsion by Mobile Terminal *IPSJ Journal* 59(2) 1-11.

F-formation and social context: How spatial orientation of participants' bodies is organized in the vast field

Yasuharu Den
Graduate School of Humanities, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, Japan
den@chiba-u.jp

Abstract

In this paper, we illustrate how participants in conversations conducted in the vast field spatially orient their bodies to each other depending on the environments and the contexts they are in. In particular, we focus on the way in which body arrangements in *F-formations* are influenced by social contexts, such as social relationships among participants and their roles in the activity. A detailed analysis of the video data from our fieldwork at Nozawa-Onsen Dosojin festival shows that participants develop various body arrangements such as the circular, the side-by-side, and the 'horseshoe' formations, with or without outsiders. We discuss dynamic social contexts, i.e., *membership categories* relevant to the ongoing activity, play an important role in organizing these spatial-orientational arrangements.

Keywords: Spatial orientation, F-formation, body arrangement, membership category, fieldwork

1. Introduction

When people engage together in conversation with each other, they often enter into a distinctive spatial-orientational arrangement. Kendon (1990) proposed the notion of *F-formation*, in which participants actively cooperate to sustain a shared inner space, called *O-space*, where the main activity takes place. In the case of talk while standing among three or more participants, the conversational group is organized typically as an F-formation in circular arrangement.

The notion of F-formation has been extended in subsequent studies. McNeill (2006) distinguished *social* and *instrumental* F-formations; the former is the Kendon's original version, while the latter is the space in which two or more people gaze at, point to, or operate on a commonly focused object. Kendon (2010) illustrated various kinds of body arrangements in F-formations including circular, side-by-side, and 'horseshoe' arrangements, with or without commonly focused objects. F-formation, and its tightly related notions, have also been investigated in various situations such as poster presentations (Bono et al., 2004), archaeological field (Goodwin, 2003), garden lessons (Mondada, 2012), guided tours (De Stefani and Mondada, 2014), and communication at a science museum (Makino et al., 2015).

In the fields of the previous studies, the space for the formation is relatively small.[1] In the field of the present study, on the other hand, the space is vast. We have been conducting, for six years, fieldwork at Nozawa-Onsen village, located in the northern part of Nagano Prefecture in Japan, in which we video-record and analyze a huge number of people working together for the preparation of the Dosojin festival, one of the biggest fire festivals in Japan (Enomoto and Den, 2015). The festival site is extensive, about 40 meters square, and people often talk referring to a distant object that is tens of meters away. In such a situation, conversational participants create various kinds of spatial-orientational arrangements.

In this paper, we illustrate how participants in conversations conducted in the vast field spatially orient their bodies to each other depending on the environments and the contexts they are in. In particular, we focus on the way in which body arrangements in F-formations are influenced by social contexts, such as social relationships among participants and their roles in the ongoing activity.

2. Data

2.1. Overview

The materials are video recordings of the preparatory works for the Nozawa-Onsen Dosojin festival. The Nozawa-Onsen Dosojin festival is one of the three greatest fire festivals in Japan, and is designated as a significant intangible folk cultural asset. Major preparatory works for the festival begin in October, when the trees to be used for building a huge wooden structure, or shrine pavilions, are cut down in the mountain. Two of the five sacred trees, which have been left halfway up the mountain, are brought down through the village on January 13 prior to being made into the shrine. By the afternoon of January 15, the shrine is constructed without using heavy machinery. The festival takes place in the evening of January 15 every year, where a "fire-setting battle" between the guards and the torch bearing villagers is being performed for a couple of hours and ends up with setting fire to the shrine.[2]

Two or more (up to 8) researchers made video-recording of various activities concerning the preparatory works for the festival with roving cameras. The data for the present study is a video clip recorded in the morning of January 12, 2017, lasting about 30 minutes. In that morning, as many as 40 people were working together at the festival site, divided into several groups according to the tasks. We focus on conversations conducted by the most central group, which occurred ubiquitously in the vast field.

[1] These fields, e.g., garden, museum, etc., could be vast, but the space for a formation at a particular moment in the activity is small, involving commonly focused objects nearby participants.

[2] See more details of the Dosojin festival at, e.g., `https://nozawa-onsen.com/nozawa-fire-festival/`.

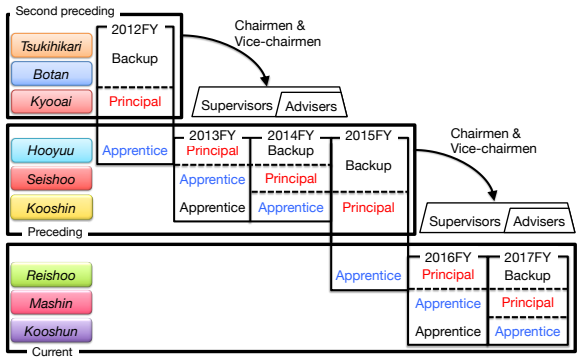


Figure 1: *San'yako* (Three-nights scheme)

Figure 2: F-formation with three participants (PSM, CSM, and PM) in circular arrangement and two outsiders standing side by side (CC and NC)

2.2. Participants

The festival and its preparation are managed by a group of men, called *San'yako* (literally, 'three-nights scheme'), consisting of about 100 people at three consecutive ages (Figure 1). Each sub-group in *San'yako*, consisting of people at the same age, has a unique team name, such as *Hooyuu* and *Reishoo* (see the leftmost column in Figure 1). In *San'yako*, the 42 year-old men, at a climacteric age, serve as principal members, the younger men as apprentice members, and the elder men, if any, as backup members. In particular, the 41 year-old apprentices are working together with the principal members all the time in order to learn the knowledge and the skills that will be required when they become the principal members in the following year. The chairman of the principal members commands the whole group and has the strongest authority.

On a three-year cycle, the members of *San'yako* are replaced by people of the next generation. The three chairmen and the three vice-chairmen in the preceding *San'yako* form *Hozonkai* (literally, 'preservation association') and supervise the *San'yako* of the next generation. The eldest chairman in the supervisors becomes the shrine master, who supervises the development of the festival site and the construction of the shrine pavilions, which will be burnt in the end of the festival. In the 2016 FY's (from April 2016 to March 2017) festival, the chairman of the *Hooyuu* team took control as the shrine master for the first time (see Figure 1). He learned the knowledge and the skills required for a shrine master last year from the preceding shrine master, who is the eldest chairman in the second preceding *San'yako* (the chairman of the *Tsukihikari* team). The main participants of the study are the following four persons: i) the current shrine master (CSM; from *Hooyuu*), ii) the preceding shrine master (PSM; from *Tsukihikari*), iii) the current chairman (CC; from *Reishoo*), and iv) the next chairman (NC; from *Mashin*).

3. Analysis

On the festival day, the festival site is blanketed by snow. In fiscal 2016, however, there was shortage of snow. The *San'yako* members brought snow from various parts of the village into the festival site, and bulldozed the site. CC was commanding the whole group at the site, and NC, as an apprentice, was always acting with CC. CSM came here to give instruction to CC concerning the development of the site. Because this was the first time for CSM to act as a supervisor, PSM also came here to give advice to CSM. PSM commented on the level of the snow surface and where to build the shrine pavilions.

In this section, we demonstrate four distinctive spatial-orientational arrangements of participants' bodies while conducting conversations in this vast field.

3.1. Case 1: Circular arrangement with outsiders

The four participants, CSM, PSM, CC, and NC, enter into the festival site, and walk forward into the back of the site. PSM finds a member of the preceding *San'yako* (PM in Figure 2), who is there for manipulating a loading shovel, and talks to him, getting into a chat. These two men and CSM, another member of the preceding *San'yako*, get into an F-formation in circular arrangement as in Figure 2. Interestingly, the other two participants, CC and NC, stand outside the circle, in the R-space of the F-formation. Kendon (2010) argues that such outsides usually exhibit an orientation either to entry into or to passing the F-formation. In this excerpt, however, CC and NC stay there to sustain this twofold arrangement.

A possible factor behind this spatial-orientational arrangement seems to reside in the social relationship among these people. Both of CSM and PM belonged to the preceding *San'yako*, and PSM supervised them as the preceding shrine master. CC and NC, on the other hand, do not have a direct relation with PSM or PM. In other words, there are two distinguishable sub-groups, or *membership categories* (Sacks, 1972), as to whether or not they have direct relation to the preceding *San'yako*.[3] This social context is manifested as the twofold spatial-orientational arrangement that is sustained through the conversation.

3.2. Case 2: Side-by-side arrangement in two rows

PSM talks to CSM about the level of the snow surface, referring to the view in front. They are in side-by-side

[3]In fact, CC, a member of the *Reishoo* team, acted with the preceding *San'yako* last year as an apprentice, but this relationship seems not in effect here. This relationship may become relevant only through his direct superiors, i.e., the *Kooshin* members.

Figure 3: Side-by-side arrangement in two rows, each consisting of two participants (PSM and CSM in the front row, and CC and NC in the back row)

Figure 4: F-formation with four participants (PSM, CSM, CC, and NC) in 'horseshoe' arrangement, in which no outsider is present

arrangement, watching the front view. Interestingly, again, CC and NC stay at the back of them, watching the same view. The four participants, thus, form a side-by-side arrangement in two rows, as shown in Figure 3.

A similar social factor as in Case 1 operates here, but in this case, the relevant category that distinguishes two sub-groups may not be the preceding *San'yako* but the shrine master. PSM, the preceding shrine master, is giving advice to CSM, the current shrine master. They are engaged in an activity of handing skills of a shrine master on the next generation. Although the land development of the festival site is also concerned with the task of the chairmen, in this membership categorization, CC and NC belong to a different sub-group from PSM and CSM; hence, two-row side-by-side arrangement emerges.[4]

3.3. Case 3: 'Horseshoe' arrangement with no outsider

As illustrated by Kendon (2010), people sometimes produce a kind of compromise between the side-by-side and the circular form, i.e., 'horseshoe' arrangement. In Figure 4, the four participants are in this formation. The 'horseshoe' arrangement enables participants to easily switch from a business talk to a more casual talk, and vice versa. Right before this excerpt, PSM was sitting down on the ground and showing the desired snow level to the other three participants, with his extended left arm. He stands up and starts joking to CC, now entering into the 'horseshoe' arrangement shown in Figure 4. The four participants sustain the formation during a chat.

Note that there is no outsider, or 'double standard,' in this formation. Unlike Case 2, the activity here is not necessarily considered as an activity of handling skills of a shrine master from PSM to CSM. Rather, PSM's depiction, with his arm, of the snow level is addressed to all of the other three participants. In this sense, there is no distinguishable sub-group. The equality of status among the four participants becomes further obvious when the activity shifts from a business talk to a casual talk. Unlike Case 1, where PSM's initiation of a chat with a part of the participants is driven by his encounter with PM, in Case 3, there is no event that can separate the participants into different categories. Rather, the chat is initiated by PSM's joke directly addressed to CC, thereby PSM deliberately invites CC to the same group as he belongs to.

This example clearly shows that spatial-orientational arrangement of participants' bodies is determined not merely by *static* social factors, such as hierarchical relationship based on age or official position, but by *dynamic* social contexts, i.e., the membership categories considered, by the participants, as relevant to the ongoing activity. PSM/CSM and CC/NC are regarded as belonging to different groups in Cases 1 and 2, where contrast between two categories, i.e., member vs. non-member of the previous *San'yako* in Case1 and person fulfilling vs. not fulfilling a role as a shrine master in Case 2, is implicated by the activity they engage in. By contrast, they are all members of the same, single group in Case 3; co-worker or chat partner is only relevant category in this situation, and no alternative is relevant.

3.4. Case 4: Side-by-side arrangement with one headliner

Further evidence for insufficiency of static social factors is spatial-orientational arrangement shown in Figure 5, in which three participants, PSM, CC, and NC, are standing side-by-side at the back and one headliner, CSM, at the front. PSM is the eldest in this group of people, and is in a position of giving advice to CSM. Thus, it is somewhat odd, at least in terms of hierarchical relationship based on age or official position, that CSM alone is standing ahead of the other three, in particular PSM.

Arrangements with one participant in a distinctive position are widely observed in activities such as lectures, classroom interactions, performances, and so on (Kendon, 2010). Giving an explanation to other participants is another example (Makino et al., 2015).[5] In the current case, however, CSM is not engaged in such an activity. He is giving CC and NC

[4]This account is further evidenced by an observation that when CSM gives CC a brief instruction about the level of the snow surface, he tentatively stands back, leaning a little closer to CC, but keeps his body oriented to the front. In doing so, CSM treats the interaction with CC as a *side involvement*, which is distinguishable from the interaction with PSM, the *main involvement* (Goffman, 1963).

[5]Prior to this excerpt, PSM gave an explanation of why the edges of the festival site should be raised above the level of the central part, facing to the other three participants, who were standing side-by-side in a row.

Figure 5: Side-by-side arrangement with three participants (PSM, CC, and NC) in the back row and one headliner (CSM) in front

instruction about how to complete the land development of the festival site, referring to the view in front of them.

The difference of this activity from lecture-like activities is also visible in CSM's body orientation; CSM's body is primarily facing to the same direction as the other three are facing to, which is never observed in lecture-like activities. CSM occasionally turns his head towards CC and NC when he talks to them, but his body stays facing to the front (Schegloff, 1998), suggesting that his main involvement is kept in an activity involving some object or view in front of him, not in a talk with men at the back. In this respect, it is similar to Case 2, shown in Figure 3. There is, however, a significant difference between Cases 2 and 4. In Case 2, the main activity is CSM's learning skills of a shrine master from PSM, while in Case 4, CSM is not engaged in a learning activity but in an instructing activity. Thus, his social role as the *current* shrine master, who supervises *San'yako*, not as an apprentice shrine master, is most relevant here. This membership categorization leads to the spatial-orientational arrangement with one headliner standing alone in front.

4. Discussion

We illustrated how participants in conversations conducted in the vast field spatially orient their bodies to each other. In particular, we focused on the way in which body arrangements in F-formations are influenced by dynamic social contexts, i.e., membership categories relevant to the ongoing activity. The significance of the present study can be summarized in the following three points.

First, in contrast to relatively small spaces for formations investigated in previous studies, the present study examined a vast field of about 40 meters square, and illustrated how participants in this vast field enter into various spatial-orientational arrangements. The participants often talk referring to a distant object that is tens of meters away, getting into suitable spatial-orientational arrangements such as the side-by-side and the 'horseshoe' arrangements. Importantly, the same group of people reconfigure the F-formation depending on the environments and the contexts they are in.

Second, we demonstrated the way in which body arrangements in F-formations are influenced not only by physical environments but also by social contexts. In particular, we showed that spatial-orientational arrangement of participants' bodies is determined not merely by *static* social factors, such as hierarchical relationship based on age or official position, but by *dynamic* social contexts. Employing the CA's notion of membership categories, which refer to social categories considered, by the participants, as relevant to the ongoing activity, we described how twofold, two-row, and headlined body arrangements emerge from the relevant categories in a particular context.

Third, we suggested possible bidirectional relationship between F-formation and social context. As described above, body arrangement in an F-formation can be determined by a social context. However, it is also possible that the spatial-orientational arrangement elicits the relevant membership category, which, in turn, imposes some constraints on who can do what in the ongoing activity. For instance, in our Case 4, where CSM was standing alone in front of the other three participants including PSM, PSM refrained from giving advice to CSM but rather gave direct instruction to CC and NC, as if he helped CSM act as a supervisor. The modest behavior of PSM, which is rarely observed elsewhere in the video data being analyzed, might be a result of this distinctive spatial-orientational arrangement, which could impose some constraints on how he behaves.

One of the remaining issues to be addressed would be micro-analysis of how body arrangements of participants are constituted, maintained, and transformed. De Stefani and Mondada (2014) provided a detailed analysis of how participants' bodies are reoriented in mobile situations. In particular, they demonstrated multimodal practices through which various kinds of participants (the "guide" and the "guided" of a tour) initiate a reorientation of the group. In our field as well, various kinds of participants can initiate a reconfiguration of the formation, and the way in which the reconfiguration is initiated may affect how the formation is sustained through the activity. Such dynamic aspects of spatial orientation of participants' bodies should be addressed in future research.

In summary, F-formation is tightly related to social context. Investigation into real-life interaction shed new light on our bodily behavior in everyday situations. We have just made a small step in this new direction. Further research should target broader situations and more participants with various social backgrounds.

5. Acknowledgments

The work was supported by JSPS KAKENHI Grant Number 15H02715, led by Mika Enomoto. We would like to thank the current and previous representatives of Nozawa-Onsen village and the members of *Hozonkai* and *San'yako* for their cooperation with our fieldwork.

