Content and Context in Conversations: The Role of Social and Situational Signals in Conversation Structure

Francesca Bonin

Thesis submitted for the Degree of Doctor of Philosophy

School of Computer Science & Statistics

Trinity College

University of Dublin

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. Wherever there is joint published or unpublished work included, it is duly acknowledged in the text.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Summary

The increasing growth of multimodal material creates, nowadays, a renewed interest in innovative approaches to information extraction from meetings and multiparty conversations; those approaches make use of various multimodal sources such as video, audio and resulting transcripts from which information can be derived to produce a richer semantic analysis of the social interaction.

Social interactions are co-constructed by a combination of contextual and linguistic aspects. However, while the social context of interaction has previously been considered in relation to the emotional level of communication, fewer studies have been devoted to exploring the relationship between the context and the linguistic level of interactions.

This thesis proposes a novel view of context in communication, exploring its effect on the discourse structure, independently of the affective and emotional spheres of communication. I explore the influence of two concepts of context on human-human communication: the social context (ensemble of the social signals exchanged by the participants) and the situational context (situation in which the conversation takes place), and show the extent to which both, social and situational contexts, have a discourse

function. In both cases, results show that the information carried by their timing can be exploited in the detection of discourse events.

Acknowledgement

I would like to express my deepest gratitude to all the persons that have guided, supported and inspired me during this wonderful path towards the PhD. First of all I would like to thank my two advisors, Carl Vogel and Nick Campbell, for everything I have learnt from them, for supporting, for trusting, and for guiding me during the last four years.

I would like to thank Carl, for always being there for me for any questions, any doubts, any concerns. At many stages in the course of this research project I benefited from his advice, particularly so when exploring new ideas. His positive outlook and confidence in my research has been a source of inspiration and support. He provided me with the possibility to explore many ideas, but has always been there to show me the path when I was feeling lost.

I would like to thank Nick for pushing me to take on new challenges and responsibilities. He has shown a continuous trust in me as a person and a researcher since the very beginning. He has pushed me to open my ideas for discussion, and has taught me to widen my horizons to other fields and applications. I would like to thank him for his contagious enthusiasm and unfailing trust.

To both, who, in different ways, have helped me to grow as a person and as a researcher, I present my deepest gratitude.

Besides my supervisors, many others have helped me in several different ways, during this work. I would like to thank Jose San Pedro and Nuria Olivier, for being wonderful mentors during my experience at Telefonica Research. Jose has been a patient and incredibly valuable teacher. I value and conserve every advice he provided me, and will carry them with me for my entire career.

To Nuria also goes my most sincere gratitude; working with her resulted in an injection of enthusiasm and inspiration, that, as a woman and as a researcher, I will always remember.

I would like to thank both my reviewers, Saturnino Luz and Kristina Jokinen, for their insightful comments and the interesting discussions. I would not be here, writing this acknowledgment, if it wasn't for my former supervisors: Alessandro Lenci, Raffaella Bernardi, Simonetta Montemagni and Massimo Poesio. To all of you: thank you. Thank you for believing in me even when I had lost confidence, for giving me the opportunity to embrace research. Without you, I would have not walked this path.

But a PhD student's life is also made by the amazing people who populate your office, your school, and who make those unforgettable years. So thank you Liliana, Alfredo, Martin, Erwan, Roman, Oscar, Hector, Gerard, Stephan, and Anne for welcoming me when I arrived in Dublin, and for making me feel at home from the very first day. Thank you to Brian, Celine, Noor, Frank, and John for all the time, the laughs, and the discussions we shared in the lab. But thank you also to all the colleagues who arrived after me and brought new laughs to the school: Grace, Carmen, Joao, Christian, Christy, Shane, Akira, Eamonn, Ketong, Fasih, and Fahim. Thank you to the all the school staff for being always so helpful.

I would like to reserve a special thank you to Loredana Cerrato who has read almost all the versions of this thesis with unfailing patience, and has enormously helped me

in structuring my ideas. She has always provided me with constructive feedback and interesting critiques. I am really grateful to have had her support in my last year of the PhD, as a colleague and as a friend. Thank you also for reminding me how wonderful a coffee break is with a real espresso.

Thanks to Emer Gilmartin, who shared with me late evening discussions, conference trips, and weekends on papers. A special acknowledgment also to my office-mate Kevin Doherty. Not only for being a wonderful officemate, but also for the enormous help he gave me in proofreading this thesis. His careful editing contributed enormously to the production of this dissertation.

Finally, I would like to thank all my co-authors. From all of you I learned a lot, to each of you I owe a part of the person and the researcher I am.

Thank you especially to Juan Pablo Carrascal and Rodrigo Oliveira for opening a window into the HCI world, to Alessandro Vinciarelli for interesting discussions, to Ronald Boeck for having shared with me the adventure of organizing two workshops, to Ronald Poppe, Asif Ekbal, Sriparna Saha, Fabio Cavulli, Felice Dell'Orletta, Giulia Venturi, Domenico Carbotta, Eduard Barbu, Egon Stemle, and all the others.

Thank you to all my colleagues at the Institute of Computational Linguistics group in Pisa for introducing me to research, to all my colleagues at the University of Trento, for their uninterrupted enthusiasm and contagious passion for this work, and to all the colleagues who made my experience at Telefonica Research so memorable.

Four years of PhD research, far from home, would not have been the same without all my 'Dublin friends': thank you for the special moments we lived together.

Finally thank you to my entire family. My uncles, aunts, cousins, and grandparents. A special thought to my grandmother Anna. Nonna Anna, thank you for all the time spent with me until your last day.

My deepest, warmest thanks goes to my beloved parents: Claudio e Cinzia. Thank you for having always supported me, for having guided me through life with love and

for being always beside me, in every difficult moment. I owe to you who I am, and everything I have achieved. Grazie di cuore.

Michele, my husband, deserves my deepest thanks for his love, his unfailing support, and continuous understanding. You have supported, motivated, and encouraged me since the first day we met. You have been my pillar and my joy. You have seen me through the ups and downs of this journey, you have shared with me all the successes and the defeats of this trip always at my side with unconditional love. For all this thank you. This thesis is dedicated to you.

This work has been supported by the Innovation Bursary of Trinity College Dublin and the School of Computer Science and Statistics, Trinity College Dublin.

Publications

Related Publications

Journals

1 Francesca Bonin, Nick Campbell, Carl Vogel, *Time for laughter,* Knowledge-Based Systems, Volume 71, November 2014, Pages 15-24, ISSN 0950-7051.

2 Alessandro Vinciarelli, Anna Esposito, Elisabeth Andre, Francesca Bonin, Mohamed Chetouani, Jeff Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, Dirk Heylen, Rene' Kaiser, Maria Koutsombogera, Alexandros Potamianos,Steve Renals, Giuseppe Riccardi, Albert Ali Salah. *Open Challenges in Modeling, Analysis and Synthesis of Human Behaviour in Social Interactions*, accepted in Journal of Cognitive Computation. March 2015

3 R.W. Poppe, R. Boeck, F. Bonin, N. Campbell, I.A. de Kok and D. Traum *From multimodal analysis to real-time interactions with virtual agents.* Editorial introduction

Journal on Multimodal User Interfaces, 8(1):1-3, 2014

Related International Conferences and Workshops

4 Francesca Bonin, Nick Campbell, and Carl Vogel. *The discourse value of social signals at topic change moments.* Accepted in INTERSPEECH, 2015.

5 Emer Gilmartin, Francesca Bonin, Loredana Cerrato, Carl Vogel, Nick Campbell. *What's the Game and Who's Got the Ball? Genre in Spoken Interaction.* Turn-taking and Coordination in Human-Machine Interaction, AAAI 2015 Spring Symposium, April 2015. Accepted.

6 Francesca Bonin, Jose San Pedro and Nuria Oliver, *A Context-Aware NLP Approach For Noteworthiness Detection in Cellphone Conversations.* In proceedings of COLING 2014, Dublin, Ireland, August 2014, pp. 25-36.

7 Bonin, Francesca; Vogel, Carl; Campbell, Nick, *Social sequence analysis: temporal sequences in interactional conversations*, Cognitive Infocommunications, CogInfoCom, 2014 IEEE, vol., no., pp.403,406, 5-7 Nov. 2014

8 Francesca Bonin, Emer Gilmartin, Carl Vogel, Nick Campbell, *Topics for the future: Genre differentiation, annotation and linguistic content integration in interaction analysis*, Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges workshop. ICMI 2014, Istanbul, Turkey

9 Emer Gilmartin, Francesca Bonin, Nick Campbell, Carl Vogel. *Exploring the Role of Laughter in Multiparty Conversation*, In Proceedings of the SEMDIAL 2013 (DialDam), Amsterdam, Netherlands, December 2013, pp 191-193.

10 Emer Gilmartin, Francesca Bonin, Carl Vogel and Nick Campbell. *Laughter and Topic Transition in Multiparty Conversation*, Proceedings of the SIGDIAL 2013, August 2013, Metz, France, pp.304-308.

11 Francesca Bonin, Nick Campbell, and Carl Vogel. *Laughter and topic changes: Temporal distribution and information flow.* In Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on, pages 53-58, 2012.

12 Francesca Bonin, Nick Campbell, Carl Vogel, *Temporal distribution of laughter in conversation*, in Proceedings of the Third Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalisations in Speech, October 2012, Dublin, Ireland, pp: 25-26.

Other Publications of the author

Edited Volumes

13 *Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction.* Proceedings of MA3HMI 2014, satellite workshop of Interspeech 2014. Boeck, Bonin, Campbell, Poppe editors. Springer. In press. ISBN: 978-3-319-15556-2.

14 *Special Issue: From multimodal analysis to realtime interactions with virtual agents, Journal of Multimodal User Interfaces (2014)*, Poppe, Boeck, Bonin, Campbell, de Kok, Traum editors. 8(1), 2014.ISSN:1783-7677

15 *Joint Proceedings of the IVA 2012 workshops: MA3 and RCVA (2012).* Boeck, Bonin, Campbell, Edlund, De Kok, Poppe, Traum: editors. ISBN: 978-3-940961-83-9.

Journals

16 Asif Ekbal, Francesca Bonin, Sriparna Saha, Egon Stemle, Eduard Barbu, Fabio Cavulli, Christian Girardi, and Massimo Poesio, *Rapid Adaptation of NE Resolvers for Humanities Domains using Active Annotation*, in Journal for Language Technology and Computational Linguistics (JLCL), 26(2):39-51, 2011.

17 Bonin F., Dell'Orletta F., Venturi G., Montemagni S. *Singling out Legal Knowledge from World Knowledge: An NLP-based Approach.* In: Informatica e Diritto. Rivista Internazionale diretta da Costantino Ciampi, vol. XXXVI annata -Seconda Serie -Vol.XIX (1-2) pp. 217-232. Edizioni Scientifiche Italiane S.p.A.

International Conferences and Workshops

18 Ans Alghamdi, Francesca Bonin, Asif Ekbal, Sriparna Saha, Fabio Cavulli, Sara Tonelli, and Massimo Poesio, Active Expert Learning for the Digital Humanities, *Semantic technologies for research in the humanities and social sciences workshop*, STRiX 2014. Gothenburg, Sweden.

19 Noor Alhusna Madzlan, JingGuang Han, Francesca Bonin and Nick Campbell. *Automatic Recognition of Attitudes in Video Blogs - Prosodic and Visual Feature Analysis.* In Proceedings of INTERSPEECH 2014, Singapore, Singapore. September 2014, pp. 1826-1830.

20 Noor Alhusna Madzlan, Jingguang Han, Francesca Bonin and Nick Campbell, *Towards Automatic Recognition of Attitudes: Prosodic Analysis of Video Blogs.* In Proceedings of SPEECH PROSODY 2014, Dublin, Ireland, pp.91-94

21 Francesca Bonin, Celine De Looze, Sucheta Ghosh, Emer Gilmartin, Carl Vogel, Anna Polychroniou, Hugues Salamin, Alessandro Vinciarelli and Nick Campbell, *Investigating fine temporal dynamics of prosodic and lexical accommodation.* Proceedings of INTERSPEECH 2013, 26-29 August 2013, Lyon, France, pp. 539-543.

22 Francesca Bonin, Ronald Boeck and Nick Campbell, *How do we react to context? Annotation of individual and group engagement in a video corpus.* In Proceedings of the CBAR Workshop, SocialCom 2012, September 2012, Amsterdam, The Netherlands, pp: 899-903.

23 Francesca Bonin, Fabio Cavulli, Massimo Poesio, and Egon W. Stemle, *Annotating Archaeological Texts: An Example of Domain-Specific Annotation in the Humanities,* in Proceedings of the Sixth Linguistic Annotation workshop, ACL 2012, Jeiu, Republic of Korea, 134-138.

24 Massimo Poesio, Eduard Barbu, Francesca Bonin, Fabio Cavulli, Asif Ekbal, Egon Stemle, and Christian Girardi, *The Humanities Research Portal: Human Language Technology Meets Humanities Publication Archives*, in Proceedings of Supporting Digital Humanities: Answering the unaskable, SDH2011, Copenhagen, DK.

25 Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. *Contrastive filtering of domain-specific multi-word terms from different types of corpora.* MWE 2010 workshop, In 23rd International Conference on Computational Linguistics, COLING 2010 p. 77. 2010.

26 Francesca Bonin, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi. *Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio.* In: SLI-2010 - XLIV Congresso Internazionale di Studi della Societa' di Linguistica Italiana (Viterbo, Universita' degli Studi della Tuscia, 27-29.

27 Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni *Singling out Legal Knowledge from World Knowledge. An NLP-based approach.* In Proceedings of LOAIT 2010 - the 4th Workshop on Legal Ontologies and Artificial Intelligence Techniques (European University Institute, Fiesole, Florence, Italy, July 7th 2010), pp. 39 - 50. Enrico Francesconi, Simonetta Montemagni, Piercarlo Rossi, Daniela Tiscornia (eds.). CEUR, 2010.

28 Francesca Bonin, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi *A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora.* In Proceedings of LREC'10 - Seventh International Conference on Language Resources and Evaluation (Valletta, Malta, 17-23 May 2010), pp. 3222 - 3229.

29 Raffaella Bernardi, Francesca Bonin, Diego Calvanese, Domenico Carbotta and Camilo Thorne, *English Querying over Ontologies: E-QuOnto,* R. Basili and M.T. Pazienza (Eds.): AI*IA 2007, LNAI 4733, pp. 170-181, 2007. Springer-Verlag Berlin Heidelberg 2007.

Contents

List of Tables

List of Figures

List of Abbreviations

R conventions for significance thresholds:

ns	$P > 0.05$
*	$P \leq 0.05$
**	$P \leq 0.01$
***	$P \leq 0.001$
****	$P \leq 0.0001$

Transcriptions convention

(.)	Noticeable pause
(.3), (2.6)	Examples of timed pauses
word [word	Square brackets aligned across adjacent lines denote the start of overlapping talk.

Statistical symbols

$\bar{x}_d$	mean
sd	standard deviation
$\tilde{x}_d$	median

CHAPTER 1

Introduction

1.1 Multimodal and multifunctional interactions

Participation in dialogue is a complex activity that involves sharing and understanding information at many levels, as well as performing actions for pursuing a certain goal.[1] Dialogue participants constantly '*evaluate whether and how they can (and/or wish to) continue, perceive, understand and react to each other intentions*', [Allwood et al., 1997, pp. 20]. They share information about the processing of each other's messages, elicit feedback, manage the use of time, take turns, and monitor contact and attention. These actions are performed in a continuous social interplay constituted by something more than just words [Argyle, 1983]. Participants in a conversation express themselves with **verbal** and **non-verbal** signals in a dance of mutual exchanges, shared with each others at a well defined **timing**.

The results of this interplay are **multifunctional** and **multimodal** dialogue utter-

[1] Goffman [1974] had raised the point that mere co-presence invokes a reaction. A natural question is if this reaction is a conscious or reflective. In this thesis only conscious reactions are considered.

ances. They are multifunctional [Bunt, 2011] because they communicate both linguistic and affective states, and multimodal because participants use all available modalities in order to get their messages across. Besides speech, face-to-face interactions incorporate gestures, facial expressions, head orientation, posture, touch and many other signals. While participants create the interaction, they shape it by using modalities other than the lexical content (tone, intonation, gazes, gestures, etc.) and a complete dialogue model has to take into account the contribution of each of these modalities, as well as their integration. An in-depth investigation of a multidimensional approach to the complexity of natural human dialogue having communicative functions in several dimensions has been reported by Petukhova and Bunt [2009]. The multidimensional view of an interaction embraces a multimodal nature where participants use several types of signals (verbal vocalizations, non verbal vocalizations, gestures, gazes, etc.) to deliver both propositional content (words, sentence, utterances, etc.) and affective states.

As detailed in Chapter 2, many researchers have shown interest in the analysis of conversational interactions from several points of views. Computational linguists have investigated the nature of discourse in conversations, elaborating novel discourse theories, ([Grosz and Sidner, 1986; Poesio and Traum, 1997] among others), and dialogue models [Bunt, 1999]; conversational analysts have explored conversational structures and dynamics [Sacks et al., 1974; Sacks, 1995; Schegloff, 2007]; sociologists have considered conversations as forms of language in interactions [Gumperz and Hymes, 1972; Goffman, 1967, 2008]; others have explored the relation between non-linguistic events and the emotional states they might evoke [Schuller and Batliner, 2013].

In this wide range of disciplines, the linguistic content has usually been regarded in relation to the semantic function of the conversation as conveying the message, while the non-linguistic content has been studied more in relation to the emotional state of the conversations[2]. In other words, there is a separation between the **semantic**

[2]The dichotomy semantic/emotional, is here used to distinguished the propositional function of the

framework/content (words, dialogue utterances, lexical item, linguistic content) and the **social framework/context** (intonation, hesitation, laughter, gazes, gestures, etc.) where the former has been conceived as the semantic channel, while the latter as the social channel. Although, non-linguistic signals are certain source of information on feelings, mental states and personality, in this thesis, it is argued that there is space for an **integrated view**, where content and context are integrated and contribute at the same time to the semantic and emotional level of the conversations.

In line with Petukhova and Bunt [2009], I consider conversations as multidimensional constructs, where the same signal can play different roles and contributes to both the semantic and the emotional level of the conversations. Therefore, I investigate whether signals belonging to the contextual sphere (for example laughter, overlaps, silences, locations, characteristics of the speaker, time of the day), as well as delivering emotional information, are also used by participants as a means to structure the information flow of the conversation.

1.2 Content and Context: real world scenario

To clarify the concept of content and context integration, consider the following scenarios. Student A and student B are talking in the classroom, complaining about the difficulties of the exam they have just taken.

Scenario 1: A, who is facing the door, sees the Professor entering the room and *whispers:* "the Professor is here, let's change topic", to invite B to change the topic.

Scenario 2: A, who is facing the door, sees the Professor entering the room and says *loudly*: Good morning Professor Y, to invite B to change the topic.

conversation from the series of paralinguistic signals that enrich the propositional content with affective functions. However, a clarification over this terminology to distinguished these levels of the conversation is needed. Current work is focusing on discussing this terminology.

Scenario 3: A, who is facing the door, sees the Professor entering the room; he *coughs* and directs a quick *gaze* towards B, to invite him to change the topic.

These three scenarios represent a common situations in our everyday lives. However they differ in the strategies used by the participants to convey information. Common alternatives are used to deliver the same underlying piece of information: *B, the professor is here, let's change topic.* In the first case, the linguistic channel is used in a direct way ("let's change the topic"). However the lower intensity (*whispers*) marks this utterance as different from normal, and the listener will give it special attention. In the second case, there is a sudden change of topic, emphasized by a change of intensity (*loudly*). No evident linguistic information is used to convey the message. In the third, the same message is delivered in an even more subtle way, by means of non-linguistic signals such as coughs and gaze. No linguistic content is uttered by A, however the timing of the non-linguistic content conveys the indirect meaning: *B, the professor is here, let's change topic.* From scenario one to scenario three, progressively less linguistic/semantic information is used, while non-linguistic signals become more and more important in delivering the message.

1.3 Main concepts

In section 1.1, I presented the main concepts that will be key in this work: content, context. The first represents the propositional content of a conversation, semantic framework, while the second, the contextual framework. In this section I provide a brief overview of these two concepts, which are discussed in-depth in Chapter 2.

1.3.1 Content

By content I here refer to the propositional content (words, sentences, conversational turns, etc.). In other words the lexical elements of the conversations' transcripts.

Referring to the six elements according to which verbal communication is exchanged [Jakobson, 1961], the content corresponds to the message exchanged between sender and receiver, as shown in Fig. 1.1. More generally, the message is sent from n participants (where n can be, in principle, indefinitely large)[3] to a groups of receivers as in [Harrah, 1984].

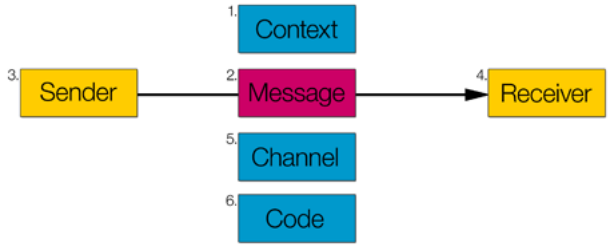


Figure 1.1 – Jakobson's six factors for effective communication.

