Just Talking – Casual Talk among Humans and Machines

Workshop Programme

09:00 – 09:10 Introduction

09.10 – 10.30 Paper Session 1

Jens Allwood and Elisabeth Ahlsen, *Small talk and its role in different social activities - a corpus based analysis*

Stefan Olafsson and Timothy Bickmore, *That reminds me...": Towards a Computational Model of Topic Development Within and Across Conversations*

Hanae Koiso, Yayoi Tanaka, Ryoko Watanabe and Yasuharu Den, *A Large-Scale Corpus of Everyday Japanese Conversation: On Methodology for Recording Naturally Occurring Conversations*

Saturnino Luz, Nick Campbell and Fasih Haider, *Data Collection Using a Real Time Feedback Tool for Non Verbal Presentation Skills Training*

10:30 – 11:00 Coffee break

11.00 – 12.00 Paper Session 2

Katri Hiovain and Kristiina Jokinen, *Different Types of Laughter in North Sami Conversational Speech*

Kevin El Haddad, Huseyin Cakmak, Stéphane Dupont and Thierry Dutoit, *Laughter and Smile Processing for Human-Computer Interactions*

Emer Gilmartin, Ketong Su, Yuyun Huang, Kevin El Haddad, Christy Elias, Benjamin R. Cowan and Nick Campbell, *Making Idle Talk: Designing and Implementing Casual Talk*

12.00 – 12.45 Discussion – Getting Natural Talk

12.45 Closing

Editors

Emer Gilmartin	Trinity College Dublin
Nick Campbell	Trinity College Dublin

Organizing Committee

Nick Campbell	Trinity College, Dublin
Emer Gilmartin	Trinity College, Dublin
Laurence Devillers	LIMSI, Paris
Sophie Rosset	LIMSI, Paris
Guillaume Dubuisson Duplessis	LIMSI, Paris

Workshop Programme Committee

Nick Campbell	Trinity College, Dublin
Emer Gilmartin	Trinity College, Dublin
Laurence Devillers	LIMSI, Paris
Sophie Rosset	LIMSI, Paris
Guillaume Dubuisson Duplessis	LIMSI, Paris

Table of Contents

Author Index

Introduction

This workshop focusses on the collection and analysis of resources, novel research, and applications in both human-human and human-machine casual interaction. A major distinction between different types of spoken interaction is whether the goal is ‘transactional’ or ‘interactional’. Transactional, or task-based, talk has short-term goals which are clearly defined and known to the participants – as in service encounters in shops or business meetings. Task-based conversations rely heavily on the transfer of linguistic or lexical information. In technology, most spoken dialogue systems have been task-based for reasons of tractability, concentrating on practical activities such as travel planning. However, in real-life social talk there is often no obvious short term task to be accomplished through speech and the purpose of the interaction is better described as building and maintaining social bonds and transferring attitudinal or affective information – examples of this interactional talk include greetings, gossip, and social chat or small talk. A tenant’s short chat about the weather with the concierge of an apartment block is not intended to transfer important meteorological data but rather to build a relationship which may serve either of the participants in the future. Of course, most transactional encounters are peppered with social or interactional elements as the establishment and maintenance of friendly relationships contributes to task success.

There is increasing interest in modelling interactional talk for applications including social robotics, education, health and companionship. In order to successfully design and implement these applications, there is a need for greater understanding of the mechanics of social talk, particularly its multimodal features. This understanding relies on relevant language resources (corpora, analysis tools), analysis, and experimental technologies.

This workshop provides a focal point for the growing research community on social talk to discuss available resources and ongoing work.

Small talk and its Role in Different Social Activities – A Corpus Based Analysis

Jens Allwood, Elisabeth Ahlsén
SCCIIL Interdisciplinary Center
Department of Applied IT, University of Gothenburg, 41296 Gothenburg, Sweden
E-mail: jens@ling.gu.se, eliza@ling.gu.se

Abstract

This study investigates the occurrence and role of small talk in a number of different social activities, based on video-recorded corpus data from the GSLC (The Gothenburg Spoken Language Corpus) which represents a broad range of different social activities. The study builds on findings from studying communication in different social activity types and compares them with respect to the occurrence, content and role of small talk. The purpose is (i) to describe the characteristics of small talk in general and (ii) to investigate whether and, in that case, in what respects the nature of small talk varies depending on the social activity where it occurs. General types of small talk are found, which can occur in most activity types, for example talk about the weather, the family or the activity at hand. Other types of small talk depend on activity specific factors, such as how formal or informal the activity is, the background and activity roles of the participants, factors in the environment and typical interaction patterns for the activity.

Keywords: small talk, social activity, activity based communication analysis

1. Introduction and Purpose

1.1 Introduction

This study investigates the occurrence and role of small talk in a number of different social activities, based on videorecorded corpus data.

1.2 GSLC – The Gothenburg Spoken Language Corpus

GSLC – The Gothenburg Spoken Language Corpus – is based on audio and video recordings, as a part of several different research projects. The main purpose for the construction of the corpus was to represent a broad range of different social activities and, thus, the corpus contains recordings which are as far apart as sermons, court proceedings, auctions, dinner conversations, patient-doctor consultations and shopping. Totally, the corpus contains about 360 recordings, distributed on 25 different types of activity (see below). The transcriptions are made according to the GTS (Göteborg Transcription Standard, see Nivre 2004) and the orthographic standard MSO6 which, in principle, is standard Swedish orthography modified to represent major features of Swedish spoken language (see Nivre 1999 for details). The corpus is described in more detail in Allwood 1999, 2000 and 2003 and in Allwood and Ahlsén 2012a). For more information about the corpus, see http://www.ling.gu.se/projekt/tal

1.3 Small-talk in Different Social Activities

The study analyzes small-talk in a number of social activities. Earlier studies based on the GSLC corpus have pinpointed the role of small talk in, for example, doctor-patient interaction, talk during a coffee break, dinner talk and talk in shops and markets. The role of including possibilities for small-talk in communication aids, such as picture based VOCAs for non-speaking people, has also been studied e.g. Ahlsén, Allwood and Nivre 2003, Berbyuk 20108, Allwood and Ahlsén 2012b, Ahlsén and Berbyuk Lindström 2013). The study builds on earlier findings from studying communication in different social activity types and compares them with respect to the occurrence, content and role of small talk. The purpose also is to investigate whether the nature of small talk varies depending on the social activity where it occurs.

2. Method – Analysis

2.1 Analyzed Activity Types

We have analyzed both activities where small talk has a dominant role and activities where it has a more ancillary role.

Activity types in the GSLC, primarily focused on goal directed transactions of some kind that were analyzed with respect to small talk, are:

- bus driver/passenger communication,
- consultations,
- formal meetings,
- conversations in shops and markets,
- therapy
- travel agency.

Analyzed activity types in the GSLC, where small-talk in interaction is expected to be dominant, are:

- dinner talk

- factory conversation

- informal conversation

- informal meeting
- interview.

2.2 Activity-based Communication Analysis

As a basis for the study an Activity based analysis of the social activity at hand describes influencing background factors related to the activity type (goals, roles, physical and psychological circumstances), the individual participants (goals, roles, physical, biological, psychological and social circumstances).
It further describes the interaction patterns occurring in the particular activity, such as typical interaction sequences, patterns of turn distribution and feedback, whereby contact, perception, understanding, acceptance and other attitudes are. ACA is, in a simplified way, illustrated by figure 1.

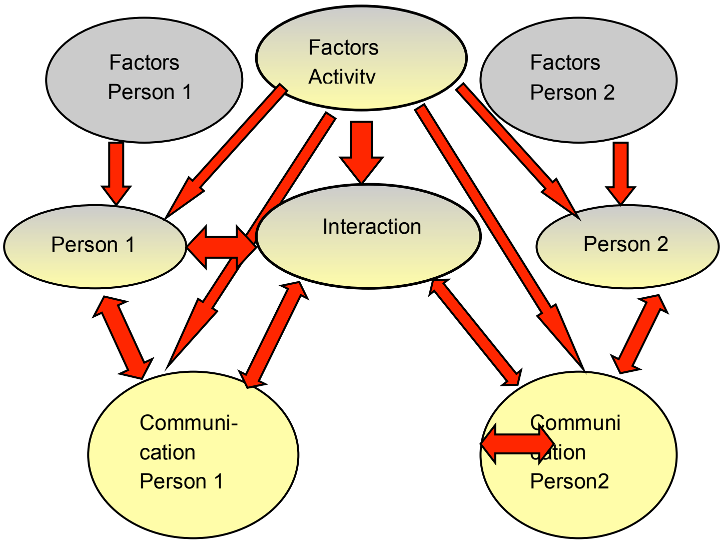


Figure 1: Simplified model of Activity-based Communication Analysis

Figure 1 illustrates how different factors affect each other at a given time in an interaction. Given that interactions are on-going during some time, however, the preceding interaction can dynamically alter the factors, so that feedback loops also occur in the model. For example, if Person 1 reveals something about his/her background during the conversation or interacts in a specific (perhaps unexpected) way, this will feed back into the Factors of Person 1, that will thereafter affect how Person 2 interacts with Person 1. In this study, the focus is on how the different factors affect patterns of interaction related to small talk and how small talk as a form of interaction affects the particular communicative activity and interaction in general.

2.3 Analysis of Small-talk in Particular

The analysis of small-talk in particular, consists of:

1) an estimation of how much small-talk occurs
2) which topics are discussed
3) how small-talk is introduced and by whom
4) how small-talk is terminated and by whom
5) the role of small talk

2.4 Further Analysis in Relation to Activity Types and ICT Applications

We present instances of what we consider to be typical small-talk for different social activity types and discuss them in relation to the Activity based analysis.
We also discuss whether the importance of small-talk varies between different activity types.
Finally, we relate the results to social talk to applications in communication technology (cf. studies by Rydeman, 2010, Ferm et al 2013, Thunberg et al 2011).

3. Results

3.1 General Characteristics – Where Does Small-talk Occur?

Results point to fairly short sequences of small talk in transactional, goal directed speech, which are, for example, inserted (i) quite early in an interaction, to establish rapport and a more informal atmosphere, (ii) at some critical point, where a potential misunderstanding or disagreement is ”in the air in order to release tension or change the topic”, and (iii) when the more transactional conversation comes to a standstill, for some reason, such as waiting for something or a non-speech action going on. Different types of stimuli/distracting events etc. in the environment also elicit small talk. Small talk also seems to be important for keeping a balance between talking too much and being too short, which can be interpreted as impolite. Recognizing the conversation partner at a more personal level seems important as a politeness strategy in many mainly transactional interactions.

3.2 Topics in Small-talk Related to Different Social Activities

Topics in small-talk occurring in mainly transactional activities, are, besides the weather, personal things that the speaker thinks that he/she may have in common with the conversation partner (for example relating to family, pets, hobbies), attempts to make a joke, things and events in the surrounding, mutual acquaintances and topics related to the transactional activity in some way, but not central to the actual transaction.

3.3 Who initiates Small-talk in Different Activities and for What Reason?

It is not obvious who will initiate small-talk in a transactional activity. The reasons for doing so can be related to the activity roles of the participants and sometimes an ambition to ”take the upper hand” or achieve a ”more equal footing”. They can also relate to the degree of tension or nervousness, as in the case of trying to fill silent periods, or they can just be elicited by the environment or by something that is said or done.

3.4 Influence of the Type of Social Activity on the Characteristics of Small talk

To what extent and in what respects small talk in different transactional activity types is similar or different is discussed in relation to typical examples

In interactions where small talk is dominant, the situation is quite different from the more transactional interactions. There are many studies of informal conversations and they are, in this study, mainly addressed in comparisons with small talk in more transactional interactions. One obvious difference is the length of small-talk topics and interaction sequences, other differences relate to patterns of turn taking, feedback and the structures of typical sequences. The Activity-based Communication Analysis also reveals substantial differences in factors related both to the activity itself and to the roles of the participants. Still, some similarities can be noted and some features of small talk that can be generalized over different types of activities.

3.5 Application in Information and Communication Technology

There are many possible applications of findings from this study in designing ICT tools. The task of designing communication aids for communicatively challenged persons has used similar studies made on the same corpus with the purpose of providing means for communication suitable for different social activities, while also providing possibilities for small-talk that can be both general and activity specific.

In designing interfaces to different services, for example using Embodied Communicative Agents, small-talk can be a sensitive issue. On the one hand, it can make the interaction more natural, on the other hand, it can appear very odd and unmotivated. This study points to some aspects that should be considered in ICT design of communicative agents.

4. Conclusions

General features of small talk, such as taking about the weather, about health and family and commenting on things and persons in the environment and the on-going activity can occur in most informal conversations and also occur in fairly formal interaction types indicating that there is a strong social convention to insert small talk and it is expected from most participants.

There is also a strong social convention to respond to small talk.

Activity factors that facilitate small talk are informality and, to some extent, familiarity between the participants. Nevertheless, small talk is also very important, but usually in shorter inserted sequences, when strangers interact and in more formal situations.

There are individual variations in the personality which influence how much small talk occurs.

Triggering factors in the situation can be silence that feels awkward, tension and nervousness, the assumption or detection of similarities between the participants, such as a common background, common interests etc.

The role of small talk for establishing rapport between participants seems very important. It is even so strong that small talk is attempted even when it is hard to accomplish, for example, when there are communication difficulties caused by differences in proficiency in the language used for the interaction or pathological communication disorders.

Small talk can be inserted in the same interaction sequence as the main on-going conversation or occur as “islands” of a slightly different interaction patterns, which are inserted into the interaction, usually seamlessly.

5. Bibliographical References

Ahlsén, E.; Allwood, J. and Nivre, J. (2003). Feedback in Different Social Activities. In P. Juel-Henrichsen (ed.) *Nordic Research on Relations between Utterances. Copenhagen Working Papers in LSP, 3.* 2003, pp.9-37.

Ahlsén, E.; Berbyuk Lindström, N. (2013). Multimodal communication in intercultural health care interactions. *Allwood, J,. Ahlsén, E., Paggio, P. (2013). Proceedings of the Fourth Nordic Symposium on Multimodal Communication, Nov 15-16, University of Gothenburg. NEALT Proceedings Series No. 21.*. Linköping: Linköping University Electronic Press. 39-46.

Allwood, J. (1999). "The Swedish Spoken Language Corpus at Göteborg University". In Proceedings of Fonetik 99, *GPTL 81*, Univ of Göteborg, Dept of Linguistics.