6. Bibliographical References

Bono, M., Suzuki, N., and Katagiri, Y. (2004). An analysis of participation structures in multi-party conversations: Do interaction behaviors give clues to know your interest? (in Japanese). *Cognitive Studies*, 11(3):214–227.

De Stefani, E. and Mondada, L. (2014). Reorganizing mobile formations: When "guided" participants initi-

ate reorientations in guided tours. *Space and Culture*, 17(2):157–175.

Enomoto, M. and Den, Y. (2015). Speech act theory in the field: Interactiveness, concurrency, and situatedness in achieving the preparatory conditions of "commanding" (in Japanese). *Cognitive Studies*, 22(2):254–267.

Goffman, E. (1963). *Behavior in public places: Notes on the organization of gatherings*. Free Press, New York.

Goodwin, C. (2003). Pointing as situated practice. In S. Kita, editor, *Pointing: Where language, culture and cognition meet*, pages 217–242. Lawrence Erlbaum, Mahwah, NJ.

Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge University Press, Cambridge.

Kendon, A. (2010). Spacing and orientation in co-present interaction. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, and A. Nijholt, editors, *Development of multimodal interfaces: Active listening and synchrony*, volume 5967 of *Lecture Notes in Computer Science*, pages 1–15. Springer, Berlin.

Makino, R., Furuyama, N., and Bono, M. (2015). Spatial-orientational behavior for the narrative in a field: A case study on science communicators' standing position employed to display the readiness to start giving an explanation of the exhibit (in Japanese). *Cognitive Studies*, 22(1):53–68.

McNeill, D. (2006). Gesture, gaze, and ground. In S. Renals and S. Bengio, editors, *Machine learning for multimodal interaction*, volume 3869 of *Lecture Notes in Computer Science*, pages 1–14. Springer, Berlin.

Mondada, L. (2012). Garden lessons: Embodied action and joint attention in extended sequences. In H. Nasu and F. C. Waksler, editors, *Interaction and everyday life: Phenomenological and ethnomethodological essays in honor of George Psathas*, pages 293–314. Lexington Books, Plymouth, UK.

Sacks, H. (1972). On the analyzability of stories by children. In J. Gumperz and D. Hymes, editors, *Directions in sociolinguistics: The ethnography of communication*, pages 325–345. Holt, Rinehart, and Winston, New York.

Schegloff, E. A. (1998). Body torque. *Social Research*, 65(3):535–596.

Temporal coordination of facial expressions and head movements in first encounter dialogues

Patrizia Paggio[*†], **Costanza Navarretta**[*]
[*]University of Copenhagen, [†]University of Malta
{paggio, costanza}@hum.ku.dk
patrizia.paggio@um.edu.mt

Abstract

This paper deals with the temporal coordination between facial expressions and co-occurring head movements in a multimodal corpus of first encounter conversations. In particular, we look at how the onset of facial expressions is coordinated with the first overlapping head movement, in other words which of the two modalities precedes the other and why. We find and discuss statistical main effects on the temporal delays between the two behaviours due to individual variation, type of head movement, and the communicative function of the multimodal signal. In particular, the analysis shows that when speakers give feedback, their facial expression becomes visible before the head starts to move, especially in the case of negative comments associated with frowning or scowling. The opposite is true when the multimodal signal is used as a comment to the speaker's own speech. The motivation for the analysis is to shed light on a less studied aspect of multimodal communication – an aspect that is relevant to the generation of natural multimodal expressions in ECAs.

Keywords: facial expressions, head movements, multimodal coordination

1. Background and goals

The coordination of different signals in human communication has been studied especially as regards gesture and speech, and there is considerable agreement that hand gestures are coordinated with prosodic events, such as pitch accents and prosodic phrase boundaries (Bolinger, 1986; Kendon, 1980; Loehr, 2004; Loehr, 2007). Experimental work has also clearly shown that people are sensitive to disruptions of the natural temporal alignment between the two modalities (Leonard and Cummins, 2010; Giorgolo and Verstraten, 2008). Coordination between head movements and speech, and how this is mediated by prosody, is discussed in Hadar et al. (1983) and (1984). More recently, Paggio (2016) and Paggio and Navarretta (2016) investigated the temporal alignment between head movements and co-occurring speech segments in multimodal data, and discussed a number of factors that affect the alignment.

Studies dealing with the relation between facial expressions and other expressive modalities have looked at the co-occurrence of several expression types. An early study found that children use eyebrow raises preceding head movements in connection with visual search (Jones and Konner, 1970). In a qualitative study, Kelner (1995) pointed out that enjoyment smiles co-occur with head movements towards the interlocutor while embarrassed smiles co-occur with head and gaze movements away from them (Keltner, 1995). Using quantitative methods, Cohnal *et al.* (2004b) studied correlations between lip-corner displacement in smiles and head or eye movements, and found that smile intensity correlates negatively with the presence of head movement in contexts involving embarrassment. Cohnal *et al.* (2004a) found that eyebrow raising is more likely to occur with forward head movements. Work where multimodal coordination of different expressions is used to model the behaviour of Embodied Conversational Agents (ECAs) include Cassell *et al.* (1999), Lee and Marsella (2006) and for emotional behaviours is discussed in Martin (2011). Finally, a study of how smiles and laughters can be generated based on the interlocutor's smiling and laughing behaviour, is in El Hadded *et al.* (2016).

In this paper, we focus on the coordination between facial expressions and head movements in cases in which there is indeed an overlap between the two modalities. In particular, we look at how the onset of facial expressions is coordinated with the first overlapping head movement, in other words which of the two modalities precedes the other and why. The motivation for the analysis is to shed light on a less studied aspect of multimodal communication – an aspect that is relevant to the generation of natural multimodal expressions in ECAs.

2. Multimodal facial expressions

The data for this study consist of 1448 facial expressions and 3117 head movements extracted from the Danish multimodal NOMCO corpus, an annotated collection of twelve first encounter dialogues involving six male and six female subjects of age 21 to 36. Each participant took part in a dialogue with a female and one with a male, for a total of about an hour of interaction. The two conversations took place on different days, and in both cases the dialogue participants had never seen each other before. The only instruction they received was to try to get to know each other. As a consequence, they spoke freely about a range of different topics. The dialogues were recorded in a studio, with the participants standing in front of each other, and were filmed by three cameras (Paggio and Navarretta, 2016).

The average duration of the facial expressions is 1.98s (sd=1.6). The spread of the duration is remarkable, with the shortest expression lasting 0.16s[1], and the longest 12.12. Smiles are the expressions showing the most variation in duration, with scowls showing the least. Head movements

[1]The expression is a short smile followed by a laughter. The annotators agreed about the two behaviours being separate expressions.

Table 1: Proportional conditional frequency of head movement types given co-occurring facial expression types

	Backward	Forward	HeadOther	Jerk	Nod	Shake	Turn	Tilt	Waggle	Sum
FaceOther	0.05	0.11	0.08	0.07	0.30	0.09	0.14	0.15	0.01	1
FrownScowl	0.10	0.13	0.11	0.02	0.14	0.12	0.18	0.16	0.04	1
Laughter	0.14	0.07	0.11	0.03	0.14	0.10	0.21	0.15	0.05	1
Raise	0.08	0.16	0.06	0.05	0.17	0.08	0.19	0.17	0.04	1
Smile	0.08	0.14	0.06	0.07	0.24	0.10	0.10	0.16	0.05	1

Table 2: Coordination of facial expression onsets with first co-occurring head movement: raw counts (proportions in parentheses)

Facial expression type	Before head	Same time as head	After head	Total
Smile	239 (.45)	35 (.06)	261 (.49)	535 (1)
Raise	122 (.40)	39 (.13)	146 (.47)	307 (1)
Laughter	63 (.45)	7 (.05)	69 (.50)	139 (1)
Frown/Scowl	39 (.38)	6 (.06)	58 (.56)	103 (1)
FaceOther	28 (.38)	6 (.08)	40 (.54)	74 (1)
Total	491 (.42)	93 (.08)	574 (.50)	1158 (1)

are shorter. Their mean duration is 0.93s (sd=0.58), with up-nods providing the shortest and least varying movements, and head shakes the longest outlier (7.08s). Head movements can be single or repeated. In our dataset there are 2315 single head movements, and 794 repeated ones. The mean duration for single movements is 0.82s (sd=0.48s), while it is 1.28 for repeated ones (sd=0.70s).

The majority of the facial expressions, i.e. 1158, or 80% of the total, co-occur with at least one head movement. Table 1 shows the proportion in which different types of head movements co-occur with the different kinds of facial expressions.

Of these, 491 (42%) start before, 93 (8%) at the same time, and 574 (50%) after the first co-occurring head movement. Frequency counts of the various facial expression types against their onset relation with the first co-occurring head movement are shown in table 2. In general, it can be concluded that there is a very high likelihood for facial expressions to be accompanied by head movements. However, whether the onset of the facial expression precedes or follows the onset of the head movement is equally likely. Nevertheless, a χ-squared test of independence showed that the type of onset delay depends on the facial expression type (χ^2=15.87, df=8, p-value=0.04429s). This dependency is mostly due to the fact that eyebrow raises (*Raise*) tend to start at the same time as the co-occurring head movement proportionally more often than the other types, while frowns and scowls (*Frown / Scowl*) tend to start after the onset of the head movement more often than the other types. There is also a slight tendency for *Smile* and *Laughter* to start before the head movement more often than expected.

These differences may well be due to different physical characteristics of the signals. For instance, eyebrow movements are quite small and their onset may therefore be more tightly coordinated with that of short accompanying head movements such as nods and turns. Conversely, smiles and laughters may imply a longer preparation phase and therefore tend to start earlier than the accompanying head movement.

3. Temporal coordination

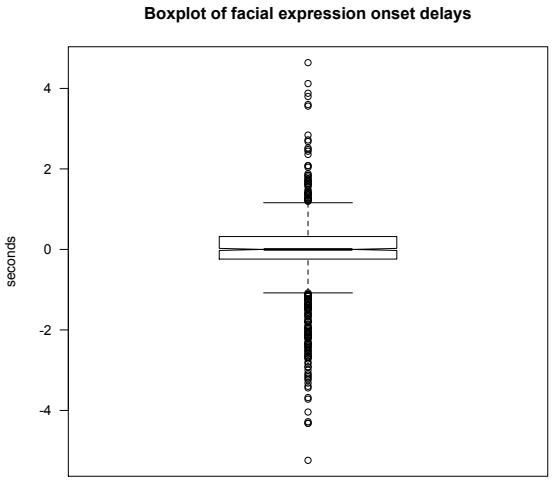


Figure 1: Boxplot of the distribution of onset delays between facial expressions and the first overlapping head movement. Positive delays indicate facial expressions starting after the co-occurring head movement.

In this section we look at the temporal coordination between the two co-occurring behaviours in a more fine-grained way. The mean onset delay between the two modalities is -0.05s (sd=0.9), indicating that the behaviours on average are almost coincidental (with a tiny likelihood for the face starting to move before the head), but that there is also considerable variation. The plot in figure 1 shows the distribution of the duration of the onset delays between facial expressions and the first overlapping head movement. Most of the delays are in the area between –1s (facial expression starting before the onset of the head movement), and +1s (facial expression starting after the onset of the head movement). There are, however, quite a number of outliers in both negative and positive ranges so that the data do not conform to the normal distribution (Shapiro-Wilk normality test, W=0.85783, $p < 0.001$).

Statistical tests show a main effect of individual speaker variation (Kruskal-Wallis: χ^2=44.002, df=11, p-value$<$0.001), an effect of head movement type (Kruskal-

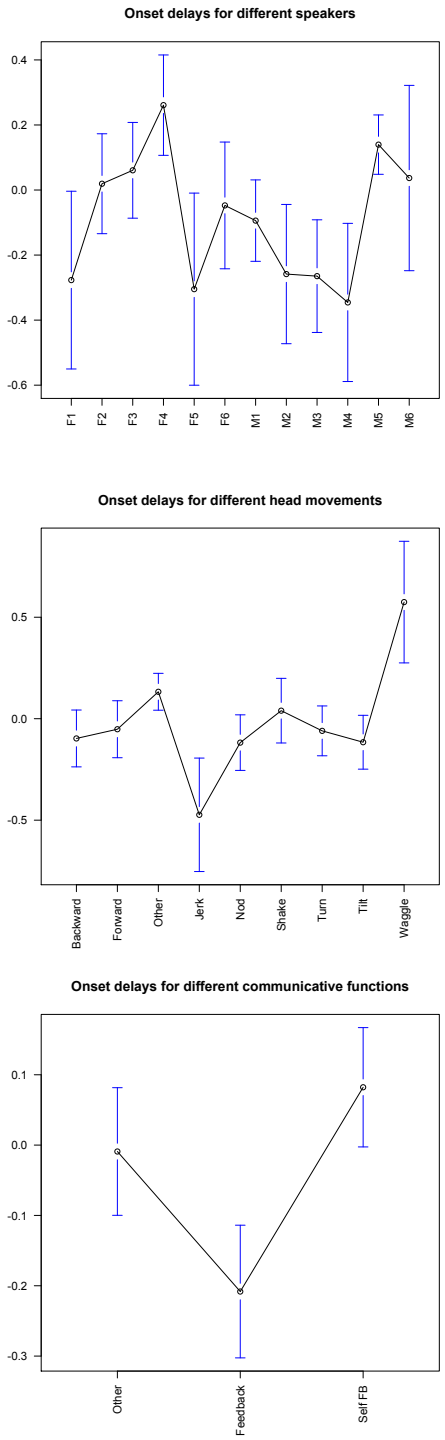


Figure 2: Mean values and confidence intervals for the temporal coordination between onsets of facial expressions and co-occurring head movements according to individual speakers (plot on top), associated head movement (plot in the middle), and function of the signal (plot below). Positive values indicate that facial expressions start after the onset of the head movement.

Wallis: χ^2=39.689, df=8, p-value<0.001), and an effect of function (Kruskal-Wallis: χ^2=22.802, df=3, p-value<0.001) on the distribution of the start delays. The effect of facial expression type, on the contrary, does not reach significance in spite of the results of the χ^2 test on the figures in table 2.

As can be seen from the topmost plot showing mean values and confidence intervals for different speakers in figure 2, only four of the speakers (F2, F3, F6 and M1) display an average delay around 0s, whilst the rest of the speakers have either a positive or a negative mean delay onset. Most of the significant differences involve F4 and M5[2]. As for the head movement type (middle plot in figure 2), negative delays are seen especially together with *Jerk* (up-nod) and positive ones with *Waggle*. Up-nods imply a backward movement of the neck which may physically be slightly more demanding than a forward movement, and have the effect of the movement becoming visible after the onset of the co-occurring facial expression. Conversely, waggles tend to precede the associated facial expressions. Waggles are rather complex and relatively long on average (mean duration=1.2s), characteristics which may explain why they are initiated early in the multimodal contribution. Most of the significantly different pairwise comparisons predictably involve *Waggle*, but the comparisons between delays involving *Jerk* and *Other* as well as *Jerk* and *Shake* also show a significant difference. This is not surprising since shakes are similar to waggles in being complex movements in which the head moves repeatedly in different directions.

A particularly interesting effect on the temporal coordination between facial expressions and head movements is the one relating to the communicative function assigned to the multimodal signal. Such dependence could in fact be exploited in the generation of facial expressions and head movements in ECAs. In this paper we distinguish between three function types: *CPU*, which stands for Contact, Perception and Understanding, for signals eliciting or giving feedback; *Self Feedback* for signals used by the speaker to comment their own contributions; and any other function[3]. The lowest plot in figure 2 shows that feedback to others and self feedback behave quite differently, with self feedback signals displaying a delay of about 1s on average, and feedback signals showing delays in the other end of the scale (about -0.2s on average) . In other words, when speakers react to their own speech, they tend to move the head first. When they give feedback, they tend to move the face first. This difference is statistically significant.

Finally, in the plot in figure 3 we show the combined effect

[2]All pairwise comparisons after the Kruskal-Wallis tests were done using the Dunn test with the Benjamini-Hochberg p-value adjustment method.

[3]Other functions relate to turn taking, discourse structuring, information stucture, etc. as defined in the MUMIN coding scheme (Allwood et al., 2007). Note that some of the functional categories in our annotations have a direct correspondence with discourse act categories in the ISO 24617-2 standard (`https://www.iso.org/standard/51967.html`). This is for example the case for the Auto- and Allo-Feedback dialogue acts, which have the same semantics as the MUMIN's SelfFeedback and FeedbackGiving attributes.