Specifically, in this work I refer to the *content* of a conversation as the set of messages (box number 2 in fig. 1.1) exchanged by senders and receivers (the participants of the conversation) throughout the course of the conversation in a continuous co-construction of meaning and interaction.

1.3.2 Context

By context of the conversation I refer to the set of non-linguistic verbal signals shared by the participants and the ensemble of situation, background and situated information in which the conversations takes place. Aware of the challenges posed by the concept of context, this thesis defines and restricts the analysis to two kinds of contexts in line with two different perspectives of context in conversation: context as social interaction (called **social context**) and context as situation called **situational context**.[4]. With this distinction, I intent to make a first step towards a taxonomy of the different definitions of context.

[3] Although, with n greater than 5, one could imagine a lecture type conversation.

[4] This can be seen in line with the Jakobson's *context* box (box number 1 in fig. 1.1), defined as the other verbal signs in the message, and the world in which the message takes place.

Social context is the ensemble of non-linguistic verbal signals exchanged by the participants during the conversation. Situational context is the the ensemble of situated information deriving from the environment in which the conversation takes place (location, time of the day, time of the week, but also the relationship between participants, dominance, purpose of the conversation, etc.). An important distinction between these two kinds of context is *timing*: while social context is expressed by signals that are transient and unfold during the conversation, a more permanent information is expressed at the situational level (location, gender, etc.)[5]

Chapter 2 discusses how the first definition is an elaboration of the conversational analysts' concept of context as the social interactions unfolding within a conversation, and the second an elaboration of critical discourse analysts' concept of context as the environmental conditions in which a conversation takes place. For further insights I refer to Chapter 2.

1.4 Content and Context: an integrated view

While the conversation unfolds, participants exchange their speech utterances and gestures, gazes, non-linguistic vocalizations. Observing a human-human conversation as an external observer, one can distinguish two levels in the background: the propositional message exchanged and the context. As previously noted, the context is reflected in two different forms *1)* the social interaction among the participants and *2)* the environmental situation in which this takes place. An attempt to provide a visualization of this structure is made in Fig. 1.2.

In this view the three elements (propositional content, social context and situational context) are integrated in a unique interplay from which the conversation is constructed. The integration of these elements informs what an external observer would

[5]Time of the day is listed among the situational signals. However the time of the day has a different granularity from the time in milliseconds that unfolds during the conversation, hence we can assume that time of the day remains constant during the conversation.

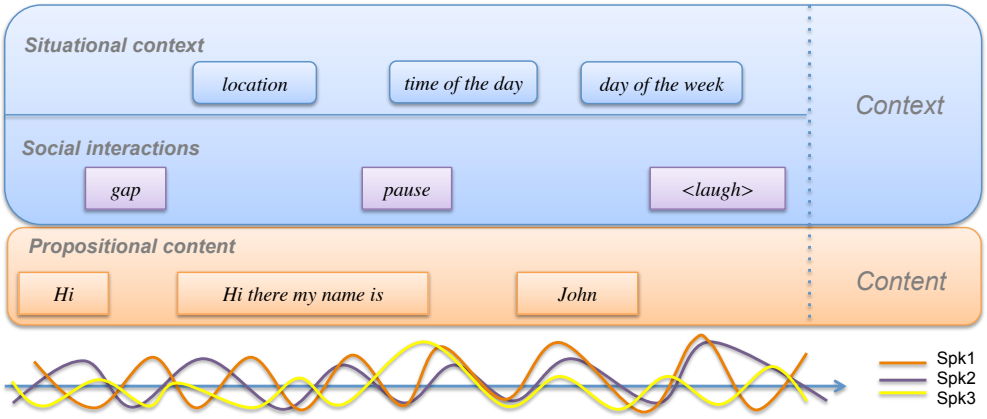


Figure 1.2 – Visualization of the conversation flow with content, social context and situational context viewed from an external observer. The arrows represents the time unfolding during the conversation. The three intertwined lines represents the contribution of the three speakers (spk1,spk2,spk3) who co-construct the conversation by fusing the three levels: propositional content (orange) and the social and situational context (blue).

view as the 'meaning' of the conversation: he would note not just the propositional content of the utterances but also the social impact, fusing this information in an unique message. This view is a generalization of the work of Healey and Vogel [1997] where an investigation of the dialogue's meaning according to the participants and an external observer is studied. While there, only the propositional content is taken into account, here it is argued that the observer's meaning is given by the propositional and the contextual level.

In this perspective, the usual distinction between propositional content having a semantic function and non-linguistic content having a social function leads the way to a novel view in which those levels contribute to both functions. In other words, the multifunctionality of a dialogue utterance is not conveyed through distinct modalities, but several modalities contribute to different functions.

This view brings to examine the conversation from a different perspective than the traditional syntax, semantics, pragmatics distinction, which allows to concentrate the analysis from the perspective of an external observer.

While the focus of attention, i.e. distinguish the part of the context that the participant directs their attention to, is an important research topic for cognitive processing, in this thesis I consider only the conversation as it can be interpreted from an external observer, without considering the perception that the participants of the conversation have from an internal point of view. In line with this, I choose in this work, to study only observable signals rather than those signals which are more prone to interpretation (like goals and intentions).[6]

1.5 Motivation and contributions

The increasing growth of multimodal material has brought to innovative approaches to information extraction from meetings and multiparty conversations. One way to make information from multimodal sources more accessible is to perform **some sort of structural segmentation** of the conversation.

Information retrieval or speech recognition (by means of a more precise indexing of documents), as well as in speech summarization and topic detection tasks, to cite a few, benefit from segmenting the discourse according into different types of units (for example topic, event segmentation). While many works have provided segmentation methodologies for semi-structured, scripted forms of interaction (such as broadcast news), less efforts have been applied to solving the problem of segmentation of spontaneous conversations. Understanding the structure of spontaneous conversations is the focus of several disciplines such as Conversational Analysis, Interactional Linguistics, Computational Linguistics as well as Human-Computer Interaction and Human-Robot Interaction. All these disciplines have investigated conversational structure from different points of view: segmentations based on the pure linguistic exchange (*what has been talked about*), segmentation based on the speaker's role inside the group of partic-

[6] Similarly, the speaker's individual goals and purposes are not considered in this thesis, as they belong to the individual sphere of the participant. Intuitively, they could also form another level of context, an intentional context, which is left to future studies.

ipants and segmentation based on the moment of the interaction (opening vs closing vs body). However, to the best of my knowledge, it is still not clear how diverse co-existing modalities that participants use in spontaneous communication co-participate to define the structure of the conversation, and how those different modalities can provide information on the structural segmentation. While the content of the conversation remains the main indicator and cue for determining the semantically coherent parts of the interaction, contextual factors may also provide useful information.

This thesis shows how social signals and situational information convey a discourse value that can be exploited to derive information about the structure of the conversation.

The main contributions of this thesis are:

i) analyzing the interaction of context and content in spontaneous conversation.

ii) exploring the existence of a discourse function of context (social and situational) in two ways:

1ii) Context as social interactions. I explore whether the social aspect of the conversation (the social context) has a discourse value. In other words, whether speakers use non-linguistic verbal signals to structure the conversation (**discourse function of social context**).

2ii) Context as situation. I explore whether situational aspects (e.g. location, time, speaker relation, etc.) influence the way speakers structure a conversation (**discourse function of situational context**).

1.6 Objectives and Research Lines

The first objective of this thesis is to understand how propositional content, social interactions and situational context fuse together and co-participate in structuring the discourse. The second objective is to investigate whether and how elements belonging to the social level and elements belonging to the situational level provide information

about the conversation's structure. To investigate this, I consider two research lines (hereafter RL) exploring the contribution of:

RL_1: social context on topic segmentation;

RL_2: situational context on event segmentation.

1.6.1 RL_1: Social context and topic segmentation

The first research line investigates whether social context provides useful information on the presence of topic boundaries. One of the ways of determining the structure of a conversation is finding the boundaries between fragments of conversation about the same *topic*. This topic segmentation task has been widely explored in Computational Linguistics and related fields, in particular regarding broadcast news. However, it remains a challenging task when it is related to spontaneous conversations for two reasons. Firstly, the definition of *topic* in spontaneous conversation is less sharp than in broadcast news and, secondly, the noisy nature of the data poses challenging hurdles to pure linguistic based segmentation techniques. In Chapter 2 an extended overview of the state of the art covering topic definition and different topic segmentation algorithms is provided. Chapter 4 investigates whether a correlation exists between social interactions, specifically a subset of social signals (laughter, overlaps, backchannels, silences) and topic changes. It will be shown how significant correlations are found between these non-linguistic vocalizations and the topic boundaries.

1.6.2 RL_2: Situational context and event segmentation.

The second research line investigates whether situational context provide useful information for event segmentation. Another way to determine the structure of a conversation is to segment events in conversation. Specifically, I refer in this thesis to the segmentation between relevant and non-relevant events in the sense of noteworthiness (defining noteworthy as worthy to be annotated/remembered). With this in mind,

I investigate whether information about the situation in which the conversation takes place (location, time, relation among participants, etc) is relevant to the detection of relevant/noteworthy information. Chapter 5 presents an analysis of the correlation between situational information and relevant chunks of conversations, and shows that the use of situational information improves the performance in automatically detecting relevant events from conversations.

1.7 Applications

Multimodal natural-language based dialogue systems are increasingly becoming feasible and attractive human-machine interfaces. Such interfaces offer a model of interaction that has certain similarities with natural human communication, since they use a range of input and output modalities which people normally employ in communication, such as speech, gesture, gaze direction and facial expressions. One of the directions underpinning natural language interfaces is the incorporation of multimodality into virtual environments, for example embodied artificial conversational agents. The design of dialogue systems that exhibit interactive behavior may be expected to benefit from a good understanding of human dialogue behavior and from the incorporation of social mechanisms that are key in human communication. This thesis proposes a novel view of the contextual aspects of communication, exploring their relationship with the discourse structure, rather than with the affective and emotional sphere of communication. When it is socially appropriate to laugh, to interrupt, to maintain a long silence, are all pieces of information that can improve the social intelligence and responsiveness of artificial agents. In Chapter 6, I discuss in more depth the possible applications that could benefit from the know-how emerging from this research.

1.8 Detailed outline

Chapter 1 provides an overview of the thesis, by describing the main objectives, motivations and research questions. A high level description of the mechanisms of human-human interactions is given and the focus of this work is described. How three levels (propositional content, social interactions and situational context), intertwining, influence the discourse structure, as well as how they contribute to delivering the message of a conversation, will be the focus of the thesis. In Section 1.6, two complementary research lines are presented: RL_1: investigation of the correlation between social context and topic segmentation; RL_2: the correlation among situational context and event segmentation. Finally in Section 1.8, a short description of each chapter's content is provided.

Chapter 2 provides a description of the background behind the analyses of social interactions. First, an historical background of the different disciplines which have considered human-human interactions is given. Secondly, two main concepts (*content and context*) and the theoretical debate behind them are discussed. Finally, I describe the state of the art for the two lines of research that are the focus of this thesis: RL_1 and RL_2. An overview of the definition of topic together with the working definition in this thesis is provided, and the current state of the art of topic segmentation is described. While many approaches to topic segmentation are provided, a methodology that makes an explicit and intentional use of social signals has not been explored yet. A description of related work regarding event detection, in particular noteworthy events, is provided.

Chapter 3 provides a description of the data explored in this thesis, consisting of three corpora: the AMI corpus, the TableTalk and the Callnotes corpus. The three corpora represent an increasing spectrum of spontaneity from AMI to Callnotes, being the AMI the more controlled scenario and Callnotes the more un-controlled one.

Chapter 4 provides the analysis related to RL_1 and investigates the discourse function of social signals. The relation between social signals and a discourse phenomenon such as topic changes is explored, investigating whether social signals have a discourse function in addition to their social function. Different analyses that investigate the temporal dynamics of laughter, backchannels, silences and overlaps, are described, finding a relation between topic changes and a decrease of social signals. Specifically, it is found that immediately after a topic change there is a significant drop in social activity, defined as interactional entropy.

Chapter 5 provides the analysis related to RL_2 and investigates whether and how situational signals influence the prediction of events in social interactions. The focus is on noteworthy events, defined as events worth remembering, and the relation between situational information and the distribution of noteworthy events in telephone conversations. I propose two analyses: a correlation analysis and a classification analysis. A correlation between situational signals and noteworthy events emerges from both the analyses. I conclude, therefore, that situational signals are correlated with the distribution of noteworthy events.

Chapter 6 This Chapter summarizes the main findings of the thesis, and presents several applications that would benefit from such findings. Finally potential directions for future work are discussed.

1.9 Summary

This chapter provides an overview of the thesis, describing the main objectives, motivations and research questions. A high level description of the mechanisms of human-human interactions has been given and the interest of this work has been described in the relation of the social context and the situational context of conversation with the discourse structure. How these three levels (content, social context and situational context) intertwine and influence the discourse structure and the way in which they contribute to construct the message of a conversation is main focus of the thesis. Two complementary research lines have been presented: RL_1: investigation of the correlation between social context and topic segmentation; RL_2: the correlation between situational context and event segmentation. In the next Chapter, I provide an overview of the state of the art relevant as background to this thesis.

CHAPTER 2

State of the Art

2.1 Introduction

In this chapter, an overview of the historical background behind the analysis of social interactions is provided. The chapter is structured in three parts. The first focuses on the historical background of the different disciplines which have considered the analysis of human-human interactions; the aim of the section is to offer an overview of the historical context and of the academic discussion on the topic, rather that analyzing in detail the historical connections among the different thinkers. In the second part, a discussion of the two main concepts of this thesis in their historical framework is presented: an overview of the state of the art on discourse theory is provided in order to situate the study of content, then, the definition of context and the numerous interpretations provided by different research lines (conversational analysis, interactional sociolinguistics) is presented. As anticipated in Chapter 1, this thesis follows two research lines, aiming at answering different, though related, research questions: RL_1,

the analysis of social signals and topic changes and RL_2 the analysis of situational signals and event detection; in the third part of this literature review, I focus on these two directions; in relation to RL_1 I explore the different approaches to the definition of topic and the state of the art techniques for topic segmentation, and in relation to RL_2 I report the current state of the art on event detection, discourse summarization and in particular noteworthiness detection.

2.2 Historical Background

Social interactions, in particular human-human conversations have been the focus of interest of many different research fields, from anthropology, linguistics and ethnomethodology to conversational analysis and interactional sociolinguistics.

2.2.1 Etnomethodology

The interest in the analysis of conversations and human-human interactions grounds its roots in sociology, particularly in the work of the american sociologist ethnomethodologist Garfinkel [1967] who inaugurates the study of everyday life as a research field in its own right, moving on from the post World-War-II sociological conception that everyday lives were too random to support systematic analysis.

With the term *ethnomethodology*, Garfinkel refers to the study of methods which societies use to establish an internal social order. Looking at the day-to-day experiences of people, in their daily interactions, ethnomethodology explores the social orders that societies are able to create through natural language, considered the main form of social interaction which allows mutual understanding, mutual engagement, in other words that leads to a *social order*. Garfinkel also suggests that the knowledge of a member of a group is never de-contextualized; rather knowledge and action are deeply linked and constitutive of each other. The fact that participants are aware of the surrounding circumstances provides for the stable organization of their activities; this understanding

emerges from a continuos mutual language exchange and interpretation. His sociological research represents the premises for a novel view of conversations as part of social interactions and has influenced both interactional sociolinguistic and conversational analysis contributions.

2.2.2 Interactional sociolinguistics and ethnography of communication

The interest of examining linguistic exchanges as part of social life emerges from the work of the linguist anthropologist Gumperz and the sociologist Goffman [Gumperz, 2008]. Goffman describes how language is situated in particular circumstances of social life and how it reflects and adds meaning and structure to those circumstances. Although a clear border between different currents is not well established, and it is beyond of the scope of this thesis to provide one, the work of Gumperz, Hymes and Goffman [Gumperz and Hymes, 1972; Goffman, 1967, 2008] comprises the body of research that has been known as interactional sociolinguistics, as the research line focusing on language in its social context, the language used in interaction by closely observing a 'speech event' in a particular community. In the words of Gumperz [2008, pp. 215] :

> Interactional sociolinguistics is an approach to discourse analysis that has its origin in the search for replicate methods of qualitative analysis that account for our ability to interpret what participants intend to convey in everyday communicative practice.

In this, it is evident the influence of the ethnography of communication approach [Hymes, 1964] who fosters the study of social interactions by concentrating on the situation in which the conversation takes place. More recently interactional sociolinguistic has been continued by Goodwin [1981]; Tannen [1989]; Schiffrin [1994]. As shown in section 2.3.2.2, the concept of situational context and of situational signals exploited in this work is built on the interactional sociolinguistic concept of *talk in situation*.

2.2.3 Conversational Analysis

Conversational analysis (CA) is the field of research which focuses its attention on social interactions. According to Heritage [2008], CA inherits from Goffman the concept of *talk-in-interaction*, and from Garfinkel the idea that conversations are *ethnomethods*, defined as methods used to assess social order. However, the main focus of CA is on identifying the internal order of social interactions, namely conversations. This field of study emerged from the work of Harvey Sacks, in collaboration with Emanuel Schegloff and Gail Jefferson, in the early sixties. The first approach to conversational analysis emerged from the privately circulated lectures of Sacks [1995], followed by early publications in sociological journals [Schegloff, 1968; Schegloff and Sacks, 1973]. For this reason, the field became visible to sociologists as a continuation of Garfinkel's ethnomethodology work. While sociologists like Garfinkel were interested in the study of social order *per se*, in CA, researchers are interested in exploring the social order within human-human interactions. The structure and the unwritten laws (turn taking, overlapping, etc.) which rule human-human conversations are, thus, the focus of interest of conversational analysts. Sacks, Schegloff and Jefferson introduced a novel methodology in the analysis of social interactions, that consists of data collection (recording and transcription of conversations) and a rigorous study of the rules implicit in such transcriptions, such as the turn taking algorithm, the opening/closing pattern, etc. Heritage [1984, pp. 241] points out the main assumptions which form the basis of conversational analysis:

1) interactions are structurally organized (see the turn organization in Sacks et al. [1974]);

2) contributions to interaction are contextually oriented (participants are aware of the situation);

3) the first two properties make it possible for conversations to provide a social order.

Conversational analysts explore the structure of conversations by looking at rules, patterns, common behaviors in the empirical conduct of speakers. This thesis is framed within the context of conversational analysis, in its aim of understanding the conversations' structure and organization that the participants create by intertwining different modalities during social interaction.

2.2.4 What is a conversation?

Despite being the center of analysis of many research fields, defining a conversation is not an easy task. Goodwin [1981] starts his *Conversational organization: Interaction between speakers and hearers* by trying to define what a conversation is. He refers to Goffman [1976] who discusses two different approaches to the definition of conversation: *casual talk in everyday setting* or the *equivalent of spoken interaction.* However, Goodwin notices how conversation is also a special case of what Goffman [2008] had previously defined *focused interaction* as the kind of interaction that occurs when persons gather close together and openly cooperate to sustain a single focus of attention. This is in contrast with unfocused attention created by mere co-presence of participants. Goffman also notices that, though the conversation is defined in terms of talk, it can include behaviors other than talk, linguistic and non-linguistic behaviors, verbal and non-verbal.

2.2.5 Verbal vs non-verbal and linguistic vs non-linguistic events in conversations

In social interactions, a distinction between verbal and non-verbal signals has to be made and requires an in-depth reflection. Verbal signals refer to everything expressed through the audio channel (words, laughs, backchannels where audible, coughs), while non-verbal signals refer to all the other signals (facial expressions, gestures, postures, gazes, etc.). Within these categories, a further classification is necessary, as both verbal

and non-verbal signals can be linguistic and non-linguistic. Linguistic verbal signals are the sounds participants utter with an intentional linguistic meaning (e.g. at a syntactic level: the words), non-linguistic signals are the sounds participants emit which do not have a linguistic value (e.g. a sneeze).

Similarly, non-verbal signals can be linguistic (such as gestures having an intentional linguistic meaning, e.g. those used in substitution of words) or non-linguistic (such as a gesture or a posture which does not convey any linguistic meaning) [McNeill, 2008]. These four categories play a crucial role in the overall meaning of the interaction and clarity regarding their definition is therefore necessary. However, many different names have been used. Schuller and Batliner [2013] talking about non-linguistic signals refers to paralinguistics, intended as everything that belongs to the acoustic channel (viz verbal), and extra-linguistics,intended as everything belonging to the non acoustic signals (viz non-linguistic non-verbal signals).

> [...] defining paralinguistics as the discipline dealing with those phenomena that are modulated onto or embedded into the verbal message, be this in acoustics (vocal, non-verbal phenomena) or in linguistics (connotations of single units or of bunches of units). Thus, I restrict the term to everything that can be found in the speech signal, e.g., in telephone speech or in audio recordings, which cannot be described only in strictly phonetic and/or linguistic terms. [Schuller and Batliner, 2013, pp. 5]

In this thesis I consistently use the terminology: (non-) linguistic and (non-) verbal, to refer to these four categories. However, a differentiation between verbal, non-verbal and linguistic, non-linguistic is not always easy. As noted by Rosch [1975, 1999] a fringe example is constituted by filled pauses, that, although always considered as verbal non-linguistic phenomena, normally follow the phonotactics of the native language, and cannot be placed everywhere. They respect thus, linguistic rules, and they are considered as words in automatic speech recognition.