Allwood, J. (2000). An Activity Based Approach to Pragmatics". In Bunt, H., & Black, B. (Eds.) *Abduction, Belief and Context in Dialogue: Studies in Computatio-nal Pragmatics*. Amsterdam, Benjamins, pp. 47-80.

Allwood, J. (2008). Multimodal Corpora. In Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics*. *An International Handbook.* Mouton de Gruyter, Berlin. 207-225.

Allwood, J.; Ahlsén, E. (2012). Incremental collection of activity.based multimodal corpora and their use in activity-based studies, *Proceedings of LREC 2012, May 2012, Istanbul, Turkey*.

Allwood, J.; Ahlsén, E. (2012). On speaker change. *Heegård, J. & Henrichsen P. J. Speech in Action. Proceedings of the 1st SJUSK Conference on comtemporary speech habits, Copehnhagen Studies in Language 42.* Frederiksberg: Sanfundslitteratur Press. 243-261.

Allwood, J.; Björnberg, M. and Grönqvist, L. et al. (2000). The Spoken Language Corpus at the Dept of Linguistics, Göteborg University. *FQA - Forum Qualitative Social Research*. 1 (3).

Berbyuk, N. 2008. Intercultural communication in health care. Non-Swedish physicians in Sweden. Gothenburg Monographs in Linguistics, 36. University of Gothenburg, Department of Linguistics.

Ferm, U.; Ahlsén, E.and Björck-Åkesson, E. (2013). Spontaneous communication with Blissymbolics between a mother and her daughter at home: What do they talk about and how?. *Norén, N., Samuelsson, C. & Plejert, C. (eds): Aided Communication in Everyday Interaction.* Guildford: J&R Press Ltd., pp. 281-313.

Nivre, J. (1999). Modified Standard Orthography, Version 6 (MSO6). University of Gothenburg, Department of Linguistics.

Nivre, J. (2004). Gothenburg Transcription Standard (GTS), V 6.4. University of Gothenburg, Department of Linguistics.

Rydeman, B. (2010). The Growth of Phrases – User Centered Design for Activity Based Voice Output Communication Aids. Gothenburg Monographs in Linguistics 41. University of Gothenburg, Depanrtment of Linguistics.

Thunberg, G.; Ahlsén, E. and Dahlgren Sandberg, A. (2011). Autism, communication and use of a speech-generating device in different environments – a case study. *Journal of Assistive Technologies*. 5 (4) pp. 181-198.

“That reminds me…”: Towards a Computational Model of Topic Development Within and Across Conversations

Stefan Olafsson, Timothy Bickmore
Northeastern University
365 Huntington Ave. Boston MA
E-mail: stefanolafs@ccs.neu.edu, bickmore@ccs.neu.edu

Abstract

We aim to increase user engagement with dialog systems over long periods of time by developing a computational model of longitudinal topic development. Examining a corpus of interactional talk made it clear that shifts and changes in topic are not random and have different manifestations depending on how long the participants have known each other. We developed and evaluated an annotation scheme based on interactional dialog theories that will inform the creation of a computational model of topic development across conversations.

Keywords: Social chat, corpus analysis, annotation scheme, topic development, dialog systems

1. Introduction

In many applications, such as long-term health behavior interventions and educational applications, user engagement and retention spanning multiple interactions over long periods of time is a prerequisite for successful outcomes. When the medium for these applications is dialog, human strategies for establishing and maintaining long-term relationships can be modeled as mechanisms for promoting engagement. Several strategies and techniques for increasing engagement over multiple interactions have been tested in virtual agent dialog systems, including variability in dialog structure, the surface form of utterances, and aspects of the interface appearance (Bickmore et al., 2010; Cafaro et al., 2013), the use of co-constructed storytelling (Battaglino & Bickmore, 2015), and social-emotional relationship-building strategies (Bickmore et al., 2011).

Social chat is another mechanism that people use to establish and maintain relationships, both within purely interactional conversations and in transactional conversations, in which social chat can be used to perform a wide range of additional conversational functions, including greeting and farewells, intimacy management, and trust building. Social chat, including first person storytelling by a computer agent, has been shown to lead directly to increased enjoyment and engagement, measured by number of voluntary conversations with an agent over time (Bickmore et al., 2009).

The models of social chat developed for conversational agent interventions designed to increase long-term engagement either use fully-deterministic models of topic development (Battaglino & Bickmore, 2015; Bickmore & Picard, 2005), or do not incorporate discourse models at all (Cassell & Bickmore, 2003). There are conversational systems created specifically for social chat that build on non-deterministic discourse models designed to decide *what* to say next (Nio et al., 2014; Banchs & Li, 2012). In this work, however, we seek to develop a computational model of longitudinal topic development to give a conversational agent the ability to decide *how* the current topic will shift, continue, or change. The aim is that this will increase both naturalness and flexibility, in order to provide more engaging long-term interactions within and across conversations.

2. Corpus Description

A corpus of multiple conversations between 2 dyads, 15.5 minutes long on average, and spaced a few days apart, was collected. Participants were recruited via fliers hung around Northeastern University and through Craigslist. They were paired up at random and did not know each other prior to the start of the study. They were instructed to chat and get to know each other for roughly 15 minutes, or until a research assistant came into the room. Conversations were video recorded and transcribed. Table 1 summarizes the corpus.

	Conver-sation	Duration (min:sec)	Approx. word count	No. of turns
Dyad 1	1	16:25	2,600	172
	2	15:30	2,700	141
	3	18:00	2,900	87
	4	15:46	2,450	119
	5	17:02	2,500	68
Dyad 2	1	16:28	3,100	136
	2	16:05	3,200	115
	3	16:42	3,300	86
	4	17:05	3,400	109

Table 1: Longitudinal social chat corpus

3. Theories of Interactional Dialog

There are several theories of topic development in the literature that do not appeal to task structure, and are thus appropriate for purely interactional dialog.

Svennevig (1999) defined four major types of topic introductions: *setting*, *encyclopedic*, *self-oriented*, and *other-oriented*. A setting topic introduction is when a participant starts a new topic with material drawn from her physical surroundings. An example from our corpus is that the participants were two strangers sitting in a room and asked to talk about whatever they wanted for 15-20 minutes, the first thing speaker A came up with was how strange their circumstances were. An encyclopedic topic is when a participant makes reference to culturally relevant objects, e.g. A asking B whether she takes the train to work. It is the cultural norm that people get to work somehow and for that particular city the train is a likely choice. A self-oriented introduction, is when the speaker references herself, e.g. A tells B that she took the train to work yesterday, while the other-oriented introduction when one references the other, e.g. A asks B what she did over the weekend.

Gardner (1984), proposed a taxonomy of topic development (Figure 1). His primary insight is that people do not simply shift or change topics, rather they do so in nuanced ways depending on their intentions. For example, shading occurs when a particular topic is being expanded upon and fading when one prepares for a new topic. Both, however, are instances of a topic shift.

Finally, Schneider (1988) described the use of "frames" with facets as a mechanism for representing topic structure. For example, a party frame includes the facets atmosphere, drink, music, participants, and food; thus a discussion where a party is the topic can involve these elements.

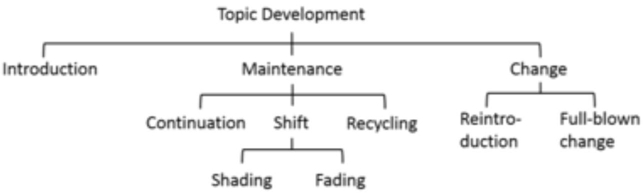


Figure 1: Topic development in spoken interaction (Gardner, 1984, Fig. 2)

Schneider also outlined three options for topic selection by the participants in an interactional conversation. The immediate situation is the most neutral, it is the situation at hand, e.g. the party, the café, etc., and is always the starting point for social chat. The external situation is an extension of the immediate situation, or 'supersituation', and is the least limiting of the three with regards to topic availability. The third is a subset of the immediate situation and involves the participants themselves, wherein topics could be safe ones, such as where they work, or dodgier, like marital status. As a result, the nature of the participants' relationship, how long they've known each other and so forth, impacts which type of situation should the next topic be drawn from (Schneider, 1988).

4. Corpus Analysis

We developed a turn-by-turn annotation scheme to capture several of the above notions of topic development. In particular, we extended Gardner's taxonomy of topic development with a new development type that reflects topic reintroduction from a prior conversation, which we feel is particularly important for annotating multiple conversations. Each tag represents a possible action that the agent can take to either introduce, maintain, or change a topic at a given turn of dialog.

		Concept	Definition	Agent Intent	Tag
		Introduction	A new topic is introduced	Initiate a topic	<intro>
Maintenence		**Continuation**	The topic continues or is being extended	Continue current topic	<cont>
	Shift	**Shading**	A shift that introduces a new aspect of the topical unit	Start a sub-topic	<shade>
		Fading	A shift away from the current topic	Abandon topic	<fade>
		Recycling	Returning to an earlier point in the same topical unit	Go back to the topic at hand	<recycle>
Change		**Reintroduction**	Moving to a new topical unit based on a topic that has been abandoned	Start a topic based on previous discourse in current conversation	<reintro>
		Reminding	Moving to a new topical unit based on a topic from prior conversation	Start a topic based on discourse in prior conversation	<remind>
		Full change	Moving to a new topical unit	Abandon topic	<fullchange>

Table 2: Annotation scheme concepts, definitions, possible underlying agent intents, and tags.

We evaluated our annotation scheme by having two annotators tag 70 turns of dialog in the social chat corpus. We found strong inter-rater reliability, with Kappa=0.76. Annotating social chat corpora in this manner will allow us to train models for automated tagging, which will ultimately lead to training software that automatically outputs the appropriate tag, or action, at any given turn of dialog.

4.1. Example

A brief example of topic development from the corpus is shown in Table 3, from the 1st and 2nd conversations of dyad 1 in the study, illustrating several examples of topic development tags.

Turn	Tag
1st Conversation	
...	
P2: So you‘re from Germany originally?	<fullchange> <intro>
P1: Mhmm	<cont>
...	
P1: Okay, and you‘re from the U.S.?	<reintro>
P2: Yeah, I‘m from New Hampshire originally [...]	<cont>
...	
P2: Have you been anywhere besides Boston since you've been here?	<intro>
P1: No I'm going to New York next weekend […]	<cont>
P2: That sounds pretty fun (laughs) [...]	<cont>
P1: Oh! Do you know in New York do they have [...] the train ticket [...] do they have three day passes?	<shade>
P2: Umm (laughs) I mean i'm sure they do [...]	<cont>
P2: I haven't been to New York in a while but I mean every time I go its fun	<recycle>
P2: but the subway system is like (short pause) a lot worse than Boston [...]	<fade>
P2: You take the T into work right?	<fullchange> <intro>
...	
2nd Conversation	
...	
P1: Hm, (short pause) yeah I've tried Back Bay train	<remind>
...	

Table 3: Example topic development from the corpus

5. Conclusion and Future Work

We are developing a computational model that can emulate natural topic development, as described in the three theories above. In our model, Schneider’s communication situations and topic frames (e.g. FrameNet cf. Baker et al., 1998) set the scene, our extended version of Gardner’s model provides the actions that the agent can take, and Svennevig’s topic types govern what form that action will take. We plan to evaluate this model in a within-subjects experiment comparing the model to ablated versions with length of conversation as a behavioral measure of engagement (as in Battaglino, 2015).

6. Main References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86-90). Association for Computational Linguistics.

Banchs, R. E., & Li, H. (2012, July). IRIS: a chat-oriented dialogue system based on the vector space model. In Proceedings of the ACL 2012 System Demonstrations (pp. 37-42). Association for Computational Linguistics.

Battaglino, C., & Bickmore, T. (2015). Increasing the engagement of conversational agents through Co-Constructed storytelling. URL http://www.aaai.org/ocs/index.php/AIIDE/AIIDE15/paper/view/11581

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *12*(2), 293-327.

Bickmore, T., Schulman, D., & Yin, L. (2009). Engagement vs. deceit: Virtual humans with human autobiographies. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. Vilhjálmsson (Eds.) *Intelligent Virtual Agents*, vol. 5773 of *Lecture Notes in Computer Science*, (pp. 6-19). Springer Berlin Heidelberg. URL http://dx.doi.org/10.1007/978-3-642-04380-2_4

Bickmore, T., Schulman, D., & Yin, L. (2010). Maintaining engagement in long-term interventions with relational agents. *Applied artificial intelligence: AAI*, 24 (6), 648-666. URL http://dx.doi.org/10.1080/08839514.2010.492259

Bickmore, T., Pfeifer, L., & Schulman, D. (2011). Relational agents improve engagement and learning in science museum visitors. In H. Vilhjálmsson, S. Kopp, S. Marsella, & K. Thórisson (Eds.) *Intelligent Virtual Agents*, vol. 6895 of *Lecture Notes in Computer Science*, (pp. 55-67). Springer Berlin Heidelberg. URL http://dx.doi.org/10.1007/978-3-642-23974-8_7

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T. W., Heylen, D., & Schulman, D. (2013). First impressions in user-agent encounters: The impact of an agent's nonverbal behavior on users' relational decisions. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, (pp. 1201-1202). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. URL http://portal.acm.org/citation.cfm?id=2485142

Cassell, J., & Bickmore, T. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13 (1-2), 89-132. URL http://dx.doi.org/10.1023/a:1024026532471

Gardner, R. (1984). Discourse analysis: implications for language teaching, with particular reference to casual conversation. *Language Teaching*, *17* , 102-117. URL http://dx.doi.org/10.1017/s0261444800010545

Nio, L., Sakti, S., Neubig, G., Toda, T., Adriani, M., & Nakamura, S. (2014). Developing non-goal dialog system based on examples of drama television. In Natural Interaction with Robots, Knowbots and Smartphones (pp. 355-361). Springer New York.

Schneider, K. P. (1988). *Small talk: Analyzing phatic discourse* (Vol. 1). Hitzeroth.

Svennevig, Jan. (1999). *Getting acquainted in conversation: a study of initial interactions*. Vol. 64. John Benjamins Publishing.