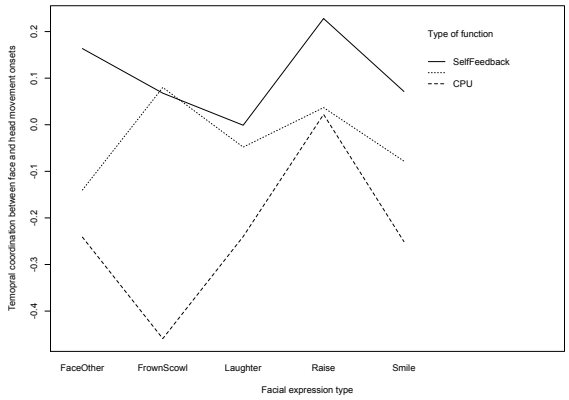


Figure 3: Interaction plot showing the combined effect of communicative function and facial expression type on the temporal coordination between facial expressions and co-occurring head movements.

of function and facial expression type. We see that the tendency for feedback behaviours (CPU in the figure) appearing in the negative end is stronger in the case of frowns and scowls, whereas in the case of eyebrow raises the different functions do not affect the direction of the delay much.

4. Discussion and conclusion

In general, our data clearly show that facial expressions have a strong tendency to co-occur with head movements and to be aligned with them at the onset. There are, however, delays in both directions. We have found interesting patterns concerning how the delays are distributed depending on the facial expression type. Thus, the onset of eyebrow raises is more tightly coordinated with the onset of the first co-occurring head movement, whereas both smiles and laughters tend to be initiated slightly earlier. These differences, however, do not reach significance in our data.

Significant effects on the temporal coordination between co-occurring behaviours in the two modalities, on the contrary, were found for head movement type and function of the signal in addition to individual variation. The effect due to head movement type can be explained at least partially in terms of the physical characteristics of the movements, with complex movements such as waggles showing a tendency to be initiated before the co-occurring facial expression. More interestingly, whether the onset of a facial expression (slightly) precedes or follows the onset of the first co-occurring head movement also depends on the function of the multimodal behaviour. In particular, we have found that when speakers give feedback, their facial expression becomes visible before the head starts to move, especially in the case of negative comments associated with frowning or scowling. Conversely, when the multimodal signal is used as a comment to the speakers' own speech contribution, the head movement tends to be noticed first. Since facial expressions are one of the strongest signals of attitudinal and emotional states, these results seem to indicate that in the case of a comment to the interlocutor's contributions, facial reactions are more immediate than feedback expressed by movements of the head. Head movements and facial expressions in our data have the same communicative function, that is were reinforcing each other, or have a function of repetition using the terminology by Poggi and Caldognetto (1996). The temporal relation between the two behaviours in other cases, for instance contradiction, should be investigated in different data.

Interactions between the various variables involved in our analysis are difficult to test statistically because of the non-normal distribution of the data, and were only illustrated graphically in this paper. In future, we intend to explore such interactions by applying machine learning techniques to the problem of predicting the alignment between facial expressions and head movements from the formal and functional factors discussed in this study. Linear mixed effects models could also be applied to investigate the interactions between the various factors.

To test the generality of our findings, it would be interesting to conduct similar analyses using data from different communicative situations as well as produced by speakers from different cultural backgrounds. We would also be interested in verifying if other patterns of behaviour than those found in the corpus would seem unnatural when implemented in an ECA.

A relevant and interesting issue we have not investigated, is how facial expressions are structured internally, and whether they contain a phase comparable to the stroke in hand gestures, see e.g. Kipp (2004). If or when they do, it is reasonable to assume that the onsets of facial expressions and head movements will be coordinated in such a way as to ensure that the strokes of the two behaviours are aligned. A related issue also not dealt with here is what happens when a protracted facial expression – for example a smile – overlaps with several distinct head movements. The way in which the temporal coordination between the two modalities should be described in such cases is far from clear, and will be left for future research.

5. Acknowledgements

We would like to thank the three reviewers for their useful comments.

6. 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. In Jean-Claude Martin, et al., editors, *Multimodal Corpora for Modelling Human Multimodal Behaviour*, volume 41 of *Special issue of the International Journal of Language Resources and Evaluation*, pages 273–287. Springer.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Stanford, CA: Stanford.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 520–527.

Cohn, J. F., Reed, L. I., Ambadar, Z., Xiao, J., and Moriyama, T. (2004a). Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, volume 1, pages 610–616, Oct.

Cohn, J. F., Reed, L. I., Moriyama, T., Xiao, J., Schmidt, K., and Ambadar, Z. (2004b). Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 129–135, May.

El Haddad, K., Çakmak, H., Gilmartin, E., Dupont, S., and Dutoit, T. (2016). Towards a listening agent: A system generating audiovisual laughs and smiles to show interest. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 248–255, New York, NY, USA. ACM.

Giorgolo, G. and Verstraten, F. A. (2008). Perception of 'speech-and-gesture' integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, pages 31–36.

Hadar, U., Steiner, T., Grant, E. C., and Rose, F. C. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2):117–129.

Hadar, U., Steiner, T., and Rose, F. C. (1984). The timing of shifts of head postures during conversation. *Human Movement Science*, 3(3):237–245.

Jones, N. G. B. and Konner, M. (1970). An experiment on eyebrow-raising and visual searching in children. *Journal of Child Psychology and Psychiatry*, 11(4):233–240.

Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality & Social Psychology*, 68:441–454.

Kendon, A. (1980). Gesture and speech: two aspects of the process of utterance. In M: R. Key, editor, *Nonverbal Communication and Language*, pages 207–227. Mouton.

Kipp, M. (2004). *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com.

Lee, J. and Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 243–255. Springer.

Leonard, T. and Cummins, F. (2010). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10):1457–1471.

Loehr, D. P. (2004). *Gesture and Intonation*. Ph.D. thesis, Georgetown University.

Loehr, D. P. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7(2).

Martin, J.-C., Devillers, L., Raouzaiou, A., Caridakis, G., Ruttkay, Z., Pelachaud, C., Mancini, M., Niewiadomski, R., Pirker, H., Krenn, B., Poggi, I., Caldognetto, E. M., Cavicchio, F., Merola, G., Rojas, A. G., Vexo, F., Thalmann, D., Egges, A., and Magnenat-Thalmann, N., (2011). *Coordinating the Generation of Signs in Multiple Modalities in an Affective Agent*, pages 349–367. Springer Berlin Heidelberg, Berlin, Heidelberg.

Paggio, P. and Navarretta, C. (2016). The Danish NOMCO corpus multimodal interaction in first acquaintance conversations. *Journal of Language Resources and Evaluation*, pages 1–32.

Paggio, P. (2016). Coordination of head movements and speech in first encounter dialogues. In *Proceedings of the 3rd European Symposium on Multimodal Communication*, Linköping Electronic Conference Proceedings, pages 69–74.

Poggi, I. and Caldognetto, E. M. (1996). A score for the analysis of gestures in multimodal communication. In *Proceedings of the Workshop on the Integration of Gesture and Language in Speech Applied Science and Engineering Laboratories*, pages 235–244.

Expert-Novice Interaction: Annotation and Analysis

Soumia Dermouche, Catherine Pelachaud
CNRS UMR 7222, ISIR, Sorbonne Universités
75005 Paris, France
{soumia.dermouche, catherine.pelachaud}@upmc.fr

Abstract

In this demonstration, we present the *NoXi* corpus of expert-novice interactions, our annotations and analysis. To analyze the data we apply HCApriori, a Temporal Sequence Mining algorithm to extract relevant behavior sequences for both expert and novice. NoXi provides over 25 hours of dyadic interactions recorded in different languages, mainly English, French, and German. The annotation tool, NOVA, developed by (Baur et al., 2015) allows annotating data using discrete and continuous schema. We use NOVA to manually annotate non-verbal behaviors (discrete annotation) and engagement levels (continuous annotation).

Keywords: Non-verbal behavior; Engagement; Sequence Mining; Virtual Agent

1. Introduction

This work is part of the H2020 project ARIA-VALUSPA (*Artificial Retrieval of Information Assistants - Virtual Agents with Linguistic Understanding, Social skills and Personalized Aspects*). In this project, a corpus of dyadic interactions, named NoXi, has been collected (Cafaro et al., 2017). NoXi is available to the research community from the website: `https://noxi.aria-agent.eu/`. During the interaction, participants exchanged through a large screen in different rooms. One participant assumes the role of an expert on a given topic and the other the role of a novice for this topic. NoXi is composed of 84 sessions recorded in three different countries France, Germany and UK and discussing 58 topics like video games, sports, cooking, etc. In the following sections, we describe our coding scheme for NoXi annotation and how we use the NOVA tool. We have manually annotated several non-verbal behaviors and engagement levels of both expert and novice. The use of sequence mining allowed us discovering relevant patterns for different engagement levels.

2. Annotation

NOVA[1] is an open-source annotation tool developed by (Baur et al., 2015) that we use to annotate the NoXi corpus. NOVA overcomes the limitations of existing annotation tools by exploring richer data like face streams or skeleton and by proposing two annotation schemas at time: discrete and continuous. Moreover, NOVA is a collaborative platform in which annotators from different sites can combine and share their annotations. Discrete annotation schema can be used to label behaviors that can be classified into a set of categories (e.g. gaze direction). Discrete annotation characterizes the starting and the ending time of behaviors. On the other hand, a continuous scale could be more appropriate for describing continuous dimensions, such as, engagement which is expressed all along the interaction. Figure 1 shows one session of NoXi viewed with NOVA. Audio-visual as well as skeleton and face streams of both expert and novice are opened. Using NOVA, continuous and discrete annotations can be visualized at same time.

In this work, we use NOVA to annotate the French part of NoXi database which is composed of 21 sessions. The total duration of all these sessions is 7 hours and 25 minutes. We use a discrete annotation schema to label six non-verbal behavior types: head direction and movement, smile, eyebrow movement, gesture and hand rest positions. Continuous scale is adapted for engagement annotation. In order to avoid content biases from the the verbal stream and prosody when annotating engagement, we have filtered it out, for both expert and novice. According to (Yannakakis et al., 2017) that suggest ordinal annotation for affect modeling, we annotate engagement over five levels: strongly disengaged, partially disengaged, neutral, partially engaged and strongly engaged. One evaluator was asked to rate and associate the engagement level of expert and novice over these levels. Table 1 illustrates the manual annotations that we realize so far. These annotations have been realized by three evaluators: one for engagement annotation, one for gesture annotation and the last one dealt with the remaining annotations. For each modality, we indicate label of annotated signals,the number of annotated sessions, their duration, and the number of annotations for expert and novice.

3. HCApriori Algorithm

Human behaviors are naturally multimodal. Human states, attitude, engagement level, etc, can be displayed through sequences of behaviors (Burgoon and Dunbar, 2006). In order to extract a meaningful multimodal sequences from NoXi, we rely on HCApriori, a temporal sequence mining algorithm (Dermouche and Pelachaud, 2016). This algorithm aims at finding frequent patterns (frequent subsequences) hidden in set of sequences. HCApriori takes as input: the sequence dataset, a minimum threshold (f_{min}), i.e. only patterns that hold within this threshold are considered as frequent, dissimilarity measure like CityBlock and dissimilarity threshold called ϵ.

HCApriori operates in two steps: (1) hierarchical clustering in which signals are grouped into the same cluster if and only if their temporal distance is less than ϵ. Temporal distance between two signals is evaluated using a dissimilarity measure such as CityBlock. At the end of this step, the

[1] `https://github.com/hcmlab/nova`

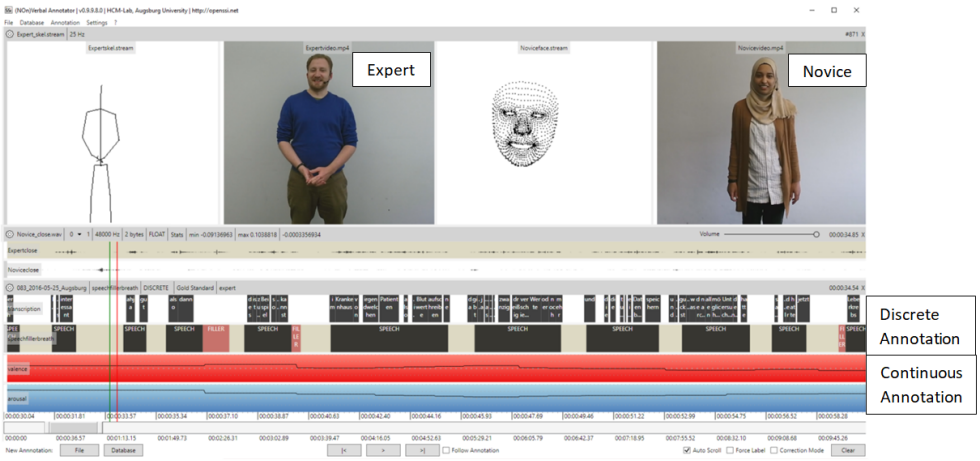


Figure 1: A screenshot of NOVA interface: videos of expert and novice, expert skeleton and novice's face tracking (top). Discrete and continuous annotations tracks are shown (bottom).

Table 1: Number of manual annotations of each non-verbal modality for expert and novice.

Modality	Label of annotated signals	Annotation number	
		Expert	Novice
Head direction and movements	Nod, Shake, Forward, Back, Up, Down, Side, Tilt	72	337
Smiles	Smile	153	157
Eyebrow movements	Frown and Raised	147	44
Gestures	Iconics, Metaphorics, Deictics, Beats, and Adaptors	1223	293
Hand rest positions	Arms crossed, Hands together, Hands in pockets, Hands behind back, and Akimbo	1317	612
Engagement	Strongly disengaged, Partially disengaged, Neutral, Partially engaged, Strongly engaged	1481	1679

cluster centroid represents a pattern of length one. (2) Taking as input the results of the previous stage, Apriori-like procedure is adapted to generate longer temporal patterns.

For NoXi analysis using HCApriori, we can, for example, explore the relationships between non-verbal behavior and engagement perception. For this purpose, we prepared the input dataset of HCApriori by collecting all sequences of non verbal behaviors that appear during a given engagement level. Table 2 presents the number of sequences we obtained for each engagement level for expert and for novice. Then, we have applied HCApriori to extract temporal patterns of nonverbal signals expressing the five engagement levels.

Our demo will consist of a presentation of the data collection, experimental setup of NoXi, as well as the annotation tool used for the manual annotation of various behaviors. It will also provide, based on HCApriori, the data analysis and the investigation of the sequential behavios of both expert and novice.

Table 2: Number of sequences of each engagement level for both expert and novice.

	Level 1	Level 2	Level 3	Level 4	Level 5	Total
Expert	48	373	373	561	126	1481
Novice	116	432	509	558	64	1679

4. Acknowledgements

Funded by European Union Horizon 2020 research and innovation programme, grant agreement No 645378.

5. Bibliographical References

Baur, T., Mehlmann, G., Damian, I., Lingenfelser, F., Wagner, J., Lugrin, B., André, E., and Gebhard, P. (2015). Context-Aware Automated Analysis and Annotation of Social Human–Agent Interactions. *ACM Transactions on Interactive Intelligent Systems*, 5(2):1–33.

Burgoon, J. K. and Dunbar, N. E. (2006). Nonverbal expressions of dominance and power in human relationships. *The SAGE Handbook of Nonverbal Communication*, (September 2014):279–298.

Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres, M. T., Pelachaud, C., Andr, E., and Valstar, M. (2017). The NoXi Database : Multimodal Recordings of Mediated Novice-Expert Interactions. In *ICMI'17,*, pages 350–359, Glasgow, Scotland. ACM.

Dermouche, S. and Pelachaud, C. (2016). Sequence-based multimodal behavior modeling for social agents. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, pages 29–36, Tokyo, Japan. ACM.

Yannakakis, G. N., Cowie, R., and Busso, C. (2017). The Ordinal Nature of Emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 248–255.