The difficulty in marking an exact boundary between linguistic and non-linguistic signals is a fertile ground in which the following hypothesis can be explored. If verbal non-linguistic signals follow linguistic rules as well as phonotactic rules of the language of the speaker, they also might follow discourse/syntactic rules (e.g. laughter to mark a topic change, longer pauses to mark a topic change, gazes to mark relevant parts of the conversation). Thus, the boundary between linguistic and non-linguistic elements of a conversation can be considered as a blurred demarcation line. In Chapter 4, I investigate the discourse function of non-linguistic verbal signals, showing how they do play a role at the discourse level.

2.3 Historical background of content, social context and timing

In this section I provide an overview of the theoretical background behind two main concepts of this work: *content and context.*

2.3.1 Definition of content

Van Dijk [1977], define discourse as :

> [...]utterances of natural language may be theoretically reconstructed as sequences of sentences, in which morpho-ponological, syntactic and semantic properties of a sentence are accounted for in relation to this of other sentences of the sequence. [...] The sequence is also been studied in its own right, viz as discourse. [Van Dijk, 1977, pp. IV]

In the recent years, there has been a wide interest in discourse analysis, and more recently in discourse analysis of speech, particularly at how discourse emerges beyond the sentence, how humans are able to create coherent discourse by combining pieces of information, how topics flow within a discourse, etc. As noted by Schiffrin [1994], one

can distinguish two main trends in discourse analysis, **the formalist** approach, that looks at discourse as language beyond the sentence level (but still structured), and the **functional one** which deals with actual language use (and takes utterances as the basic unit). In the first case, there has been a primary interest in understanding the structure of the conversation and how sentences (or speech production) are connected in order to create a coherent discourse. Many theories have been developed to understand the discourse structure,[1] such as Hobbs [1985] the Rhetorical Structure theory by William and Thompson [1988], the Grosz & Sidner theory [Sidner and Grosz, 1986], the DRT on a more formal semantic level [Kamp et al., 2011].

With respect to the use of language, many works have been conducted in psycholinguistics and sociolinguistics in order to better understand discourse production, comprehension and acquisition.[2] Anthropology, the study of rhetoric, sociology and literary scholarship, have also produced interesting works related to the analysis of discourse. Anthropology has paid attention within the field of ethnography of speaking analyzing discourse types in different cultures and societies [Maranda, 1972; Van Dijk, 1977]. Finally, social psychology has examined the effect of discourse and its *content* in relation to beliefs and behaviors of individuals in society in the frameworks of mass media messages [Hovland, 1966; Van Dijk, 1977].

One of the methods of analyzing the discourse structure is by segmenting it into coherent fragments, but degrees of freedom in the definition of coherence can lead to different fragmentations. A reasonable fragmentation would segment the discourse into topically coherent segments, by detecting topic changes. In Chapter 4, within RL_1, I discuss whether and how social context generated by non-linguistic verbal signals (viz social signals) has a discourse function in marking topic changes, and can therefore aid in their prediction.[3] Another method of analyzing discourse is by detecting relevant

[1] For a wider overview please refer to Ghosh et al. [2012].
[2] For an overview over this literature refer to Balota and Marsh [2004]
[3] The fuzzy definition of topics is discussed in section 2.4.1.

events. In Chapter 5, within RL_2, I investigate whether and how situational context (where, when, between who the conversation takes place) influences the way participants structure the discourse, and whether it could thus help with the detection of relevant noteworthy events.

2.3.2 Definition of context

Context is a multifaceted concept that has been studied across different research disciplines, including computer science (in particular with artificial intelligence) cognitive science, linguistics, philosophy, psychology, and organizational sciences [Adomavicius and Tuzhilin, 2005]. Each discipline tends to narrow down the standard dictionary definition of context to *conditions or circumstances which affect something*, [McKechnie, 1983]. Therefore, there exist many definitions of context even within specific subfields of these disciplines. Bazire and Brézillon [2005] present and examine one-hundred and fifty different definitions of context from different fields.

In linguistics, the relation between language and context has been a key concept both in pragmatics and in ethnographic studies [Goodwin, 2003], and scholars in the last twenty years have led towards a more dialogically conceived notion of contextually situated talks. The attention devoted towards context started with anthropological linguists in the mid sixties[4], when scholars began to consider languages in the social context as a form of social engagement. Similarly, Ochs and Schieffelin [2001], have shown how the process of language acquisition in children has to be considered as language socialization, as it is not only learning a language, but learning a set of social rules within the language. Language, thus, has often been studied in context, or to say: in a **situated discourse**. In this framework, as stated by Duranti [1999, page 2]:

> Providing a formal definition of a concept can lead to important analytic insights. [...] However it might not be possible at the present time to give a

[4]See Section 2.2.2 for further details.

> single precise technical definition of context, and eventually I might have to accept that such a definition might not be possible.

While there is a general consensus regarding the importance of context in linguistics, researchers often disagree on what should be considered as context. In fact, the way in which context is treated, distinguishes different research traditions [Tracy, 1998]. Conversational analysts, for example, consider context those elements that belong to the sequential unfolding of the interaction (Section 2.3.2.1), and believe that other factors like gender, social background, location, political context, should only be taken in consideration if they are made relevant within the interaction. Critical discourse analysts, on the other hand, argue that it is the connection between social background, location, time, gender/political/cultural aspects, and the linguistic interaction that needs to be examined most explicity (Section 2.3.2.2).

From a different point of view, context has also been a key notion within the context-aware computing literature. In an early work, Schilit and Theimer [1994] define context as *location and the identity of nearby people and objects.* Only several years later Dey moved from a definition-by-example to a more abstract definition [Dey, 2001, pp. 3]:

> Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application

This is probably the first definition to be broadly adopted within the computational sciences. Dourish [2004] further analyzes the definition of context in computational environments, distinguishing two main views of context: **representational** and **interactional**. The representational view separates context from action. In some sense context defines an action and provides some form of information about it. This view also assumes that context could be defined before an action and has a single interpretation across various activities. On the opposite hand, the interactional view defines context as

a relational property. Moreover, the scope of the context is defined dynamically, therefore no enumeration of contextual conditions is possible beforehand. Dourish [2004] explains that context arises from the activity, it is actively produced, maintained and enacted in the course of the activity at hand. In many context aware computation systems (eg. recommendation systems) the representational view is adopted. Contextual conditions are enumerated beforehand by a system architect or expert of the domain. This provides a much simpler and computationally feasible approach, compared to the interactional view.

A parallelism emerges. The conversational analysis perspective on context is modeled by the interactional view, as it better represents the dynamic nature of the interaction. The context emerges dynamically from the interaction among participants, hence, no prior enumeration of the contextual information is possible. On the other hand, the Critical Discourse analysis perspective finds a good fit with the representational view, as the environmental aspects are separated by the action (in this case the conversation) itself, but provide information about it.

On the basis of this linguistic/anthropologic discussion, as well as on the computational literature, lie the two interpretations of context defined here: the social context (closer to the conversational analysis definition) of signals exchanged by participant during the interaction, and the situational context as the ensemble of geospatial, temporal, social, political, cultural settings in which the interaction takes place (closer to the critical discourse analysts definition.). In detail:

Social context is the ensemble of the social signals exchanged by participants, where by social signals I refer to non-linguistic signals (laughter, gaze, head movements, gestures, backchannels, hesitations, etc.) performed during the interaction. The timing of social context is the discourse timing in a continuous unfolding of events. Social signals are described in Section 2.3.2.1.

Situational context is the ensemble of the environmental conditions in which the conversation takes place (location,time, relation between participants, gender, age, sociopolitical conditions etc.). These signals reflect the social situation, the environment in which the conversation takes place as in the interpretation of context made by interactional sociolinguists and critical discourse analysts. Situational signals will be described in 2.3.2.2.

Relating with Dourish [2004]'s distinction, the social context is represented in an interactional sense, while the situational context is represented in a representational sense. In Chapter 4 and 5, subsets of social signals and of possible situational information are considered.

2.3.2.1 Social Context through Social Signals

Conversational analysts consider context as the set of signals evoked in the sequential unfolding of the conversation, these non-linguistic signals (both verbal and non-verbal) that participants exchange in interactions. The study of both verbal and non-verbal non-linguistic signals has been of interest to many research fields during the years, such as pragmatics, ethnography of communication, linguistic anthropology, sociology of language, sociolinguistics. Recently it has also attracted the interest of a new research line which goes under the name of **social signal processing**, that defines the novel concepts of social signals, or behavioral signals.

Social Signal Processing refers to social signals as the verbal and non-verbal non-linguistic signals exchanged in communication (for example: laughs and coughs, as well as gazes, gestures etc). The field finds an early predecessor in Malinowski [1947], who, for the first time, used the term phatic communication in linguistics, referring to those expressions which are expressed to construct a social bond with the interlocutor. The first main introduction to pragmatics is provided by Levinson [1983], but only Mey [2001] introduces, for the first time, social aspects to the study of pragmatics. In 2007,

Pentland [2007] describes the use of social signals to better understand human-human interactions. He does not provide a definition of social signals, but intuitively refers to the gestures, facial expressions, gazes, vocal prosody etc, that participants exchange during an interaction. Pentland distinguishes between two signaling frameworks in which social interactions have been addressed. The first comes from cognitive psychology and focuses on emotions: people perceive others' emotions through stereotypes facial expressions, tone of voice etc, so social signals can be useful in understanding participants' emotions. However, participants are able to hide their emotions, so their measurement is not easy. The second framework comes from linguistics and treats social interactions, where vocal prosody and gestures are features of the linguistic channel. Pentland argues that an alternative computational framework is social signaling in which social signals are used to convey participants' intentions and attitudes. This is different from the linguistic case because it considers purely non-linguistic elements of conversation, and is different from the emotional one in that it does not relate to the inner emotion of the speaker, but to the intention that the participant wants to convey. Pentland's distinction between emotions, considered as inner feelings, and attitudes, considered as the intentions that the speaker wants to convey (which is not necessarily his inner feeling), has been exploited in Madzlan et al. [2014b,a].

Pentland [2007, pp. 108]'s words present of social signals as contributing to the understanding of a conversation at the same level as the discourse:

> Social signaling is what you perceive when observing a conversation in an unfamiliar language and find that you can still see someone taking the charge of a conversation or establishing a friendly conversation.

A systematic overview of what has then become a proper research field known as Social Signal Processing (SSP) is given by Vinciarelli [2009] and by Pantic et al. [2011]. Social signal processing is here defined as a technological domain that aims to provide computers with the ability to sense and understand human social signals. During so-

cial interactions, non-linguistic behaviors convey information for each of the involved individuals and determine the nature and the quality of the social relations. This is achieved by means of a wide range of behavioral cues that are displayed and perceived almost unconsciously. Vinciarelli [2009] provides a taxonomy of behavioral cues and their related social signals. In particular, the author lists: *physical appearance*, *gesture and posture*, *face and eye behaviors*, *vocal behavior*, *space and environment*. Table 2.1 lists the signals related to each category.[5]

Behavioral cue	**Social signals**
Physical Appearance	*Height*
	Attractiveness
	Body shape
Gesture and Posture	*Hand gesture*
	Posture
	Walking
Face and Eye Behavior	*Facial expression*
	Gaze Behavior
	Focus of attention
Vocal Behavior	*Prosody*
	Turn Taking
	Vocal outburst
	Silence
Space and environment	*Distance*
	Seating arrangements

Table 2.1 – Behavioral cues Vinciarelli [2009] - reduced table.

In this work, the term social signals will be used hereafter to indicate the verbal non-linguistic phenomena that, in my hypothesis, can have a function at the discourse (hence, in fact, linguistic) level. I direct the reader to Chapter 4 for a deeper description of the signals analyzed in this work.

2.3.2.2 Situational Context through: Situational Signals

Tracy [1998] describes a view of context at a more extended level, namely as the so-

[5]The original table presented in [Vinciarelli, 2009] presents the behavioral cues in relation with the related social behaviors and technologies.

cial/political/cultural situation in which the conversation takes place.

Once again, an early predecessor is Malinowski, who [Malinowski, 1947, pp. 306] introduces the concept of *context in situation*, indicating the conditions under which a language is spoken. From an anthropological perspective, in the seventies, Gumperz and Hymes [1972] propose that in order to understand a culture, it is necessary to pay attention to the *speech events*. It is thus stated that, from social interactions, it is possible to understand not only grammatical competence, but also insights on the cultural and social context in which the conversation is settled. Interesting is the SPEAKING grid by Hymes [1972] an acronym to help analyst to remember the key elements to take into consideration in analyzing speech events: Setting, Participants, Ends (as goals), Actions, Key (as manner and tone) and Instrumentalities (as channel of communication used), social and cultural Norms of the interaction, and the Genre (as text type) of the text. Later on, Kendon [1990] introduces the idea of context as the (back)ground against which a text becomes interpreted. However, this background is a quite fuzzy concept, that fosters the discussion in many directions [Kendon, 1990; Duranti and Goodwin, 1992]. These works underline three key concerns, that can be summarized as follows: firstly, they recognize a recurrent methodological problem in determining at what level to define the situation one studies; secondly, they underline the existence of an increasingly complex understanding of context; thirdly, they stress the important role played by the discussion of context in understanding the relation among language structure, social organization and culture [Duranti and Goodwin, 1992]. Others have tried to put order in the context definition using layers; Fetzer [2004] distinguishes a linguistic context (genre, intonation), social context (participants, roles, situations), socio-cultural context (organizational dimensions) and cognitive context (memory, prior knowledge, mental representations).

Finally, it is worth reporting the concept of situations in Natural Language Semantics [Barwise and Perry, 1981; Kratzer, 2014]. Situation semantics was developed as an

alternative to possible worlds semantics. In situation semantics, linguistic expressions are evaluated with respect to partial, rather than complete, worlds, in other words with respect to a situation. The scope of an expression is restricted to a particular domain, or situation.

In this works situational context is studies through situational signals. These include, for example: the gender of the participants (if known); the location in which the conversation takes place (outdoor/indoor or the exact geo-spatial location if known); the number of the participants, the time of the call, the day of the call, the relation among the participants in a telephone conversation belong to the situational context. In Chapter 5, this definition is used to explore to what extent the situational context influences the conversation structure.

2.4 State of the art of topic segmentation and event segmentation

The following section presents the background on which the research lines RL_1 and RL_2 (described in Section 1.6) are based. In relation to RL_1, this section provides an in-depth analysis of the definition of *topic* over time, related to the field of study (Section 2.4.1) and it provides a definition of topic in social interactions as applied in this thesis.

Thus, in Section 2.4.3, an overview of the state of the art of discourse segmentation is provided. In relation to RL_2 an overview of discourse noteworthiness detection studies is addressed in Section 2.4.4.

2.4.1 Definition of Topic

There are few terms in linguistics that have been as problematic as topic, and few terms have had such an influence in a wide range of disciplines from linguistics, to artificial intelligence, text analysis, discourse analysis, etc. Different disciplines have approached the problem from various perspectives, tackling the concept of topic at different levels,

exploiting a wide variety of phenomena such as referentiality, predication, information structure [Sornicola, 2006]. The aim of this section is to give a structured description of the different theories and different definitions of topic, developing a conceptual map which would facilitate a definition of topic that better suits our framework.

2.4.1.1 Beyond an intuitive notion of topic

Fragments of a discourse or conversations are the working material and source of interest for any discourse or conversational analysts. To determine those fragments of conversation a segmentation process is implicitly necessary. Although segmenting a conversation on some form of coherence is an essential step for structuring the conversation or the discourse, it is difficult to decide what constitutes a satisfactory unit for the analysis. The type of decision that is typically made is to rely on an intuitive notion of topic. Participants stop talking **about** X, and start talking **about** Y. This notion of topic is merely an arbitrary way to unify a chunk of text or discourse as being about something, and the next chunk about something else. An in-depth formalization of the concept of aboutness is provided by Ginzburg [1995a], who assesses an extensive theory of aboutness by introducing a semantic theory of questions.

2.4.1.2 Topic and discourse analysis

In the notion of topic one can distinguish *sentence* topic, what is predicated about an entity in a sentence and *discourse* topic, what a part of a discourse is about, [Van Dijk, 1981], with respect to the unit under consideration. An early definition of topic was provided by Hockett [1958, pp. 201] in the context of American Structuralism and can be recalled in terms of *topic* opposed to *comment*, ie. as **what is being talked about** opposed to **what is being said about what is being talked about**. This defines topicality in terms of aboutness. The same line is followed by Lambrecht [1994], whose definition of *topic as about* is a milestone in the entire discussion on topicality.

> A referent is interpreted as the topic of a proposition if in a given situation the proposition is construed to be **about** this referent, i.e. as expressing information which is relevant to and which increases the addresses's knowledge of this referent. [Lambrecht, 1994, pp. 131]

In his contribution, Lambrecht states that for an element to be referred to as topic it has to be referential, Topic-of a proposition, where the proposition is what is talked about (the comment). In this view, the existence of topic is strictly related to the existence of comment. In many approaches, a correspondence, even if not an exact correlation, has been assumed between Topic-Comment and Subject-Predicate, [Lambrecht, 1994; Strawson, 1974]. Yet this association is problematic, as together with obvious similarities, sentence topics can be expressed in different ways; for example in different syntactic movements, such as passivisation, left dislocation and special syntactic construction. E.g. 2.4.1-a and 2.4.1-b.

2.4.1-a ***The new president*** *(subj+topic) has been strongly criticized for his foreign policy.*

2.4.1-b *The people (subj) criticized* ***the new president*** *(topic) for his foreign policy.*

2.4.1-c *For his foreign policy (topic), the* ***the new president****(subj) was criticized*

Sornicola [2006] underlines the differences between those two concepts as belonging to distinct levels of analysis: the couple Topic-Comment anchored to the pragmatic dimension and the pair Subject-Predicate to the syntactic one. An example, given by Sornicola [2006]:

2.4.1-c *John loves the sea.*

2.4.1-d *S[NP[John] VP[V[loves]] [NP[Det[the] NP[sea]].*

2.4.1-e *As for John, I am telling you, John loves the sea.*

The simple utterance in Ex:2.4.1-c can be described at the syntactic level as in Ex:2.4.1-d (John being the subject), or at an informal pragmatic level as in Ex:2.4.2-e (John being the topic, and 'John loves the sea' being what I say about the topic). Topic-Comment belongs to the pragmatic level, even if there may be a connection with the syntactic level. In other words one cannot state the equivalence *topic == subject*, but can exploit this connection in the particular syntactic cases in which the relation holds.

Also interesting is the definition by Chafe [1976] who represents topic in terms of temporal and spatial frame of a sentence. In this view, topic is a unit which sets a spatial, temporal, or individual framework for the main predication (i.e. the comment), and which is partly integrated into the sentence structure. In other words, the topic is a set of **domain** and **context** specific constraints for the comment. This view is of particular interest, and anticipates, at a sentence level, notions that will be made explicit at the conversational level.

> What the topics appear to do is limit the applicability of the main predication to a certain restricted domain. . . . The topic sets a spatial, temporal, or individual framework within which the main predication holds.[Chafe, 1976, 464]

The definition given by Chafe underlines that topic (even sentence topic) can be considered as a discourse notion, that serves to define the center of attention of the sentence (Chafes's functional role of the topic). Finally, Grobet [2002] defines topic as an anchor between discourse units. Her definition lies at the micro level of the sentence, but it strictly relates with the discourse level. In both Chafe's and Grobet's definitions, the notion of topic emerges at the sentence level, for finding a more complete realization at the discourse level.

At the discourse level, topic has been the center of attention of both discourse theories and conversational theories. In discourse analysis, many have tackled the problem of topic and content representation of a text. The first approach to discourse topic is

the one by Keenan [1976]. This study introduces the concept of discourse topic emphasizing the fact that discourse topic is not a simple NP, but a proposition about which some claim has been made. Focusing on children's speech, they define the topic in terms of *question of immediate concern* expressed by a sentence. Their implication is that for each fragment of a discourse there must be an idea which represents the entire topic of the fragment. Brown and Yule [1983] introduce the notion of topic framework, that makes a characterization of a topic on the base of the context. A topic framework can be described as a the set of features required to answer the question *what is the text talking about* and it depends on both the speaker/writer and on the hearer/reader (what the hearer/reader knows about the speaker/writer). In fact, their notion of topic framework is related to why the speaker said what he/she said in a particular context and to a particular hearer. Finally, Givón [1983] analyzes the notion of topic and comment and the information structures of the discourse taking into account the coherence among multiple sentences of the discourse. Givón believes that the definition of topic as an atomic singular and discrete unit at the sentence level (the so-called micro topic) should be replaced by a model that represents the topic of multi-sentence paragraphs and their continuity through the text.