A Large-Scale Corpus of Everyday Japanese Conversation: On Methodology for Recording Naturally Occurring Conversations

Hanae Koiso† Yayoi Tanaka† Ryoko Watanabe† Yasuharu Den‡,†

† National Institute for Japanese Language and Linguistics, Japan
‡ Faculty of Letters, Chiba University, Japan

Abstract

In 2016, we set about building a large-scale corpus of everyday Japanese conversation—a collection of conversations embedded in naturally occurring activities in daily life. We will collect more than 200 hours of recordings over six years, publishing the corpus in 2022. Before building such a corpus, we have conducted a pilot project, whose purposes are i) to establish a corpus design for collecting various kinds of everyday conversations in a balanced manner, ii) to develop a methodology for recording naturally occurring conversations, and iii) to create a transcription system suitable for precisely and efficiently transcribing natural conversations. This paper focuses on the second issue. We first describe two types of methods for recording various kinds of conversations embedded in naturally occurring activities in everyday situations. Next, we show recording devices we use and sample data. Finally, we discuss ethical and other issues, including the collection of consent forms and questionnaires.

Keywords: corpus of everyday conversation, naturally occurring activities, recording methods, ethical issues

1. Introduction

Since the 1990s, many corpora of Japanese conversations have been constructed. Table 1 lists major corpora of Japanese conversations. Most corpora are biased towards conversations among friends or families on the telephone or in artificially created settings in terms of topics and recording situations. However, in order to capture the diversity of everyday conversations and to observe the endogenous organization of social activities in their ordinary settings, we must record conversations embedded in activities that naturally occur in daily life, without the exogenous intervention of researchers imposing topics and tasks or displacing the context of action (Mondada, 2012). In addition, the existing corpora do not provide video data. Our social activities are organized not only by verbal utterances but also by bodily resources such as gaze and gesture (Streeck et al., 2014). Thus, video data is essential for understanding the mechanism of our real-life social conduct.

In 2016, we set about building a large-scale corpus of everyday Japanese conversation—a collection of conversations embedded in naturally occurring activities in daily life. Before building such a corpus, we have conducted a pilot project, whose purposes are i) to establish a corpus design for collecting various kinds of everyday conversations in a balanced manner, ii) to develop a methodology for recording naturally occurring conversations, and iii) to create a transcription system suitable for precisely and efficiently transcribing natural conversations.

As for the first issue, we conducted a survey of everyday conversational behavior, with about 250 Japanese adults, in order to reveal how diverse our everyday conversational behavior is and to build an empirical foundation for corpus design. Koiso et al. (2016) reported an overview of this survey study, and discussed how to design a large-scale corpus of everyday Japanese conversation on this basis.

In this paper, we focus on the second issue, discussing how to record various kinds of conversations embedded in naturally occurring activities in everyday situations. We first describe two types of methods for recording such conversations. Next, we show recording devices we use and sample data. Finally, we discuss ethical and other issues, including the collection of consent forms and questionnaires.

2. Two types of recording methods

In order to capture the diversity of everyday conversations and to observe the endogenous organization of social activities in their ordinary settings, we must record conversations embedded in naturally occurring activities in daily situations, without imposing topics and tasks or displacing the context of action (Mondada, 2012).

The British National Corpus (BNC), constructed in the former half of the 1990s in the U.K., provides a methodology for such purpose. The BNC is comprised of one-hundred million British English words (Crowdy, 1995; Burnard and Aston, 1998). While the majority of it contains written language, approximately 10% of the words (ten million) are from spoken language. This spoken language part of the BNC is composed of the following two data groups:

Spoken demographic: Recorded with a portable tape recorder over the course of seven days by 124 informants who were chosen so as to avoid bias in terms of age, sex, social class, and region.

Spoken context-governed: Spoken language that many people listen to (e.g., broadcasts and lectures). Divided into four categories: educational, business, public/organizational, and leisure.

The first of the above BNC methods is suited for recording conversations embedded in naturally occurring activities. With their approach in mind, we decided to record everyday conversations using the following two methods.

Individual-based method: We choose a set of informants balanced in terms of sex, age, etc., provide them

Table 1: Major existing corpora of Japanese conversations

Corpus Name	Size	Contents
Multilingual Corpus of Spoken Language by Basic Transcription System (BTS)	294 conversations 66 hours	chats among friends, professor-student mentoring, telephone conversations, etc. (audio files available only for some portion)
Sakura Corpus	18 conversations	chats among four undergraduate students (topics assigned)
Chiba Three-Party Conversation Corpus	12 conversations 2 hours	chats among three undergraduate/graduate students on campus (initial topics assigned)
CALL HOME Japanese	120 conversations 20 hours	telephone conversations between Japanese living in the U.S. and their families/friends in Japan
CallFriend Japanese	31 conversations	telephone conversations between Japanese living in the U.S.
Meidai Conversational Corpus	161 speakers 100 hours	chats among friends (audio files unavailable)
Women's Language at the Workplace Men's Language at the Workplace	21 speakers each	natural conversations in formal and informal situations at the workplace (audio files unavailable)

portable recording devices for approximately one to two months, and have them record conversations in their daily activities.[1] In principle, the project members do not mediate their field recordings.

Situation-specific method: We select specific situations in which recording based on the individual-based method is technically and/or ethically difficult, e.g., exchanges with store employees, meetings at workplaces, regional activities, public events, etc., and record conversations occurring in these situations. Although the project members coordinate recording settings, only conversations in these naturally occurring activities are recorded.

In what follows, we focus on how to record conversations based on the individual-based method.

3. Recording devices

Non-verbal behaviours such as gazes and gestures play a significant role in face-to-face conversations. In daily situations, we often have conversations while conducting some activities such as eating and cooking. The perspective of such *multi-modalities* and *multi-activities* has been increasingly focused on in conversation studies (Streeck et al., 2014; Haddington et al., 2014). Video recordings are essential for analyzing such conversations embedded in daily activities. Since in the individual-based method, informants themselves carry portable recording devices and record their everyday activities in a variety of situations such as at home, at a restaurant, and outdoors, it is preferable to use small, light, and easy-to-operate recording devices. Taking these conditions into consideration, we are planning to record conversations using the following recording devices.

3.1. Video recording

The following two types of compact action cameras are used when recording indoors:

- Kodak PIXPRO SP360 4K[2] (weight: 102g, setting: 2880 × 2880, 30fps): SP360 is used for recording an inside-out image of the conversation field. It can shoot a 360-degree global view.

- GoPro Hero3+[3] (weight: 74g, setting: 1920 × 1080, 60fps): One or two GoPro cameras are optionally used for recording a normal flat image.

Figure 1 shows video images of a recording of a four-party conversation at a Japanese-style restaurant. Two men and two women sit across from one another at a table. An SP360 on a 5cm-high-stand was located in the center of the table.[4] The 360-degree global view (the left image of the figure), recorded by the SP360, includes all four conversants. Two GoPro cameras on tabletop tripods were placed diagonally to each other on the corner of the table. The top- and bottom-right images, recorded by the GoPro cameras, enable us to observe intuitively the positional relation of the conversants.

By using software provided by KODAK, a 360-degree global view, captured by SP360, can be converted into several modes, such as a 360-degree panoramic expanded image, a nearly-flat image of an extracted portion, and an image divided by two or four extracted portions (See Figure 2). We are currently developing a software with similar functions, which will be publicly available to users of our corpus in the future.

Cameras for recording outdoors are under consideration.

[1] One may complain that the mere recording of an activity disrupts the "naturalness" of the data. Due to the presence of recording devices, conversants may refer to them and bring them up as a topic of an ongoing discourse. However, we do not think it causes a big problem as long as the "naturalness" of an ongoing *activity* is preserved. "Unnatural" digression at the linguistic level does not necessarily lead to "unnatural" ways of organizing a social activity.

[2] `http://kodakpixpro.com/Americas/cameras/actioncam/sp3604k/`

[3] `http://gopro.com/`

[4] In this recording, a previous version of SP360 was used.

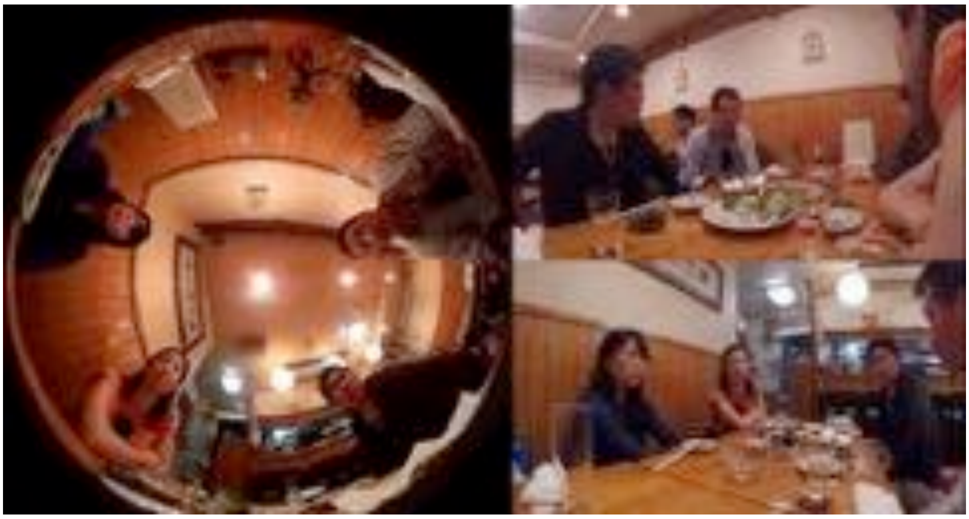


Figure 1: Video images of a four-party conversation at a Japanese-style restaurant. Left image: recorded by SP360 located in the center of the table. Top- and bottom-right images: recorded by two GoPro cameras placed diagonally to each other on the corner of the table.

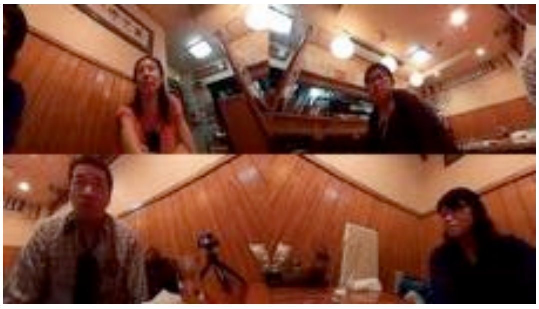
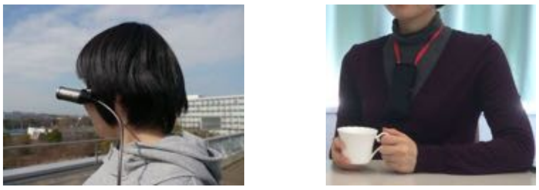


Figure 2: Four-division image of the left image in Figure1



Wearable camera

IC recorder

Figure 3: Images of wearing a camera and an IC recorder

A wearable camera, Panasonic HX-A500[5] (weight: 31 + 128g, setting: 1920 × 1080, 60fps) is one of the candidates. In order to grasp a conversational situation, a wearable camera is fixed on the conversant's head or on the shoulder (See the left image in Figure 3), capturing what he/she is watching from his/her viewpoint.

Figure 4 shows a video image of recording a conversation while participating in a festival.[6] A woman is carrying a wearable camera fixed on the shoulder strap of her bag, and is recording the festival and a conversation with her companions.

Figure 4: Video image recorded at a festival by a wearable camera, which was fixed on the shoulder strap of a woman companion's bag.

3.2. Audio recording

When the number of conversants are 6 or fewer, all conversants wear IC recorders (Sony ICD-SX734[7], weight: 81g, setting: linear PCM, 44.1kHz/16bit). A conversant inserts a recorder into a holder around his/her neck and adjust the holder so as to locate the microphone unit about 15cm below his/her mouth (See the right image in Figure 3). When more than 6 conversants attend, all conversants' voices are recorded by one or two IC recorders[8].

Figure 5 shows a speech sample of the four-party restaurant conversation shown in Figure 1. The figure contains a speech waveform, a spectrum, and a pitch counter of speaker A, and transcriptions of all four speakers. Speaker A's waveform shows that although A's IC recorder picks up a bit of other conversants' voices, it catches speaker A's own voice more clearly.

[5] `http://www.panasonic.com/ca/consumer/cameras-camcorders/active-style-camcorders/hx-a500.html`

[6] In this recording, a variant of HX-A500, i.e., HX-A1H, was used.

[7] `http://www.sony-asia.com/electronics/voice-recorders/icd-sx734/`

[8] According to the results of the survey on conversational behavior (Koiso et al., 2016), large-party conversations are very few (more-than-six-party conversations account for only 10% of the total), and most are small-party conversations (two- or three-party conversations, 75%).

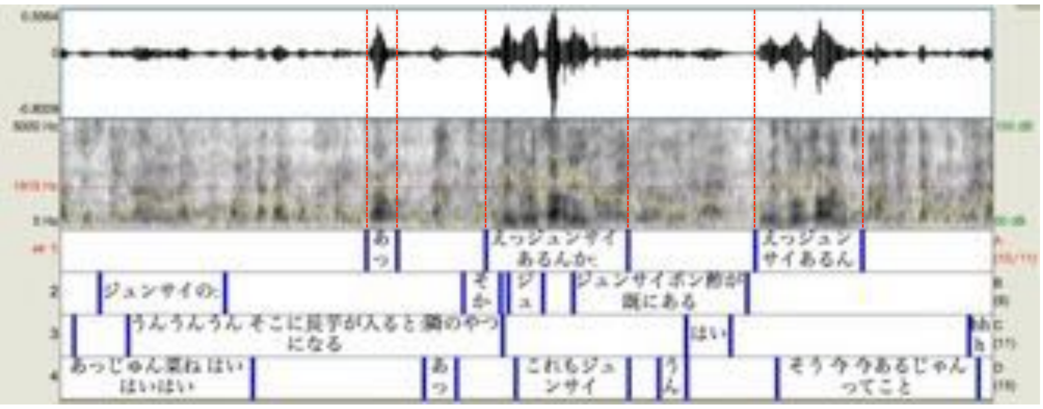


Figure 5: Speech sample recorded at a restaurant. A speech waveform, a spectrum, and a pitch counter of speaker A, and transcriptions of all four speakers.

3.3. Synchronization of video and audio data

In order to synchronize multiple video and audio sources after recording, conversants clap hands at the beginning (and possibly at the ending) of a conversation as a clue for the data synchronization. We synchronized video and audio sources of the four-party restaurant conversation, finding that the maximum gap among four IC recorders was as low as 30 milliseconds in two-hour recording, while gaps between IC recorders and GoPro cameras become 200 milliseconds in maximum.[9] These results suggest that we may want to synchronize video and audio sources not only at the beginning but also in the middle of the conversation when it lasts very long.