Studying Japanese Distal Demonstrative 'are' Using Video Corpus

Saori Daiju, Tsuyoshi Ono
University of Alberta
3-31 Pembina Hall
University of Alberta
Edmonton, AB T6G 2H8
Canada
daiju@ualberta.ca, tono@ualberta.ca

Abstract

We examine a test version of the Corpus of Everyday Japanese Conversation (CEJC), currently being built at the National Institute for Japanese Language and Linguistics in Tokyo (NINJAL), Japan. By focusing on non-verbal behaviors, we highlight the multimodal nature of the use of the Japanese distal demonstrative *are*. In particular, video data allows us to observe a previously unrecognized type of *are* where the speaker uses *are* while gazing towards and pointing fingers at the cell phone placed near her in order to refer to the photographs digitally stored on it. This use is not anaphoric, as its referent is newly introduced in the conversation through the combined use of *are* and the speaker's non-verbal behaviors. It is not spatial either, as that would have resulted in the employment of the proximal demonstrative *kore* and because the potential referent is not visible on the screen at the time of speech. Instead, it is used to indirectly refer to the digital data stored in the cell phone. Due to a shared understanding that cell phones digitally store photographs, the speaker's use of *are* along with simply gesturing at the cell phone makes such a reference possible.

Keywords: Japanese distal demonstrative *are*, non-verbal behavior, multimodality, everyday conversation, video, Japanese

1. Introduction

Japanese is known for its three-way demonstrative system where the *ko*-series is said to refer to an object close to the speaker, the *so*-series to the addresee, and the *a*-series far in space from both the speaker and the addresee (Kuno 1973; Martin 1975; Iwasaki 2013; Hasegawa 2015). Among the *a-series*, *are* 'that' has long been discussed with regard to its spatial and anaphoric use both in linguistics and language teaching (Kuno 1973; Martin 1975; articles in Kinsui and Takubo 1992; Iwasaki 2013; Banno et al. 2011; Hatasa et al. 2011; Tohsaku 2006). In terms of spatial use, for example, *are* is introduced in Banno et al. (2011) with a picture of a woman talking to a man. She points to a pen, which is held by another person far from both her (the speaker) and the addressee, and says:

(1) **are** wa watashi no pen desu
that TOP I GEN pen COP
'**That** is my pen.'

(Banno et al. 2011: 62)

This type of *are* is used when the referent is physically available and possibly visible to both the speaker and the addressee.

From the perspective of anaphoric use, a referent is first introduced in the discourse and the demonstrative *are* subsequently refers back to it. In the following example constructed by Kuno (1973), A is talking about a fire which s/he saw the other day:

(2)
1 A: watashi mo choodo Harvard Square no
I also exactly Harvard Square GEN
2 soba ni ite
near in COP
'I also happened to be in the Harvard Square area and'
3 **sono kaji** o mimashita.
that fire ACC saw
'saw **that fire**.'
4 **are** wa hidoi kaji deshita ne.
that TOP terrible fire COP.PAST PTCL
'**That** was a terrible fire, wasn't it?' (Kuno 1973)

A says to the addressee *watashi mo choodo Harvard Square no soba ni ite* ***sono kaji*** *o mimashita* 'I also happened to be in the Harvard Square area and saw that fire' in lines 1-3. Then s/he comments on it in line 4, saying ***are*** *wa hidoi kaji deshita ne* 'That was a terrible fire, wasn't it?' This *are* 'that' refers to *sono kaji* 'that fire' in line 3. Although the uses of *are* have been discussed quite extensively, most of the research is based on constructed sentences like (1) and (2) above.

More recently, however, the availability of and interest in language use data have allowed researchers to uncover previously unidentified functions of this demonstrative (Hayashi 2004; Daiju 2017, etc.). For example, Hayashi (2004), based on the examination of audio recorded conversation, highlights its cataphoric use where he suggests that *are* can serve as a 'dummy' to project a subsequent specification. In the example below, A is talking about gas pipes:[1]

(3)
1 A: sono= saikin **are** na n desu yo.
uh recently that COP NOL COP PTCL
'Uh, recently (it)'s been **that**.'
2 ano=, gasu kan aru ja nai desu ka=.
uhm gas pipe exsist COP not COP PTCL
'Uhm, you know there are gas pipes, right?'
3 **are zenbu ima purasuchikku ni naritsutsu aru**
that all now plastic to is becoming exist
4 n desu yo=.
NOL COP PTCL
'**They've all been changing to plastic pipes now**.'

(Hayashi 2004)

In line 1, A begins by saying *sono= saikin* ***are*** *na n desu yo* 'uh, recently (it)'s been **that**'. Then A introduces gas pipes in line 2 by saying *ano=, gasu kan aru ja nai desu*

[1] In the examples used in this paper, an equal sign (=) indicates lengthning, an at sign (@) laughter, and square brackets ([ ]) overlapped speech.

ka= 'uhm, you know there are gas pipes, right?' Then in lines 3-4, he continues *are zenbu ima purasuchikku ni naritsutsu aru n desu yo*= 'they've all been changing to plastic now'. According to Hayashi (2004), the phrase in line 1 ***are*** *na n desu yo* '(it)'s been **that**' projects the subsequent specification of *are*. That is, the addressee is "instructed" that its specification is coming. In line 3, the speaker says *are zenbu ima purasuchikku ni naritsutsu aru* 'they've all been changing to plastic pipes now' to specify the *are* from line 1. Please note that *are* in line 3 is anaphoric; it refers back to *gasu kan* 'gas pipes' in line 2.

We have broadened the study of the demonstrative *are* by examining video-recorded everyday speech data, which has become more available in recent years with the advancement of digital technology. Specifically, we used the test version of the Corpus of Everyday Japanese Conversation (CEJC), currently being built at the National Institute for Japanese Language and Linguistics (NINJAL) in Tokyo, Japan.

It should be noted that studies of languages other than Japanese have examined demonstratives in actual use. The pioneering work by researchers such as Auer (1984), Hanks (1992, 2005), Himmelmann (1996), Enfield (2002, 2003), and Sidnell and Enfield (2017), with a focus on non-verbal aspects in more recent studies, are particularly noteworthy. We hope to contribute to this ongoing discussion by adding Japanese audio and video data from CEJC, which just became available.

2. Analysis

Our examination of the use of *are* in CEJC has resulted in a number of striking examples which give insight into the situated nature of its actual use, mainly because the video portion of the corpus provides access to non-verbal aspects of everyday speech. We will give a preliminary observation of some of these examples in this section.

One example involves a husband who, while making a drink, says ***are*** *nai no* 'Don't (we) have **that**?'. The wife immediately responds with *n* 'huh?', which he quickly follows with the right index finger in stirring motion.

Figure 1: The husband is making a stirring motion with his right index finger.

This was apparently successfully communicated as the wife then says *a kakimawasu no aru* 'Oh (we) have (a) stirring one'. That is, the referent negotiation of *are* in this example can be only understood by taking a multimodal perspective which CEJC allows.

Another example, also taken from a drinking situation, again highlights the multimodal construction of the referent of *are*. In bringing out a bamboo-made cup to serve sake (Japanese alcohol beverage) to the guest, the speaker says *demo ne chotto* ***are*** *na n da yo ushiro ga* 'But (it) is a little bit **that**, the bottom (is a little bit **that**)' while showing the bottom of the cup to her.

Figure 2: The host is showing the bottom of the bamboo-made cup to the guest.

The guest has no trouble understanding what ***are*** '**that**' refers to and immediately says *aa ii yo betsuni zenzen* 'Oh, fine. No problem at all'. The exact referent of *are* was not verbalized throughout the conversation, but *are* along with the showing of the bottom of the can seems to have create a mutual understanding between the speakers, perhaps aided by the common knowledge that products made out of natural resources like bamboo are sometimes deformed or might even be damaged.

The rest of the paper focuses on one particular example which highlights a more intricate connection between *are* and non-verbal behaviors in the specification of the intended referent where the role played by knowledge shared by the speakers appears to be even more critical. In the interaction the segment below is taken from, M is talking about the new cabinet where she placed her printer:[2]

(4)
1 M: dakara= maa purintaa wa oite atte=
so um priter TOP put exist:and
'So um (the) printer is put (on the cabinet), and'
2 purintaa wa tsukatteru wa[ke].
printer TOP use:exist PTCL
'(I) am using (the) printer.'
3 A: [a=].
oh
'Oh.'
4 M: ano=
5 A: yoku [ne]?
often PTCL
'often, right?'

[2] We corrected transcription errors which we identified in the test version of CEJC. We also made minor changes in the transcript to increase the readability of the example. Our examples have been romanized based on the Japanese original along with slightly different transcribing conventions described in the last note.

6 M: [**are**] ni ne. <gazing towards and pointing
that in PTCL her fingers at her cell phone>
'for **that**, right?'
7 soo.
yes
'Yes.'
8 shashin toka <@ insatsu @> suru kara sa.
photograph etc. print do beause PTCL
'because (I) print photos etc.'

(CEJC: K001-004; 13 min)

In lines 1-2, M says 'So um (the) printer is put (on the cabinet), and (I) am using (the) printer.' After A's contribution 'Oh, (you are using the printer) often, right?' in lines 3 and 5, M produces ***are*** *ni ne* 'for **that**, right?' in line 6. This is a type of 'increment' (Couper-Kuhlen and Ono 2007) in that it can be understood to combine with the utterance in line 2 and results in a syntactically well-formed string [***are*** *ni ne*] *purintaa wa tsukatteru wake* '(I) am using (the) printer [for **that**, right?]'. Notice that due to the word order of Japanese, this would take the form of insertion, placing ***are*** *ni ne* 'for **that**, right?' at the beginning. This increment shifts the original understanding '(I) am using (the) printer.' in line 2 to a new understanding '(I) am using (the) printer for that, right?'

The demonstrative *are* apparently refers to the printing of photographs as can be seen M's utterance 'because (I) print photos etc.' in line 8. Without video, one might suggest that *are* in line 6 is another example of cataphoric *are* which projects the specification of its referent in the upcoming interaction, in fact accomplished with *shashin* 'photos' in line 8 (Hayashi 2004). An examination of the video recording of the segment, however, reveals a more intricate process in identifying the referent of *are*. Intriguingly, as M produces *are ni ne* 'for that, right?' in line 6, she gazes towards and points her fingers at her cell phone as shown in figure 1. This is not a spatial use of the demonstrative; due to the proximity between M and her cell phone, the spatial use would have resulted in the employment of the proximal demonstrative *kore*. Equally importantly, we see a blank screen on her cell phone; there is no photograph which M is pointing towards.

Figure 3 : M is gazing towards and pointing her fingers at the cell phone.

What seems to be happening, instead, is that M is relying on a shared understanding among current Japanese speakers that cell phones digitally store photographs. This understanding allows M to make reference to photographs just by gazing towards and pointing her fingers at the cell phone. If the listener were to only consider the speaker's actions from the spatial perspective, they may incorrectly interpret these non-verbal behaviors as referring to the cell phone instead of the photographs. However, *are* in line 6 is not used spatially. Instead, our shared knowledge of how cell phones work makes it possible for the listener to understand that *are* here refers to the digital data present within the machine, which is cataphorically made more explicit in line 8.

3. Conclusion

Overall, the current study underscores the importance of the study of linguistic form in actual use. In particular, video recordings allow researchers to examine the non-verbal aspects of how people interact as they produce language. The increasing availability of video data accomplished by video corpora such as CEJC by NINJAL gives a critical edge to our efforts to understand how language is actually used and what language itself is.

4. Acknowledgements

We would like to thank the National Institute for Japanese Language and Linguistics in Tokyo, especially Hanae Koiso, for making their test version of the Corpus of Everyday Japanese Conversation available to us. We would also like to thank Kanza Tariq and Maggie Camp for their assistance with the preparation of this paper.

The work reported in this presentation was supported by the NINJAL collaborative research project 'A Multifaceted Study of Spoken Language Using a Large-Scale Corpus of Everyday Japanese Conversation'.

5. References

Auer, Peter. (1984). Referential Problems in Conversation. *Journal of Pragmatics* 8: 627-648.

Banno, E., Ikeda, Y., Ohno, Y., Shinagawa, C., and Tokashiki, K. (2011). *Genki I: An integrated course in elementary Japanese*, 2nd edition. Tokyo: The Japan Times.

Couper-Kuhlen, E. and Ono, T. (2007). 'Incrementing' in conversation. A comparison of practices in english, German and Japanese. *Pragmatics* 17 (4): 513-552.

Daiju, S. (2017). Not saying exactly what it is is sometimes good enough: the unspecified use of demonstrative *are* in Japanese everyday talk. Proceedings of the Canadian Association of Japanese Language Education 2017. 44-52. http://www.cajle.info/wp-content/uploads/2017/09/06CAJLE2017Proceedings_DaijuSaori.pdf

Enfield, N. J. (2002). 'Lip-pointing': A discussion of form and function and reference to data from Laos. *Gesture* 2, 185–211.

Enfield, N. J. (2003). The definition of *WHAT-d'you-call-it*: semantics and pragmatics of recognitional deixis. Journal of Pragmatics 35: 101–117.

Hanks, W. F. (1992). The indexical ground of deictic reference. In A. Duranti and C. Goodwin (eds.), *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press. 43-76.

Hanks, W. F. (2005). Explorations in the Deictic Field. *Current Anthropology*, 46 (2): 191-220.

Hasegawa, Y. (2015). *Japanese: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Hatasa, Y. A., Hatasa, K., and Makino, S. (2011). *Nakama 1: Japanese communication, culture, context*, 2nd edition. Boston, MA: Heinle Cengage Learning.

Hayashi, M. (2004). Projection and grammar: Notes on the 'action-projecting' use of the distal demonstrative are in Japanese. *Journal of Pragmatics* 36 (8), 1337-1374.

Himmelmann. N. P. (1996). Demonstratives in Narrative Discourse: A Taxonomy of Universal Uses. In B. Fox (ed.), *Studies in Anaphora*. John Benjamins, Amsterdam. 205–254.

Iwasaki, S. (2013). *Japanese*, *Revised edition*. Amsterdam, Philadelphia: John Benjamins.

Kinsui, S. and Takubo Y. (eds.). (1992). *Shijishi* [*Demonstratives*]. Tokyo: Hituzi Syobo.

Kuno, S. (1973). *The Structure of the Japanese Language*. Cambridge, MA: The MIT Press.

Martin, S. E. (1975). *A Reference Grammar of Japanese*. New Haven: Yale University Press.

Couper-Kuhlen, E. and Ono, T. (2007). 'Incrementing' in conversation. A comparison of practices in english, German and Japanese. *Pragmatics* 17 (4): 513-552.

Sidnell, J. and Enfield, N. J. (2017). Deixis and the interactional foundations of reference. in Y. Huang (ed). The Oxford Handbook of Pragmatics. 217-239.

Tohsaku, Y. (2006). *Yookoso!: An invitation to Contemporary Japanese*, 3rd edition. New York: McGraw-Hill Companies.

Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon's ALEXA

Ingo Siegert[1], Julia Krüger[2], Olga Egorow[1], Jannik Nietzold[1], Ralph Heinemann[1], Alicia Lotz[1]
[1]Institute for Information and Communications Engineering, Cognitive Systems Group,
Otto-von-Guericke University, 39106 Magdeburg, Germany,
[2]Department of Psychosomatic Medicine and Psychotherapy, Otto von Guericke University, 39120 Magdeburg, Germany

Abstract

A new conversation corpus in the area of human-computer interaction is introduced. It consists of conversations between one and two interaction partners with a commercial voice assistant system (Amazon's ALEXA) in two different settings. The fundamental aim for building up this corpus is to investigate how humans address technical systems. Thereby, two different scenarios, a formal and an informal one, are designed. The scenarios are conducted by the participants alone and with an accompanying person. Furthermore, questionnaires are used to get a self-evaluation of the participants in terms of their experience of the interaction and their conscious changes in voice and behaviour while addressing a technical system. Additionally, also their experience with technical systems and the evaluation of the utilized commercial voice assistant is retrieved via questionnaires. The corpus consists of high-quality microphone recordings of 27 German speaking subjects, all students at the University Magdeburg.

Keywords: Corpus, Addressee Detection, Speech Assistant, Multi-Scenario, Multi-User, Speaking-Style

1. Introduction

Human-computer interaction (HCI) still receives increased attention, the market for commercial voice assistants is rapidly growing. Besides making the operation of technical systems as simple as possible, voice assistants should enable a natural interaction. Therefore, one aspect that still needs improvement is to automatically recognise the addressee of a user's utterance.