The notion of discourse topic in the framework of discourse analysis implies taking into account the notion of sentence topic (what is the sentence about) and the continuity of this topic in a coherent discourse. In all previous approaches, the sense of aboutness, formalized in Ginzburg [1995a,b], plays a main role in the definition of topic, but its meaning is enriched in the context of text coherence: topic is referred to as a chunk of text which shows coherence in term of aboutness, where coherence on "what we are talking about", can be obtained in different ways. Halliday and Hasan [1976] talk about cohesion in English as a means to coherence. According to their view, lexical cohesion is created by the use of lexically cohesive relations (repetition of the same word, the use of a synonym for a word, the use of a superordinate for a word and the use of

collocations). Building on this concept, Morris and Hirst [1991] showed that the lexical chains in a text tend to reflect the discourse structure of that text. That is, the pattern of topics and subtopics in a document is similar to the pattern of occurrences of different elements of the lexical chains in that document. In this sense topic can be seen as emerging from a metaphorical chain lexical elements that evolve throughout the text.[6] Finally, it is worth noticing the distinction underlined in [Jokinen and Wilcock, 2006] between topic of a conversation, in the Ginzburg's sense of aboutness, and *newsby*, as the new information about the topic.

2.4.1.3 Topic and conversational analysis

In conversations, the concept of topic presents different challenges. First of all topics are constructed by interaction between participants. Topics are constructed, not pre-existent, and things a participant wanted to say may not get said because the topic flow has taken a different direction from that of our previous intervention. Here the attention on a possible sentence topic has given way to that of the *speaker's topic*, and then to the notion of the mutual construction of a discourse topic in conversation by two or more participants (Chafe [1997] among others). In Sacks' words:

> Talking topically doesn't consist in blocks of talk about a topic. When you present a topic you can be assured that others will try to talk topically with what you've talked about, but you can't be sure that the topic you intended was the topic they will talk to.[Sacks, 1995, pp. 762]

Sacks [1995] underlines how topicality in spontaneous conversations involves both a cognitive and an interactive process. It is not something predetermined and already known to the participants in the conversation, but an achievement worked out by their mutual negotiation of the topic. The focus on interaction is visible in a shift from the

[6] The concept of lexical chains will influence also one of the first topic segmentation algorithm [Hearst, 1997].

what perspective to the *how* perspective: topic is not only represented in terms of *aboutness*, but also in how it is manifested and signaled.

To conclude this Section, in Fig. 2.1, a figurative conceptual framework of the definition of topic is provided.[7]

2.4.2 Working definition of topic in this study

As emerged from the discussion in the previous sections, topic is a complex term whose definition has been tackled by many scholars.

In this work, I consider topic in conversational interactions, taking from Sacks [1995] and his definition of topicality which establishes the importance of the context, situation and timing together with aboutness. Sacks [1995] underlines how topicality in spontaneous conversations involves both a cognitive and an interactive process. A topic is not something predetermined and already known by the speakers, but it is an achievement worked out by their mutual negotiation. A shift from the *what* perspective to the *how* perspective occurs: topic is not only represented in terms of *aboutness*, but also in how it is manifested and signaled. In fact, moving from a discourse to a conversational perspective, interaction and timing become two fundamental notions that help construct the definition of topic, together with aboutness. *What* we are talking about is not sufficient anymore, but, the topic of a discourse will be determined by the conjunction of *what* we are talking about together with *when* and *how*.

A topic is hence the result not only of the *aboutness*, what the participants are talking about, but also of the interaction among the participants. In what follows, I provide an example of spontaneous conversation where the situational and the social context play a role at the content level, changing the course of the conversation and the topic.

Speaker A and Speaker B are talking about *X*. Speaker B sneezes and Speaker A, after the social politeness formula (*bless you*), stops talking about X to assess the health

[7] The Figure is at the end of this chapter.

condition of Speaker A, with a sentence like: *are you ok? are you getting a cold?*, or *shall I close the windows?* Speaker B accepts the new topic, updating Speaker A about his/her recent cold. In other words, in conversations, differently from texts, both the content and the interaction among participants contributes to the co-construction of the conversation, intended as the co-construction of the sequential topic of the conversation. In this case the trigger signaling a topic change is a non-linguistic verbal event. *When* and *how* the sneeze happens, influence the evolution of the conversation. Therefore, for the present work, I have considered topics as: **chunks of conversation showing coherence in sense of aboutness and social context.** In other words, for a chunk of conversation to belong to the same topic, it is not sufficient to talk about the same thing, but it is necessary that no external *break* is present. By external break I intend a break in the conversation due by an external event, than trigger the conversation towards a new topic (such as a new person joining, something breaking, a sneeze occurring such that it changes the flow of the conversation with the the assessment of the physical condition of one of the participants). In Section 4.2, I will show how the datasets used in this thesis reflect this definition.

2.4.3 Discourse segmentation

From a methodological point of view, the simplest answer to the question: *what is topic?* is *the subject of a conversation.* Sometimes this answer is extremely clear. In a newspaper each article could be seen as tackling a different topic, as well as in meetings, each item in the agenda could be represented as a different topic, [Purver, 2011]. However, sometimes, this is not so clear. If one considers a meeting with a single agenda item, then the subject of the meeting will be broadly topically coherent, but, during the meeting there may be different phases that represent different activities (presentations, round table and decision making, etc.). Those activities represent a different way to topically segment the meeting, referring to a paralinguistic level (pragmatic in this case) more than

a linguistic level (that would involve the concept of aboutness, described above). This is a completely different point of view, but it is also a way of fragmenting the discourse which goes under the definition of discourse segmentation. In terms of Passonneau and Litman [1997], these different activities are different *intentions* of the conversation. Including these as one topically coherent segment or treating them as separate units is an open question, probably depending on the application purpose. Gruenstein et al. [2005] pointed out the difficulties of defining topic among annotators by asking annotators to mark topic shifts in the ICSI Meeting Corpus, [Shriberg et al., 2004]. Results of this research show that segmentation can be a hard task also for humans, in particular when the subject of the conversation is not constrained and the discourse structure is not well defined (Kappa agreement among two annotators around 44.6 - 46.5%). On the other hand it appears in [Banerjee and Rudnicky, 2006] that agreement improves if the annotators are familiar with the topic of the conversation.

Three main approaches can be distinguished in topic segmentation algorithms: *1)* content based approaches, *2)* boundary-detection based approaches, and *3)* a combination of the two. In the first case, the aim is to detect a change in the lexical content, in the second, it is to identify cues of the boundary among two topics.

2.4.3.1 Content based approaches

Approaches based on changes of content rely on the fact that people talk about different topics in different ways: they use different words, and refer to different things. Discussing a particular set of concepts, people will use words relevant to those concepts; and discussion on particular entities, objects or places will involve a relevant set of names and related referring expressions. These notions are based on the work of Halliday and Hasan [1976]. Repeated mentions of the same concepts will therefore be associated with repeated reference, whether by using the same words or phrases or by using co-referent or anaphoric terms [Morris and Hirst, 1991]. Conversely, a change

in topic will be associated with the introduction of new vocabulary. Hence the vocabulary (and/or the set of referring expressions) used in a text of conversation remains relatively constant during the discussion of each topic, but changes markedly when we move between topics. Regions with relatively small changes should then correspond to topic segments, with large changes at the segment boundaries. The same may be true for features of the non-linguistic content, depending on the domain: in multi-party dialogue one may find that different speakers are more or less active during the discussion of different topics, or that some particular gestures are typical of inter or intra topic segments [Eisenstein et al., 2008].

In this framework, lexical cohesion has been the most exploited measure: Hearst [1997] introduces a metric to measure the difference in lexical cohesion between neighboring sections, where this would indicate a new topic. Reynar [1994] used clustering to group together neighboring sentences which appear very similar to each other until he builds up a set of topic clusters which cover the whole discourse. In all those methods, topics are associated with content and therefore characterized by a particular set of words, concepts and referents. The above mentioned algorithms make use of the notion of aboutness. If this is a reasonable assumption for text based situations, the case of speech may be different.

In all these works, in fact, the discourse segmentation has been influenced by the text segmentation literature, and dialogues have been considered as pure text. While this relation may hold for broadcast news or controlled speech, it is more difficult to use the same parallelism in dialogue segmentation where the interactional level plays a fundamental role. Weinstein [2009] adapts the measure of similarity of Hearst [1997] for speech, analyzing a corpus of monologues and song lyrics, but he does not take into consideration the interactional level.

In addition to purely lexical based techniques, many have used content based methods, exploiting other forms of content with combinations of feature: conversational

features such as vocalization, dialogue structure, Galley et al. [2003], video features, Dielmann and Renals [2007] or speakers role, Hsueh and Moore [2007]. An interesting recent approach by Luz [2012] exploits very simple and robust content-free information for detecting topic boundaries that the authors annotated in medical meetings. He uses a Naive Bayes classifier, trained to manually annotated corpus and exploiting pauses, individual vocalizations and group vocalizations.

2.4.3.2 Boundary detection based approaches

The second main approach to topic segmentation exploits distinctive features of topic boundaries. When switching from one topic to another, speakers tend to signal this shift. There are various cue words and phrases (discourse markers) that directly provide clues about discourse structure, Grosz and Sidner [1986], Hirschberg and Grosz [1992]. In certain domains there can be very specific cues, eg. mentions of the next item on the agenda in formal meetings, and reporters' name and network identifier in news broadcasts. Boundary detection methods have also exploited some prosodic features: before moving to a new segment, it is common to pause for longer than usual, as well as when starting a new segment, speakers then tend to speed up, speak louder and pause less [Passonneau and Litman, 1997].

2.4.4 Event segmentation: the case of noteworthy events

In RL_2, the aim is to investigate how the situation in which the conversation takes place can influence and, therefore, help to understand, the structure of the discourse itself, what is important, which events are worthy to extract. In Chapter 5, I concentrate on the relation between situational factors and the detection of relevant events, where relevant is here considered as noteworthy (worthy to take note of). Extracting relevant events from a discourse, as well as from a conversation, is the first step towards creating a summary of the discourse itself. In fact event detection, in this case

noteworthy event detection, can be considered as a particular case of discourse summarization: the aim is to summarize the conversation maintaining only the chunks that the participants wishes to recall. The task is thus preserving information that may be worth annotating for later recall. Although related, the main distinction between automatic summarization and detection of noteworthy information lays in the notion of *relevance*. The aim is filtering pieces of information that the user considers noteworthy. The concept of *relevant* in noteworthiness detection does not include informative fragments of content, which would be part of a summary, unless those fragments are also worth remembering for future recall. To the best of my knowledge not many scholars have investigated the possibility as well as the necessary knowledge for automatically detecting noteworthy elements in a conversation. The reason might be that judging which pieces of information are noteworthy is a very subjective task. Galley [2006] investigates inter annotator agreement in summarization and finds that a low inter annotator agreement is one of the challenges of meeting summarization. Different people have different ideas on what should go into a summary of the same meeting. If this is true for summarization, it might be true also for noteworthiness detection. In fact, different persons might have different ideas on which notes should be taken from the same conversation, depending on what they are more afraid to forget, on what they consider more relevant. A low inter annotator agreement in meeting noteworthiness detection, in fact, has been shown to be a main challenge also in noteworthiness discovery by Banerjee and Rudnicky [2009]. Banerjee *et al.* investigate the feasibility of discovering noteworthy chunks in meetings, exploring if a notes-suggestion task can be accomplished by a human being. To this aim they investigate whether it is possible for a human to identify noteworthy utterances in a meeting such that: (a) For at least some fraction of the suggestions, one or more meeting participants agree that the suggested notes should indeed be included into their notes, and (b) The fraction of suggested notes that meeting participants find noteworthy is high enough that, over a

sequence of meetings, the meeting participants do not learn to simply ignore the suggestions. To answer this question they conduct a pilot Wizard of Oz study where a human Wizard was asked to listen to the conversation in a meeting and to suggest to participants chunks to be inserted in their notes. In order to imitate the condition of an automatic system, the Wizard could only suggest sentences as they resulted from the automatic transcription of the meeting (without summarizing them). In addition, since the system would have had no understanding of the content of the meeting, they chose a Wizard who had no-prior knowledge of the content of the meeting. The experiment was conducted by the same Wizard over nine meetings, and the results report a Wizard's Precision of 0.35 and Recall of 0.41 over the manual annotations of the participants of the meeting. In [Banerjee and Rudnicky, 2008], the same authors apply techniques developed in extractive meeting summarization for automatically identifying noteworthy information from meetings. Towards this end, they have recorded sequences of weekly project meetings where participants take notes and manually detect those utterances in the meeting that are most closely related to these notes, and label them as noteworthy. As a consequence, every utterance in the meeting sequence is labeled as either noteworthy or not. They extract several lexical features such as n- grams, term frequency and inverse document frequency and information about who has uttered the current and next utterance. They use this dataset to train a Decision Tree classifier and they achieve precision of 0.15, recall of 0.12 and f-measure of 0.14.

To the best of my knowledge [Banerjee and Rudnicky, 2008] is the closest prior work in noteworthiness detection, which shows the challenges of the noteworthiness detection task in meetings. Conversations, in particular telephone conversations, present even more challenges such as: noisy environments, noisy transcriptions. However, they also present an ensemble of situational information that could bring an interesting orthogonal knowledge with respect to the lexical one used by Banerjee and Rudnicky [2008]. In Chapter 5, I investigate how situational signals such as location, time of the

day, time of the week, gender of the speaker, can provide a strong contribution to the detection of noteworthy events.

2.5 Summary

This Chapter provides a description of the background behind the analyses of social interactions. First, an historical background of the different disciplines which have considered human-human interactions has been given. Secondly, the three main concepts (*content, context* and *timing*) and the theoretical discussions behind them have been discussed. Finally I have described the state of the art of the two lines of research that are the focus of this thesis: RL_1 and RL_2. Definitions of topic and the current state of the art of topic segmentation has also been described. While many approaches to topic segmentation have been provided, a methodology that makes explicit and intentional use of social signals has not been explored yet. The next Chapter describes the three datasets analyzed in this thesis.

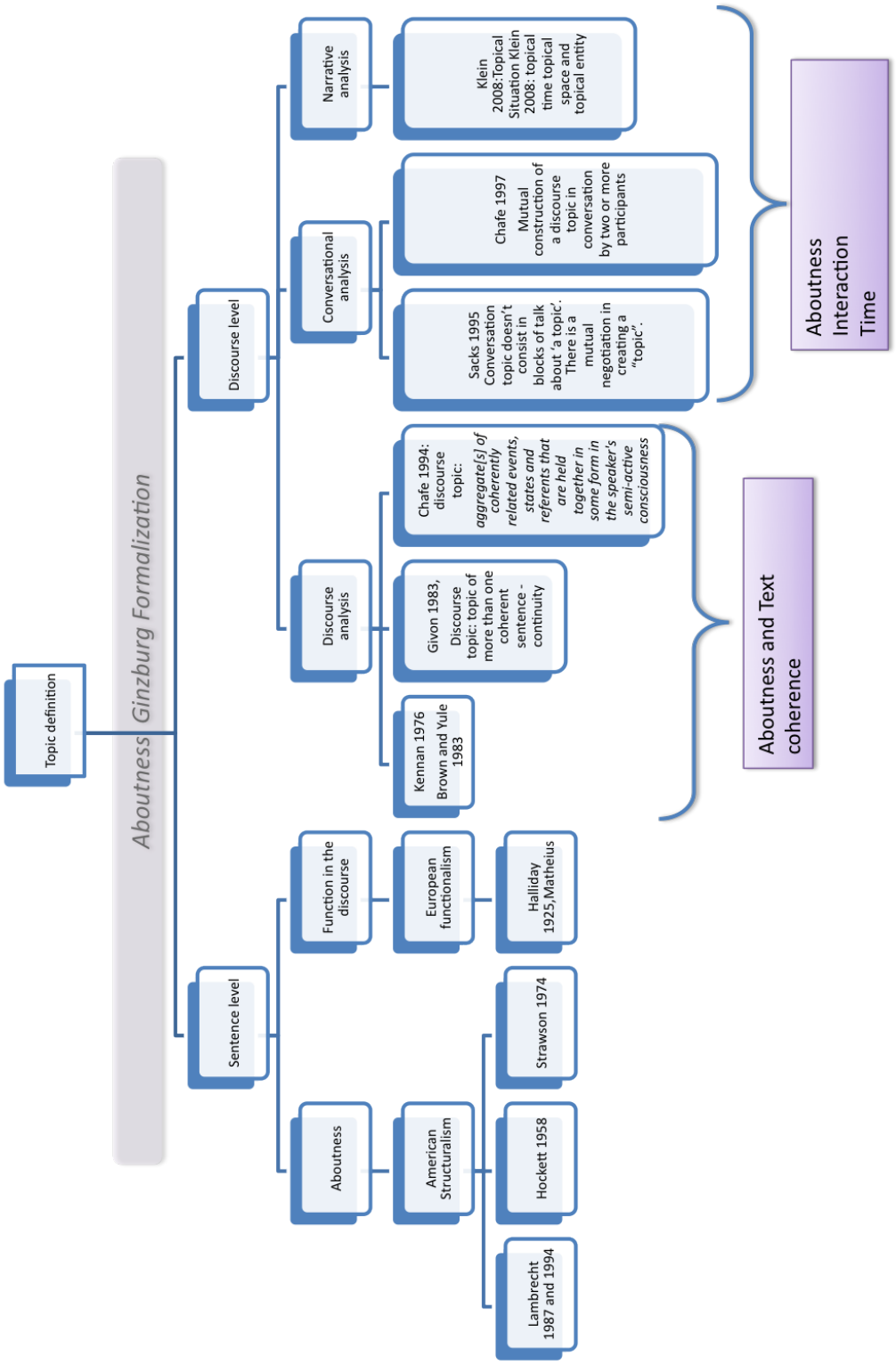


Figure 2.1 – Conceptual Map of the approach to *topic* in the Literature.

CHAPTER 3

A spectrum of dialogue corpora

In this Chapter I present the three datasets used to carry out the analyses of this thesis. Collections of multimodal interactions cover a wide spectrum of verbal, non-verbal and social phenomena. However, the majority of current resources do not address all aspects of social interactions at once, but focus on the investigation of specific contexts and settings; for example some datasets represent informal spontaneous interactions and other task based situations.

Aware of the wide range of the different "*speech-exchange systems*" [Sacks et al., 1974], in this work I consider three datasets that represent three different speech exchange systems, investigating the correlation between the content and the context of a conversation in different scenarios:

- working/task-based scenario;
- spontaneous chit-chat;
- telephone spontaneous conversations.

The working-task based scenario is represented by the AMI (Augmented Multi-party Interaction) Meeting Corpus [Mccowan et al., 2005], a collection of meetings recorded in an in-vitro experiment at the University of Edinburgh. The spontaneous conversation dataset is represented by the TableTalk corpus, a collection of three days' chats among friends, over coffee. The telephone conversations are represented by the Callnotes corpus Carrascal et al. [2012], a collection of real users' phone calls, collected within a project conducted by Telefonica Research.[1]

3.1 AMI corpus

3.1.1 General information on the corpus

The AMI (Augmented Multi-party Interaction) Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings [Mccowan et al., 2005]. The dataset is derived from real meetings, as well as scenario-driven meetings designed to elicit several realistic human behaviors. The recordings use a range of signals synchronized to a common timeline. These include close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. During the meetings, the participants also had special pens available to them that recorded what was written. The meetings were conducted in English using three different rooms with different acoustic properties, and included mostly non-native speakers.

The AMI corpus comprises scenario based meetings and natural meetings, where the former are task based conversations, while the latter are free chats over different topics (usually working discussions among speech researchers). Part of the richness of a corpus stands in the annotation provided on the data. The AMI Meeting Corpus

[1]The AMI corpus is available under Creative Commons Attribution ShareAlike Licence, the TableTalk is freely available at `http://sspnet.eu/2010/02/freetalk/`. The Callnotes corpus is not publicly available, but it was available to this study by reason of a collaboration with Telefonica Research.

includes manually produced orthographic transcriptions for each speaker, with word-level timing. In addition to this, it also contains a wide range of other annotations, not just for linguistic phenomena but also detailing behaviors in other modalities. These include dialogue acts; topic segmentation; summaries of the meeting (for abstractive summarization); named entities; the types of head gesture, hand gesture, and gaze direction that are most related to communicative intention; movement around the room; emotional state; and where heads are located on the video frames. All these annotations are provided for the scenario based meetings, while some are missing for the natural meetings.

3.1.2 Scenario Based meetings

The majority of the corpus has been collected by having participants playing different roles in a fictitious design team that takes a new project from kick-off to completion over the course of a day. Participants were recruited within the University of Edinburgh students, researchers or other professionals. The day starts with training for the participants about what is involved in the roles they have been assigned (industrial designer, interface designer, marketing, or project manager) and then contains four meetings, plus individual work to prepare for them and to report on what happened. Their work also includes web pages, email, text processing, and slide presentations, and all these data are provided together with the recording, transcripts and annotation of the corpus. Although the controlled-scenario clearly reduces the spontaneity of the discussion, it provides other advantages, such as the possibility of detecting the outcome of the meetings, in terms of success rate, as well as finer control of the behavioral dynamics of the different roles. On the other hand, there is no guarantee that the participants are engaged in the task/conversation and that the outcomes are comparable with a spontaneous interaction.