4. Ethical and other issues

4.1. Consent form and questionnaire

In recording based on the individual-based method, we ask informants to record conversations embedded in naturally occurring activities in their daily situations, and the project members do not mediate their field recordings. Therefore, informants have to i) explain the purpose of the recording to other conversants, ii) obtain their consent to publish the recorded conversations, iii) have them fill in face sheets including their birthday, residence, birthplace, sex, occupation, and relationship to the informant, and iv) write recording date and time, overview of conversations and activities, and layout of conversants and recording devices.

4.2. Ethical issue

Recording everyday conversations and publishing them require careful consideration from an ethical perspective. Since informants themselves record their conversations in everyday situations, they need to judge, for instance, whether the place where they converse is permitted for recording. We prepare flyers explaining the aim of recording everyday conversations and how to publish them, which would be helpful when informants need to get permission.

Personal information including conversant names as well as parts of recordings which conversants have not given permission to be made public will be replaced by anonyms or turned letters in the transcripts, and the corresponding regions of the audio files will be made inaudible. When the video data contains faces of third parties, those parts will be modified by means of image effects, except in case where face images are enough small or indistinct to identify.

5. Concluding Remarks

In this paper, we reported a methodology for recording various kinds of conversations embedded in naturally occurring activities in everyday situations. We will collect more than 200 hours of recordings over six years, publishing the corpus in 2022. We believe that our corpus will contribute not only to basic research on conversational interaction but also to improving applications such as social robotics.

6. Acknowledgements

This work is supported by Grant-in-Aid for Collaborative Research Project of NINJAL "A multifaceted study of spoken language using a large-scale corpus of everyday Japanese conversation" led by Hanae Koiso.

7. References

Burnard, L. and Aston, G. (1998). *The BNC handbook*. Edinburgh University Press, Edinburgh, U.K.

Crowdy, S. (1995). The BNC spoken corpus. In Leech, G., Myers, G., and Thomas, J., editors, *Spoken English on computer: Transcription, mark-up and application*, pages 224–235. Longman, Harlow, U.K.

Haddington, P., Keisanen, T., Mondada, L., and Nevile, M., editors. (2014). *Multiactivity in social interaction: Beyond multitasking*. John Benjamins Publishing Company, Amsterdam.

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

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

Streeck, J., Goodwin, C., and LeBaron, C., editors. (2014). *Embodied interaction: Language and body in the material world*. Cambridge University Press, New York.

[9] The newest SP360 4K camera gives a similar value as GoPro.

METALOGUE: Data Collection Using a Real Time Feedback Tool for Non Verbal Presentation Skills Training

***Fasih Haider*[†], *Saturnino Luz*[‡], *Nick Campbell*[†],**

[†] School of Computer Science and Statistics, Trinity College Dublin, Ireland
[‡]Usher Institute of Population Health Sciences & Informatics, University of Edinburgh, UK
{haiderf,nick}@tcd.ie[†], S.Luz@ed.ac.uk[‡]

Abstract

This paper describes the data recording setting and systems used in the first pilot data collection activity for the METALOGUE project. The objective of this activity is to provide an opportunity for participants to use a 'real time presentation trainer' feedback tool during debate. In total 2 sessions have been recorded and the data is made available to all participants of the project.

Keywords: Multimodal Corpora Collection, Presentation Quality, Educational System

1. Introduction

The goal of the METALOGUE[1] project is to develop a Multiperspective Multimodal Dialogue System with Metacognitive Abilities (METALOGUE) which is able to provide instructional advice to users (in action and about action feedback) (Alexandersson et al., 2014). This feedback can be on the speech content, prosodic characteristics and body language of the user. Several manuals on public speaking highlight the stated characteristics for good presentation skills (Stassen and others, 1993; Lamerton, 2001; Grandstaff, 2004; DeCoske and White, 2010). In METALOGUE, a metacognitive skill is defined as the ability of a participant in an interaction to understand, control and modify his own cognitive process. Such skills are believed to be useful in real life learning and training processes, in particular for debating skills (Tumposky, 2004).

The system overview is shown in Figure 1. As can be seen in the diagram, the the system requires dialogue or multiparty interaction data for training of the input modules. Although there are multiple datasets of multiparty interaction available (e.g. IFA, MIMLA and AMI meeting corpora (van Son et al., 2008; McCowan et al., 2005; Ochoa et al., 2014)), they do not exactly fit into the METALOGUE project scenario, in that they either do not target learning contexts, or lack instructional and real time feedback elements. This paper describes the process of recording a suitable corpus for the METALOGUE project.

2. Recording Settings and Environment

The METALOGUE project consists of two scenarios, targeting respectively: training students for debating and presentation skills, and training call centre agents. The data collection activity reported in this paper focuses on the first scenario, in which two students are standing in front of an audience and debate over a social issue (a proposed ban on smoking), with one in favour of the ban and the other against it.

In a pre-pilot data collection activity, we recorded around 2 hours of audio-visual data along with skeleton information (tracked by Kinect) simulating the Hellenic youth parliament scenario. In total 11 pre-pilot sessions have been recorded with each session lasting between 10–15 minutes. Although most of the settings remain the same as in the pre-pilot data collection activity (Haider et al., 2016) — the same quiet room, students, situation (there aren't any windows behind participants), microphones, Kinect and video cameras— in the pilot 1 recordings a real time body language feedback tool is presented to the students (users). The tool gives feedback to the student about their body posture (e.g. standing straight, crossing their legs, hand movements etc). Refer to Figure 2 for a systematic representation of the equipment and recording set-up.

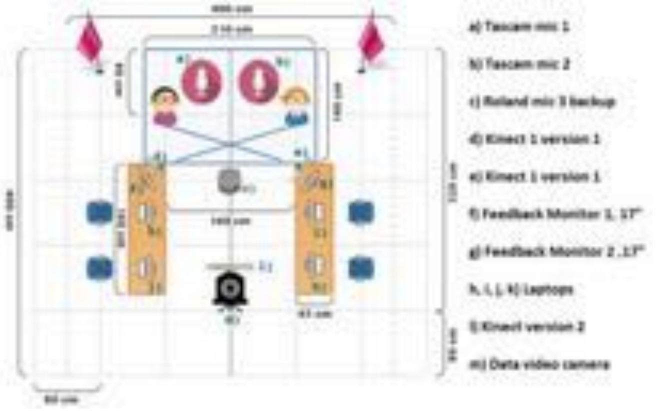


Figure 2: A systematic representation of recording settings.

3. Presentation Trainer Tool

This tool is developed using Kinect SDK[2]. It helps users to improve their non verbal behaviour (body language) for presentations and provides them a real time feedback as shown in Figure 4 (Schneider et al., 2014). The tool provides feedback to the user through a 17-inch screen (item *f* and *g* as shown in Figure 2). We have also time stamped and saved feedback in EAF files (ELAN (Wittenburg et al., 2006) annotation format).

[1] http://www.metalogue.eu/

[2] https://dev.windows.com/en-us/kinect

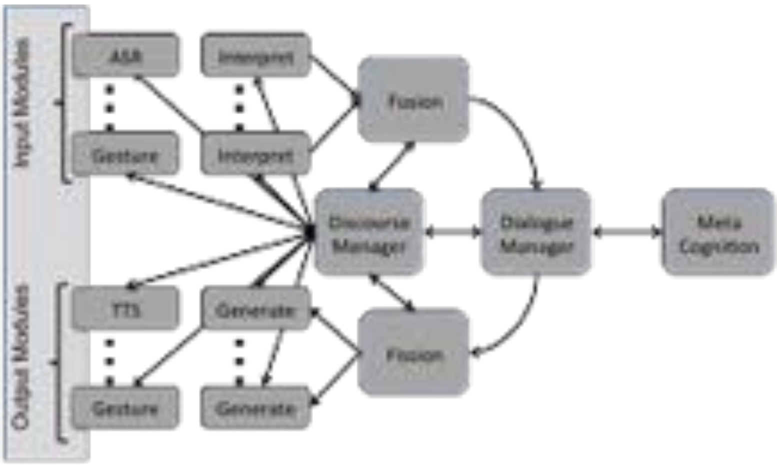


Figure 1: System Overview (Alexandersson et al., 2014)

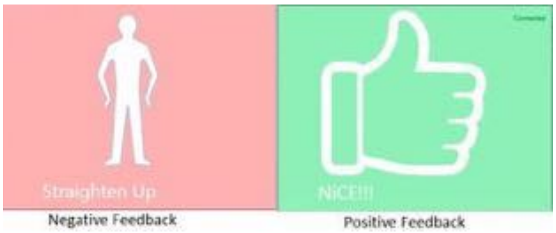


Figure 3: Feedback of presentation trainer in the form of red (body posture) and green lights.

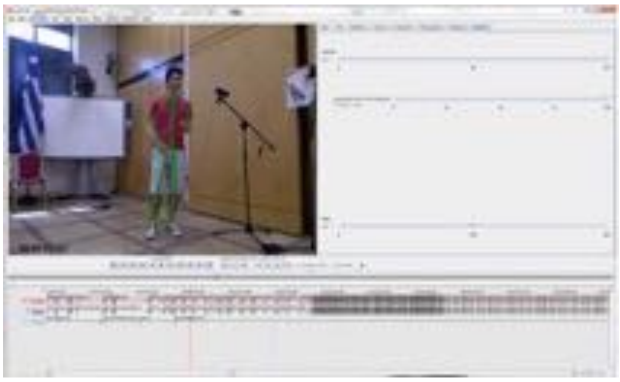


Figure 4: A snapshot showing the in-action feedback on an ELAN time line.

4. Collected Data

In total two sessions have been recorded and each one lasts around 15 minutes. The same students as the pre-pilot recordings (Haider et al., 2016) take part in this data collection activity and the characteristics of the participants are as follow:

- Young school students aged between 17 to 20 years.
- Previously presented in the annual Hellenic youth parliament sessions.
- Greek speakers of English who know each other.

The audio-visual data along with full skeleton information (tracked by Kinects) is recorded using the equipment described in Section 2.. The loudness level of the participants' speech is also detected in real time using openSMILE (Eyben et al., 2013) and time stamped for further analysis. However, this information was not available to the students as feedback.

5. Conclusion and Future Work

The collected corpora (pilot 1 data collection) using the settings and procedures described in this paper has been made available to other project members using our private cloud server. The data will be evaluated against the pre-pilot data (Haider et al., 2016) to see if there are any improvements in the body language of the students. We wish to assess whether the tool helps improve the user's body language, and whether the users are comfortable with the form and frequency of the feedback. This will help us in improving the feedback tool. For future data collections, new sensor technologies will be tested, including Myo gesture tracking armband and Intel RealSense depth camera (which are also able to track finger movements). An introduction of real time feedback tool for prosody will also take place. For the call centre scenario of the METALOGUE project, which is going to be addressed in the next phase of the project, call centre companies have some concerns about the privacy of their customers and have not yet agreed to provide us real world data. Therefore a data collection activity which simulates the customer service scenario is needed. Moreover we are also exploring options to use other data sets which fit into the context of call centre scenario like map-task data (where one person has full information and the other has less information).

6. Acknowledgements

This research is supported by EU FP7-METALOGUE project under Grant No. 611073 in Speech Communication Lab (SCL) at School of Computer Science and Statistics, Trinity College Dublin.

7. Bibliographical References

Alexandersson, J., Girenko, A., Spiliotopoulos, D., Petukhova, V., Klakow, D., Koryzis, D., Taatgen, N., Specht, M. M., Campbell, N., Aretoulaki, M., et al. (2014). Metalogue: A multiperspective multimodal dialogue system with metacognitive abilities for highly adaptive and flexible dialogue management. In *Intelligent Environments (IE), 2014 International Conference on*, pages 365–368. IEEE.

DeCoske, M. A. and White, S. J. (2010). Public speaking revisited: delivery, structure, and style. *American journal of health-system pharmacy*, 67(15):1225–1227, August.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM.

Grandstaff, D. (2004). *Speaking as a Professional: Enhance Your Therapy Or Coaching Practice Through Presentations, Workshops, and Seminars*. A Norton Professional Book. W.W. Norton & Company.

Haider, F., Luz, S., and Campbell, N. (2016). Data collection and synchronisation: Towards a multiperspective multimodal dialogue system with metacognitive abilities. In *Proceding of International Workshop on Spoken Dialogue Systems, IWSDS*.

Lamerton, J. (2001). *Public Speaking. Everything you need to know*. Harpercollins Publishers Ltd.

McCowan, I., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005). The AMI meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*.

Ochoa, X., Worsley, M., Chiluiza, K., and Luz, S. (2014). MLA'14: Third multimodal learning analytics workshop and grand challenges. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 531–532, New York, NY, USA. ACM.

Schneider, J., Börner, D., van Rosmalen, P., and Specht, M. (2014). Presentation trainer: a study on immediate feedback for developing non-verbal public speaking skills. *Bulletin of the IEEE Technical Committee on Learning Technology*, 16(2/3):6.

Stassen, H. et al. (1993). Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of Psychiatric Research*, 27(3):289–307.

Tumposky, N. R. (2004). The debate debate. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 78(2):52–56.

van Son, R., Wesseling, W., Sanders, E., and van den Heuvel, H. (2008). The IFADV corpus: a free dialog video corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, page 5th.

Acoustic Features of Different Types of Laughter in North Sami Conversational Speech

Katri Hiovain, Kristiina Jokinen
University of Helsinki
Siltavuorenpenger 1 A
E-mail: firstname.lastname@helsinki.fi

Abstract

This paper describes how various laughter types differ from each other acoustically in a North Sami conversational speech corpus collected and annotated within the DigiSami project. The laughter annotations were done with Praat and included two tag types, the first of which indicated if the laugh was a free laugh or speech-talk (laughing speech), and the second one indicating more specific laughter type. In our study, pitch, duration and intensity were extracted for laughter bouts representing every laughter type, and the paper describes the first analysis of the data. The conversational speech of the Sami languages has not yet been systematically studied, so our analysis can be compared with the results of laughter studies conducted in other languages, while also contributing to empirical observations of the North Sami language.