Today, multiple solutions are implemented to detect if a system should react to an uttered speech command, in particular used are push-to-talk and keywords. Besides this unnaturalness in the interaction initiation, especially the currently preferred keyword method is error-prone. It can result in users' confusion, e.g., when the keyword has been said but no interaction with the system was intended by the user. Therefore, technical systems should be able to perform an addressee detection. Various aspects have already been investigated in this field of research, however most of the studies dealing with speech-enabled technical systems utilize datasets either with one human and a technical system (Lee et al., 2013), groups of humans (mostly two) interacting with each other and a technical system (Shriberg et al., 2012; Vinyals et al., 2012) or teams of robots and teams of humans (Dowding et al., 2006). These studies are mostly done using one specific scenario (Shriberg et al., 2013), just a few researchers analyse how people interact with technical systems in different scenarios (Lee et al., 2013). In these studies, the technical system is either a robot (Dowding et al., 2006; Katzenmaier et al., 2004), a research system (Shriberg et al., 2012; Vinyals et al., 2012), or a Wizard-of-Oz (WOZ)-experiment (van Turnhout et al., 2005). To the best of our knowledge, a current commercial system has not been used so far to examine addressee detection in HCI. Furthermore, previous research concentrated on analysing observable users' speech characteristics in the recorded data. The question whether users themselves recognise differences or even perhaps deliberately change their speaking style when interacting with a technical system has not been evaluated. The fact that users can be aware of speaking differently with technical systems than with humans has been described in (Frommer et al., 2017).

To address these issues, we designed the Voice Assistant Conversation Corpus (VACC) based on the interaction with a commercial voice assistant (Amazon's ALEXA). VACC further includes users' self-reports on their experiences during the interaction with the system, especially regarding their speaking style.

2. Study Design

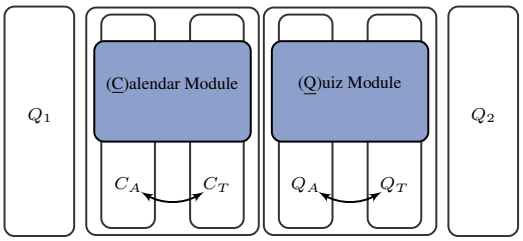


Figure 1: A sketch of the experimental procedure. Q_1 and Q_2 are the two questionnaire rounds. The order of the the scenarios (Calendar Module and Quiz Module) is fixed. A and T denote the experimental conditions (a)lone and (t)ogether with an confederate respectively. They can be interchanged.

The recorded corpus can be used for various analyses. However, we created it based on the following research questions: 1) How do humans speak with current speech-based assistant systems? 2) Which differences in the speaking style during the interaction with the technical system can be observed when they are alone or with an confederate? 3) Do they themselves recognise differences in the interaction? 4) Do the differences in the observed and/or reported interaction style differ between a formal and an informal interaction setting?

VACC consists of recordings of interaction experiments accompanied by various questionnaires presented before and after the experiment (see Sec. 4. and Fig. 1). The initial

instruction of the experiment entailed information about the basic capabilities and the keyword-based addressing of ALEXA. Two experimental modules followed, arranged according to their complexity level. There were two conditions for each module, which were permuted for different participants. Thus, each experiment contained four "rounds". A round was finished when the aim was reached or broken up to avoid frustration if hardly any success could be realized. Although, the proscribed role of the confederate is distinct from that of ALEXA, we decided for such an attempt to gather natural interactions, as they would occur in daily life when using speech-enabled assistants.

Module 1 ("Calendar Module"): In this more formal interaction the participant had to make appointments with a the project partner. He/she was instructed that ALEXA could give information about the confederates' calendar for request including exemplary commands. In condition C_A ("without the confederate") the participant only got written information about his/her partners' available dates. In condition C_T ("with a confederate") the project partner entered the room and could give this information himself. Thus, appointments could now be made by the help of both, ALEXA and the project partner. The confederate was part of the research team and was instructed to interact only with the participant, not with ALEXA.

Module 2 ("Quiz Module"): In this more informal interaction the participant had to answer questions of a quiz (e.g., "How old was Albert Einstein?"). He/she was instructed that ALEXA was not able to give the full answer, but could offer support by solving partial steps to get it. In condition Q_A the participant had to answer the questions on his/her own. In condition Q_T the participant and the project partner built up a team supported by ALEXA. The confederate (here again only interacting with the participant, not with ALEXA) was instructed to make command proposals to the participant if frustration due to failures was imminent. The quiz in Q_T was more sophisticated than in Q_A to force cooperation between the two speakers and ALEXA.

3. Recording Setup

The recordings took place at the Institute of Information and Communication Engineering, Cognitive Systems Group, University Magdeburg. They were conducted in a living-room-like surrounding, see Fig. 2. The aim of this setting was to enable the participant to get into a natural communication atmosphere (in contrast to the distraction of laboratory surroundings). The participant sat on the sofa (right side of the photo in Fig. 2) and interacted with the voice assistant system, placed on the table in the middle. The informed second speaker – Jannik – present only in the two-person variants of each scenario, sat on the armchair (left side of the photo in Fig. 2). The positions were identical for each experiment to ensure comparability.

As voice assistant system, we used the Amazon ALEXA Echo Dot (2nd generation). We opted for a commercial system to allow a fully free interaction with a currently available system. We decided against developing a voice assistant system or using a WOZ(-technique), because we wanted to meet the abilities of current commercial voice assistant systems and did not want to pretend having further capabilities. For this dataset, we declined to do video recording as we wanted to use commercial systems as they are – they do not support video or gaze analyses. Furthermore, the awareness of video recording has the danger that participants behave differently and thus distorting our primary and only analysis modality, the speaking style.

Figure 2: A snapshot of the data collection setup. The informed second speaker – Jannik – (left side) and the participant (right side) are sitting around a table, where the voice assistant (Amazon ALEXA Echo Dot) is located.

To conduct the recordings, we used two high-quality neckband microphones (Sennheiser HSP 2-EW-3) to capture the voices of the participant and the informed second speaker as well as one high-quality shotgun microphone (Sennheiser ME 66) to capture the overall acoustic and especially the output of the voice assistant. The recordings were stored in WAV-format with 44.1 kHz sample rate and 16 bit resolution.

4. Questionnaires

Several psychological questionnaires accompanied the experiment: *Before the experiment*, a short form of a self-defined questionnaire used in (Rösner et al., 2012) was utilized to obtain socio-demographic information as well as the participants' experience with technical systems (Q_1).

After the experiment, some self-defined computer-aided questionnaires were applied (Q_2). The first two of them focused on participants' experiences regarding a) the interaction with the voice assistant and the second speaker in general, b) possible changes in voice and speaking style while interacting with the voice assistant and the second speaker. According to the so-called principle of openness in examining subjective experiences (Hoffmann-Riem, 1980), the formulation of questions developed from higher openness and a free, non-restricted answering format (e.g., "If you compare your speaking style when interacting with ALEXA or with Jannik – did you recognise differences? If yes, please describe the differences when speaking with ALEXA!") to lower openness and highly structured answering formats (e.g., "Did your speed of speech differ when interacting with ALEXA or with Jannik? Yes or No? If yes, please describe the differences!"). This structure allows to examine the degree of participants' awareness of changes in the their voice and speaking style: If they already describe changes in some features (e.g. melody or speed) according to the open, initial questions, a higher degree of awareness

is indicated than if they report about differences regarding these features only when they are explicitly asked for in the closed questions.
A third questionnaire focused on previous experiences with voice assistants. Furthermore, AttrakDiff (Hassenzahl et al., 2003) was used to supplement the open questions on self-evaluation of the interaction by a quantifying measurement of the quality of the interaction with the voice assistant (hedonic and pragmatic quality).
The answering of all questionnaires takes about 20 minutes.

5. Dataset Characteristics

VACC contains recordings of 27 German speaking participants, all students at the Otto von Guericke University Magdeburg. The sex is nearly balanced with 13 males and 14 females, the age ranges from 20 to 32 years (24.11 $\pm$ 3.32 y). The data collection took about 60 minutes (40 minutes recording and 20 minutes questionnaires) per participant. Table 1 summarises the dataset characteristics. The participants came from different study courses including computer science, engineering science and humanities. Thus, this dataset is not biased towards technophilic students. Regarding the experience with voice assistants, all

Subjects/Experiments	27
Sex	Male 13 / Female 14
Total Recorded Data	17 h 07 min
Experiment Duration	Mean: 31 min
Age	Mean 24 (Std: 3.32) Min: 20; Max: 32
Language	German
Annotation	Transcription, Speaker Events
Supplementary self-reports	evaluation of interaction, AttrakDiff, speaking style, experiences in interacting with voice assistants

Table 1: Dataset Characteristics

participants had known Amazon ALEXA before. When asked about experience with ALEXA, only six participants specified that they had used ALEXA prior to this experiment. Five of them used ALEXA rarely for testing, only one participant specified that he uses ALEXA regularly – for playing music. Regarding experience with other voice assistants, additional ten participants indicated prior use. As voice assistants, they indicated Apple SIRI, GOOGLE NOW, or Microsoft CORTANA. Seven of them used these voice assistants seldom, just to try. Only three used them regularly, e.g. for programming a timer. In total, 18 out of 27 participants have prior experience with voice assistants. The nine participants not using any voice assistant before mistrusted the necessity of voice control and expressed data protection concerns when asked for reasons.
Furthermore, we analyzed the participants' technique affinity by asking how often the participant installed new software. We identified a clear distintion of 15 users who familiarise with new software at least once a quarter and 12 users familiarising with new software only 1-2x per year or less often. Interestingly, there is no significant difference between these groups in terms of the joy of computer work (JOY), the easement of work by the help of computers (EASE), weekly computer work (HOURS), or the age of the first use of computers (AGE), see Fig. 3. Comparing technique affinity and prior experience with voice assistants, seven out of nine participants having no prior experience with voice assistants also have less affinity to technology.

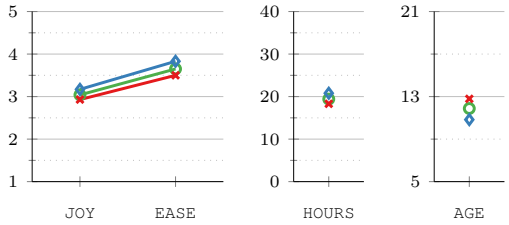


Figure 3: Evaluation of technique affinity regarding joy of computer work (JOY), the easement of work by the help of computers (EASE), weekly computer work (HOURS), or the age of the first use of computers (AGE) for all participants (—○—) as well as technology experienced (—◆—) and technology unexperienced (—✖—).

AttrakDiff is employed to understand how participants evaluate the usability and design of interactive products (Hassenzahl et al., 2003). It distinguishes four aspects (pragmatic quality (PQ), hedonic Quality (HQ) – including the subqualities identity (HQ-I) and stimulation (HQ-S), as well as attractiveness (ATT)). Regarding the experience with Amazon ALEXA, PQ, HQ-I, and ATT are perceived as neutral. Thus, it can be assumed that ALEXA provides useful features, it allows participants to identify themselves with ALEXA, and it has a kind of attractiveness. But all of these aspects leave room for improvements. Regarding HQ-S, a slightly negative assessment can be observed, showing that the support of the own needs was inappropriate. This can be justified by the difficulties of the calendar task where ALEXA has deficits. For all four aspects, no significant difference between technology experienced and technology unexperienced participants could be observed.
Furthermore, the participants filled out questionnaires dealing with their experiences of the interaction with ALEXA and the second speaker in general, regarding possible changes in their voice and speaking-style during the interaction with both as well as regarding their previous experiences with voice assistants (see Sec. 4., Q2). Besides the structured part of these questionnaires (e.g., "Have you ever worked with voice assistants aside from ALEXA? Yes or No?"), there were more open and unstructured ones, which had to be answered by using free text fields. For this, the participants used headwords and sentences to describe their experiences and evaluations. These texts made up a total number of 43307 characters. Regarding their speaking style in interacting with ALEXA compared to interacting with the second speaker, a first unsystematic analysis suggest, that participants are aware of differences e.g., in the length of

sentences or the accentuation.

As stated in Sec. 2., the two scenarios (Calendar Module, Quiz Module) are either conducted alone or together with an informed speaker. Regarding the duration of the different sequences, it can be stated that for the calendar task, the duration of the first round is remarkably longer together with the informed speaker (submodule C_T). This can be attributed to the effect that in this case, the second speaker is frequently asked about the operation of ALEXA. Regarding the Quiz Module, the submodule condition Q_T (together with the informed speaker) took longer no matter of the order. This was expectable due to harder questions. Surprisingly, if Q_T was conducted after Q_A, it took remarkably longer in comparison to the case when Q_T was conducted before Q_A.

Although aimed at analysing the speaking styles for different scenarios in single and multi-user HCI among the same participants, this dataset can be used for a variety of applications Besides the mentioned characteristics, VACC is a fruitful resource for realistic and natural HCI. It contains different communication phenomena, for instance off-talk, overlaps, laughter, engagement, and emotional reactions. This additional information is currently being annotated using listening evaluations and will be distributed as EXMARaLDA transcripts (Schmidt, 2004).

6. Conclusion

In this paper, a new dataset on natural single- and multi-user HCI is proposed. The focus if this dataset is on the interaction with a commercial voice assistant system and the speaking style while addressing the system. Within the course of the recorded interactions, participants face two different situations with and without a second supportive speaker. Furthermore, the participants' socio-demographic characteristics, their self-assessment of the interaction, their speaking style, as well as a quantifying measurement of the quality of the interaction was gathered via questionnaires. Therefore, VACC captures both, the objectively measurable voice characteristics as well as their subjective assessment. Thus, it allows to correlate voice characteristics and subjective assessments in different situations. In total, 27 subjects took part in the experiment. The mean recording time per person is about 31 m, resulting in 17 hours of recorded material. The dataset will be enriched with additional information gained from post-processing (off-talk, overlaps, laughter).

As VACC aims to represent the same participants in two different scenarios with and without an accompanying speaker and furthermore represents a naturalistic HCI, it allows to analyse the problem of addressing the technical system in these different scenarios. Furthermore, this dataset enables comparisons of user behaviour in general in different scenarios for human-computer interaction and human-human interaction.

Availability

The Voice Assistant Conversation Corpus is available for research purposes upon written request from the authors.

Literature

Dowding, J., Clancey, W. J., and Graham, J. (2006). Are you talking to me? dialogue systems supporting mixed teams of humans and robots. In *AIAA Fall Symposium Annually Informed Performance: Integrating Machine Listing and Auditory Presentation in Robotic Systems*, Washington, DC; United States, October.

Frommer, J., Rösner, D., Andrich, R., Friesen, R., Günther, S., Haase, M., and Krüger, J., (2017). *LAST MINUTE: An Empirical Experiment in User-Companion Interaction and Its Evaluation*, pages 253–275. Springer International Publishing, Cham.

Hassenzahl, M., Burmester, M., and Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003*, volume 57 of *Berichte des German Chapter of the ACM*, pages 187–196. Vieweg+Teubner, Wiesbaden, Germany.

Hoffmann-Riem, C. (1980). Die Sozialforschung einer interpretativen Soziologie – Der Datengewinn. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 32:339–372.

Katzenmaier, M., Stiefelhagen, R., and Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. of the 6th ACM ICMI*, pages 144–151.

Lee, H., Stolcke, A., and Shriberg, E. (2013). Using out-of-domain data for lexical addressee detection in human-human-computer dialog. In *Proc. NAACL*, pages 221–229, Atlanta, USA, June.

Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., and Otto, M. (2012). LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In *Proc. of the 8th LREC*, pages 96–103, Istanbul, Turkey.

Schmidt, T. (2004). Transcribing and annotating spoken language with exmaralda. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*. ELRA. EN.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Heck, L. (2012). Learning when to listen: Detecting system-addressed speech in human-human-computer dialog. In *Proc. of the INTERSPEECH'12*, pages 334–337, Portland, USA, September.

Shriberg, E., Stolcke, A., and Ravuri, S. (2013). Addressee detection for dialog systems using temporal and spectral dimensions of speaking style. In *Proc. of the INTERSPEECH'13*, pages 2559–2563, Lyon, France, August.

van Turnhout, K., Terken, J., Bakx, I., and Eggen, B. (2005). Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proc. of the 7th ACM ICMI*, pages 175–182.

Vinyals, O., Bohus, D., and Caruana, R. (2012). Learning speaker, addressee and overlap detection models from multimodal streams. In *Proc. of the 14th ACM ICMI*, pages 417–424.