The scenario is played over 4 meetings of around 20 min. The participants play

Figure 3.1 – AMI screenshot - © AMI website: https://www.idiap.ch/dataset/ami/ (URL last verified on May 2015).

the roles of employees in an electronics company and they have to develop a new type of television remote control. They are told they are joining a design team whose task, over a day of individual work and group meetings, is to develop a prototype of the new remote control. Each scenario maintains the same participants playing the role of the interface designer, the project manager, the marketer and the industrial designer, and is composed of a kick off meeting, a functional design meeting, a concept design meeting and a detailed design meeting. The participants were chosen not to be expert in the role they were given and not to be familiar with each other.

3.1.3 The AMI corpus in this study

The analyses in this thesis are based only on the scenario based meetings recorded in Edinburgh (ES meetings) which represent a total of 60 meetings out of 138. However only the 57 having topic annotations were selected. The analyses are based on the transcription and annotation of the dataset Version 0.2, for a total of 717239 word tokens. Conversations are all in English, but participants are not all English native speakers (91 of 187 are English native speakers, the rest is divided among 27 other nationalities[2]).

[2]Arabic, Czech, Dutch, English, Estonian, Finnish, French, German, Greek, Hindi, Italian, Konkani, Malayalam, Mandarin, Persian, Polish, Portuguese, Russian, Spanish, Swedish, Swiss, Tamil, Telugu, Vietnamese, Wolof, Romanian.

The meetings have an average length of 29.3 minutes (*standard deviation* 9 minutes) and 5509 turns (*median* 5740, *standard deviation* 2113.6).

AMI transcription provides annotation of start and end times at the word level and transcriptions of many nonverbal outbursts under the label of *vocalsounds*. Table 3.1 reports the list of all the vocalsounds annotated and their frequency in the considered meetings.

As one can see, in the subset of AMI here analyzed, there are 3687 instances of laughter, 294 of coughs and 2623 *others*, which represents the underspecified label. In addition to this, to the purpose of this thesis other information has been extracted at the utterance level, specifically: overlaps, silences, lexical and nonlexical backchannels. In this work I consider lexical backchannels and non-lexical backchannels, the former being a series of standard backchannels such as *yeah, yep, right, alright*, the latter being *Mm-hmm, ah ah, uhm, uuhm*. Table 3.2 shows the frequency of laughter, overlaps, silences, backchannels in the corpus.

3.1.4 Problems in AMI annotation and solutions proposed

As noted in Gilmartin et al. [2013a], AMI presents some peculiarities regarding the laughter annotation: these include, missing end times or end time equal to zero, start time equal to zero, end time smaller than start time. However these cases appear to be quite rare in the dataset. In particular, over the 57 Edinburgh meetings with topic annotation (used in this corpus) there are:

- 22 laughs with a negative duration
- 17 vocal sounds (not laughter) with negative duration

Since, over a total amount of 3687 laughs and 2623 other vocal sounds these cases amount to a very small percentage (0.6%), the rows corresponding to such events have been removed from the analysis. Since the timing annotation in AMI is at the word level, the nearby elements of the dialogue were not affected by this.

Vocalsound	Frequency
clicking-noises	1
clicks-fingers	1
cough	294
do-d–do-sounds-imitating-whistling	1
fricative-exhalation-indicating-displeasure	1
fricative-noise	1
hums	1
imitates-background-typing-sound	1
imitates-rigorous-unidentified-sound	1
laugh	3687
loud-exhale	5
makes-buzzing-sound	1
makes-clicking-sounds-with-tongue	1
makes-machine-noises	1
makes-straining-sound	1
makes-tapping-sounds	1
makes-tutting-sounds	1
nasal-sound	1
other	2623
random-noise	1
random-sound	1
rhythmic-exhalation	1
sound-imitating-an-alarm	1
sound-imitating-devouring-something	1
sound-imitating-pager	1
sound-imitating-vibration	1
sound-imitating-zapping	1
sound-indicating-fatigue	1
sounds-imitating-joystick	1
sounds-imitating-machine-noises	1
sounds-imitating-winding-up	1
stammering-sound	1
vocal-noise	4
weird-noises	1
whispers	1
whistling	5
yawn	1

Table 3.1 – Types of vocalsounds annotated in AMI

Signal	Freq.
Laugh	3687
Overlaps	44016
Silences	21992
Lexical Backchannels	6577
Non lexical Backchannels	3470

Table 3.2 – Frequencies of social signals in the considered AMI corpus

3.2 TableTalk

3.2.1 General information on TableTalk

TableTalk[3] is a corpus of free flowing natural conversations, recorded at the Advanced Telecommunication Research Labs in Japan. It is a multi-modal corpus of conversations among five individuals [Campbell, 2009; Jokinen, 2009b]. In order to collect as natural data as possible, neither topics of discussion nor activities were restricted in advance. The recordings were made in an informal setting over coffee, by three female (Australian, Finnish, and Japanese) and two male (Belgian and British) participants. The recordings are carried out over three sessions, of different lengths ranging from 35' to 1h and 30', recorded on consecutive days. The conversations are fully transcribed and segmented for topic, and also annotated for affective states of participants and for gesture and postural communicative functions using MUMIN [Allwood et al., 2007].

TableTalk has been analyzed in terms of engagement and laughter [Bonin et al., 2012a,c; Gilmartin et al., 2013a,b], lexical accommodation [Vogel and Behan, 2012], gestures and gaze [Jokinen, 2010, 2009a].

3.2.2 The TableTalk corpus in this study

In this thesis, the analyses are based on transcripts of the entire corpus: about 3h 30, 31523 tokens and 5980 turns. Specifically day one lasts 34.33 minutes, day two 82 min-

[3]Freely available at: http://sspnet.eu/2010/02/freetalk.

utes and day three 83 minutes. In this case the entire corpus has been used, as the entire corpus presented topic annotation. Transcriptions have been carried on by non-native labelers, and the timings at the utterance level have been marked. Backchannels can be retrieved from the transcription as they have been transcribed as *Mm-mm, yeah, yep* as well as laughter, that have been marked with the convention *@w*. In addition, to the purpose of the analysis in this dissertation, overlapping utterances and silences have been extracted.

Figure 3.2 – TableTalk screenshot.

Table 3.3 reports the frequency of these signals within TableTalk.

Signal	Freq.
Laugh	680
Overlaps	1423
Silences	1750
Lexical Backchannels	861
Non lexical Backchannels	607

Table 3.3 – General Figures of TableTalk

3.3 Topic annotation in TableTalk and AMI

The choice of these two corpora is due to the following reasons: they represent two examples of task based (AMI) and spontaneous (TableTalk) corpora, and they are pro-

vided with topic annotation. AMI provides the annotation of top-topics and subtopics. Top-level topics refer to topics whose content reflects the main meeting structure, while subtopics reflects small digressions inside the core topics.

In TableTalk, topics have been annotated manually by two labelers with no distinction is made between core topics and subtopics.

For the following analysis in this thesis, I rely on the annotation provided with the corpora, focusing on the core topic segmentation of AMI, as more similar to the coarse grain topic segmentation presented in TableTalk, and in line with the definition of topic provided in Chapter 2.

However, the AMI annotation of topics presents some gaps between different topics, since there are words that are not included in a particular topic. Specifically, analyzing the dataset, it emerged how the annotators seemed to have a clear idea on when a topic would start, but, not about when a topic would end. In particular in the chitchats, transitions among topics present a gray zone of turns belonging nor to the previous neither to the second topic. Since in my analysis, topics are to be considered as contiguous, I apply, on the entire dataset, the assumption that words outside a topic belong to the previous contiguous topic. This is in line with the fact that it is more clear when a topic starts rather than when a topic ends. Therefore, if there is topic *A*, followed by topic *B*, and the end of topic *A* is marked before the start of topic *B*, leaving out some turns, the end of topic *A* is extended in order to include those turns.

In particular a word belongs to the topic if it:

- is fully included in the topic
- starts before the topic and ends within the topic
- starts within the topic and ends after the topic
- starts after the end of a topic but ends before the start of the next topic

3.4 The Callnotes Corpus

3.4.1 General information on the Callnotes corpus

The last dataset analyzed in this dissertation is the so-called Callnotes corpus. Such corpus is the result of a human-computer interaction research conducted by Telefonica research and described by [Carrascal et al., 2012]. The corpus is constituted by cellphone conversations of Spanish speakers. In the study of Carrascal et al. [2012], a large sample of mobile phone conversations was recorded and semi-automatically transcribed, as participants were given the opportunity to revise the automatic transcriptions during the annotation phase. The calls were recorded from real everyday conversations that the participants were making during a set period of time from their cellphones. Participants were recruited among people who had offered to participate to the experiment and they could choose whether to use the default calling application of their phone or the *callnote* application. By choosing the callnotes application they knew that their call would have been recorded and transcribed, but, as reward they were getting the call for free. A message warning that the call was being recorded was played both on the caller and the receiver side. After the call the participant could log in to a web system, see their list of calls and delete the calls they they did not want to share for the study. The study spanned for 64 days, in which 796 mobile phone conversations from 62 volunteering subjects (20 female) were recorded. All the participants were Spanish native speakers, having an average age of 31.5 years (s=7.52, min=20, max=51). Hence, all conversations were recorded and transcribed in Spanish.

In addition to the call, also metadata about the call (*e.g.* duration, date, time) was stored along with the actual conversation and its transcript. Finally, participants were asked to annotate and provide contextual information about the call. They have to annotate the parts of their calls that they would like to take a note of: *i.e.* relevant or noteworthy fragments of conversations. To this end, they used the Web interface to access their calls and highlighted with the mouse the parts of the transcript that they

considered to be worth keeping for future reference. The participants were asked to fill out a questionnaire after annotating each call, which was used to collect contextual information, including: location of the call (at work, at home, while commuting, while doing shopping, while exercising), and category of the call (discussing a topic, taking an appointment, giving/receiving information, asking a favor, social). Finally participants were presented a series of contextual questions related to each phone call: 1) Relationship with callee; 2) Who was with the caller at the time of the call; 3) Location of the caller at the time of the call; 4) Objective of the call; 5) Level of importance of the call; 6) Level of importance of the notes; and 7) General questions about sound and transcription quality. Due to the sensitivity of the data collected the dataset is not publicly available, and it is part of this work as result of a collaboration with Telefonica Research.

3.4.2 The Callnotes corpus in this study

The original conversation collection consisted of a total of 796 conversations, of an average length of 178 seconds ($sd = 384$ sec.). This original set was pre-filtered to exclude calls with problems in the transcript (*e.g.* empty transcript, only one speaker audible, etc). Out of the entire corpus a subset was finally selected of 659 conversations. Even if participants annotate only chunks of conversations, they usually tended to annotated entire turns, or the 70% (in number of word tokens) of a turn. For this reason the turn has been chosen as the unit of analysis. This dataset comprises $22,474$ turns, with an average of 34.10 ($sd = 45$) turns per conversation. From these, only 671 are annotated as relevant (2.98%), which represent an average of 1.02 turns ($sd = 1.803$) per call.

Table 3.4 provides the general information on the dataset.

	# Calls	Turns		Annotated Turns	
		Total	$avg.$ per call	Total	Fraction
	659	22,4	34.1 ($sd = 45$)	671	2.9%

Table 3.4 – General statistics on the Callnotes dataset.

3.5 Characteristics of the three corpora: a spectrum of spontaneity

The three corpora analyzed in this thesis represent a wide range of human speech exchange systems, specifically, they show a increase in their spontaneity. One can say, in fact, that from the AMI to the Callnotes corpus passing through TableTalk, the interactions range from task based scenario to a spontaneous/into the wild scenario. The AMI corpus represents a key resource for the study of meetings, and task-based speech exchange systems;[4] nevertheless, according to the corpus designers, in two thirds of the dataset the participants are engaged in the task, while in one third of the conversations they are captured in free conversations. However, even if speakers enter in their role and act without a predefined script, they are still within a predefined scenario in which they are *working*, not just chatting.

The TableTalk corpus is an example of spontaneous multiparty conversation corpora. It represents the second degree of spontaneity in the scale from AMI to Callnotes. The natural setting over a coffee, around a table try to reproduce the most natural situation for a free chat. The freedom of the speakers of talking about anything guarantees spontaneity in the conversation, and may also bring to the situation in which the topic naturally exhausts, that is less likely when the participants have to accomplish tasks. However, the participants are still within an unnatural setting surrounded by cameras and microphones.

The Callnotes corpus overcomes this issue. There are no microphones, or cameras (in fact there is no video signal recorded in the dataset), but participants are chatting

[4] In this study only the task-based meetings in AMI are considered; cfr. 3.1.

from their usual cellphone. From a linguistic and conversational point of view this corpus represents a rare example of real free and spontaneous conversation. Although some could argue that the speaker's awareness of being recorded could influence their spontaneity, it is also true that the speakers could talk from their phone in daily situations, without the use of external devices, microphones, cameras, or other non-natural settings. The natural setting, and the fact that the participants find themselves in different situations and context (differently from what happens in the previous corpora) make this corpus a good datasets for the study of our RL_2, the interaction between conversations and situational context. On the other hand, a fixed situational context, but a variable social context makes AMI and TableTalk good dataset for the study of RL_1, allowing also to explore the difference between task based and non-tasked based conversations.

Despite the effort made in this thesis to cover speech exchange systems of different nature, it is worth noting that the range of spoken interactions humans engage in is enormous. It is not certain if a generalization from different speech exchange systems is possible, since it is not known, at present, whether even basic mechanisms such as turn-taking, vary with the type of interaction. In other words, it is not clear whether observations made over certain data generalize to the entire range of human interaction and, if yes, to what extent [Bonin et al., 2014b]. The best effort stands in analyzing individually speech exchange systems and proceed with careful generalizations. To this extent, not being possible to investigate all possible speech exchange systems, in this work I consider three different kinds of interactions and report results over these.

3.6 Summary

This chapter provides an overview of the datasets analyzed in this thesis, consisting of data selected from three corpora: the AMI corpus, the TableTalk and the Callnotes corpus. While TableTalk and AMI are investigated in Chapter 4 to address RL_1, the

Callnotes corpus is the focus of Chapter 5 to address RL_2. TableTalk and AMI provide a fine annotation of the social signals and an uniform situational context for all the conversations, in Callnotes the situational context varies within the dataset. The three corpora represent also an increasing climax of spontaneity from AMI to Callnotes: where AMI is the more controlled scenario, while Callnotes the more un-controlled one.

CHAPTER 4

Social Context and topic segmentation

4.1 Social Context within the conversation

This chapter focuses on the research line RL_1 presented in Chapter 1: investigating whether social signals can be considered not only from an affective, but also from a linguistic point of view as signals with a specific discourse function. In order to answer this question I explore the correlation between social signals and a specific type of discourse event: the topic change. I investigate whether a relationship exists between four different classes of social signals (laughter, overlaps, silences, backchannels) and topic changes.[1]

This chapter is organized as follows: in section 4.2 I recall the working definition of topic used in the work at hand as defined in 2.4.1; in section 4.3 I describe social signals that are taken in consideration and in section 4.4 the methodology used to explore the distribution of these signals in meeting and in spontaneous conversations. In section

[1]Part of the analyses and results of this Chapter have been published in [Bonin et al., 2012c, 2014a; Gilmartin et al., 2013a,b; Bonin et al., 2014d, 2015] and is reported here with co-authors' permission.

4.7, I provide a deeper study of laughter, which accounted as an interesting social event with a specific behavior in relation to topic changes.

4.2 Topic definition and annotations for this study

As discussed in Chapter 2, topic is a complex term whose definition has been reconsidered by several scholars. Some have examined topic at the sentence level, some at the discourse level. Definitions of topic has developed from considering only the coherence among consecutive sentences, to the mutual negotiation of a topic in a conversation as in Conversational analysis. Also, a narrative perspective of topic has been given, as the topical situation in which events of a story take place. Those are all further discussed in Chapter 2. In this work, I refer to the definition of topic provided in 2.4.1, which refers to topic as **chunk of conversation showing coherence in sense of aboutness and social context.**

I rely on the annotation of topic change as made in the two corpora described in Section 3.1 and Section 3.2, and this choice is meant to avoid circularity in the present work and to base my analyses on an external and objective ground truth rather than on a personal annotation of topic. In addition this provides me with results comparable with other works. The annotation used reflects the definition of topic used in this thesis. In fact, in both AMI and TableTalk topics are not only indicated by changes in the conversation's content, but also by events occurring during the conversation (i.e. a new person joining, or something breaking). To better explain this, I report here the annotation specifications of the AMI and TableTalk corpora.

In the AMI corpus a list of topics in the scenario meeting is identified and comprises *content related* topics and *social related topics*, such as *chitchat, openings of the meeting, and closing of the meeting*. The content related topic are associated with the different parts of the meeting evolution (for example discussion of a design), while the social related content are relate to the social events. I report here the description of the topic

chitchat, from the AMI topic segmentation guidelines [Weiqun Xu and Karaiskos, 2005, pp. 4]:

> Sometimes during a meeting the participants just chat aimlessly, usually about social matters. This especially happens after the microphones have been switched on but before the beginning of the meeting proper, and again at the end. It can also happen in the middle of a meeting, for instance, when a projector breaks, or simply if someone drags the group off-topic. When this happens, divide the meeting so that these areas form their own segments, but label them with the special topic description, chitchat.

Interestingly, the authors report a real world scenario (as when the *projector breaks*) to describe a moment of social interaction, which has to be marked as a chitchat topic. Although in AMI a primary role is played by content-related topics sometimes the social context shifts the conversation towards social chitchats; the example reported by the authors is, in fact, a real world event like *the projector breaking*, a social event that has the potential to shift the topic of the conversation from the task to a social chitchat.

TableTalk presents a different situation. Content-related and social-related aspects of a conversation are intertwined as the corpus consists of a completely free conversation. The conversation can change drastically from a discussion of *almonds* to some coffee dropped on the table, in relation to the social activity which emerges. Annotators have marked every change in the content (*about*). When a social event (*dropping coffee*) causes a topic change (i.s. water-proof devices) this is also marked as a topic change.

Topic changes (hereafter: T) are the annotated time points where topic shifts in conversation.

4.3 Social signals analyzed in this study

Five different categories of social signals are described: changes in physical appearance, gesture and posture, face and eye behavior, linguistic and non-linguistic vocalizations and space and environment. The non-linguistic vocalizations, also known as vocal outbursts, include non-verbal sounds such as laughing, sobbing, crying, whispering, groaning, and similar, but not necessarily accompanying words,

Relying on the classification of Vinciarelli et al. [2008], the non-linguistic vocalizations include prosody, turn taking, silences and vocal outbursts. In the current work, I take in consideration vocal outburst such as laughs, and backchannels (lexical and non-lexical), and silences. In addition, I examine the dynamics of overlap as the complement of a silence.[2] Gesture and postures lie outside the scope of this thesis. However it is worth reminding how pointing gestures have been shown playing a role in structuring the information flow of a conversation in Jokinen [2010].

In the rest of this section, I will address each of the signals in detail.

4.3.1 Laughter

While laughter has long being studied as the vocalization of mirth, in this work I start from the hypothesis that sometimes laughter can have a different function, unrelated to the presence of a joke or a mirthful event. Bea and Marijuán [2003] studied the characteristics of controlled and uncontrolled laughter noticing a difference in the internal structure of the laughter: controlled laughter does not exhibit random structure but repetitions; uncontrolled spontaneous laughter has been found to have random internal structure.

Laughter can be understood as a joint activity: one interlocutor may laugh alone, or a number may join the laughter. Previous authors [Holt, 2011] have described laughter as an action which may be independent from the presence of humor. In this context,

[2] When two consecutive utterances are not separated by a silence, they might be in the situation of a no-gap-no-overlap or overlap case.

laughter has been seen as a highly ordered phenomenon, internally and externally. In this sense, it is also relevant to explore the timing of laughter with respect to other elements of interaction in dialog, such as topic changes. While the timing of mirthful laughter is effectively random, given the distribution of potential triggers, the timing of non-mirthful laughter might be related to the conversation's structure.

The hypothesis is that when laughter functions as a social signal, its timing is structured and conveys information about the underlying discourse structure. Previous works have explored other non-verbal features that can be predictive of discourse structure [Luz, 2012]. Luz [2012] investigates the potential of non-verbal signals such as silences (between two speakers' vocalizations as well as within the same speaker's turn) and overlaps in predicting topic changes in meetings. Results show that pauses and overlaps on their own are good estimators of the topic structure of meetings' conversation, reaching performance comparable with lexically-based methods.

4.3.2 Silences and Overlaps

Silences have long been studied in conversational analysis and they play a fundamental role in vocal behavior [Zellner, 1994]. Linguists first showed interest in silence from two different perspective: first, from a functional perspective [Jensen, 1973], under the influence of philosophy and literature; then from an acoustic perspective, and only by this route was silence introduced as a subject of study. In the acoustic paradigm silence developed along two paths. One was the chronometric analysis of speech, where quantitative chronometric data on speech rates were collected to show the ratios of speech to non-speech, etc., in isolation or in relation to personality variables, as early as Chapple [1939]; Goldman-Eisler [1958]. Something (speech) and nothing (the spaces, or the silences, between words) were counted. Such studies produced quantitative predictions, such as the constant ratio between vocalization and silence in spontaneous speech Crown and Feldstein [1991]. The second path was discourse analysis. Sacks

et al. [1974] perceive silences as the interactive locus of turn-taking, allocating the floor, during discourse. More recently silence has also been studied as an important part of interaction. Three kinds of silences have been distinguished: hesitation silence, psycholinguistic silence, and interactive silence [Richmond et al., 1991]. The first arises when a speaker is hesitant in completing a sentence. It is mainly an intra-speaker silence. Psycholinguistic silences take place when the speaker needs time to encode or decode the speech, in particular at the beginning of a turn, when the participants need to think about how to respond, or of what he is going to say. Interactive silences are the silences due to turn taking: one participants is paying attention to the other's contribution by listening in silence. Recent works [Heldner and Edlund, 2010] have distinguished two kinds of silence: gaps and pauses, where gaps are the silences between two speakers' contributions and pauses are the silences within the same speaker turn; I do not consider this distinction between gaps or pauses, I consider silence every timespan in the conversation reported as a lack of a vocal event (linguistic or non-linguistic) by any of the participants.