Keywords: laughter, speech-talk, conversation, North Sami

1. Introduction

In this paper we describe how speech interactions are organized with respect to laughter in North Sami dialogues, and try to pin down the function of various types of laughter in dialogues. Our corpus is the North Sami dialogue corpus collected within the DigiSami project (Jokinen et al. 2016), and the analysis is the first study of its kind with this language. The ultimate goal of the study is to examine the use of laughs in interaction management, and to provide a starting point for conversational studies on North Sami in general. The analysis can be compared with the analyses conducted in other languages. The model can also be used for enabling natural interactions with situated robot agents which are sensitive to affective signals of the user.

Laughter is usually related to joking and humour, but it also occurs in connection to socially critical situations such as signalling relief of embarrassment. We assume that laughing is, first of all, a means of creating common understanding and rapport among the participants, i.e. an effective feedback signal that the participants use to show benevolent contact and their willingness to continue the interaction. However, we also hypothesize that laughing used as an interaction strategy to distance oneself from the partner and from the discussed topics, i.e. it is an acceptable way to disassociate oneself from the conversation.

The paper is structured as follows. We first give a brief overview of the previous research on laughter in interaction in Section 2. We then describe the data in Section 3, and present the analysis is presented in Section 4. Conclusions and future work are discussed in Section 5.

2. Previous Research

Laughter is commonly related to joking and amusement, and it has been studied in humour studies (Chafe, 2007). However, laughter does not occur in humorous contexts only, but in potentially face-threatening situations where it is a sign of politeness and socially acceptable behaviour. In sociolinguistic research, laughter is regarded as a typical social phenomenon, described as serving a broad range of interactional functions. Goffman (1974) talks about situated interactions and the bursts of laughter that break the ordinary interactional frames. In a seminal work on the organization of laughter in talk with the Conversation Analysis, Jefferson (1984) focussed on interactional consequences of laughter, and pointed out that the partner may choose to be silent in which case laughing and silence are two systematic possibilities for joke completions.

Speech research has focussed on the acoustic analysis of laughter and on the categorization of various forms of laughter for the purposes of emotion recognition or speech synthesizer (Trouvain, 2003; Truong and van Leeuwen, 2007; Owren, 2007; Bachorowski et al., 2001; Tanaka and Campbell, 2011). Acoustic properties of laughter vary a lot between speakers and within a speaker, but it is generally concluded that F0 formant is much higher in laughter than in speech, and that the ratio of the length of unvoiced to voiced parts is greater for laughter than for speech (Bachorowski et al., 2001; Truong and van Leeuwen, 2007).

Classifications of laughter often distinguish between free laughter and speech-laugh, i.e. laughter which is synchronous with speech. Nwokah et al. (1999) found that up to 50% of laughs overlap with speech in their corpus of mother and child communication, and also that the duration of speech-laugh was significantly longer than that of only laughter (1.24s vs. 1.07s). Tanaka and Campbell (2011) used a four-way classification, with the most common distinction between the spontaneous mirthful laugh and polite laugh, which together apparently account for 80% of the laughs.

In recent studies on social and situational signals and their correlation with the interactional context (Bonin et al., 2014; Bonin, 2016) it is shown that when laughter functions as a social signal, its timing is structured and conveys information about the underlying discourse structure. Higher amounts of laughter occur in topic transition moments than in topic continuation moments and when the temporal distance from the topic boundary increases, laughter becomes more likely to occur. Gilmartin et al. (2013) studied laughter and engagement and noted that a significant change in the amount of laughter occurs at fifteen seconds around the topic changes.

3. DigiSami Corpus and its Annotations

The DigiSami Corpus of spoken North Sami has been collected in the areas traditionally inhabited by the Sami people: in Enontekiö, Utsjoki, Inari and Ivalo in Finland, and in Kautokeino and Karasjok in Norway (see the map in Figure 1). North Sami belongs to the Fenno-Ugric language family, and is one of the Sami languages spoken in Northern Scandinavia, Finland and Kola Peninsula (Seurujärvi et al., 1997). The corpus includes read and conversational speech, and the conversations are both recorded and videotaped. All the speakers are native speakers of North Sami, and their age vary between 16 and 65 years. The data are thus versatile, including informants from two different countries and of different ages. See more about the data collection in Jokinen (2014), Jokinen and Wilcock (2014).

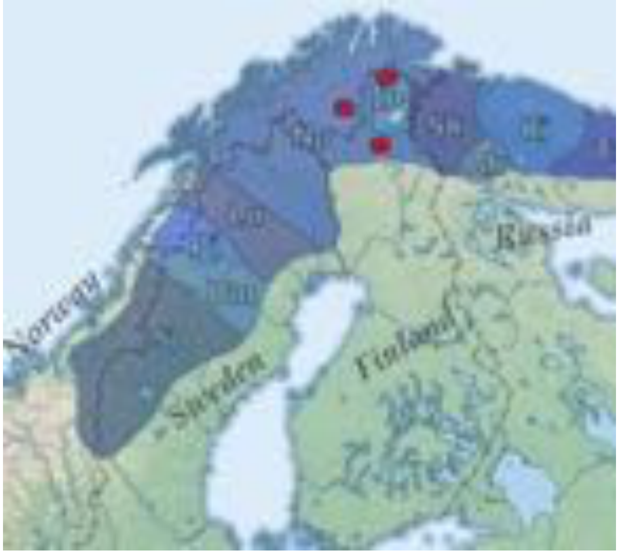


Figure 1. The Sami languages and the data collection

In this paper we focus on the North Sami conversational speech data. Although the data is not huge (195 minutes of annotated data), it is valuable because it is the first North Sami conversational corpus. Conversations concern the participants' everyday life, and their styles differ depending on the age of the speaker and their social status. The topics among young students concern the next vacation, driving school, and cars, while two adult men, mutually acquainted with each other, converse about Sami translation and other technological tools for writing North Sami. The conversations between a pupil and a teacher are fairly formal, and the topics stick to the forthcoming task, i.e. things that one could write a Wikipedia article about.

For the purposes of measuring engagement and to see how laughs function as part of conversations, we annotated the data with laughter features using Praat. Following the previous research, the laughter annotation included the markers for the two laughter types free laugh (fl) and speech-laugh (st), and for the more specific characterization: 'm' – mirth, 'e' – embarrassed, 'b' – breath, 'p' – polite, 'd' – derision and 'o' – other. Table 1 presents the laughter types with explanations.

fl	free laughter	laughter without speaking simultaneously
st	speech-laugh	laughter and speech combined
b	breath	heavy breathing, smirk, sniff; unvoiced, glottal sounds and sibilants
e	embarrassed	speaker is embarrassed, confused, uncertain; disassociating
m	mirth	fun, humorous, real laughter, occurring when telling jokes etc.
d	derision	mocking the partner
p	polite	polite laughter showing positive attitude towards the other speaker
o	other	laughter that doesn't fit in the previous categories; acoustically unusual laughter

Table 1. The annotated laughter types

The total number of laughter occurrences was 341 in 8 different conversations. Two of these conversations were recorded in Karasjok, Norway, and the rest in Ivalo and Utsjoki, Finland. There were 19 conversation informants altogether – some of the conversations had 2 and some 3 participants. 11 of the participants were female and 8 were male. Altogether, 59% (201) of the laughter occurrences were performed by a female informant, while 41% (140) were performed by a male informant. Table 2 shows the number of different laughter types in different conversations.

4. Laughter Types

The basic statistics are shown in Table 2. Free laughter occurs 58% of the laugh occurrences while speech laugh occurs 42% (see discussion below). Three of the specific laughter types occur significantly more frequently than the other types: mirth 29%, embarrassed 49%, and breath 19%, of the total occurrences, and can be called basic laughter types. The laughter bouts annotated as derision, polite and other together only account for 3% of the total occurrences, and can be considered marked types of laughter.

The differences between different conversations can be seen when the laugh activity is normalized with respect to the time. In our data, the average number of laughs is 4.8 per minute, but this varies from almost three times more in 02_V to almost one eight in 07_SX. Qualitative analysis of the conversations shows that the frequency and types of laughter are linked to how well the participants know each other, how nervous they are, and what kind of relationship they have with each other. For example, in the conversation 02_V in which the participants laugh and chuckle the most, they know each other very well, whereas in 07_SX where only a handful of laughs occur, the speakers' relationship is asymmetrical and the whole interaction more formal.

When studying the most laugh-active and engaging conversation 02_V more closely, we notice that the relative count of free laughter is 79% and that of speech-laugh 21%, i.e. the percentage of free laughters is almost four times more than speech-laughs. A closer analysis shows that half of the laughter instances are mirthful or embarrassed laughs, and the other half breathy sounds. This is in contrast with the other conversations where laughters seem to be either mirthful or breathy.

It appears that 02_V is an exception among the conversations in other respects, too: its free laughs account for two thirds of the free laughs in the whole data and it also has most of the embarrassed laughter occurrences. In fact, laughing in 02_V seems to function quite unlike laughing in the other conversations: it signals uncertainty, confusion or embarrassment. This is supported by observations that conversation topics change very fast and have long silences in them, and that the speakers seem nervous in general.

Conv.	fl	st	m	b	e	d	p	o	Tot.	/min
02_V	138	36	28	87	59	0	0	0	174	13.14
06_PS	0	10	4	4	0	0	2	0	10	8.70
05_TP	18	37	27	27	0	0	0	1	55	6.60
08_VV	12	22	19	15	0	0	0	0	34	2.76
01_S	25	18	14	20	6	2	0	1	43	2.64
04_S	3	4	2	3	1	0	0	1	7	2.16
03_V	2	12	6	8	0	0	0	0	14	1.74
07_SX	0	4	0	2	0	0	2	0	4	0.60
Total	198	143	100	166	66	2	4	3	**341**	
%	58	42	29	49	19		3		**100**	

Table 2. The counts of different laughter categories in the DigiSami conversation and the laughter per minute values

Ignoring laughs in 02_V, we notice that free laughing is reduced to only one third of all the laughing occurrences (60/167), i.e. laughter simultaneously with speech seems to be more common than free laughing, and can obviously be used as an effective signal of the speaker's engagement and attitude. On the other hand, 02_V seems to exemplify that laughing is also used as an effective strategy to relieve stress and confusion, besides indicating the speaker's personal characteristics and conversational roles.

In general, we can hypothesise that in natural conversations where people know each other and show no overt nervousness, the basic laughter types occur in two situations: when the participants have real fun, i.e. when telling jokes or funny stories, or when they provide breathy feedback to the partner to signal their engagement in the conversation. However, if the conversational situation creates nervousness, this can be signalled by two extremes: by excessive laughter, or by lack of laughter. The former is common among peers who can thus jokingly share their confusion, uncertainty and embarrassment, while the latter is common among strangers and participants who have asymmetrical power relations and thus markedly signal their non-sharing: laughing automatically creates closeness and in-group feeling which makes the partners more equal.

5. Acoustic Analysis

Following the previous research, we hypothesize that the different laughter types in our data differ in their acoustic properties, such as pitch, formants and intensity, and also duration. In the following acoustic analysis, only the basic and most common laughter types, mirthful (m), breath (b) and uncertain/embarrassed (e) are included, occurring either in free laughters (fl) or in speech-laughs (st). The analyses have been made with different Praat scripts and further processed for min/max/ave/std values.

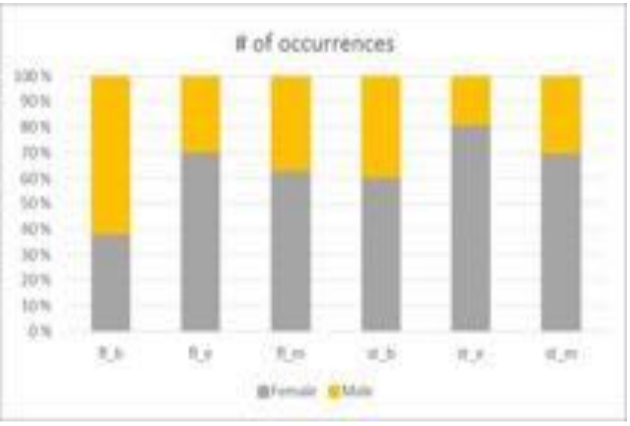


Figure 2. Comparison of the number of laughter occurrences between male and female informants

Our analysis of the acoustic features showed various differences in the investigated laughter types. To compare the results of the acoustic analyses, we calculated average values from the 3 most common types (breath, embarrassed, mirth) of male and female informants separately.

The most common laughter types are shown in Figure 2. As can be seen, female participants produce more laughter signals than men, usually about twice as many. An exception is free laughing breath types where the ratio is the other way round: this is the typical laugh type for the men in our data. It is also interesting that females produce embarrassed and uncertain speech-laughs about four time as many as male participants, being the most typical laugh-type for women in our data.

Figure 3 shows f0 (pitch) values which were extracted with different ranges for male (75-400Hz) and female (100-500Hz) informants; thus comparison of male and female average values is not adequate. However, it was clear that the f0 in free laughter types was higher than f0 in speech-laughter for both male and female informants, which accords e.g. with Truong and van Leeuwen (2007).

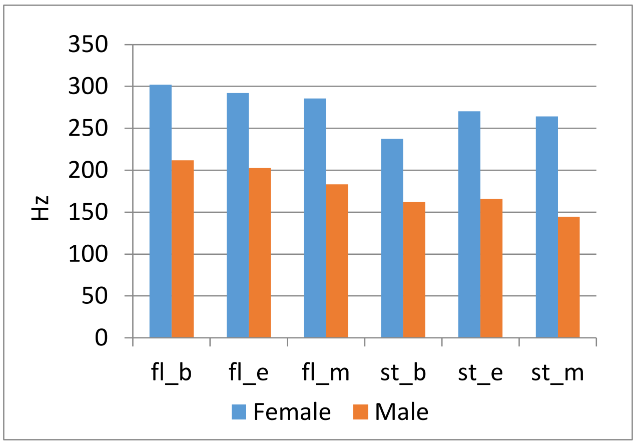


Figure 3. The average f0 values (Hz)

Figure 4 depicts the average duration of the laugh types. There were big differences in duration between the laugher types: durations of embarrassed laughs were significantly longer (2.1s – 3.2s) than all other types for both male and female informants, and breath laughs were the shortest (1.1s –1.4s). However, our data did not support the findings of Nwokah et al. (1999) since free laughter in our data was not significantly longer than speech-laughs. This may be due to the different interaction activities: our data records people conversing in fairly equal situations compared with a mother and child care-giving interaction.