Alignment in a Multimodal Interlingual Computer-Mediated Map Task Corpus

Justine Reverdy, Hayakawa Akira, Carl Vogel
ADAPT Centre
Computational Linguistics Group
School of Computer Science and Statistics, Trinity College Dublin, Ireland
{reverdyj, campbeak, vogel}@tcd.ie

Abstract

This work presents an assessment of interlocutor alignment using a semi-automated method in the context of multimodal interlingual (English-Portuguese) computer-mediated interactions. We study the adaptation phenomenon (also known as convergence behaviour and alignment behaviour) by looking at verbal repetition at different levels of linguistic representation. Since alignment behaviour has already been analysed in direct human-to-human and in human-to-agent dialogues, one may wonder whether the same behaviour is observed in interlingual computer-mediated communication. First, we compare repetitions patterns in task-oriented dialogues of human-to-human communication (HCRC Edinburgh Map Task corpus) and interlingual computer-mediated human-to-human communication (ILMT-s2s corpus), for eye-contact and no eye-contact scenarios. Secondly, we study the relation between the cognitive state of the subject, and the alignment process in interlingual computer-mediated communication settings. Results show that above chance repetitions, signalling verbal alignment, are present in both direct human-to-human communication and interlingual computer-mediated interactions, and that interlingual computer-mediated setting yields significantly more self-repetitions than direct human-to-human interactions. Also, in interlingual computer-mediated communication, a lack of alignment cues for long sequences correlated with a high amount of negative cognitive states in the eye-contact setting, implying a potential lack of mutual understanding.

Keywords: alignment, mutual understanding, task-oriented, computer-mediated, interlingual communication

1. Introduction

Interlocutor alignment (repetition of linguistic choices) is said to be an important part of human-to-human communication. In particular, the Interactive Alignment Model (Pickering and Garrod, 2004) has been taken as the basis of various works exploring this phenomenon. Different levels of linguistic representation reflect this alignment, for example in lexical choices and syntactic structures (Branigan et al., 2000; Reitter and Moore, 2007; Garrod and Anderson, 1987) or prosodic features (Giles et al., 1991).

The results of those studies show evidence that interlocutors tend to align their representation of the world to establish mutual understanding throughout conversation (Turnbull, 2003) that is sufficient for the purpose of the exchange (Newlands et al., 2003, p. 327). The achievement of mutual understanding is never entirely certain; however, interlocutors can achieve a state in which they lack direct evidence of misunderstanding (Taylor, 1992), i.e., achieve a level of understanding that is adequate to accomplish a given task (depending if the task type requires this achievement) (Brown et al., 1985). Repetition mechanisms are central in the alignment process and hold multiple functions (Tannen, 2007). They can signal understanding and by contrast, in other cases, can express a misunderstanding that will induce repair. The presence of repetitions is also an indicator of involvement or engagement in an interaction.

In two previous studies we conducted (Reverdy and Vogel, 2017a; Reverdy and Vogel, 2017b) using the data of the HCRC Edinburgh Map Task corpus (Anderson et al., 1991), we reported that in task-based interaction, repetitions that occur 'above chance' have an impact on task-success, results which are consistent with other findings in direct human-to-human communication (Branigan et al., 2000; Reitter and Moore, 2007; Nenkova et al., 2008).

Using the map-task setting, a study comparing face-to-face and video-mediated interactions (O'Malley et al., 1996, p. 177) suggested that "when speakers are not physically co-present, they are less confident in general that they have mutual understanding [...], and therefore over-compensate by increasing the level of both verbal and non-verbal information". Other studies about the alignment process with a virtual agent reported evidence of exaggerated alignment when the speakers thought they were talking to a machine (Branigan et al., 2010; Dubuisson Duplessis et al., 2017). Previous experiments have also found that dialogue acts used by the subjects during task-oriented computer-mediated communication differ substantially from direct human-to-human communication, with backchannel utterances (acknowledging understanding) reduced significantly in computer-mediated interlingual communication (Hayakawa et al., 2016b).

Another study examined alignment in machine-translated communication, but in a de-contextualized setting (Schneider and Luz, 2011), including a Wizard-of-Oz experiment where participants were asked to answer machine-translated questions. Half of the questions contained translation mistakes resembling ones an MT system would produce. Their results indicate that people align their answer and reproduce the obvious errors (translation mistakes), assuming that the speech-to-speech machine translation (S2S-MT) system would understand them better. To the best of our knowledge, alignment has yet to be studied where the communication is mediated by an S2S-MT system, between two people who are aware that they are interacting with each other, in particular in the context of a map-task, where specific lexical items need to be transmitted in order to achieve a common goal.

Therefore, we see a need to extend these studies to

computer-mediated communication to verify how alignment through repetition changes in this new communication style. In addition, the speaker's cognitive state could be an identifier of smooth or problematic communication. For example, results of a study in the context of call centres show that customers' frustration, irritation or surprise (that one could define as negative cognitive states) , have a negative impact on communication. The call centre staff would try to reduce the customers' negative emotional attitudes to ease the interaction and resolve the customers' issues (Botherel and Maffiolo, 2006, p. 3).

In this paper, we exploit two multimodal corpora to observe repetition of linguistic choices as cues of an alignment process. (i) We first compare direct human-to-human communication with interlingual computer-mediated communication to verify if alignment is exaggerated in computer-mediated interlingual communication. The results show that the method detected equivalent cues of alignment in both direct human-to-human and interlingual computer-mediated communication settings, with the latter displaying significantly more self-repetition than direct human-to-human communication. (ii) Secondly, we emphasize the possible role of repetitions in relation with the cognitive states of the subjects within computer-mediated interlingual communication. In those settings, we found that the lack of alignment cues for long sequences correlated with high amounts of negative cognitive states, pointing to possible communication problems (lack of mutual understanding).

2. Data Set

Data from two multimodal corpora that use the Map Task technique to elicit spontaneous communicative behaviour was used. For the direct human-to-human communication, we used a subset of 16 dialogues from the HCRC Map Task corpus (Anderson et al., 1991), and for the computer mediated interlingual communication, we used all 15 dialogues from the ILMT-s2s corpus (Hayakawa et al., 2016c), see Table 1. The subjects were assigned the role of Information Giver (IG) or Information Follower (IF) and each given a map containing similar landmarks. The IG had a map with a route drawn along the landmarks with a START and a FINISH indicated, and was tasked with guiding the IF through a map not displaying FINISH.

	HCRC (Subset)	ILMT-s2s	
Language	English	English	Portuguese
Tokens	22,106	13,761	12,671
Turns	3,790	2,310	2,236
SELF REP	2,448	3,877	2,306
OTHER REP	2,653	2,407	1,107

Table 1: HCRC Map Task and ILMT-s2s Corpora Summary; SELF REP and OTHER REP (see definition § 3.) are given for the linguistic representation level token only.

2.1. The HCRC Map Task corpus

The HCRC Map Task corpus consists of 128 English dialogues of direct human-to-human task based interactions. The recordings were split in two settings, with half the subjects being able to see their interlocutor's face (i.e., with eye-contact), while the other half had screens placed between them (i.e., without eye-contact). To standardise the data, only dialogues that used the same maps (maps 1 & 7) as those used in the ILMT-s2s corpus (§ 2.2.) were kept for this study, resulting in a total of 16 out of the 128 (half male, half female in both the main corpus and the subset).

2.2. The ILMT-s2s corpus

As with the HCRC Map Task corpus, the dialogues use the map task technique, but with a difference that the subjects are located in different rooms, speak different languages to each other and communicate via a Speech-to-Speech Machine Translation (S2S-MT) system — the ILMT-s2s System. The ILMT-s2s corpus consists of fifteen dialogues between fifteen English, and fifteen Portuguese subjects (16 females, 14 males). The maps that are used are the same as the HCRC Map Task corpus, in their original version for the English speakers, and translated for the Portuguese speaking subjects. The ILMT-s2s System is a rapidly built system that uses off-the-shelf components — the Google Speech API for Automatic Speech Recognition (ASR), the Microsoft Bing translation service for Machine Translation (MT), and the Apple system voices provided with Mac OS X computers for Text-to-Speech synthesis (TTS) — to perform the S2S-MT. The corpus was annotated for the cognitive states of Frustration, Amusement,[1] and Surprise, for each speaker in all the dialogues, with the assessment made through video and audio modalities. The inter-coder agreement for the labels was calculated[2] and the results are well above .6. A user survey was also conducted to collect the user's appreciation of the system. Each question follows a 7 point Likert scale ranging from '1 – Strongly disagree' to '7 – Strongly agree', designed in such a way that the more they agreed to the statement, the more positive their experience was. Due to the push-to-talk activation method of the system, subjects did not only talk to their interlocutor (*On-Talk*), but also spoke out loud to themselves and other people in the room (*Off-Talk*) (Hayakawa et al., 2016a). To standardise the data between corpora, only *On-Talk* was used for the analysis.

3. Method

We counted the repetition of tokens of a contribution and the immediately preceding contribution, that we assimilated as a dialogue turn of each speaker (Vogel and Behan, 2012; Vogel, 2013). A REGISTER is created for each participant, containing her or his most recent contribution. For each dialogue turn, the REGISTER is populated with counts of each repetition of a token, for other-repetitions (repetition of a token uttered by the other participant — OTHERSHARED) and self-repetitions (SELFSHARED). Tokens are counted as n-grams, up to $n = 5$. The n-grams length was divided into three length types — N: n = All ($n = 5$); N1: $n = 1$; N2+: $n > 1$ (from 2 to 5, long sequences). In each dialogue, the turns are then randomly re-ordered

[1] We note that Amusement was considered negative for English speaking subjects, as it was a reaction to high word error rate utterances output (Hayakawa et al., 2017).

[2] Using the modified kappa feature of ELAN (Wittenburg et al., 2006) version 4.9.0's "Inter-Annotator Reliability…" function.

ten times. This resulted in ten randomly ordered dialogues where other and self-repetitions were counted again. In the direct human-to-human dialogues, the count was carried out between the utterances of the two human subjects. However, for the computer-mediated dialogues, the count was carried out within the same language — the utterances from the English speakers are coupled with the English translation of the Portuguese speakers utterances and vice-versa, which created two fully monolingual dialogues.

A pre-process labelling, designed to measure five different levels of linguistic repetition types, was applied: (i) Token, (ii) Lemma, (iii) Part-Of-Speech (POS), (iv) a combination of Lemma with POS, and (v) a combination of Token with POS. Data from the HCRC Map Task corpus and the English dialogues of the ILMT-s2s corpus were labelled with the TreeTagger English training set (Schmid, 1994), while the Portuguese dialogues of the ILMT-s2s corpus were labelled using the TreeTagger tagset proposed by Pablo Gamallo (Gamallo and Garcia, 2013). The aim is to observe if a significant difference is identified between the actual dialogues and the randomized dialogues, using the statistical test described below.

To verify if there was a difference in the subject repetition patterns in the two corpora, the single-step Tukey HSD multiple comparison test was performed using a general linear model with a binomial error family (Bretz et al., 2016). The null hypothesis for the test was as follows:

$H_0 : Random.Speaker.Level.N - Actual.Speaker.Level.N \geq 0$

The null hypothesis (H_0) states that the difference between the amount of repetitions in the randomized dialogues and the actual dialogues should equal (or exceed) zero if repetitions are simply due to chance. If rejected, the assumption is that a potential role in the communication could be accepted. For each dialogue, the model was computed and dialogues with repetitions 'above chance' or not were identified: (i) per speaker (IG: Information Giver, IF: Information Follower), (ii) per n-gram (All n-grams [up to length 5]; N1: $n = 1$ [length 1]; N2+: $n > 1$ [length 2 to 5]), (iii) per type of repetition (OTHERSHARED and SELFSHARED), and (iv) per linguistic Level: TOKEN (L1), LEMMA (L2), LEMMA+POS (L3), POS (L4), TOKEN+POS (L5). This allowed us to observe a rate of H_0 rejection, defined as the "number of actual rejections of the null hypothesis" over the "number of possible rejections of the null hypothesis" in each categories. We compared the rates of rejection of H_0 in the two corpora, and the combinations of those tests is the basis of our meta-analysis. Since the two corpora contained dialogues with and without eye-contact, and the ILMT-s2s corpus is annotated for cognitive states and two languages, we observed the rates of rejections in relation with those conditions.

4. Results

4.1. Human-to-Human vs Computer-Mediated

The null hypothesis (H_0), with the threshold of $p \geq 0.05$, was rejected 233 times out of 300 for OTHERSHARED and 273 times out of 300 for SELFSHARED in the ILMT-s2s corpus across all linguistic levels while in the data from the HCRC Map Task, OTHERSHARED was rejected 111 times out of 160 and SELFSHARED was rejected 25 times out of 160 (Table 2). This reveals a considerable difference in the rejection rate for SELFSHARED repetitions between the direct human-to-human dialogues of the HCRC Map Task corpus (25/160 = 0.15) and those of the ILMT-s2s corpus (273/300 = 0.91), with SELFSHARED repetitions happening 'above chance' more often in the computer-mediated corpus. A Mann-Whitney-Wilcoxons test found that across all linguistic levels, the number of SELFSHARED repetitions is significantly different ($p = 2.686e-06$) between the HCRC Map Task (with an average rejection of $\overline{x} = 2.5$) and the ILMT-s2s corpus (with an average rejection of $\overline{x} = 13.65$). However, no significant difference ($p = 0.9636$) was found between the two corpora concerning OTHERSHARED repetitions at level n-grams = All, both corpora showing a high rate of rejection of H_0. No significant difference was found between the two corpora in terms of speaker role, language spoken, and eye-contact modality at level n-grams = All.

Lng	SHARED	Role	L1	L2	L3	L4	L5	*M*
ILMT-s2s English n-grams = All								
Eng	OTHER	IG	12	12	12	11	12	11.8
Eng	OTHER	IF	12	12	13	9	13	11.8
Eng	SELF	IG	14	14	14	13	14	13.8
Eng	SELF	IF	14	14	14	11	14	13.4
H_0 rejection: 254 / 300 (OTHER: 118 / 150, SELF: 136 / 150)								
ILMT-s2s Portuguese n-grams = All								
Por	OTHER	IG	13	12	13	10	13	12.2
Por	OTHER	IF	12	12	12	6	12	10.8
Por	SELF	IG	14	15	15	14	14	14.4
Por	SELF	IF	14	14	14	9	14	13
H_0 rejection: 233 / 300 (OTHER: 115 / 150, SELF: 137 / 150)								
HCRC Map Task n-grams = All								
Eng	OTHER	IG	11	12	10	4	6	8.6
Eng	OTHER	IF	15	14	14	10	15	13.6
Eng	SELF	IG	**2**	**2**	**3**	**0**	**2**	**1.8**
Eng	SELF	IF	**4**	**2**	**4**	**2**	**4**	**3.2**
H_0 rejection: 136 / 320 (OTHER: 111 / 160, **SELF: 25 / 160**)								

Table 2: Rejection count of H_0 for levels L1 to L5 and mean (*M*) values in the ILMT-s2s corpus and HCRC Map Task corpus for all n-grams. For each dialogue at each level, the number of possible H_0 rejection is 15 in the ILMT-s2s corpus, and 16 in the HCRC Map Task corpus.

4.2. Within Computer-Mediated Interactions

No impact of 'above chance' repetition in relation to the cognitive states of the participants was found at n-grams length n = All (count listed in Table 3). However, differences appeared for OTHERSHARED repetitions of Portuguese (IF) at n-gram length n >1 (N2+) in "Eye-Contact" conditions (Table 4). While in all other settings the rate of rejections of H_0 remains high, the Portuguese IF speakers did not repeat the English speakers' token in the same proportion in the "Eye-Contact" condition.

This relation is highlighted with Pearson's standardized residuals from log-linear models in Figure 1. For long sequences of n-gram repetitions (N2+), we observe that when there is Eye-Contact, the Portuguese speakers show higher levels of negative cognitive states than expected when they are at the same time not repeating the English speaker.

Role	IF			IG			Total
Cog.	Fru	Sur	Amu	Fru	Sur	Amu	
Eng	67	57	220	103	54	263	764
Por	290	137	113	210	105	184	1039
Total	884			919			1803

Table 3: Number of Cognitive States per Subject Role (Information Follower, Information Giver), Spoken Languages (English, Portuguese) and Cognitive State Type (Frustrated, Surprised, Amused) in the ILMT-s2s corpus

Lng	SHARED	Role	L1	L2	L3	L4	L5	*M*
With Eye-Contact n >1 (N2+)								
Eng	OTHER	IG	6	6	6	6	6	6.0
Eng	OTHER	IF	6	6	5	5	5	5.4
Eng	SELF	IG	7	7	7	7	7	7.0
Eng	SELF	IF	8	8	8	6	6	7.2
Por	OTHER	IG	5	4	5	4	5	4.6
Por	OTHER	IF	**3**	**4**	**4**	**3**	**2**	**3.2**
Por	SELF	IG	7	7	7	7	7	7.0
Por	SELF	IF	7	7	6	5	6	6.2

Table 4: Rejection count of H_0 for levels L1 to L5 and mean (*M*) values. In each case the number of possible H_0 rejection is 8 (modality: eye-contact).