To the purpose of this analysis, I extract silences from the transcripts, calculating the silence intercourses between the end of an utterance and the beginning of the following utterance. No threshold is applied, every span of time between the end of an utterance and the beginning of a new one is considered a silence. While there is no general agreement on the threshold above which one should regard a period of no vocal activity to constitute a silence [Sellen, 1995; Luz, 2009], in this approach the determination of a silence is provided by the segmentation in utterances given in the annotation. Finally, no distinction between gaps or silences is done.

I start from the hypothesis that silence is more frequent in interactive moments of a conversation, characterized by shorter turns and higher number of exchanges. In fact, while in these situations there will be both gaps and pauses, in a monologue (or monologue-like part of a conversation) only pauses should be expected.

In the same context I consider overlaps. Many have looked at overlaps as indicators of conflicts, disputes or dominance display [Smith-Lovin and Brody, 1989]. In this work I intend to look at overlap as a natural event occurring in dynamic conversations. In this view, a generalization is proposed: while in many cases overlaps might represent the escalation of a conflict situation, in general they are a natural phenomenon of spontaneous interaction, which indicate a high level of interaction in the conversation.[3]

To conclude, one can imagine the conversation as possessing different degrees of social interaction. At the very bottom of the scale, low interaction, there will be moments dominated by a single speaker where few silences and few overlaps will happen. At the very top of the scale there will be high interactional moments characterized by many speakers interlacing their contributions with a higher probability of overlaps.

4.3.3 Backchannels

I consider lexical backchannels and non-lexical backchannels, the former being lexicalized chunks such as *yeah, yep* etc., the latter being non-lexicalized vocalizations such as *Mm-hmm, ah ah* etc. Other types of backchannels, as in Cerrato [2007] are not considered to the extent of this thesis. I investigate the two separately to explore potential different behaviors.

4.4 Methodology

In order to better understand the dynamics of the subset of social signals considered in relation to topic changes, four different analyses have been conducted.

Analysis 1 First of all the topics are divided into topic transition and topic continuation segments, in order to analyze the distribution of laughter, silences, overlaps and backchannels (lexicalized and non-lexicalized) among those. Operational models of

[3]Defining a taxonomy of overlaps, and studying the acoustic features of different kinds of overlaps lies outside the scope of this dissertation, but represents an interesting research line to follow up.

topic continuation segments and topic transition segments are then constructed, calling the former *wi* segments, and the latter *wo* segments.[4] These are defined as follows (see Fig. 4.1):

- *wi* segments: the central half of each topic
- *wo* segments: the final quarter of one topic and first quarter of the next topic

By construction, *wi* segments represent the core of a topic and have topic cores *within* them, while *wo* (*without*) segments do not contain the core of a topic, but do contain a transition between two topics. Both are defined in relation to the duration of a sequential pair of topics, not as absolute durations. Although arbitrary, this decomposition of conversational flow into segments of topic-core talk and topic transitions has *face validity*, in the sense the term is used in psychology to indicate that the objects used operationally relate naturally to the corresponding theoretical constructs. The *wi* segments model topic continuation and the *wo* segments topic transitions.

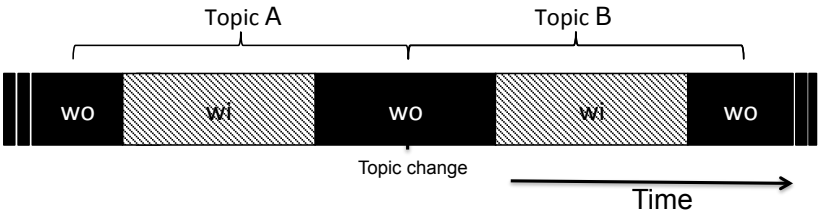


Figure 4.1 – Topic continuum and topic transition segmentation

Analysis 2 and 3 I analyze the first and second halves of a topic (called *wf* and *ws*) and, consequently the first, second and third thirds (called *w1, w2, w3*). Fig. 4.2 and Fig. 4.3 illustrate of the topic segmentation in the two cases.

Analysis 4 Having a general idea of the signal distributions within the topic change, I then consider the extremes: topic termination and topic beginning. In this case an

[4]Part of this analysis has been published in [Bonin et al., 2012c] and it is reported here with the co-authors' consensus.

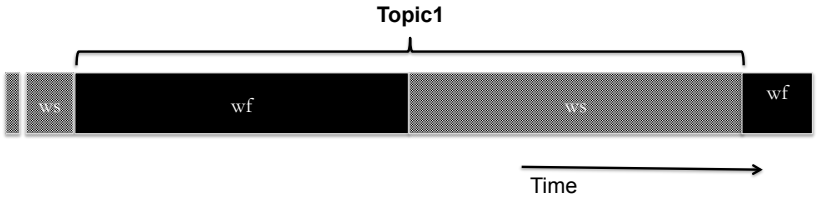


Figure 4.2 – First and second half of a topic

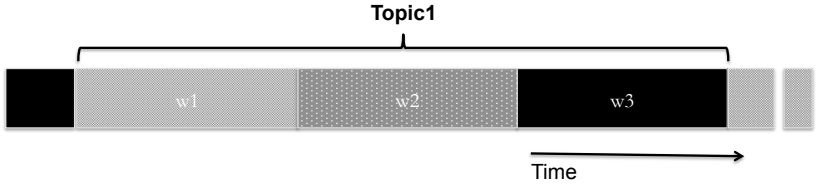


Figure 4.3 – First, second and third parts of a topic

operational model for topic termination and topic beginning is provided. As in the work of Gilmartin et al. [2013a] a threshold of fifteen seconds around topic changes is chosen as a demarkation point for topic terminations as well as for topic beginnings' segments, which are defined as follows:

- *wt* - topic termination segment: from topic change minus 15 seconds and topic change
- *wb* - topic beginnings segments: from topic change to topic change plus 15 seconds

Gilmartin et al. [2013a] counted the frequency of laughter, shared laughter, and solo laughter into 5-second bins at T minus multiples of 5 seconds (T-5, T-10, T- 15, T-20) in order to look at the laughter trend near topic termination. A meaningful threshold emerges (T-15 seconds) where a change in the laughter trend (amount of laughter increasing significantly with respect to T-20) is visible. Hence, the threshold of Ts -15 seconds and T + 15 seconds is considered.

Figure 4.4 provides a visualization of this fourth scenario.

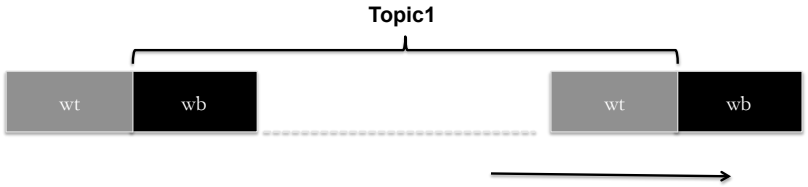


Figure 4.4 – Topic beginnings and topic terminations

4.5 Social Signals Timing

In this section I report the results of the statistical tests conducted over the different distributions in AMI and TableTalk. I refer to the topic change as a point in time, denoted T. In order to compare different distributions, I use nonparametric measures due to the non-normality of the studied variables. In particular I use the Wilcoxon Test [Bauer, 1972].[5] Histograms in Fig. 4.5 and 4.6, which show the distribution of laughter in analysis 1 and of overlaps in analysis 2, are representative of all the distributions that will be analyzed.

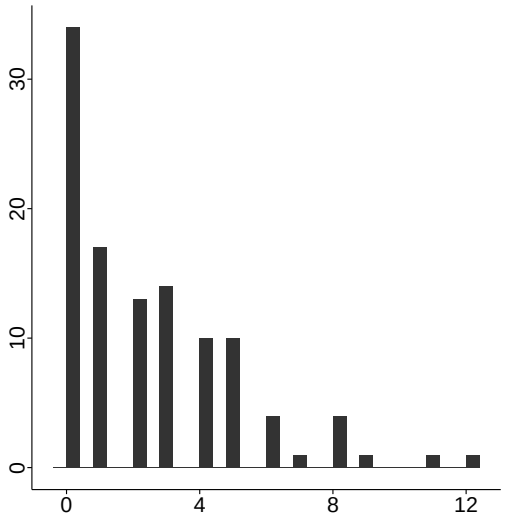


Figure 4.5 – Distribution of laughter in Analysis 1 -*wi segments*.

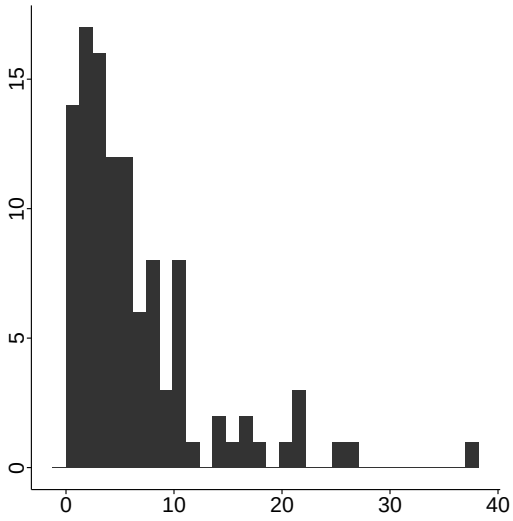


Figure 4.6 – Distribution of overlaps in Analysis 2 - *wf segments*.

In order to tackle the multiple comparison problem, I apply the Bonferroni cor-

[5] The two tailed Wilcoxon test is also know as the Mann-Whitney test.

rection to all the tests, dividing the critical value by the number of tests. Therefore, I used the formula $p < 0.05/n$, being n the number of conducted test (usually $n = 2$). In addition, I report, when a significant difference is found, the effect size of the non-parametric test used, calculated as the statistic value ($z - value$) over the square root of N, with N= the number of observations [Pallant, 2007]; I report, hereafter, Cohen's scale, that defines small effect sizes with $r = 0.2$, medium effect sizes with $r = 0.5$, and large effect sizes $r \geq 0.8$ [Cohen, 1988]. In each of the following analyses my attention will be on the differences in the distribution of the amount of social signals in several segments of the conversation.

One could consider the amount of social signals as a measure of the interaction level among the participants, which I call **interactional entropy** defined as:

> *The interactional entropy of a segment x is here defined as the number of occurrences of social signals in x.*

4.5.1 Analysis 1: topic continuation vs topic transition

I explore the distribution of the signals described in Section 4.3 among topic continuation and topic beginnings over all the conversations of AMI and TableTalk. Table 4.1 and Table 4.2 provide an overview of the descriptive statistics of the distribution between *wi* and *wo*.

Table 4.3 and Table 4.4 summarize the results of the statistical test on the distribution of the different signals in topic continuations and topic transition segments.[6] As the tables show, a significant difference in the distribution of laughter between type of segments is found in both corpora (a significantly higher amount of laughter in topic transition moments than in topic continuation moments). A Wilcoxon test, one tailed, alternative less, between the distribution of laughter in *wi* segments and in *wo* segments confirms that there are significantly more laughs in *wo* segments, hence at topic

[6] In these as in the following tables of statistics in this Chapter, for each distribution d, I report: mean $\bar{x}_d$, standard deviation sd and median $\tilde{x}_d$.

transitions, p <0.005***. No significant difference is found with respect to the other signals.

Signal	$\bar{x}_{wi}$	sd_{wi}	$\bar{x}_{wo}$	sd_{wo}	$\tilde{x}_{wi}$	$\tilde{x}_{wo}$
Laugh	0.58	1	1.85	1.76	0	2
Overlaps	1.45	1.44	2.12	1.55	1	2
Silences	2.35	1.31	2.37	1.39	2	2
Lexical Backchannels	0.86	1	1.37	1.37	1	1
Non-lexical backchannels	0.86	1.07	0.75	0.91	1	1

Table 4.1 – General Statistics of *wi* and *wo* segments in TableTalk. Mean $\bar{x}$, standard deviation sd and median $\tilde{x}$ are reported.

Signal	$\bar{x}_{wi}$	sd_{wi}	$\bar{x}_{wo}$	sd_{wo}	$\tilde{x}_{wi}$	$\tilde{x}_{wo}$
Laugh	3.62	7.29	3.87	8.080	1	0
Overlaps	43.36	78.90	42.57	69.383	7	10.50
Silences	19.85	29.86	22.38	30.61	7	10
Lexical Backchannel	6.47	12.95	6.40	11.42	3	3
Non-Lexical Backchannel	3.31	4.93	3.40	4.98	1	1

Table 4.2 – General Statistics of *wi* and *wo* segments in AMI. Mean $\bar{x}$, standard deviation sd and median $\tilde{x}$ are reported.

Signal	$\bar{x}_{wi} \neq \bar{x}_{wo}$	$\bar{x}_{wi} < \bar{x}_{wo}$
Laugh	***	***
Overlaps	ns	ns
Silence	ns	ns
Backchannels	ns	ns
Lexical-Backchannel	ns	ns
Non-lexical-Backchannel	ns	ns

Table 4.3 – Significance values resulting from the two-tailed wilcoxon test, and from the one tail wilcoxon test in AMI. Medium effect size, $r = 0.31$.

4.5.2 Analysis 2: first and second half of a topic

In the second analysis, I compare the distribution of the signals between first and second halves of topic, as in Fig. 4.2. Table 4.5 and Table 4.6 show the general statistics of the distributions over these segments.

Signal	$\bar{x}_{wi} \neq \bar{x}_{wo}$	$\bar{x}_{wi} < \bar{x}_{wo}$
Laugh	***	***
Overlaps	ns	ns
Silence	ns	ns
Backchannels	ns	ns
Lexical-Backchannel	ns	ns
Non-lexical-Backchannel	ns	ns

Table 4.4 – Significance values resulting from the two-tailed wilcoxon test, and from the one tail wilcoxon test in TableTalk. Medium effect size, $r = 0.30$.

Signal	$\bar{x}_{wf}$	sd_{wf}	$\bar{x}_{ws}$	sd_{ws}	$\tilde{x}_{wf}$	$\tilde{x}_{ws}$
Laugh	2.45	2.60	3.72	3.35	2	3
Overlaps	6.33	6.30	6.6	5.20	4	5
Silences	8.03	6.14	7.87	6.46	6.5	6
Lexical Backchannels	3.73	3.66	4.09	3.63	3	4
Non-lexical Backchannels	2.82	2.82	2.69	3.06	2	2

Table 4.5 – General Statistics of *wf* and *ws* segments in TableTalk. Mean $\bar{x}$, standard deviation sd and median $\tilde{x}$ are reported.

Signal	$\bar{x}_{wf}$	sd_{wf}	$\bar{x}_{ws}$	sd_{ws}	$\tilde{x}_{wf}$	$\tilde{x}_{ws}$
Laugh	3.135	7.31	4.29	7.65	0	1
Overlaps	33.25	64.86	52.36	83.79	4	13.50
Silences	17.39	26.65	24.66	33.33	6	10.50
Lexical backchannels	4.90	10.65	7.95	13.92	1	2
non-lexical backchannels	3.14	4.90	3.49	4.82	1	2

Table 4.6 – General Statistics of *wf* and *ws* segments in AMI. Mean $\bar{x}$, standard deviation sd and median $\tilde{x}$ are reported.

Signal	$\bar{x}_{wf} \neq \bar{x}_{ws}$	$\bar{x}_{wf} < \bar{x}_{ws}$
Laugh	***	***
Overlaps	ns	ns
Silence	ns	ns
Backchannels	ns	ns
Lexical-Backchannel	ns	ns
Non-lexical-Backchannel	ns	ns

Table 4.7 – Significance values resulting from the two-tailed wilcoxon test, and from the one tail wilcoxon test alternative less - *wf* and *ws* segments in TableTalk. Medium effect size, $r = 0.31$.

Signal	$\bar{x}_{wf} \neq \bar{x}_{ws}$	$\bar{x}_{wf} < \bar{x}_{ws}$
Laugh	***	***
Overlaps	***	***
Silence	***	***
Backchannels	***	***
Lexical-Backchannel	***	***
Non-lexical-Backchannel	***	***

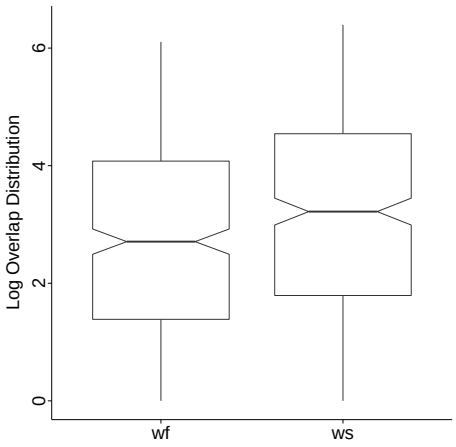
Table 4.8 – Significance values resulting from the two-tailed wilcoxon test, and from the one tail wilcoxon test alternative less - *wf* and *ws* segments in AMI. Medium effect size, with r varying between $r = 0.3$ and $r = 0.4$.

Interestingly, a pattern is found in the laughter distribution of TableTalk and AMI where the first half constantly presents a lower number of laughs with respect to the second half. Given that the null hypothesis is that no difference exists among the two distributions, one has to reject the null hypothesis as the one tailed Wilcoxon Test shows significantly fewer laughs in *wf* with respect to *ws* ($p < 0.0005^{***}$).

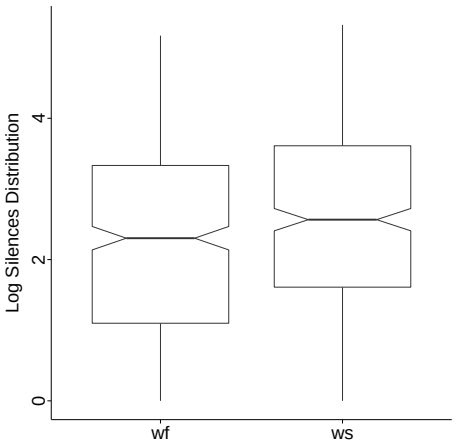
In addition, the same pattern ($wf < ws$, with significance) is found in the AMI corpus with respect to all the other signals analyzed. Fig. 4.7 provides a visualization of the distributions for overlaps, silences, lexical backchannels and non-lexical backchannels in AMI. In all these comparisons a one tailed Wilcoxon test is run and the null hypothesis (wf=ws) is rejected in favor of the alternative hypothesis $wf < ws$, $p < 0.005^{***}$.

4.5.3 Analysis 3: thirds of topic

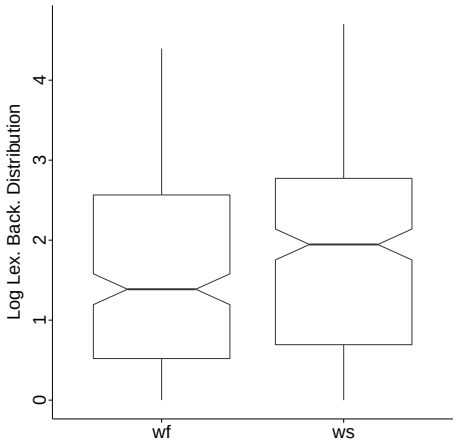
In this analysis, I compare the distribution of laughter among first, second and third segment of equal length of a topic, as in Fig. 4.3. Table 4.9 and Table 4.10 show the general statistics of the signals per window. Also in this case, the distributions are not normal. However, in order to conduct an ANOVA analysis, these distribution have been normalized extracting the logarithm, and the log function of the distribution has been analyzed. Interestingly, two different patterns are found in the laughter distribution of TableTalk and AMI. In TableTalk, an ANOVA analysis on the normalized distributions *w1, w2, w3* shows significant differences among the groups. A Tukey Honest Significant



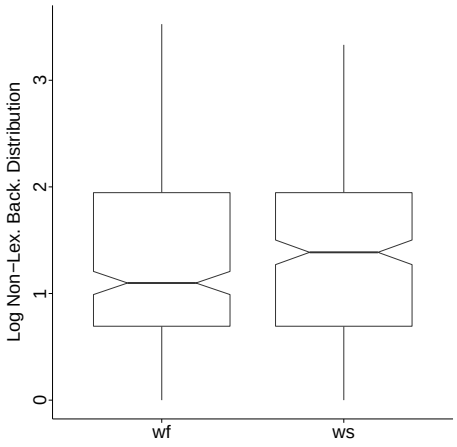
(a) Log of the **overlap** distribution in *wf* and *ws* in AMI.



(b) Log of the **silences** distribution in *wf* and *ws* in AMI



(c) Log of the **lexical backchannels** distribution in *wf* and *ws* in AMI



(d) Log of the **non-lexical backchannels** distribution in *wf* and *ws* in AMI

Figure 4.7 – The different distributions in AMI corpus between *wf* and *ws*.