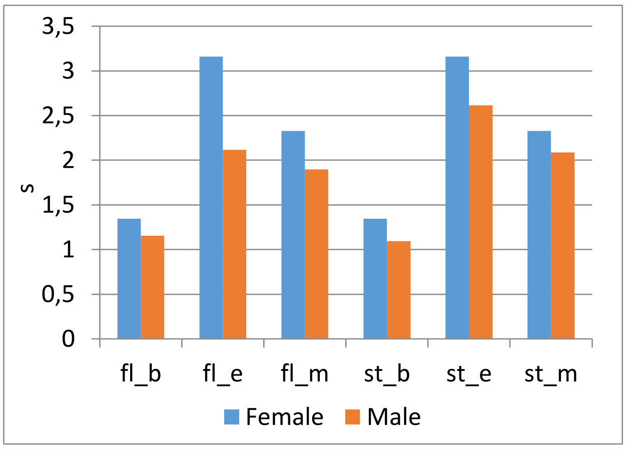


Figure 4. The average duration (s)

The intensity of different laughter types was rather similar in all laughter types, as shown in Figure 5. No significant differences occurred between the different laughter types, but the most surprising difference was that the average intensity with female informants was generally bigger than with males.

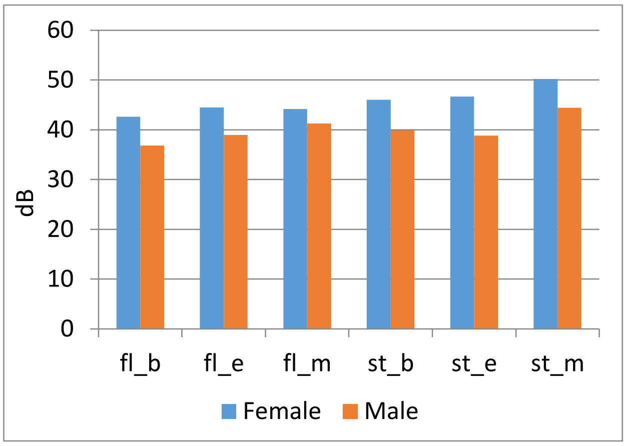


Figure 5. The average intensity (dB).

6. Conclusion

In this paper we have studied the interlocutors' laughing in the North Sami interactions and the functions of laughter in conversational engagement. We can conclude that laughter has several functions that range from fun and happiness to a relief burst of embarrassment. We observed that laughter types depend on the situation and the role of the speakers, and we hypothesize that in natural conversations, the basic laughter types occur when the participants have real fun (mirth) or when they give breathy feedback (breath). However, if the situation is embarrassing or uncomfortable to the speaker, this is signaled by two extremes of laughter frequency, which differ depending on the participants' power relation: the peers use excessive laughter so as to share the embarrassing situation with the others, whereas the partners in an asymmetrical relation indicate more formal, non-sharing behavior by the lack of laughter.

Also our hypothesis concerning the acoustic differences of different laughter types was supported. Durations of embarrassed laughs were significantly longer than the durations of all other types for both male and female informants, but we did not get support for Nwokah et al. (1999) finding of speech-laughs being longer than free laugh. As for the distinctions in intensity, this was small between the laugh types, but interestingly the women had higher intensity laughs than the men in our data.

These observations will be substantiated with deeper statistical analysis, and models for joking, laughing and generally positive attitude will be explored further so as to enable appropriate models be implemented in the SamiTalk application (Wilcock et al. 2016). A useful case is e.g. to be able to recognize the user's embarrassment or uncertainty on the basis of the amount of laughter and their role in the conversation, and alleviate such situations appropriately.

Moreover, as the collected data is multimodal, it is possible to study non-verbal as well as verbal communication. As argued in the previous research, the participants' engagement in the conversation and mutual bonding can be measured with the help of multimodal and non-verbal cues, such as the number of laughs or chuckles, or overlapping speech (Bonin, 2016). Our future studies concern the use of non-verbal information when laughing, to measure the participants' engagement in the interaction.

In addition to studying the functions and acoustic features of laughter in North Sami conversations, our aim is also to raise the visibility of this minority language. Studying North Sami conversation and laughter for the first time opens new perspectives for North Sami language studies and also for the speech community itself. The strength of our conversation corpus is that it presents the language in use and as it naturally is, instead of only focusing on e. g. the grammatical features. Although the basic functions of laughter in North Sami conversations seemed similar to the typical European conversations, it might also be useful to compare different humour types in different cultures. For example, there are Sami comedy TV shows concerning majority peoples' stereotypes about Sami people and self-irony of the Sami people. There are also cases of specific Sami humour in our corpus, which concern the differences between minority and majority cultures, and link the language use into the culture itself.

7. Acknowledgements

The work has been conducted within the Academy of Finland project Finno-Ugric Digital Citizens (grant n°270082). The Praat scripts were made by Mietta Lennes and Shigeto Kawahara (intensity), and we are grateful for their open code.

8. Bibliographical References

Bachorowski, J.-A., Smoski, M. J. and Owren, M. J. (2001). The acoustic features of human laughter. *Journal of Acoustic Society of America*, 110, pp. 1581--1591.

Bonin, F., Campbell, N. and Vogel, C. (2014). Time for laughter. *Knowledge-Based Systems*, 71, pp. 15--24.

Bonin, F. (2016). Content and Context in Conversations: The Role of Social and Situational Signals in Conversation Structure. PhD Thesis, Trinity College Dublin.

Campbell, N., Kashioka, H. and Ohara, R. (2005). No laughing matter. *Proceedings of Interspeech*, pp. 465--478.

Chafe, W. (2007). *The Importance of Being Earnest. The Feeling Behind Laughter and Humor*. Amsterdam: John Benjamins.

Gilmartin, E., Bonin, F., Vogel, C. and Campbell, N. (2013). Laughter and Topic Transition in Multiparty Conversation. In *Proceedings of the SIGDIAL 2013*, Metz, France, pp. 304--308.

Goffman, E. (1974). *Frame Analysis. An Essay on the Organization of Experience*. Boston: Northeastern University Press, reprinted 1986.

Jefferson, G. (1984). On the organization of laughter in talk about troubles. In J. Atkinson, J. Maxwell Heritage (Eds), In *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press, pp. 346--69

Jokinen, K. (2014). Open-domain interaction and online content in the Sami language. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2014).*

Jokinen, K., Hiovain, K., Laxström, N., Rauhala, I. and Wilcock, G. (2016): DigiSami and digital natives: Interaction technology for the North Sami language. In *Proceedings of Seventh International Workshop on Spoken Dialogue Systems* (IWSDS 2016). Saariselkä.

Jokinen, K. and Wilcock, G. (2014): Community-based resource building and data collection. In *Proceedings of 4th Workshop on Spoken Language Technologies for Under-resourced Languages* (SLTU-2014), pp. 201–206. St. Petersburg.

Seurujärvi-Kari, I., Pedersen, S. and Hirvonen, V. (1997). The Sámi. The indigenous people of northernmost Europe. *European Languages 5*. Brussels: European Bureau for Lesser Used Languages.

Tanaka, H. and Campbell, N. (2011). Acoustic Features of Four Types of Laughter in Natural Converational Speech. In *Proceedings of XVIIth ICPhS*, Hong Kong.

Trouvain, J. (2003). Segmenting phonetic units in laughter. In *Proceedings of XVth ICPhS* Barcelona, pp. 2793--2796.

Truong, K. P. and van Leeuwen, D. A. (2007). Automatic discrimination between laughter and speech. *Speech Communication*, 49(2), pp. 144--158.

Wilcock. G. and Jokinen, K. (2013). Wikitalk human-robot interactions. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI)*, pp. 73--74.

Wilcock, G., Laxström, N., Leinonen, J., Smit, P., Kurimo, M., and Jokinen, K. (2016). Towards SamiTalk: a Sami-speaking Robot linked to Sami Wikipedia. In *Proceedings of Seventh International Workshop on Spoken Dialogue Systems* (IWSDS 2016). Saariselkä.

9. Language Resource References

North Sami Conversational Corpus. (2014).

Laughter and Smile Processing for Human-Computer Interactions

Kevin El Haddad, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit
TCTS lab - University of Mons
31 Boulevard Dolez, 7000, Mons Belgium
kevin.elhaddad@umons.ac.be

Abstract

This paper provides a short summary of the importance of taking into account laughter and smile expressions in Human-Computer Interaction systems. Based on the literature, we mention some important characteristics of these expressions in our daily social interactions. We describe some of our own contributions and ongoing work to this field.

Keywords: laughter, smiling, Human-Computer Interaction, synthesis, recognition

1. Introduction

Computers are increasingly becoming part of our lives, making a lot of our daily tasks easier. As interactions with them increase, so too does the need to have interfaces that are more natural to use than the traditional keyboard or mouse. These interfaces included speech, which is our most common means of communication. An ideal interface would be one with which we could communicate as if we were talking to person. This would entail access to the same expressiveness and emotions as in a conversation with another person. Laughter and smiling have been the subject of several studies in the social sciences, including psychology, anthropology, and paralinguistics, because of their importance in social interaction. They are indeed multifunctional and extremely common. Therefore in order to create a natural Human-Computer Interaction (HCI) system, these expressions have to be integrated in it. However, the need to consider laughter and smiling in HCI systems may not be immediately obvious to researchers not related to this field of study in particular, or to affective computing in general. This might be because integrating emotions in general might not seem important for some, or, at first glance, laughter and smiling may seem to be no more important than any other paralinguistic expression. The goal of this paper is provide a broad overview of studies in different research fields highlighting the necessity of considering laughs and smiles into an HCI system in order to make the interaction more natural. Below, we will first discuss the relevance of these non-verbal paralinguistic expressions to HCI by providing a short survey of previous work. We will then sketch some applications in which they have been considered, and finally we will describe our own contributions to this field.

2. Laughter and Smiles in Interactions

Laughs and smiles are among the most, if not the most, important non-verbal expressions in our daily interactions and this makes them worthy to be considered in HCI systems. The reason of their importance are cited in the following paragraphs:

Frequent Occurrences in Conversations: Laughs and smiles occur frequently in conversations. Indeed, in the ICSI meeting corpus, Laskowski and Burger reported 9.5% of the total verbalizing time being laughter (Laskowski and Burger, 2007; Vogel et al., 2013). In other work, Chovile did not consider smiles in his analysis of affect in conversation as smiles were so overwhelmingly frequently present in the data compared to other expressions (Chovil, 1991). This high frequency of occurrence is the first reason to work on including smiling and laughter in HCI systems which aim to replicate human-human interactions.

Expression of Different Emotions: Laughter can express several different affective states. Although intuitively and commonly related to emotions with positive valence (generally amusement, joy and sympathy), laughter can also express other negative emotions such as disappointment, stress and embarrassment (Devillers and Vidrascu, 2007).

Smiling can also express different emotions of different valence, such as joy or embarrassment (Ambadar et al., 2008; Keltner, 1995; Frank et al., 1993).

Emotions are crucial to understanding (recognition) or creating (synthesis) a certain context or mood. Being able to automatically and accurately assimilate this dimension in a dialogue would improve the interaction by increasing its naturalness.

Social Functions: Laughs and smiles are more likely to happen with someone rather than alone (Glenn, 2003; Fridlund, 1991). They have been shown to be somewhat related to the cultural background (Soury and Devillers, 2014). In social interactions, they are used not only to express emotions, but also to apply certain social functionalies that do not really contain emotions. In fact, people do sometimes laugh and smile without really feeling any emotion.

Laughter and smiling can be used in the course of a conversation, with social functions, punctuating the dialogue with social information (Provine, 2010), expressing politeness (Hoque et al., 2011) or changing the topic (Bonin et al., 2014; Vogel et al., 2013).

Both laughter and smiling can be used as backchannels to show interest in the speaker and to encourage him or her to carry on talking (Duncan, 1972; Poggi and Pelachaud, 2000).

Being able to use these expressions with these social functionalities in dialogue systems will increase the naturalness of an agent's reaction during an interaction.

Perception of Laughs & Smiles: Laughs and smiles are contagious as shown by Provine (Provine, 2013) and Wild (Wild et al., 2003) respectively. Indeed it is likely that a subject will smile or laugh under when they are exposed to another's laughter or smiling. They can also affect the perception of a subject: viewing a smiling photograph versus a photograph of the same person with a neutral expression has been reported to result in an increased perception of characteristics such as attractiveness, trustworthiness and sociability (Reis et al., 1990).

Gelotophobia: Gelotophobia is the fear of being laughed or smiled at (Ruch et al., 2014). This disease is another example of the importance of these two expressions in our social communications and show the influence they can have on individuals.

3. Laughter and Smiling Embedded in HCI Applications

Several HCI system have already been developed which include laughter and smiling detection systems.

Mendel et. al. (Melder et al., 2007) presented a multimodal real-time HCI system with the goal to detect and elicit laughter. In this application, a user's behavior is monitored, interpreted and regulated by the system in an interactive loop. An audio laughter (Truong and van Leeuwen, 2007) detection system and visual smile recognition system were developed and contributed to assess the user's emotions state.

Some HCI experiments were also conducted in the framework of the European project Ilhaire (Dupont et al., 2016), which was dedicated to study laughter. For example, in (Pecune et al., 2015a), a laughing avatar is used to study the contribution of a virtual agent to enhancing a user's experience. A user is presented with stimuli in the presence or not of the avatar. When the avatar is present, it either copies the user's behavior of laughs at predefined times and intensities. The multimodal synthesis system developed in (Ding et al., 2014) is used here to generate laughter animation. The detection is made using a platform based on a EyesWeb XMI platform (Mancini et al., 2014).

A facial smile detection system was integrated in a Perceptual User Interface (PUI) in (Deniz et al., 2008). This PUI was used in an application to control the status and insert smile/big smile emoticons in an Instant Messaging client conversation window. The system can assess the level of the facial smile and map it to the emoticon to be inserted.

One of the goals of the European project JOKER (Devillers et al., 2015) is to study the impact an emotional social agent showing empathy and compassion might have on a user's mood during a conversation. Interfaces related to laughter and smiling are crucial for obtaining such a virtual social companion.