Meanwhile they show less frustration than expected if they repeat the English speaker for long sequences (N2+).

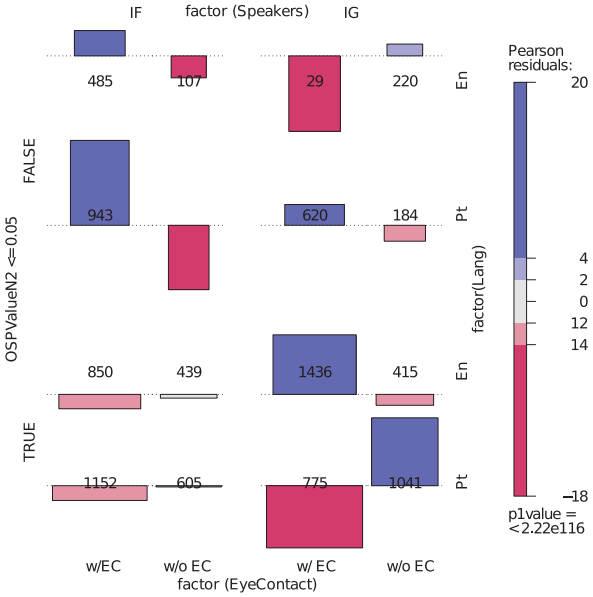


Figure 1: Association Plot of significant OTHERSHARED residuals (TRUE: p <=0.05 — FALSE: p > 0.05) for n-gram>1 (N2+), Subject Role (IG: Information Giver—IF: Information Follower), Eye-Contact (w/ EC: with Eye-Contact—w/o EC: without Eye-Contact), and Language Spoken (En: English—Pt: Portuguese)

The distributions of negative cognitive states was found significantly different between 'above chance' and non-'above chance' OTHERSHARED repetitions for the Portuguese IF speakers at n-gram>1 level (W = 883, p-value = 0.027).
The low rate of N2+ repetitions detected is echoed in the user survey conducted in the ILMT-s2s corpus. The Portuguese speakers (IF) in "Eye-Contact" conditions showed the lowest appreciation of the system (Median score = 3.0; Overall Median score = 5.0), which correlates with a high amount of negative cognitive states for those speakers.

5. Discussion

The high rate of 'above chance' OTHERSHARED repetition in the computer mediated dialogues of the ILMT-s2s corpus indicates that alignment occurs in at least the same proportion as in direct human-to-human communication. We did not find evidence of its' exaggeration with the method, as it detected equally high alignment cues in direct human-to-human communication. However, 'above chance' repetitions occurred at all linguistic levels at a high rate in the ILMT-s2s corpus, for both OTHERSHARED and SELFSHARED. This is different from the direct human-to-human dialogues where 'above chance' SELFSHARED repetitions occurred at a much lower rate. This high rate of SELFSHARED repetition could be attributed to the perceived difficulty for the speakers to have their utterance properly recognized by the ASR and correctly translated to their interlocutor, hence their tendencies to repeat themselves more. The high rate of repetition, in both types (OTHER and SELF), in this interlingual computer-mediated corpus, follows past findings that suggest strong alignment in human-computer interaction. To the best of our knowledge, this is the first time that a method of assessing alignment, by counting repetition, has been applied to dialogues of interlingual computer-mediated task-based communication.
Secondly, a relation emerged within the computer-mediated dialogues, between negative cognitive states and low 'above chance' repetitions of long sequences. Portuguese speakers in eye-contact conditions had a higher than expected negative cognitive states which also related to their low appreciation of the system. Previous work suggested that exaggerated alignment toward a system was detrimental to the interaction since the subjects also repeated translation errors (Schneider and Luz, 2011). Our findings show that the lack of alignment of long token sequences in video conditions indicates problematic interactions.

6. Conclusion

We note that even if the small size of the two corpora prevents us from making too broad a statement, the repetitions patterns detected by the automatic method present S2S-MT software design cues that constitute another step toward aiding human-to-human communication when interacting through machine translation. One might wonder if the reason that differences appeared between English and Portuguese speakers could be interpreted as a cultural difference. This could be examined in the future by comparing other language pairs and/or larger data sets.

7. Acknowledgements

This research is supported by Science Foundation Ireland through the Research Centres Programme (Grant 13/RC/2106) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin. The ADAPT Centre for Digital Content Technology is co-funded under the European Regional Development Fund.

8. Bibliographical References

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.

Botherel, V. and Maffiolo, V. (2006). Regulation of emotional attitudes for a better interaction: Field study in call centres. In *Proceedings of 20th International Symposium on Human Factors in Telecommunication*, Sophia Antipolis, France.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Branigan, H. P., Pickering, M. J., Pearson, J., and McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

Bretz, F., Hothorn, T., and Westfall, P. (2016). *Multiple comparisons using R*. CRC Press.

Brown, G., Anderson, A., Shillcock, R., and Yule, G. (1985). *Teaching talk: Strategies for production and assessment*. Cambridge University Press.

Dubuisson Duplessis, G., Clavel, C., and Landragin, F. (2017). Automatic measures to characterise verbal alignment in human-agent interaction. In *Proceedings of SIGDIAL 2017*, pages 71–81.

Gamallo, P. and Garcia, M. (2013). FreeLing e TreeTagger: um estudo comparativo no âmbito do Português. Technical report, Universidade de Santiago de Compostela.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Giles, H., Coupland, J., and Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Hayakawa, A., Haider, F., Luz, S., Cerrato, L., and Campbell, N. (2016a). Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task. In *Proceedings of Speech Prosody 2016*, pages 776–780, Boston, MA, USA. ISCA.

Hayakawa, A., Luz, S., and Campbell, N. (2016b). Talking to a System and Talking to a Human: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In *Proceedings of INTERSPEECH'16*, pages 1422–1426, San Francisco, CA, USA. ISCA.

Hayakawa, A., Luz, S., Cerrato, L., and Campbell, N. (2016c). The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of LREC 2016*, pages 605–612, Portorož, Slovenia. ELRA.

Hayakawa, A., Vogel, C., Luz, S., and Campbell, N. (2017). Perception Changes With and Without the Video Channel: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In *Proceedings of CogInfoCom 2017*, pages 401–406, Debrecen, Hungary. IEEE.

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of ACL HLT 2008*, pages 169–172. Association for Computational Linguistics.

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.

O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., and Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction. *Interacting with computers*, 8(2):177–192.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190.

Reitter, D. and Moore, J. D. (2007). Predicting Success in Dialogue. In *Proceedings of ACL 2007*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.

Reverdy, J. and Vogel, C. (2017a). Linguistic Repetitions, Task-based Experience and A Proxy Measure of Mutual Understanding. In *Proceedings of CogInfoCom 2017*, pages 395–400, Debrecen, Hungary. IEEE.

Reverdy, J. and Vogel, C. (2017b). Measuring Synchrony in Task-Based Dialogues. In *Proceedings of INTERSPEECH'17*, pages 1701–1705, Stockholm, Sweden. ISCA.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 154–164, Manchester, UK.

Schneider, A. and Luz, S. (2011). Speaker Alignment in Synthesised, Machine Translated Communication. In *Proceedings of IWSLT 2011*, pages 254–260, San Francisco, California, USA. ISCA.

Tannen, D. (2007). *Talking voices: Repetition, dialogue, and imagery in conversational discourse*, volume 26. Cambridge University Press.

Taylor, T. J. (1992). *Mutual Misunderstanding: Scepticism and the theorizing of language and interpretation*. Duke University Press.

Turnbull, W. (2003). *Language in action: Psychological models of conversation*. Psychology Press.

Vogel, C. and Behan, L. (2012). Measuring Synchrony in Dialog Transcripts. In *Cognitive Behavioural Systems. Lecture Notes in Computer Science*, volume 7403, pages 73–88. Springer Berlin Heidelberg.

Vogel, C. (2013). Attribution of Mutual Understanding. In *Journal of Law and Policy*, volume 21.2, pages 377–420.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC'06*, pages 1556–1559, Genoa, Italy. ELRA.

9. 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.

Conversational Gaze Modelling in First Encounter Robot Dialogues

Kristiina Jokinen
AIRC AIST Waterfront
Aomi, Koto-ku, Tokyo JAPAN
{ kristiina.jokinen}@aist.go.jp}

Abstract

Given the popularity of humanoid social robots which can talk with humans and maintain human-like communication patterns, an interesting question is whether the users engage themselves with such systems in a manner similar to human-human communication. If the humanoid robot is perceived as a communicative agent, it can be hypothesized that the user's engagement with the robot resembles social interaction rather than tool manipulation. This paper reports on a pilot study that explores if the hypothesis is supported in the context of a humanoid robot application which reads a digital newspaper interactively for the user. Human eye-gaze patterns are used as an objective measure of the engagement with the robot. The study found support for the hypothesis, but concludes that the interaction is socially less binding than with humans.

Keywords: conversation management, gaze modelling, eye-tracking, first encounter human-robot interaction

1. Introduction

The popularity of humanoid social robots which can talk with humans and appear having human-like communication patterns, has brought in interesting questions about whether the users engage themselves with such robots in a similar manner as they do when communicating with human partners. Although natural language interactions between humans and intelligent agents are often problematic due to limited communicative capabilities of the system, they nevertheless give rise to expectations that the system functions more like a communicating agent than a voice-controlled tool. This may be due to the people's tendency to anthropomorhisize computers and other media, i.e. treat them as if they were real people (Reeves and Nass 1996). However, when interactions are conducted with a humanoid robot which does not only speak, but also acts in a human like manner (i.e. moves and gestures), such expectations are reinforced and easily lead the robot to be perceived as an intelligent agent with near-human communicative competence. Consequently, if the users perceive a humanoid robot as a communicating agent, it can be hypothesised that their behaviour towards the robot resembles social interaction with other humans, rather than tool manipulation.

One of the fundamental characteristics of human-human conversations is that the interlocutors look at their partners' face (not necessarily straight into the eyes but in the facial area), i.e. eye-gaze is an important means for joint control and coordination of the interaction. According to Gullberg & Holmqvist (1999), gaze is fixated on the partner's face about 90% of the interaction time, and a similar fixation pattern is carried over to conversations conducted through videoconference technology, although gaze tends to wander around the screen and the overall environment especially if the partner's video is not life-size (Gullberg and Holmquist, 2006). The gaze of another person is a strong cue of where to focus one's visual attention (Friesen and Kingstone, 1998), and in developmental psychology, gaze-following and visual joint attention are regarded as social phenomena learnt through interaction with the others, and children learn them early, at the age of about 1 year (Meltzoff and Brooks 2007).

As the fundamental function of eye-gaze is related to monitoring the partner's gaze direction so as to establish joint attention, it can be assumed that also in human-robot interactions, visual attention plays an important role. Even if the robot cannot reciprocate the gaze, the users may apply visual attention to their robot partners in a similar way as they do with their human partners, i.e. their gaze patterns follow social expectations found in human interactions.

In this paper we set out to study if the hypothesis that the robot is perceived as a communicative agent rather than an interface tool can be supported by the user's eye-gaze behaviour. We report of a small experimental study that explored if the hypothesis is supported in the context of a humanoid robot application which reads a digital newspaper interactively for the user. Using eye-tracker technology, human eye-gaze patterns are detected and used as an objective measure for the user's engagement with the robot. The study found support for the hypothesis, but also concludes that the interaction is socially less binding with the robot agent than with humans. Due to the case study nature of the experiment, the results will be investigated later with a large set of participants.

The paper is structured as follows. Section 2 provides a brief overview of the background and related research in interaction studies and eye gaze as a social signal. Section 3 presents the experimental setup, and Section 4 describes the results. Section 5 concludes with future views.

2. Gaze as social signal

The role of eye-gaze as a means of social signalling has long been established, see Kendon (1967), Argyle & Cook (1976), Goodwin (1980). Its fundamental function is related to visual attention, and in communicative situations this means monitoring the partner's gaze so as to establish joint attention and enable construction of shared context and mutual understanding. Land (2006) points out that gaze is also proactive in nature since it anticipates actions: we often gather visual information from our surroundings before performing motor actions.

Conversational feedback can be effectively mediated by gaze behaviour. The partner's willingness to continue interaction can be inferred from their looking at or looking away from the partner, and in general, direct and averted gaze can signal the speaker's interest to approach or to avoid the object of attention (Mutlu et al. 2012).

Also turn-taking is coordinated by gaze: a quick shared gazing at each other, mutual gaze, is used to agree on the change of the speaker (Kendon, 1967; Brennan et al. 2008;

Jokinen et al., 2010, 2012, Mutlu et al. 2006). Levitski et al. (2012) observed different gaze patterns within a one second window at the beginning and at end of the utterance and noticed that in the beginning of the utterance, mutual gaze is quickly broken by the speaker, whereas at the end of the utterance, the speaker's gaze fixates onto the partner quite a long time before their speaking ends. As the speaker needs to focus their attention to the next speaker to facilitate smooth turn-taking, the interlocutors also fixate their eyes more often and longer in the beginning than at end of one's utterance, whereas in the middle of their speaking, the gaze wonders off since the speaker focusses on producing their own utterance.

See Jokinen (2014) for a longer description of eye-tracker and gaze research, and Ruhland et al. (2015) and Broz et al. (2015) for overviews of the work on eye gaze and human-robot interaction.

3. Experimental setup

The main hypotheses that the experiment focus on, are:

1. Majority of human eye gaze focus on the robot's head.
2. Gaze focus in the beginning of the interaction differs from the gaze focus at the end of the interaction.
3. There is more focus on the robot's face in the beginning than at the end of the interaction (the user becomes more familiar with the robot).
4. There is not much focus on the robot's gesturing.

The study used two female participants who were between 20-40 years of age and worked as researchers at the university. Neither of them had prior contact with robots and they also had neutral expectations of the interaction with the robot. Both participants had normal vision.

Eye gaze was measured using SMI Mobile eye-tracking glasses (SMI ETG 2 Wireless 60 Hz), and the data created using a Lenovo X230 laptop with Intel® Core TM i7-3520M CPU 2.90 GHz. Statistics were calculated using IBM SPSS Statistics 22.0.

Figure 1. The NAO robot in the centre of the study setting.

The robot was the humanoid NAO robot developed by Softbank/Aldebaran Robotics (Figure 1). The robot was installed with MoroTalk, a newspaper reading application for the national newspaper (Jokinen & Wilcock, 2013). This had been developed in collaboration with the Koti ('Home') project aiming to improve well-being and health care in future smart homes. The application is based on the WikiTalk technology (Wilcock, 2012) which allows interactive access to digital repositories. The robot supports open-domain conversations, and the user can shift to related topics or switch to a totally new topic by spelling the first few letters.

The setting was a brightly lit classroom and the robot stood on a table facing the participants so that the robot and the human were at similar eye level. The participants were first asked to fill in a short demographic form and survey on their previous experience with robots. Then they were fitted with the eye tracking glasses which were calibrated during a three-point calibrating session until accurate. The participants were instructed how to interact with the robot and a short description of the robot's abilities was given. The participants were told that the robot will read them news from today's newspaper and that they could select interesting news for the robot to read. The session started when the robot began its introduction speech, after which the participant began commanding the robot.

The experiment leader controlled the beginning and the end of the session, and, intervened if necessary, e.g. if the robot shut down. All interactions were videotaped using the eye tracker and two extra video cameras. The interactions took 10-15 minutes and afterwards the participants filled in a short feedback form about their experience.

The data from the eye gaze videotapes were annotated using the Elan Linguistic Annotator version 4.1.0 (Wittenburg et al. 2006). The eye movement data were coded for the duration and the location of the fixations. Five different categories for the targets of eye gaze were used:

1. gaze focused on the robot's head;
2. gaze focused on the robot's hand;
3. gaze focused on another part of the robot;
4. gaze focused on the study conductor;
5. gaze focused on background.

Annotations were done on the first and the last three minutes of each of the two robot human interaction sequences (altogether 12 minutes), so as to be able to compare the participants' gaze behaviour at the beginning and at the end of the interaction.

The data sets were annotated by two annotators, who were blind to each other's annotations. A two-minute section in the beginning of one of the videotapes was annotated by both annotators to determine consistency among the annotators. The interrater reliability was found to be Kappa = 0.42 ($p<.001$), 95% CI (0.249, 0.597). According to the scale proposed by Rietveld and van Hout (1993), these values indicate fair agreement between the two scorers.

For the analyses, three measures were of interest:

1. the amount of changes in gaze focus;
2. the length of individual fixations on the five eye gaze targets coded for;
3. the accumulated fixations time on the five gaze targets.