Differences test[7] shows a significant difference between *w1* and *w3*, but also between *w2* and *w3*. Not between *w1* and *w2*. In TableTalk, the last part of a topic shows a higher tendency to certain laughter. In order to verify these results on the original non-normal distribution, I also conduct a two by two one tailed Wilcoxon test of null hypotheses $w1 = w3$, $w2 = w3$ and $w1 = w2$. In the first and in the second case, the null hypothesis

[7] R function TukeyHSD stats, [Yandell, 1997].

is rejected in favor of $w1 < w3$ and $w2 < w3$, p$< 0.005^{***}$. Only $w1 = w2$ is verified. In other words, there is a tendency of increase of laughs in *w3*. However no significant patterns emerge with respect to the other signals. Table 4.11 and Fig. 4.9 show the results of the statistical tests and the distribution for TableTalk.

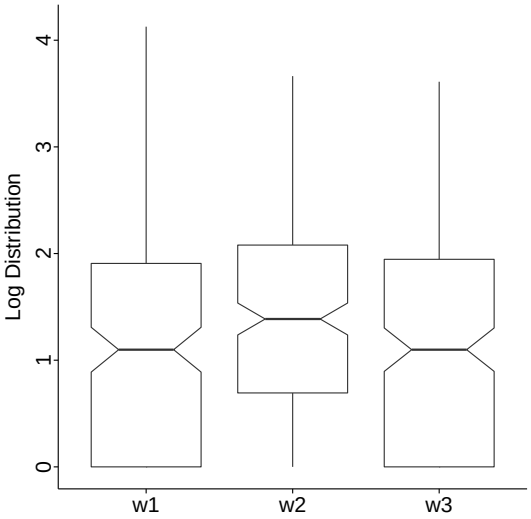


Figure 4.8 – AMI laughter distributions in $w1w2w3$. General increase of laughs in $w2$ and $w3$. Total amount of laughs: w1:1099; w2: 1278; w3: 1310.

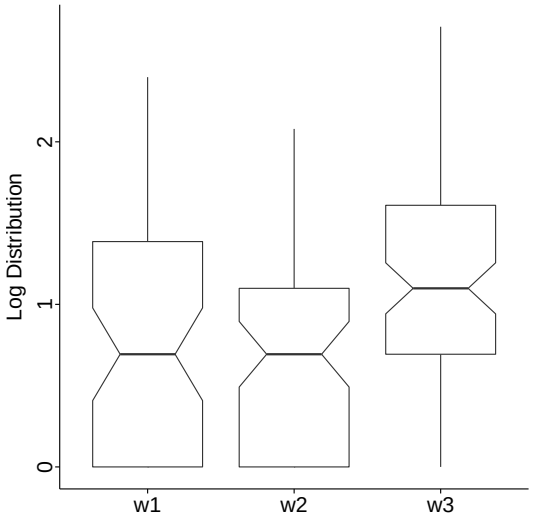


Figure 4.9 – TableTalk laughter distribution in $w1w2w3$. General increase of laughs only in $w3$. Total amount of laughs: w1:164; w2: 184; w3: 332.

Signal	$\bar{x}_{w1}$	sd_{w1}	$\bar{x}_{w2}$	sd_{w2}	$\bar{x}_{w3}$	sd_{w3}	$\tilde{x}_{w1}$	$\tilde{x}_{w2}$	$\tilde{x}_{w3}$
Laugh	1.49	1.95	1.672	1.76	3.01	3.00	1	1	2
Overlaps	3.91	4.10	4.48	4.27	4.53	3.76	3	3	4
Silences	5.32	4.11	5.2	4.32	5.38	4.57	4	4	4
Lexical Backchannels	2.43	2.63	2.65	2.79	2.73	2.63	2	2	2
Non-lexical Backchannels	1.88	1.92	1.79	2.20	1.84	2.30	1.5	1	1

Table 4.9 – General Statistics of *w1 w2 w3* segments in TableTalk. Mean $\bar{x}$, standard deviation sd and median $\tilde{x}$ are reported.

In AMI a different pattern emerges. Also in this case the analyses are conducted using ANOVA and Tukey tests on the normalized distributions and non-parametric measures on the non-normalized data. With respect to laughter, $w1$ presents a significantly smaller amount of laughter if compared with $w2$ and $w3$, while no difference

Signal	$\bar{x}_{w1}$	sd_{w1}	$\bar{x}_{w2}$	sd_{w2}	$\bar{x}_{w3}$	sd_{w3}	$\tilde{x}_{w1}$	$\tilde{x}_{w2}$	$\tilde{x}_{w3}$
Laugh	1.96	5.13	3.10	5.42	2.95	5.64	0	1	1
Overlaps	19.53	39.47	36.74	58.76	36.32	58.41	3	10	10
Silences	11.31	17.35	16.64	21.33	17.24	23.23	4	7	7
Lexical Backchannels	2.91	6.577	5.46	9.60	5.50	9.76	0	1	1
Non-lexical Backchannels	2.06	3.30	2.79	3.62	2.31	3.48	1	2	1

Table 4.10 – General Statistics of *w1 w2 w3* segments in AMI. Mean $\bar{x}$, standard deviation sd and median $\tilde{x}$ are reported.

Signal	$\bar{x}_{w1} < \bar{x}_{w2}$	$\bar{x}_{w1} < \bar{x}_{w3}$	$\bar{x}_{w2} < \bar{x}_{w3}$
Laugh	ns	***	***
Overlaps	ns	ns	ns
Silence	ns	ns	ns
Backchannels	ns	ns	ns
Lexical-Backchannel	ns	ns	ns
Non-lexical-Backchannel	ns	ns	ns

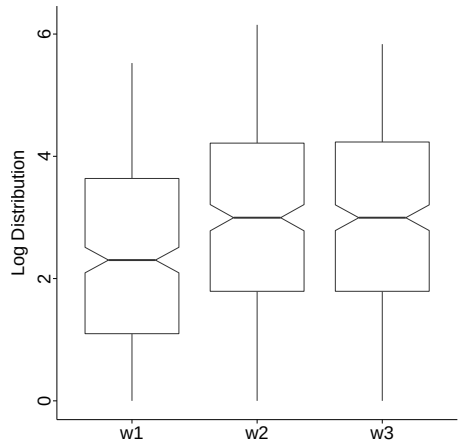
Table 4.11 – Significance values resulting from the two-tailed wilcoxon test, and from the one tail wilcoxon test alternative less - *w1 w2 w3* segments in TableTalk. Medium effect size, $r = 0.3$.

Signal	$\bar{x}_{w1} < \bar{x}_{w2}$	$\bar{x}_{w1} < \bar{x}_{w3}$	$\bar{x}_{w2} < \bar{x}_{w3}$
Laugh	***	***	ns
Overlaps	***	***	ns
Silence	***	***	ns
Backchannels	***	***	ns
Lexical-Backchannel	***	***	ns
Non-lexical-Backchannel	***	***	ns

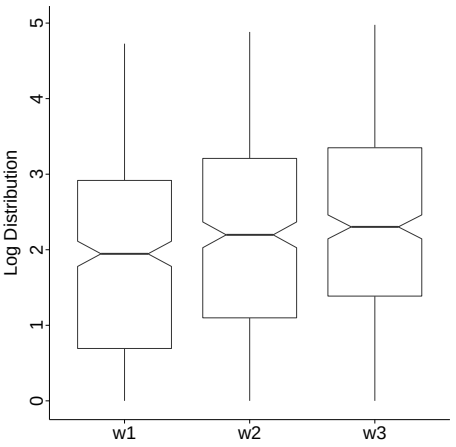
Table 4.12 – Significance values resulting from the two-tailed wilcoxon test, and from the one tail wilcoxon test alternative less - *w1 w2 w3* segments in AMI. Medium effect size, varying from $r = 0.29$ to $r = 0.4$.

is found between $w2$ and $w3$ (Wilcoxon test, one tailed alternative less, $p < 0.005^{***}$). Fig. 4.8 shows the distribution for AMI.

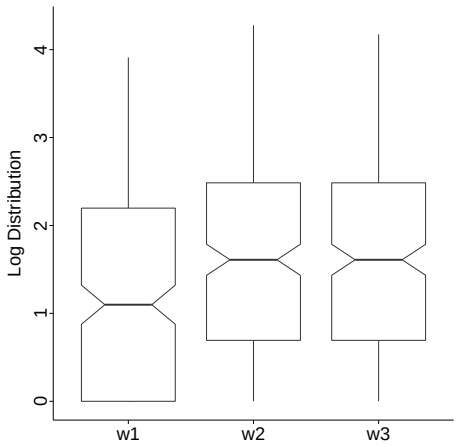
Interestingly the same pattern emerges consistently for all the other signals analyzed. As summarized in Table 4.12, a significant difference between $w1$ and $w2$, as well as $w1$ and $w3$ is found for all the signals. This emerges both from an ANOVA analysis and Tukey test on the normalized distribution, and from a two by two Wilcoxon test on the non-normalized distributions. Fig. 4.11 reports the distributions for overlaps,



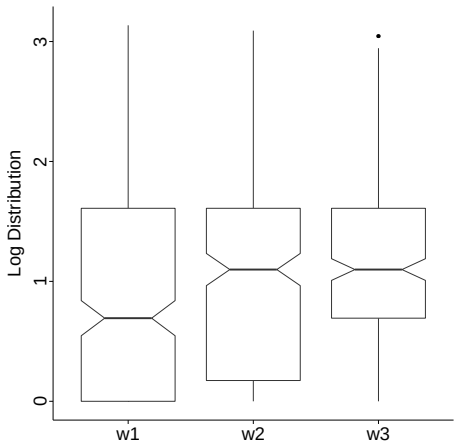
(a) Log of the **overlap** distribution in *w1, w2, w3* in AMI. Total amount of overlaps: w1:10921, w2: 15140, w3: 17955.



(b) Log of the **silences** distribution in $w1w2w3$ in AMI. Total amount of silences: w1:6323, w2: 6859, w3: 8810



(c) Log of the **Lexical Backchannels** distribution in $w1w2w3$ in AMI. Total amount of lexical back.: w1:1627, w2: 2252, w3: 2698.



(d) Log of the **non-Lexical Backchannels** distribution in $w1w2w3$ in AMI. Total amount of non-lexical back.: w1:1155, w2: 1150, w3: 1165.

Figure 4.10 – The different distributions in AMI corpus between $w1w2w3$

silences, lexical backchannels and non-lexical backchannels over the three windows in AMI.

4.5.4 Analysis 4: topic terminations vs topic beginnings

In the final analysis I explore the distributions of the social signals considered in proximity to the topic change event. As explained in Section 4.4, I consider topic terminations and topic beginnings providing an operational model for such intervals. I define a topic change, T, as the point in time in which a topic starts. Relying on Gilmartin et al. [2013a], a threshold of fifteen seconds around topic changes has been chosen as a demarkation point for topic terminations and beginning, which are defined as follows:

wt: topic termination segments from T-15 seconds to T.

wb: topic beginning segments from T to T+15 seconds.

Gilmartin et al. [2013a] analysis the probability of laughter in 5 second bins at T minus multiples of 5 seconds (T-5, T-10, T-15, T-20) in order to study laughter trends near topic terminations. A meaningful threshold emerges (T-15 seconds) where a change in the laughter trend (number of laughs increasing significantly with respect to T-20) is visible. I use this outcome to choose a threshold of T-15 seconds and T+15 seconds for defining *wt* and *wb*. Figure 4.4 depicts this operational segmentation. I then analyze the distribution of social signals among these segments with non-parametric statistical tests with Bonferroni correction; a pattern indicates a non-random relationship between topic change and the amount of social signals.

Signal	$\bar{x}_{wb}$	sd_{wb}	$\bar{x}_{wt}$	sd_{wt}	$\tilde{x}_{wb}$	$\tilde{x}_{wt}$
Laugh	0.58	1.00	1.85	1.76	0	2
Overlaps	1.45	1.44	2.12	1.55	1	2
Silences	2.35	1.31	2.37	1.39	2	2
Lexical Backchannels	0.86	1.05	1.37	1.37	1	1
Non-lexical Backchannels	0.863	1.07	0.75	0.91	1	1

Table 4.13 – General Statistics of *wb wt* segments in TableTalk. Mean $\bar{x}$, standard deviation sd and median $\tilde{x}$ are reported.

Being $\bar{x}$ the average amount of signals per segment, I analyze the distribution of $\bar{x}$ for wt and wb segments.

Signal	$\bar{x}_{wb}$	sd_{wb}	$\bar{x}_{wt}$	sd_{wt}	$\tilde{x}_{wb}$	$\tilde{x}_{wt}$
Laugh	0.28	0.79	0.62	1.38	0	0
Overlaps	2.72	4.69	4.99	6.08	0	3
Silences	2.2	2.04	3.06	2.34	2	3
Lexical Backchannels	0.32	0.76	0.70	1.13	0	0
Non-lexical Backchannels	0.45	0.70	0.37	0.65	0	0

Table 4.14 – General Statistics of *wb wt* segments in AMI. Mean $\bar{x}$, standard deviation sd and median $\tilde{x}$ are reported.

Signal	$\bar{x}_{wb} \neq \bar{x}_{wt}$	$\bar{x}_{wb} < \bar{x}_{wt}$
Laugh	***	***
Overlaps	***	***
Silence	ns	ns
Backchannels	***	***
Lexical-Backchannel	***	***
Non-lexical-Backchannel	ns	ns

Table 4.15 – Significance values resulting from the two-tailed wilcoxon test, and from the one tail wilcoxon test alternative less - wb and wt segments in TableTalk. Large effect sizes, $r > 0.5$ for laugh and overlaps, medium effect size for backchannels ($r = 0.3$).

Signal	$\bar{x}_{wb} \neq \bar{x}_{wt}$	$\bar{x}_{wb} < \bar{x}_{wt}$
Laugh	***	***
Overlaps	***	***
Silence	***	***
Backchannels	***	***
Lexical-Backchannel	***	***
Non-lexical-Backchannel	***	***

Table 4.16 – Significance values resulting from the two-tailed wilcoxon test, and from the one tail wilcoxon test alternative less - wb and wt segments in AMI. Small effect sizes, r varying between $r = 0.16$ to $r = 0.28$.

Due to the non-normality of such distribution a non-parametric test is used (Wilcoxon Test). I define H_0: $\bar{x}_{wb}$=$\bar{x}_{wt}$ and H_1:$\bar{x}_{wb} < \bar{x}_{wt}$. The null hypothesis is rejected in favor of H_1 in the majority of cases with $p < 0.001$. As one can see from Tables 4.15 and 4.16, a general trend emerges in both TableTalk and AMI: topic terminations, compared to topic beginnings, show a significantly greater presence of social signals. In AMI, for each social signal—laughter, overlap, silence, backchannels—in *wt* and *wb* the non-parametric Wilcoxon test rejects the null hypotheses of $wb = wt$ and $wb \geq wt$, and

validates the alternative hypothesis of $wb < wt$, $p < 0.001^{***}$. In TableTalk the same applies to laughter, overlaps, and lexical backchannels.

Discussion In conclusion, a constant trend emerges in both TableTalk and AMI: topic terminations show a significantly higher presence of signals if compared to topic beginnings. In AMI, among all the distributions of frequencies of laughter, overlaps, silences, lexical and non-lexical backchannels in *wt* and *wb* the non-parametric Wilcoxon test rejects the null hypothesis of $wb = wt$ and validates the alternative hypothesis of $wb < wt, p < 0.0005^{***}$. In TableTalk the same applies to laughter, overlaps, and lexical backchannels. In other words, topic terminations reveal higher interactional entropy (as defined in 4.5) than topic beginnings.

4.6 Analysis of the topic change neighborhood

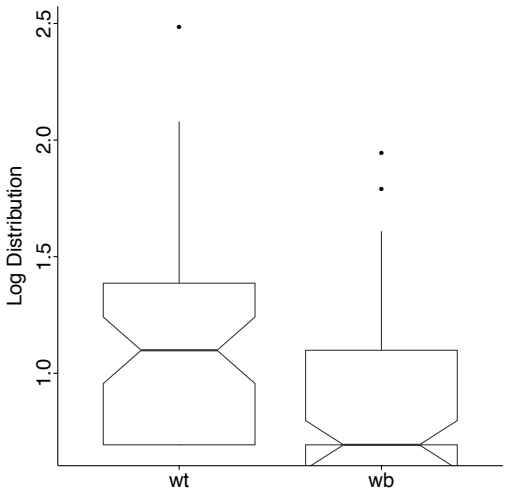
4.6.1 Social signals distribution surrounding a topic change

From the preceding analysis a consistent higher frequency of social signals in topic terminations rather than in topic beginnings emerges. However, this result does not explore in detail the temporal relation between the events opening three possible scenarios: *1)* the decrease in interaction is precedent the topic change; *2*) the topic change is precent the decrease in interaction; *iii)* both the topic change and the decrease of interaction are subsequent to a potential latent external event.

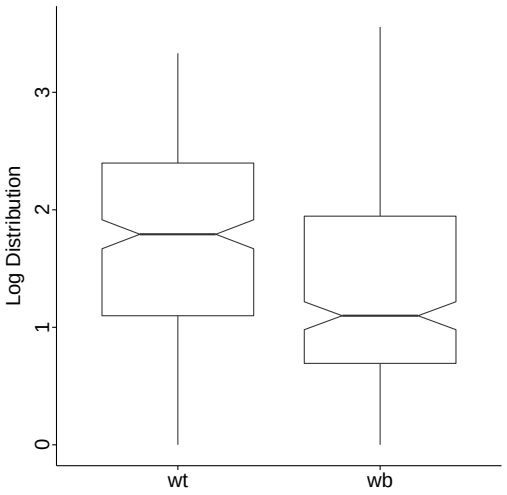
In order to have a fine grained idea, I concentrate my analysis on the topic terminations and beginnings, examining where the decrease happens.[8].

Method I split the *wt* and *wb* segments into bins of 5 seconds which go from T to T-5, T-10, T-15 seconds and similarly from T to T+5, T+10, T+15 as shown in Fig 4.12 (bottom part). I then calculate the average amount of events per bin, over all the topics

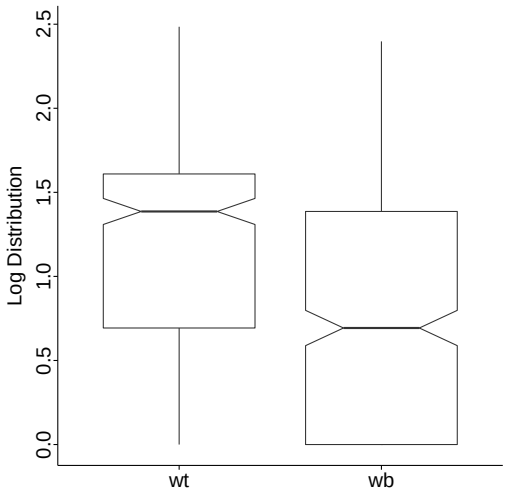
[8]Part of the analyses of this section have been published in Bonin et al. [2015]



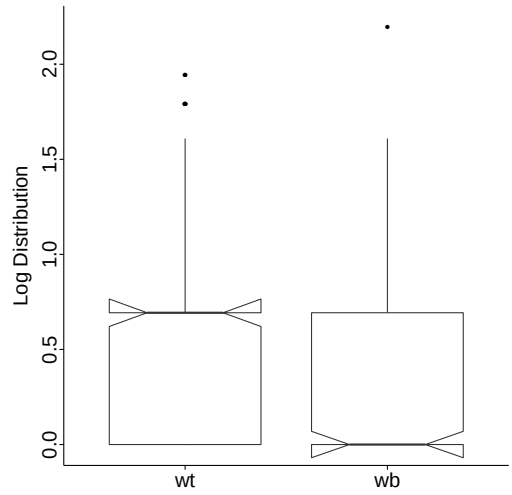
(a) Log of the Laugh distribution in wt and wb in AMI. Total amount of laughs: wt: 262; wb: 159



(b) Log of the Overlap distribution in wt and wb in AMI. Total amount of laughs: wt: 2092; wb: 1516



(c) Log of the Silences distribution in wt and wb in AMI. Total amount of laughs: wt:1284; wb: 1227



(d) Log of the Backchannels distribution in wt and wb in AMI. Total amount of laughs: wt: 451; wb: 430

Figure 4.11 – The different distributions in AMI corpus between $w1w2w3$

per AMI and TableTalk.

Results Fig 4.13 to Fig. 4.16 show the mean occurrences frequencies of signals from T-15 to T and T to T+15, in bin of 5 seconds for AMI. Fig 4.17 to Fig 4.20 refer to

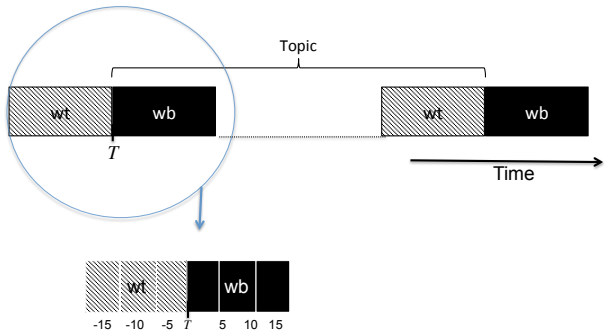


Figure 4.12 – At the top: *wt* and *wb* segments - [Analysis 4]. At the bottom: fine grain segmentation of *wt* and *wb* - [Analysis neighborhood].