4. Interfaces for HCI applications

In order to take into account laughter and smiling in HCI systems, interfaces must be developed for this task. These interfaces take care of the generation (synthesis) and detection (recognition) of laughs/smiles. This section will present our work and main contribution in this field which concern interfaces related to laughs and smiles. It will also mention interesting work of others in this field. Please also note that all the synthesis and recognition/detection modules mentioned in Section 3., even though relevant here, will not be repeated.

4.1. Synthesis Applications

Being able to synthesize laughter and smiles would, in general, increase the naturalness of an HCI and therefore make the interaction more comfortable to the user(s) as shown in (Theonas et al., 2008). Application examples of laughter and smiling synthesis systems in HCI can be, first, the control of a conversation flow by using the social functions that smiles and laughs have. It could provide the user(s) with feedbacks while he/she is speaking and thus encouraging him/her to carry on. It could, for instance, change the subject of the conversation, or express agreement or even disagreement (with a mockery laugh for example). A second example would be to influence the user's mood or emotional state. Indeed this could be used to express empathy in order to make the avatar more likable (Devillers et al., 2015), or trigger amusement by uttering amused laughs or smiles (Niewiadomski et al., 2013; Pecune et al., 2015b). Such synthesis systems could also be used for medical purposes, helping to study the phenomenon of gelotophobia and even treating it. This was one the of the purposes of the Ilhaire European project (Ruch et al., 2015; Ruch et al., 2014). It can also help reducing stress since it has been found that laughter helps reduce stress (Bennett et al., 2003).

Urbain et. al. (Urbain et al., 2014) presented a Hidden Markov Model (HMM)-based audio laughter synthesis system in which the level of the arousal intensity or of the laughter is controllable. Other work on audiovisual laughter synthesis can be found in (Çakmak, 2016). In this thesis, the author presents synthesis and evaluation of audio and motion capture cues of laughter. He also presents synchronization rules between the audio and visual cues for synthesizing laughter from a virtual agent. In (de Kok and Heylen, 2011), the authors present an attempt on predicting the types of smiles that should be generated, based on the context. But no actual synthesis is presented. In (Ochs et al., 2010), a decision tree is used to predict the type of smile to be generated. The generation system was also evaluated using a subjective perceptual test.

Our contribution in this field focused on adding smiles and laughs to synthesized speech, thus creating speech-smiled and speech-laugh. Hidden Markov Models (HMM)-based systems were used to synthesize speech-smiles (El Haddad et al., 2015e; El Haddad et al., 2015b) and control the arousal level of smiling in an utterance. A speech-laugh synthesis system was also created based also on HMM and proved to increase the naturalness perceived compared to neutrally synthesized sentences (El Haddad et al., 2015f; El Haddad et al., 2015a). In order to do this, databases were collected containing laughter and smiled speech. The next step is first to be able to synthesize in real time sentences will controlling the level of amusement in speech. This includes varying the level of smiling and adding laughter bursts. We will also work on reproducing this system in different languages. In order to do that, a multilingual database similar to the one in (El Haddad et al., 2015f) has been collected. We also intend to create the same speech-laugh/smiling synthesis systems audiovisually. This means synthesizing also motion capture speech-laugh and controllable smiling data synchronized with the synthesized audio cues.

4.2. Recognition Applications

Since smiles and laughs can express different types of emotions and can also have several social functions, their detection and recognition would help understanding the emotional state of the user(s) and therefore also the context. A context understanding will help an agent react more adequately. This would improve the quality of the interaction (Yang et al., 2015). This can also be used for user mood monitoring, for instance, to detect the level of amusement and estimate the level of stress since they are related (Bennett et al., 2003). In addition, being able to recognize/detect smiles and laughs in speech, would increase the robustness of an automatic speech recognition system by differentiating between speech and non-speech.

Knox et. al. (Knox and Mirghafori, 2007) presents an automatic audio laughter detection using a neural network.

In (Yang et al., 2015), presents a multimodal laughter and smiling recognition system to be used in a human-robot interaction with elderly people. In (Ito et al., 2005), Ito et. al. also present a laughter and smiling audiovisual detection system. This system was developed for application in natural conversation videos.

The main contribution we have in this field is work related the arousal level assessment of amusement. In (El Haddad et al., 2015c; El Haddad et al., 2015d) we defined so called Amused Speech Components (ASC), collected data and presented analyses and classification systems for them. This work is in the larger framework of assessing the amusement arousal level in a given sentence. Indeed, we aim at building an ASC detection system and then accurately assess a level of amusement arousal in the given sentence based on the detected ASC. A multimodal system will be used based on a database containing motion capture data as well as audio data.

5. Acknowledgements

This work was partly supported by the Chist-Era project JOKER with contribution from the Belgian Fonds de la Recherche Scientifique (FNRS), contract no. R.50.01.14.F. H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

6. Bibliographical References

Ambadar, Z., Cohn, J. F., and Reed, L. I. (2008). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior*, 33(1):17–34.

Bennett, M. P., Zeller, J. M., Rosenberg, L., and McCann, J. (2003). The effect of mirthful laughter on stress and natural killer cell activity. *Alternative therapies in health and medicine*, 9(2):38.

Bonin, F., Campbell, N., and Vogel, C. (2014). Time for laughter. *Knowledge-Based Systems*, 71:15 – 24.

Çakmak, H. (2016). *Audiovisual Laughter Synthesis - A Statistical Parametric Approach*. Ph.D. thesis, University of Mons, February.

Chovil, N. (1991). Discourse oriented facial displays in conversation. *Research on Language and Social Interaction*, 25(1-4):163–194.

de Kok, I. and Heylen, D. (2011). When do we smile? analysis and modeling of the nonverbal context of listener smiles in conversation. In *Affective Computing and Intelligent Interaction*, volume 6974 of *Lecture Notes in Computer Science*, pages 477–486, Berlin, Germany, October. Springer Verlag.

Deniz, O., Castrillon, M., Lorenzo, J., Anton, L., and Bueno, G., (2008). *Advances in Visual Computing: 4th International Symposium, ISVC 2008, Las Vegas, NV, USA, December 1-3, 2008. Proceedings, Part II*, chapter Smile Detection for User Interfaces, pages 602–611. Springer Berlin Heidelberg, Berlin, Heidelberg.

Devillers, L. and Vidrascu, L. (2007). Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. In *Interdisciplinary Workshop on The Phonetics of Laughter*, page 37.

Devillers, L., Rosset, S., Dubuisson Duplessis, G., Sehili, M. A., Bechade, L., Delaborde, A., Gossart, C., Letard, V., Yang, F., Yemez, Y., Turker, B. B., Sezgin, M., El Haddad, K., Dupont, S., Luzzati, D., Esteve, Y., Gilmartin, E., and Nick, C. (2015). Multimodal data collection of human-robot humorous interactions in the joker project. In *ACII*, Xi'an, China, September.

Ding, Y., Prepin, K., Huang, J., Pelachaud, C., and Artières, T. (2014). Laughter animation synthesis. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 773–780, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

Dupont, S., Çakmak, H., Curran, W., Dutoit, T., Hofmann, J., McKeown, G., Pietquin, O., Platt, T., Ruch, W., and Urbain, J., (2016). *Toward Robotic Socially Believable Behaving Systems - Volume I : Modeling Emotions*, chapter Laughter Research: A Review of the ILHAIRE Project, pages 147–181. Springer International Publishing, Cham.

El Haddad, K., Cakmak, H., Dupont, S., , and Dutoit, T. (2015a). Breath and repeat: An attempt at enhancing speech-laugh synthesis quality. In *European Signal Processing Conference (EUSIPCO 2015)*, Nice, France, 31 August-4 September.

El Haddad, K., Cakmak, H., Dupont, S., and Dutoit, T. (2015b). An HMM Approach for Synthesizing Amused Speech with a Controllable Intensity of Smile. In *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, UAE, 7-10 December.

El Haddad, K., Cakmak, H., Dupont, S., and Dutoit,

T. (2015c). Towards a Level Assessment System of Amusement in Speech Signals: Amused Speech Components Classification. In *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, UAE, 7-10 December.

El Haddad, K., Dupont, S., Cakmak, H., and Dutoit, T. (2015d). Shaking and Speech-smile Vowels Classification: An Attempt at Amusement Arousal Estimation from Speech Signals. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Orlando, Florida, US, 14-16 December.

El Haddad, K., Dupont, S., d'Alessandro, N., and Dutoit, T. (2015e). An HMM-based speech-smile synthesis system: An approach for amusement synthesis. In *International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Ljubljana, Slovenia, 4-8 May.

El Haddad, K., Dupont, S., Urbain, J., and Dutoit, T. (2015f). Speech-laughs: An HMM-based Approach for Amused Speech Synthesis. In *Internation Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, pages 4939–4943, Brisbane, Australia, 19-24 April.

Frank, M. G., Ekman, P., and Friesen, W. V. (1993). Behavioral markers and recognizability of the smile of enjoyment. *Journal of personality and social psychology*, 64(1):83.

Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, 60(2):229.

Glenn, P. (2003). *Laughter in interaction*, volume 18. Cambridge University Press.

Hoque, M., Morency, L.-P., and Picard, R. W. (2011). Are you friendly or just polite?–analysis of smiles in spontaneous face-to-face interactions. In *Affective Computing and Intelligent Interaction*, pages 135–144. Springer.

Ito, A., Wang, X., Suzuki, M., and Makino, S. (2005). Smile and laughter recognition using speech processing and face recognition from conversation video. In *Cyberworlds, 2005. International Conference on*, pages 8 pp.–444, Nov.

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

Knox, M. T. and Mirghafori, N. (2007). Automatic laughter detection using neural networks. In *INTERSPEECH*, pages 2973–2976.

Laskowski, K. and Burger, S. (2007). Analysis of the occurrence of laughter in meetings. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, pages 1258–1261, Antwerp, Belgium, 27-31 August.

Mancini, M., Varni, G., Niewiadomski, R., Volpe, G., and Camurri, A. (2014). How is your laugh today? In *Proceedings of the Extended Abstracts of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI EA '14, pages 1855–1860, New York, NY, USA. ACM.

Melder, W. A., Truong, K. P., Uyl, M. D., Van Leeuwen, D. A., Neerincx, M. A., Loos, L. R., and Plum, B. S. (2007). Affective multimodal mirror: Sensing and eliciting laughter. In *Proceedings of the International Workshop on Human-centered Multimedia*, HCM '07, pages 31–40, New York, NY, USA. ACM.

Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., Geist, M., Lingenfelser, F., McKeown, G., Pietquin, O., and Ruch, W. (2013). Laugh-aware virtual agent and its impact on user amusement. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 619–626, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Ochs, M., Niewiadomski, R., and Pelachaud, C. (2010). How a virtual agent should smile? - morphological and dynamic characteristics of virtual agent's smiles. In *10th International Conference on Intelligent Virtual Agents (IVA)*, Philadelphia, Pennsylvania, US.

Pecune, F., Mancini, M., Biancardi, B., Varni, G., Ding, Y., Pelachaud, C., Volpe, G., and Camurri, A. (2015a). Laughing with a virtual agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1817–1818. International Foundation for Autonomous Agents and Multiagent Systems.

Pecune, F., Mancini, M., Biancardi, B., Varni, G., Ding, Y., Pelachaud, C., Volpe, G., and Camurri, A. (2015b). Laughing with a virtual agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 1817–1818, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Poggi, I. and Pelachaud, C. (2000). Embodied conversational agents. chapter Performative Facial Expressions in Animated Faces, pages 155–188. MIT Press, Cambridge, MA, USA.

Provine, R. R. (2010). Laughter Punctuates Speech: Linguistic, Social and Gender Contexts of Laughter. *Ethology*, 95:291–298.

Provine, R. R. (2013). Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bulletin of the Psychonomic Society*, 30(1):1–4.

Reis, H. T., Wilson, I. M., Monestere, C., Bernstein, S., Clark, K., Seidl, E., Franco, M., Gioioso, E., Freeman, L., and Radoane, K. (1990). What is smiling is beautiful and good. *European Journal of Social Psychology*, 20(3):259–267.

Ruch, W. F., Platt, T., Hofmann, J., Niewiadomski, R., Urbain, J., Mancini, M., and Dupont, S. (2014). Gelotophobia and the challenges of implementing laughter into virtual agents interactions. *Frontiers in Human Neuroscience*, 8(928).

Ruch, W., Hofmann, J., and Platt, T. (2015). Individual differences in gelotophobia and responses to laughter-eliciting emotions. *Personality and Individual Differences*, 72:117 – 121.

Soury, M. and Devillers, L. (2014). Smile and laughter in human-machine interaction: a study of engagement. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Theonas, G., Hobbs, D., and Rigas, D. (2008). Employing virtual lecturers' facial expressions in virtual educational environments. *IJVR*, 7(1):31–44.

Truong, K. P. and van Leeuwen, D. A. (2007). Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144 – 158.

Urbain, J., Cakmak, H., Charlier, A., Denti, M., Dutoit, T., and Dupont, S. (2014). Arousal-driven synthesis of laughter. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):273–284, April.

Vogel, C., Campbell, N., Bonin, F., and Gilmartin, E. (2013). Exploring the role of laughter in multiparty conversation.

Wild, B., Erb, M., Eyb, M., Bartels, M., and Grodd, W. (2003). Why are smiles contagious? an fmri study of the interaction between perception of facial affect and facial movements. *Psychiatry Research: Neuroimaging*, 123(1):17 – 36.

Yang, F., Sehili, M. A., Barras, C., and Devillers, L., (2015). *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings*, chapter Smile and Laughter Detection for Elderly People-Robot Interaction, pages 694–703. Springer International Publishing, Cham.

Making Idle Conversation – Towards a Casual Talk System

Emer Gilmartin[1], Ketong Su[1], Yuyun Huang[1], Kevin El Haddad[3], Benjamim R. Cowan[2], Nick Campbell[1]
[1]Speech Communication Lab, Trinity College Dublin
[2]University College Dublin
[3]University of Mons, Belgium
gilmare@tcd.ie, nick@tcd.ie

Abstract

While casual conversation or talk is ubiquitous in human life, it has not been taxonomised or indeed recreated in spoken dialogue applications to the extent that more tractable task-based or practical dialogues have been. With the recent surge in interest in more sociable speaking systems for use in companionship, educational, and healthcare applications, there is increasing need for clearer understanding at several levels of what casual talk is and how it may be modelled. In this paper we describe our current work on exploring, creating, and evaluating human-machine casual social talk.