The frequencies and time durations were then compared between the first two minutes and the last two minutes of the annotated interaction sequences. The difference in the changes of gaze fixations between the end and the beginning of the interaction sequences was assessed using the Chi-Square test. Due to uneven sizes of cases categorized into the five different coding categories, the assumption of homoscedasticity for analyses of variance was not met. However, to assess whether the mean duration of the human participants' gaze focus differed between the beginning and the end of the interactions one paired t-test was conducted.

4. Results and discussion

Changes in the human participants' gaze focus between the different parts of the robot and the background is summarized in Table 1 and illustrated in Figure 2. The data on the two naïve users' eye tracking patterns suggests that there were differences between the beginning and the end of the interaction period, and overall, the three original hypotheses are corroborated by the data. First, most of the fixations in the beginning and in the end of the interactions were on the robot's face. Moreover, there were more changes in gaze focus in the beginning than at the end of the human robot interactions for both of the naïve participants. As illustrated in Figure 2, in the beginning of the interactions there were more fixations on the robots head, and other parts of the robot. In contrast, at the end of the interactions there were fewer fixations on the robot's head and body, and more fixations on the background for one of the participants. There were no differences between the beginning and the end of the interactions regarding fixation counts on the hands of the robot. Chi-Square tests indicated a statistically significant association between the target of gaze focus changes and the time during the interaction ($\chi 2$ (4, N=205) = 15.378, p = .004). That is to say, the five different gaze targets were focused on differently between the beginning and end of the interactions. Cramer's V test of the strength of association indicated a medium effect size (ϕCramer = .274).

	Beginning of interaction		End of interaction	
Gaze focus	Count	Percentage	Count	Percentage
Participant 1				
Robot's head	29	42.6	14	41.2
Robot's hands	3	4.4	3	8.8
Other part of robot	23	33.8	10	29.4
Background	11	16.2	7	20.6
Study conductor	2	2.9	0	0
Participant 2				
Robot's head	29	43.3	18	50
Robot's hands	10	14.9	10	27.8
Other part of robot	25	37.3	0	0
Background	3	4.5	8	22.2
Study conductor	0	0	0	0
Total	135	100	70	100

Table 1. Counts of changes in gaze focus in the beginning and end of the interactions; visualisation in Figure 1.

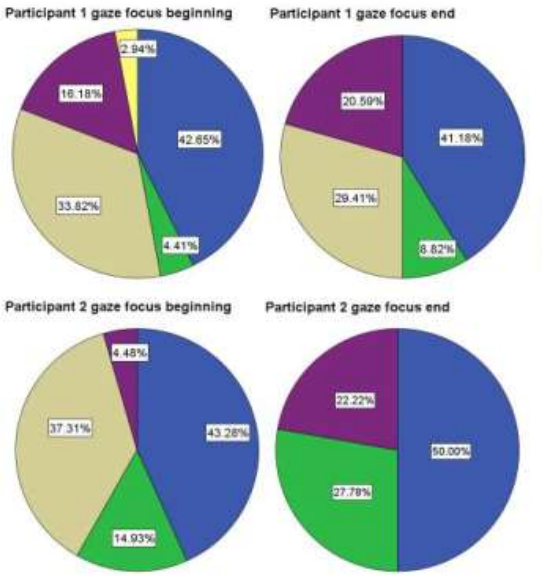


Figure 2 Gaze focus changes for the two human participants. Blue= head, green = hand, grey = other part, violet = background, yellow = study conductor.

	Beginning of interaction			End of interaction		
Gaze focus	Mean	Std. Deviation	Sum	Mean	Std. Deviation	Sum
Participant 1						
Robot's head	2.88	4.31	83.60	2.60	2.86	36.45
Robot's hands	0.63	0.58	1.90	0.26	0.13	0.77
Other part of robot	0.74	0.72	16.95	0.51	0.44	5.05
Background	1.49	2.07	16.38	1.15	1.15	8.04
Study conductor	0.89	0.27	1.78			0
Participant 2						
Robot's head	3.42	3.73	99.15	6.18	7.81	111.21
Robot's hands	0.73	0.67	7.26	0.20	0.14	2.00
Other part of robot	0.89	1.07	22.17			0
Background	0.26	0.15	0.79	2.25	5.43	18.02
Study conductor			0			0

Table 2 Average length of gaze fixations in the beginning and end of the interaction sequences for the two human participants.

Descriptive statistics of the eye gaze durations are shown in Table 2. The mean durations of the eye gazes on the five different targets are roughly similar between the end and the beginning of the interactions. An exception is that for participant 2 the fixations on the robots head are on average longer in the end compared with the beginning of the interaction (means 3.42 and 6.18 respectively) like the average fixations on the background (means .26 and 2.25 respectively). For most categories the mean gaze durations are inconsistent between the participants (one participant has longer durations in the beginning when the other has longer durations in the end) or there are too few cases for comparison (e.g. only participant 1 has fixations on the study director) However, the mean duration of fixations on the robot's hands is longer in the beginning of the interactions. A paired t-test found that this difference in mean gaze duration on the robot's hands was significant (t(12)=2.811, p=.016).

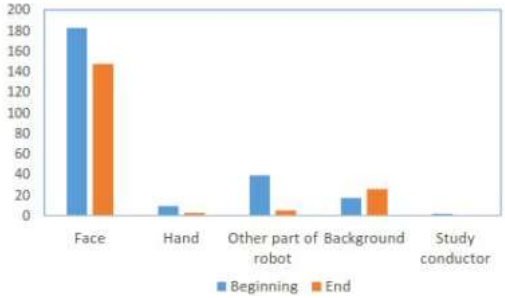


Figure 3 Totals of gaze durations between the end and beginning of interactions.

Figure 3 shows the accumulated sums of the length of fixations on the five different gaze targets between the beginning and end of the interactions. As can be seen from Figure 3, overall the longest duration of time was spent on focusing on the robot's face. When comparing the beginning and end of the interactions, the participants spent less time on focusing on the robot and more time focusing on the background in the end of the interaction.

5. Conclusions and future work

This experimental study provided support for the hypothesis that human-robot interaction is social interaction as opposed to interaction with a tool, and the results were in line with previous study results that have found human robot interaction to resemble that of human to human interactions (Jokinen et al, 2012; Yonezawa et al., 2007; Yu et al., 2012). Overall, the counts of eye gaze fixations and the duration on fixations was largest for the robot's head at all times during the interaction and for both the study participants, which supported our hypothesis (1)

in Section 3. Furthermore, the pilot study provided support for the hypothesis (2) that adapting to the robot changes human's gaze fixations. There were more changes in gaze focus in the beginning than at the end of the human robot interactions. Chi-square analysis indicated that the targets of the gaze fixations differed significantly between the end and the beginning of the interactions. Finally, the pilot study found support for the hypothesis (3) that when the robot becomes more familiar to the human, there is less focus on the robots head and gaze starts to wonder elsewhere. There were more fixations on the robots head, and other parts of the robot in the beginning than during the end of the interactions. The total length of the fixations on the robot's head and other parts of the robot's body was also longer in the beginning of the interactions. In contrast, there were more fixations on the background and the total length of fixations on the background was longer in the end as compared to the beginning of the interactions.

An interesting finding is that while the counts of fixations on the hands of the robot were the same in the beginning and the end of the interactions, the duration of these hand fixations were longer in the beginning than in the end. This can mean that in the beginning of the interaction the user found the robot's gesture behaviour novel and focused attention longer on the gestures to gather more information about them, whereas at the end of the interaction the user had already got familiar with the robot's gesturing and did not need to spend so much time on them.

The study is an experimental study with small sample size of naïve participants, which makes the results less generalizable. More data with more participants and more interactions with the robot are needed to fully investigate how people learn to interact with humanoid robots in the future. However, the strength of the study was the novel topic of study and its explorative nature. The results give a good indication of how people new to humanoid robots may react to them. Based on averaged and subjective estimations, it seems that our initial hypothesis regarding the fixation points and durations was partially correct, although not sufficiently accurate.

Regarding the practical setup, care should be paid to the eye-tracking glasses. They were not always held correctly during the experiments, but slipped down the bridge of the participants' noses, in particular when the user laughed. In this experiment it did not seem to present a major issue but should be taken into attention in future studies.

The purpose of the study was to investigate how humans without any previous experience of humanoid robots begin to interact and adapt their gaze behaviour when they first meet and interact with a humanoid robot. The results suggest that humanoid robot interaction is social, but it is not as encaptivating and smooth as interaction between humans. Naïve participants instantly focused mainly on the robot's head and perhaps learned to ignore hand gestures as the interaction progressed (although hand gestures were designed to support presentation and rhythm of the robot's utterance). After the novelty of the beginning was worn out there were less changes in gaze fixations.

While the Nao robot is a cute humanoid robot, its facial expressions are limited to flash lights. According to Media Equation Hypothesis this does not prevent the user to bond with the robot and interact in a natural manner since people's interactions with computers and new media are "fundamentally social like interactions in real life" (Reeves and Nash 1996). On the other hand, the robot's human-like appearance is known to have impact on the interaction and the participants' social behaviour (e.g. industrial robots are not designed to arouse affection or social effects, so social gaze in industrial robots does not create affective and emotional effects (yet supports floor management and makes the users feel more responsible for the task, Fisher et al. (2013)). It seems obvious that human-like appearance as such does not guarantee agenthood, since this is a complex phenomenon and requires the robot to exhibit human-like behaviour as well, i.e. the robot's appearance needs to conform to the robot's level of social competence. The view of an automated system as an intelligent agent can be related to *affordance*, the concept originally discussed by Gibson (1979), applied to HCI by Norman (1988), to robotic control by Chemero and Turvey (2007) and suggested by Jokinen (2010) to account for the flexible use of natural language dialogue systems: the system's communicative competence affords natural language interaction and lends itself to the intuitive use of the system where the system is communicative agent, not just a tool.

However, in the case of Nao, participants commonly perceive it likable, intelligent and safe, and the gaze fixations onto its face may thus indicate the human partners' initial attraction and benevolence towards the robot and its face in general, rather than "agenthood". To replicate the experiment using a robot with a more human-like, expressive head and compare the results along the robot's perceived agenthood and appearance will be an interesting future study: we may be able to infer how the user's engagement in interaction, as measured by eye-gaze behaviour, is related to the humanoid's appearance and communication skills. This task also has implications to the famous Uncanny Valley hypothesis (Mori 1977), according to which the artefact's increasing human-likeness will, at some point close to the real resemblance, cause the user's acceptance of the artefact suddenly drop. Moore (2015) explains the Uncanny Valley effect on the basis of category boundaries and the uncomfortable feeling that humans experience when typical or normal boundaries are crossed. A humanoid robot may cause uncomfortableness as it is not a typical member of either the classes "human" or "robot", and its accommodation into the existing world requires that a new category is created. The uncomfortable feeling can be overcome by more regular encounters with the untypical object, and thus autonomous and communicating robots can become more acceptable as the audience have more interactions with them, and as their social communication capability increases.

6. Acknowledgements

The main research was carried out when the author was affiliated with University of Helsinki. Thanks go to the students and colleagues for useful discussions and taking part in the experiments, and to Trung Ngo Trong and to Graham Wilcock for their contributions to the work. The work was supported by the Academy of Finland project Digital Citizens. The author would also like to thank the Japanese Nedo project for support in the final editing phase of the paper.

7. Bibliographical References

Argyle, M. and Cook, M. (1976). Gaze and Mutual Gaze. Cambridge University Press, Cambridge

Brennan S., Chen X., Dickinson C., Neider M., and Zelinsky G. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. Cognition, 1465-1477.

Broz, F., Lehmann, H., Mutlu, B., and Nakano, Y. (Eds.) (2015). Gaze in Human-Robot Communication. John Benjamins Publishing Company.

Chemero, A., Turvey, M.T. (2007). Gibsonian Affordances for Roboticists. Adaptive Behavioiur 15(4): 473-480.

Fischer K., Lohan K.S., Nehaniv C., & Lehmann H. (2013). Effects of Different Kinds of Robot Feedback. In: Herrmann G., Pearson M.J., Lenz A., Bremner P., Spiers A., Leonards U. (eds) Social Robotics. ICSR 2013. Lecture Notes in Computer Science, vol 8239. Springer, Cham.

Friesen, C.K. and Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonomic Bulletin & Review. 5(3): 490–495.

Gibson, J.J. (1979). The Ecological Approach to Visual Perception. Houghton Mifflin.

Goodwin, C. (1981). Conversational Organization. Interaction between Speakers and Hearers. New York, London

Gullberg, M. and Holmqvist, K. (2006). What speakers do and what addressees look at. Visual attention to gestures in human interaction live and on video. Pragmatics and Cognition, 14(1), 53-82.

Gullberg, M. and Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. Pragmatics and Cognition, 7(1), 35-63.

Jokinen, K. (2014). Eye-trackers and Multimodal Communication. In: Paggio, P. and Wessel-Tolvig, B. (eds.) Proceedings from the 1st European Symposium on Multimodal Communication, University of Malta, Valletta, October 17–18, 2013. Linköping Electronic Conference Proceedings 101. Linköpings universitet.

Jokinen, K. 2010. Rational communication and affordable natural language interaction for ambient environments. In: Spoken Dialogue Systems for Ambient Environments: Second International Workshop on Spoken Dialogue Systems Technology (IWSDS 2010), pp.163-168, Springer (2010)

Jokinen, K. (2009). Constructive dialogue management – Speech interaction and rational agents. John Wiley & Sons.

Jokinen, K., Furukawa, H, Nishida, M., Yamamoto, S. (2012). Gaze and Turn-taking behaviour in Casual Conversational Interactions. Special Issue on Eye Gaze in Intelligent Human-Machine Interaction, ACM Trans. Interactive Intelligent Systems.

Jokinen, K., Harada, K., Nishida, M., Yamamoto, S. (2010). Turn alignment using eye-gaze and speech in spoken interaction. Proceedings of Interspeech 2010. Makuhari, Japan.

Jokinen, K. and Wilcock, G. (2013). Multimodal Open-domain Conversations with the Nao Robot. In: Mariani, J., Devillers, L., Garnier-Rizet, M., Rosset, S. (eds.) Natural Interaction with Robots, Knowbots, and Smartphones – Putting Spoken Dialog Systems into Practice. Springer Science+Business Media

Kendon, A. (1967). Some functions of gaze direction in social interaction. Acta Psychologica, 26, 22–63.

Land, M.F. (2006). Eye movements and the control of action in everyday life. Progress in Retinal and Eye Research 25, pp. 296-324.

Meltzoff, A.N. and Brooks, R. (2007). Eyes wide shut: The importance of eyes in infant gaze following and understanding other minds. In R. Flom, K. Lee, & D. Muir (Eds.), Gaze following: Its development and significance (pp. 217-241). Mahwah, NJ: Erlbaum.

Moore, R. (2014). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3499759/

Mori M. (1970). The Uncanny Valley. Energy, 7(4), 33–35. English translation by Karl F. MacDorman and Takashi Minato.

Mutlu B., Hodgins J. and Forlizzi J. (2006). A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. Proeedings of the 6th IEEE-RAS International Conference on Humanoid Robots HUMANOIDS'06.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. ACM Transactions on Interactive Intelligent Systems (TiiS), Vol.1, Issue 2.

Norman, D.A. (1988). The Psychology of Everyday Things. Basic Books: New York.

Reeves, B. and Nass, C. (1996). The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places, Cambridge University Press.

Ruhland, K., Peters, C., Andrist, S., Badler, J., Badler, N., Gleicher, M., Mutlu, B., McDonnell, R. (2015). A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception. Comp Graphics Forum 34(6): 299-326.

Rietveld, T. and Van Hout, R. (1993). Statistical Techniques for the Study of Language and Language Behaviour. Berlin: Walter de Gruyter.

Stratou, G. and Morency, LP. (2017). MultiSense—Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case. In: IEEE Transaction on Affective Computing, 8(2).

Wilcock, G. (2012). WikiTalk: A spoken Wikipedia-based open-domain knowledge access system. In Proceedings of the COLING-2012 Workshop on Question Answering in Complex Domains. Mumbai, India, 57-69. 2012.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), 1556-1559.

Yonezawa, T., Yamazoe, H., Utsumi, A., and Abe, S. (2007). Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. Proceedings of the 9th ICMI, pp. 140-145.

Yu, C., Schermerhorn, P., Scheutz, M. (2012). Adaptive eye gaze patterns in interactions with human and artificial agents. ACM Transactions on Interactive Intelligent Systems (TiiS), 1(2), 13.