Keywords: casual conversation, human-machine interaction

1. Introduction

Spoken interaction is everywhere in human life. Much of our speech pertains to instrumental or practical transactions and communication, where the linguistic content of the utterance is the 'currency' of the exchange. However, not all talk has a clear, short term goal, and spoken interaction in the form of casual conversation has been described as a medium through which social bonds are built and maintained (Dunbar, 1998). This 'talking just for the sake of talking'(Eggins and Slade, 2004) includes smalltalk, gossip and conversational narrative, and may well be the most fundamental form of spoken interaction (Malinowski, 1923). Until recently, spoken dialogue applications have been applied to practical exchanges in task-based domains, such as travel bookings, sales and order fulfilment, or direction giving. These domains were regarded as more tractable in early spoken dialogue work (Allen et al., 2000). Recently, technological advances and particularly the advent of ubiquitous computing have led to greater interest in the creation of applications where social or friendly conversation is implemented.

Building such systems leads to a number of challenges. There is a need for analysis of casual talk in terms of structure, content, and timing. This analysis should extract design parameters in the form of ground truths from corpora of human-human casual conversation. Many available English language corpora of human spoken interaction do not meet the requirements of this task, as they are based on transactional or task-based dialogues - comprising artificial tasks (e.g. Maptask (Anderson et al., 1991)), or real or staged workplace meetings (e.g. ICSI (Janin et al., 2003), AMI (McCowan et al., 2005)). More 'social' data is often telephonic (e.g. Switchboard (Godfrey et al., 1992)) and may not extrapolate well to face to face casual conversation. Although there are some corpora of English language human-human face-to-face casual spoken interaction, including parts of the British National Corpus (BNC-Consortium, 2000), the ICE corpus (Greenbaum, 1991), and the Santa Barbara Corpus (DuBois et al., 2000), the data are not multimodal.

It has long been theorised in the pragmatics literature that the timing and the 'feel' of a conversation is important to the success of a casual or social conversation (Abercrombie, 1956; Hayakawa, 1990). We speculate that prosodic elements in artificial social talk will need to be controlled far more accurately than is necessary in task-based talk. This is because the goal of social talk conversation is to maintain a social co-presence rather than to exchange information in order to complete a clearly defined short term task, and thus simply providing the right answers to queries will not suffice for dialogue success. For example, in timing terms, a short gap may make the conversation seem rushed or uncomfortable, whereas an overlong silence might give the impression of boredom. In addition, the neutral voice quality found in state of the art text to speech (TTS) may not be sufficient to create the affective element necessary.

We are currently working on a number of areas of social talk. We are working on establishing parameters for gap length and exchange structure using human-human data. We have built a dialogue system which gives us freedom to manipulate all aspects of interaction. We are using the system in automatic and in Wizard of Oz (WOZ) mode to investigate the effects of varying interspeaker gap, the feasibility of introducing laughter and amused voice to dialogues, and to provide data for studies on engagement of participants in casual talk. Below, we briefly describe the CARA dialogue system, developed at the Speech Communication Lab, Trinity College, Dublin, and how we are using the system as a testbed for experiments and to collect data.

2. CARA Dialogue System

To investigate and implement social talk, we need prototype dialogue systems which chat to, joke with, and indeed tease interlocutors. The JOKER project aims to build dialogue systems with social communication skills including humour, empathy, compassion, charm, and other informal socially-oriented behaviour. As part of this project we have been developing a dialogue system, CARA, which gives us the freedom and control necessary to experiment with parameters of the dialogue in order to create interactions which more closely resemble natural casual talk. CARA is a multimodal distributed spoken dialog system implementation using various Java technologies. It consists of the following modules: a browser-based user interface, a Voice Activity Detection (VAD) enabled recorder, a local ASR decoder, a real-time synthesiser and a Finite State Machine (FSM) based dialogue management core. Below, we explain the messaging strategy adopted by the system to facilitate distributed-ness, and give a brief introduction to the technical architecture and the functionalities of each of the modules. Communication between different components is handled by Java Message Service (JMS). This mature message oriented middleware allows messages to be transmitted in a fully decoupled and asynchronous manner. In terms of the current version of the system used in this experiment, all modules were run locally and efficiently on a CORE i7 Windows workstation. The various latencies introduced across the execution flow were mainly due to the specific implementation of the individual module, rather than computational power available. However, in order to minimise those undesirable system generated lags, more sophisticated design is required for further versions, which subsequently demands more resources. The JMS communication layer is added to realise the full potentials of a highly distributed version of the system since it is designed to bridge physically separate and heterogeneous systems together.

In addition, from the programming point of view, it provides a pluggable interface for additional input sensors or analytical units to be trivially injected into the dialogue decision-making process. As the mentioned above, open-domain ASR is both time and resource consuming. However, with the aid of JMS, an off-the-shelf, cloud-based, possibly commercial ASR module can be easily integrated into the system, saving the tremendous effort required to implement and maintain it locally. The facade of the system is a web application hosted on Glassfish server, which takes advantage of the new Web Real-Time Communication (WebRTC) introduced by World Wide Web Consortium. WebRTC offers a simple API for retrieving the local audio and video input streams within the browser environment. Similar to the exploitation of JMS, this web-based interface makes the system accessible cross-platform and cross-device, and more importantly without any prerequisites apart from a modern browser application.

The recorder is responsible for receiving the live audio input through the JMS messaging layer in the format of raw audio segments, and performing a volume-based VAD, finally storing the detected speech as WAV files for further processing. Sphinx4 is integrated as the ASR decoder using its latest Java API.

For this study, a live instance of the recogniser is maintained with a limited grammar throughout the execution to minimise the processing delay. As a result, the current configuration alleviates the latency to a negligible range. In general, ASR is still the major timing bottleneck compared to others when open-domain recognition is added into the equation. Since one of the primary objectives of our work is to demonstrate that responsiveness is essential for natural and interactive social conversations, the system also has cloud-based ASR integrated for the open-domain case. We are currently testing the quality of IBM and ATT's cloud services for the case of casual talk. We have incorporated CereProc's Caitlin voice as the synthesiser, which is a commercial grade product with a high quality Irish-accented English voice. As a result, the synthesis of medium-sized utterances happen on the fly and no latency is introduced. However, as we have control of the coding of the system, other voices can be easily incorporated for different experimental goals.

3. Current Work Using the CARA system

3.1. Data Collection for Teasing Social Talk

As part of the JOKER project, team members in France and Ireland built French and English speaking casual dialogue system prototypes which were used to collect social talk data. For this task, the domain was dyadic talk about food, and dialogues consisted of two phases – a 'blague' or 'joshing' stage where the system engaged the user in a short chat about themselves and about food, while producing puns and teasing, and a 'defi' or guessing game, where the user attempted to guess the system's favourite dish. The English speaking trials were carried out over two sessions per participant in order to record the dialogues in two conditions, human-machine and human-human. In the human-machine condition, the same two-phase dialogue was performed by each participant in two separate sessions – in automatic mode, and in WOZ mode where a human chose WHEN to make the next utterance but not WHAT to say. A human-human condition was added for 'gold-standard' human conversational data, and to allow researchers to contrast human and human-machine social talk. The content of the human-human sessions was similar to the human-machine condition – pairs of naive subjects were instructed to chat together and then to play 'Guess my favourite food'.

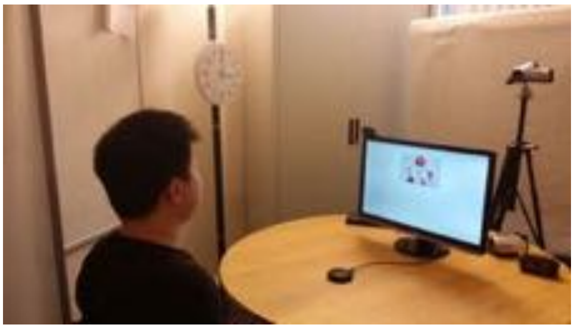


Figure 1: Human Machine Setup

For the human-machine conditions, the subject was seated at a table opposite a screen showing an image of a robot as in Fig. 1. In the WOZ case, the experimenter controlled the timing for all participants, and could not control WHAT was said by the system, but only press a button to play the next utterance. For the human-human recordings, pairs of subjects sat opposite each other, with HD video cameras facing each of them as in Fig. 2.

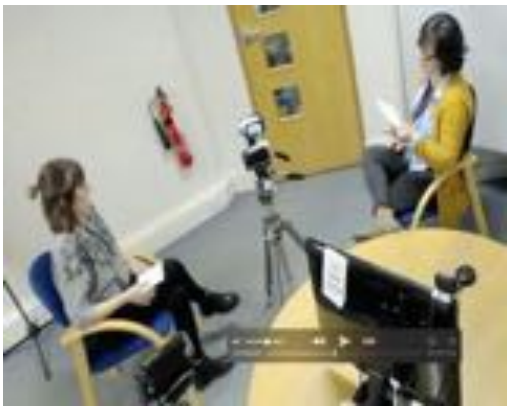


Figure 2: Human Human Setup

The audio and video recordings have been segmented and synchronized, and are being transcribed. The human-machine automatic dialogues were of mean length 170 seconds, while the WOZ dialogues were of mean length 194 seconds. The human-human dialogues were of mean length 13 minutes and 7 seconds, with the game section having a mean of 198 seconds and the chat section having a mean of 9 minutes and 49 seconds. A second cycle of recordings is planned.

3.2. Investigating Engagement in Social Talk

The current corpus is being used to investigate levels of engagement in social talk dialogues. Engagement is defined by Sidner as *The process by which two (or more) participants establish, maintain and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in the interaction, evaluating whether to stay involved and deciding when to end the connection.* (Sidner and Dzikovska, 2002, p. 141). Knowledge of the level of an interlocutor's engagement in an interaction would provide a measure of dialogue success, particularly in social talk where particular 'answers' do not necessarily provide the information that the dialogue is succeeding to the same extent as they do in task based dialogues. Such knowledge would be very useful in dialogue management. Many existing casual human-human corpus do no have high enough resolution video quality for frontal face analysis, or the interactions are too short and it is difficult to find disengagement segments for analysis. There is also a need for corpora of human-machine interaction. Our current and future collections of human-human data and similar human-machine data mediated by the CARA system will provide high quality data in a suitable domain for use in our ongoing work on engagement detection. We have annotated the recordings described above for engagement, and have used deep-learning methods to model engagement using audio and video features, with very promising results.

3.3. Evaluating Amused Voice Synthesis in Dialogue

We are also currently working with colleagues from UMONS to evaluate the impact of amused speech synthesis on dialogue. We are creating voices which can display varying levels of amusement, which can then be evaluated in a realistic context using the CARA dialogue system described above. The HMM-based speech-laugh synthesis approach used is based on the system described by El Haddad et al. (El Haddad et al., 2015a). In brief, unified models for the acoustic features of pitch (f0), spectrum coefficients and phoneme durations are created during a training step (Yoshimura et al., 1999), using Gaussian Mixture Models (GMM). Previously trained models can be transformed into adapted models of the target voice through the Constrained Maximum Likelihood Linear Regression (CMLLR) method (Digalakis et al., 1995) (a description of the CMLLR adaptation algorithm can be found in Yamagishi et al. (Yamagishi et al., 2009)), which has been used here to adapt to a target person's neutral speech and speech-smile voices. For our initial implementation, the CMU Arctic database RMS male voice is used to create the main neutral voice. Neutral and smiled speech were recorded from a male and a female speaker asked to read a subset of the ARCTIC sentences, giving approximately 10 minutes of material. This material is being used. Using the adaptation and interpolation techniques described in El Haddad et al. (El Haddad et al., 2015b), we created amused and neutral voices for the male speaker, and interpolated between them to create intermediate levels of amusement, to allow us to control the system's output. We are currently doing the same for the female speaker. French voices created using this method have produced positive evaluations in perception tests (El Haddad et al., 2015a), and we are currently implementing recording protocols to test their performance in a social dialogue.

4. Conclusion

We have given an overview of work in progress on several aspects of artificial social dialogue creation, including the creation of a dialogue platform, the collection of a corpus of human-machine and human-human data, and trial implementation of amused and neutral synthetic speech. We hope that the data collected and conclusions drawn will be of use to the research community

5. Acknowledgements

This work is supported by the CHISTERA-JOKER project and by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre at Trinity College Dublin

6. Bibliographical References

Abercrombie, D. (1956). *Problems and principles: Studies in the Teaching of English as a Second Language*. Longmans, Green.

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2000). An architecture for a generic dialogue shell. *Natural Language Engineering*, 6(3&4):213–228.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus. *Language and speech*, 34(4):351–366.

BNC-Consortium. (2000). British national corpus. *URL http://www. hcu. ox. ac. uk/BNC.*

Digalakis, V. V., Rtischev, D., and Neumeyer, L. G. (1995). Speaker adaptation using constrained estimation of gaussian mixtures. *Speech and Audio Processing, IEEE Transactions on*, 3(5):357–366.

DuBois, J. W., Chafe, W. L., Meyer, C., and Thompson, S. A. (2000). *Santa Barbara Corpus of Spoken American English. CD-ROM. Philadelphia: Linguistic Data Consortium.*

Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Harvard Univ Press.

Eggins, S. and Slade, D. (2004). *Analysing casual conversation*. Equinox Publishing Ltd.

El Haddad, K., Cakmak, H., Dupont, S., , and Dutoit, T. (2015a). Breath and repeat: An attempt at enhancing speech-laugh synthesis quality. In *European Signal Processing Conference (EUSIPCO 2015)*, Nice, France, 31 August-4 September.

El Haddad, K., Cakmak, H., Dupont, S., and Dutoit, T. (2015b). An HMM Approach for Synthesizing Amused Speech with a Controllable Intensity of Smile. In *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, UAE, 7-10 December.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520.

Greenbaum, S. (1991). ICE: The international corpus of English. *English Today*, 28(7.4):3–7.

Hayakawa, S. I. (1990). *Language in thought and action*. Houghton Mifflin Harcourt.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., and Stolcke, A. (2003). The ICSI meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–364.

Malinowski, B. (1923). The problem of meaning in primitive languages. *Supplementary in the Meaning of Meaning*, pages 1–84.

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

Sidner, C. L. and Dzikovska, M. (2002). Human-robot interaction: engagement between humans and robots for hosting activities. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 123–128.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(1):66–83.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*, Budapest, Hungary.