TableTalk. The bars *wt15, wt10, wt5* represent the bins before a topic change (light grey), the bars *wb15, wb10, wb5* represent the bins after a topic change (dark grey). The stars indicate significant differences in the distributions. With the exception of silences, a trend emerges: a drop in the frequency of events occurs after the topic change. This is also confirmed by a series of statistical tests aimed at testing the differences in the distributions among consecutive windows. Once again the test applied is a non-parametric test (Wilcoxon, two tailed) due to the non-normal nature of the distribution. Table 4.17 and Table 4.18 summarize the results of these tests for AMI and TableTalk respectively. A significant drop in the frequency of signals is visible between wt_5 and wb_5 (signaled by the stars in the figures), but not, generally, between the other bins. In the light of these results, it is reasonable to suppose that in the datasets examined, the decrease of social interaction, signaled by a decrease in social signals, is a consequence of topic change, as it occurs immediately after the topic change.

Discussion In general, in the light of these results, it emerges that, in the dataset examined, the decrease of interaction happens immediately after the topic change. Given this temporal sequence, a possible conjecture is that the topic change causes of the decrease in interaction, and not the opposite.

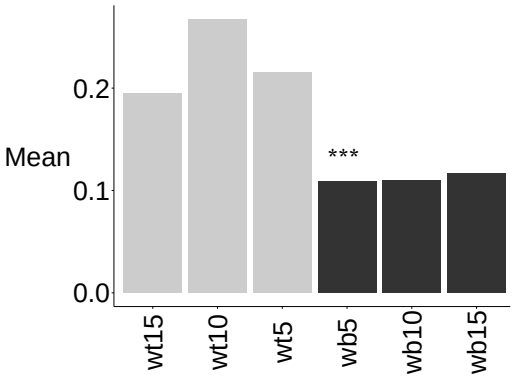


Figure 4.13 – AMI: laughter distribution

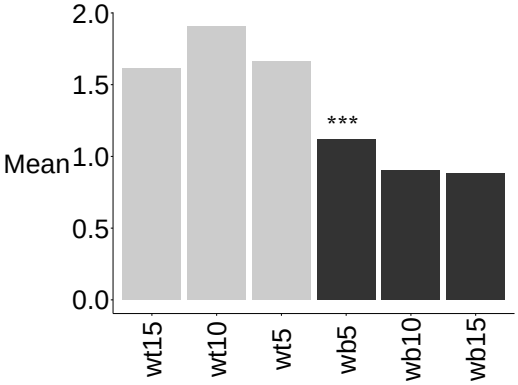


Figure 4.14 – AMI: overlap distribution

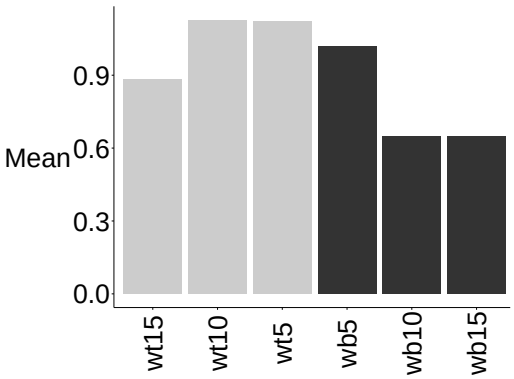


Figure 4.15 – AMI: silence distribution

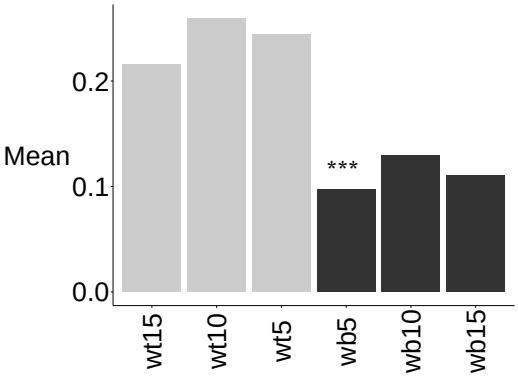


Figure 4.16 – AMI: backchannel distribution

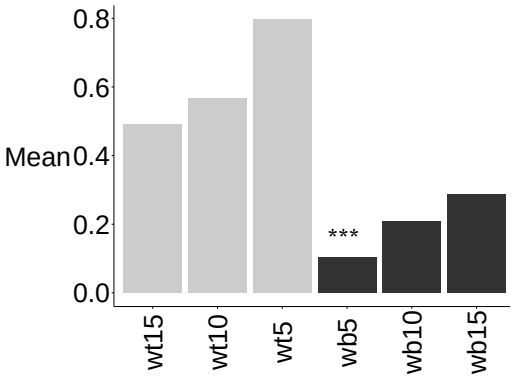


Figure 4.17 – TableTalk: laugh distribution

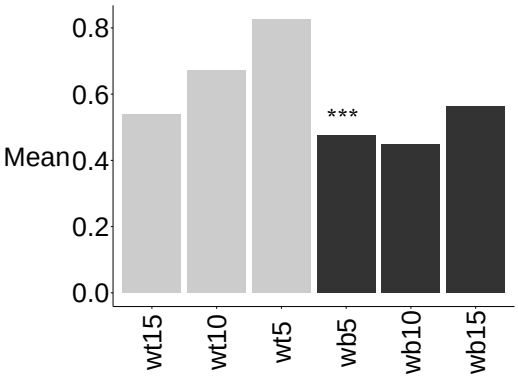


Figure 4.18 – TableTalk: overlap distribution

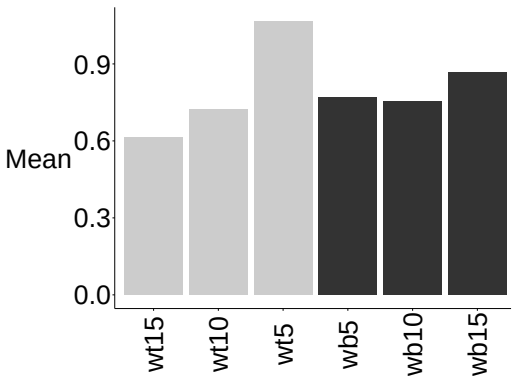


Figure 4.19 – TableTalk: silence distribution

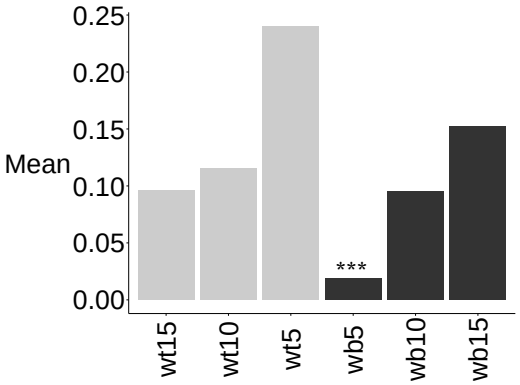


Figure 4.20 – TableTalk: backchannel distribution

4.6.2 Lexical volume distribution around a T

In the previous section I analyzed the distribution of signals around the topic change, showing how, in the analyzed data, a drop of social signals is found immediately after the topic change. This drop is a consequence of T, rather than its cause.

If the beginning of a topic shows a lower level of social interaction, two hypotheses emerge: *wb* are characterized by higher contribution in lexical content, or a greater number of silences (the latter could indicate a higher cognitive activity, additional to speech production). The second hypothesis can be rejected reflecting on the experiments in the previous section, as it is not found that *wb* segments have a significantly higher number of silences than *wt* segments (the opposite is found in AMI).

Method On the other hand, in order to explore the first hypothesis, we examine the distribution of lexical volume in wb_5 and wt_5 segments.[9]

I define lexical volume as the amount of lexical contributions in a segment excluding punctuation. Being *w* a token in an segment *S*, lexical volume of *S* is indicated as LV_S and corresponds to the amount of *w* in *S*. No normalization over the length of the segment is applied, as I always compare segments with the same length.

[9]Experiments conducted on wb_{15} and wt_{15} lead to the same result.

Signal	x	y	$\bar{x} \neq \bar{y}$	$\bar{x} > \bar{y}$
Laugh	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	ns	ns
	wt_5	wb_5	**	**
	wb_5	wb_{10}	*	ns
	wb_{10}	wb_{15}	ns	ns
Overlaps	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	ns	**
	wt_5	wb_5	**	***
	wb_5	wb_{10}	*	ns
	wb_{10}	wb_{15}	ns	ns
Silences	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	ns	ns
	wt_5	wb_5	ns	ns
	wb_5	wb_{10}	ns	ns
	wb_{10}	wb_{15}	ns	ns
Lexical Backchannels	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	ns	ns
	wt_5	wb_5	****	****
	wb_5	wb_{10}	ns	ns
	wb_{10}	wb_{15}	ns	ns
Non-lexical Backchannels	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	ns	ns
	wt_5	wb_5	ns	ns
	wb_5	wb_{10}	ns	ns
	wb_{10}	wb_{15}	ns	ns

Table 4.17 – AMI: Significance table of the distributions of social signals in bins around topic changes. The significant difference is found in a decrease in events immediately after the topic change (wb_5). Small effect sizes (Laugh: r=0.12; Overlaps: r=0.13; Lexical Backchannels: r=0.18).

Results Fig. 4.21 and Fig. 4.22 show respectively the distribution of lexical volumes per AMI and TableTalk corpora. In both cases, *wb* segments show significantly higher lexical volume than *wt* segments, as confirmed by T-student test, with H_0: $LV_{wb_5} = LV_{wt_5}$, and H_1: $LV_{wb_5} > LV_{wt5}$. The null hypothesis is rejected in favor of H_1, $p < 0.001$.

Signal	x	y	$\bar{x} \neq \bar{y}$	$\bar{x} > \bar{y}$
Laugh	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	*	ns •
	wt_5	wb_5	****	****
	wb_5	wb_{10}	ns	ns
	wb_{10}	wb_{15}	ns	ns
Overlaps	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	ns	**
	wt_5	wb_5	**	***
	wb_5	wb_{10}	ns	ns
	wb_{10}	wb_{15}	ns	ns
Silences	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	*	ns •
	wt_5	wb_5	*	*
	wb_5	wb_{10}	ns	ns
	wb_{10}	wb_{15}	ns	ns
Lexical Backchannels	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	*	ns •
	wt_5	wb_5	****	****
	wb_5	wb_{10}	*	ns •
	wb_{10}	wb_{15}	ns	ns
Non-lexical Backchannels	wt_{15}	wt_{10}	ns	ns
	wt_{10}	wt_5	ns	ns
	wt_5	wb_5	ns	ns
	wb_5	wb_{10}	ns	ns
	wb_{10}	wb_{15}	ns	ns

Table 4.18 – TableTalk: Significance table of the distributions of social signals in bins around topic changes. The significant difference is found in a decrease in events immediately after the topic change (wb_5). Cases indicated with • show significance: $\bar{x} < \bar{y}$. Medium and large effect sizes (Laugh: r=0.60; Overlaps: r=0.30; Silence: r=0.25; Lexical Backchannels: r=0.44).

Discussion The first five seconds of a topic show, therefore, a higher presence of lexical content and a lower presence of social signals. I finally investigate whether the higher lexical content in *wb* is in relation with a higher number of speakers. Interestingly, *wb* segments show a constantly lower number of speakers where compared with *wt* segments (T-test, $p < 0.001$).

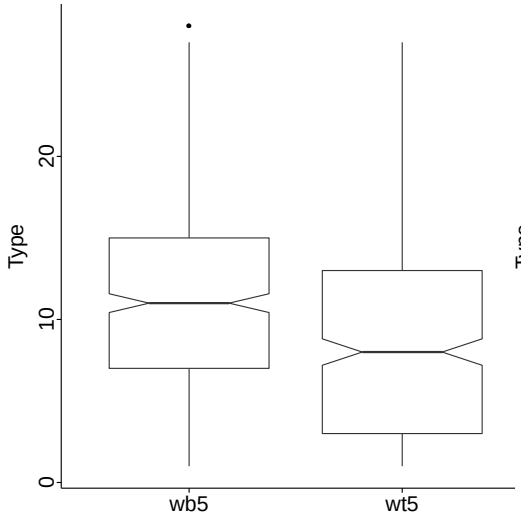


Figure 4.21 – AMI:Lexical Volume in wb_5 and wt_5

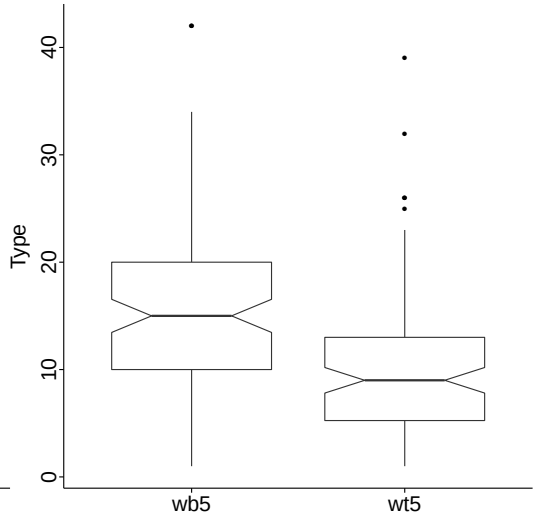


Figure 4.22 – TableTalk:Lexical Volume in wb_5 and wt_5

4.7 Fine grained analysis of laughter timing

In the previous section, we have seen that laughter maintains a distinctive behavior around T in both a meeting and a spontaneous corpus. In this section a more in-depth study on the laughter timing is presented. I consider only laughter events in relation to T. First I explore the distance between laughter in general and T, looking at the time spans between the last laugh in topic A and T (namely LT) and the T and the first laugh in topic B (namely TL). For a clearer visualization of this please refer to Fig. 4.23. Then, I will analyze the behavior of types of laughter, shared vs. solo, with respect to T. In this case, my foci are the last solo (SO) and shared (SH) laughs prior to a T (named LL: SoLL or ShLL, respectively). Fig. 4.24 provides a visualization of this.

4.7.1 Shared laughter annotation

In order to analyze the dynamics of shared and solo laughter in both corpora, an annotation of whether a laugh is isolated or shared is necessary. TableTalk and AMI do not provide such detailed annotation. Hence, a novel strategy for shared laughter annota-

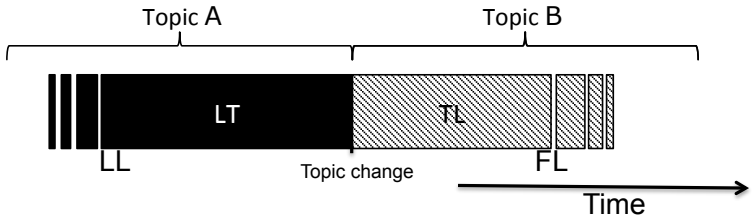


Figure 4.23 – Topic boundary neighborhood. LL and FL represent the last and the first laughs. LT and TL represent respectively a topic termination segment and a topic beginning segment.

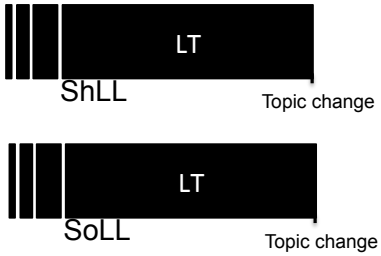


Figure 4.24 – Topic boundary left neighborhood with shared and solo last laughs (ShLL and SoLL).

tions was developed. Previous work by Bonin et al. [2012b] had defined shared laughter as overlapping laughs or consecutive laughs within 1 second distance. This was based on the intuition that consecutive laughs, if separated by a small enough distance would still be experienced and externally perceived as shared. This threshold was experimentally determined without the existence of a gold standard to refer to. Here, as in work by [Bonin et al., 2014a], I test an extreme position that only truly overlapping laughter is to be regarded as shared. Therefore, I consider shared co-occurrent laughter of different speakers, where co-occurrent indicate overlapping as well as successive laughter with no gap between them. The reason for this stand in investigating a baseline situation in which a laughter has to overlap or occur sequentially without an intervening gap, in order to be defined as a shared laugh.

I extend this annotation to the entire TableTalk and AMI and Table 4.19 and 4.20 indicate the figure regarding shared and solo laughter in AMI and TableTalk respectively.

However, as it has been noted by others [Truong and Trouvain, 2012], the annota-

	Avg	SD
Shared	25.46	19.03
Solo	48.89	41.29
Sh+So combined	81.8	105.13

Table 4.19 – Average distribution of laughs per speaker in the AMI corpus.

Speaker	Shared	Solo	Total
d	33	5	38
g*	20	9	29
k	138	56	194
n	175	59	234
y	162	56	218

Table 4.20 – Distribution of laughter among speakers - *Speaker g participated only in Day 2.

tions of the temporal aspects of laughs in AMI are partly flawed. The current analysis focuses only on the laughs which posses start times different to end times.

4.7.2 Laughter & topic: temporal distributions

In the first analysis an attempt is made to understand whether *there is a pattern in the temporal distribution of laughter with respect to topic changes in the analysed corpora.* I remind the reader to Appendix A where two excerpts, from the AMI corpus, are reported, in order to provide a deeper insight on the data that are analysed in this section.

Method I define the measure μ(x) as the distance in seconds between x and T. I consider μ(LT) and μ(TL) (Fig. 4.23), being LT the last laugh before T and TL the first laugh after T.

Results As shown in [Bonin et al., 2012b][10], analysis of TableTalk shows that LTs tend to occur at a shorter temporal distance from the T, than TLs: $\mu(\text{LT}) < \mu(\text{TL})$.[11] The

[10] These results have been presented in [Bonin et al., 2012b] and they are reported here with co-authors' consensus.

[11] One tail wilcox.test, mu=0, alternative less: p-value < 0.005.

temporal distance between the last laugh of a topic and topic boundary, is significantly shorter than the temporal distance between the topic boundary and the first laugh, and Fig. 4.25 shows this difference in distributions.[12] From the parallel analysis of these two corpora an interesting finding emerges: laughter becomes more likely to occur when the temporal distance from the topic boundary increases. Although the two corpora present a similar behavior, it is worth noticing the difference in the distance between laughs and topic boundaries. In TableTalk the first laugh after a topic change happens (median value) around 27 seconds after the beginning of a topic, while in AMI after 30 seconds. The last laugh tends to happen around 9 seconds before the end of a topic in TableTalk, and around 26 seconds before the end of a topic in AMI. Although aware of the gross nature of the median, those results may be due to the fact that TableTalk is characterized by shorter topics and a more dynamic and unstructured exchange than AMI.

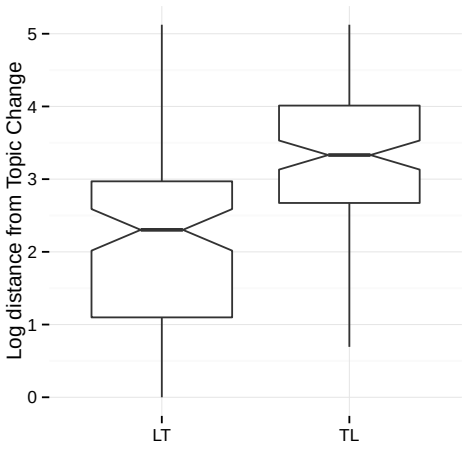


Figure 4.25 – $\mu(LT)$ vs $\mu(TL)$ comparison in TableTalk.

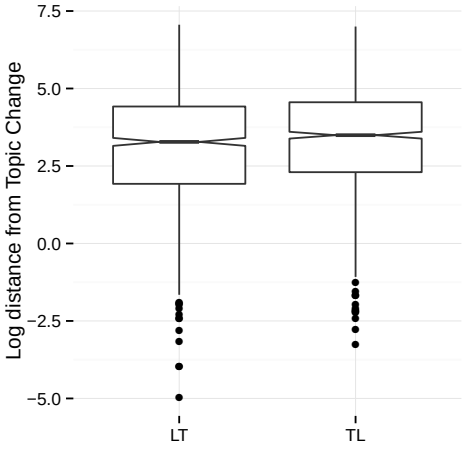


Figure 4.26 – $\mu(LT)$ vs $\mu(TL)$ comparison in AMI.

[12] In Fig. 4.25, I report the logarithm of the distribution to emphasize differences visually. A one tailed Student's T-Test on the logarithm of the distribution is in line with the Wilcox test on the raw data.

Ramification From this analysis, it emerges that laughter is more likely as the temporal distance from the start of the topic increases. This finding is not sufficient to support the fact that laughter can be considered, in isolation, a valid topic termination cue, but suggests that laughs are more likely to occur at the topic terminations, rather than immediately after a topic change (at the topic onset). The particular distribution of laughter emerging from the two corpora underlies a discourse function of laughter which could be useful information in automatic topic boundary detection (cf. Luz [2012]).

4.7.3 Shared laughter and topic termination

Having considered the laughter distribution at a coarse grain level, in this subsection I refine my analysis exploring the temporal distribution of shared and solo laughs with respect to topic changes. In the following I will examine:

(a) Distribution of shared/solo laughter at topic terminations.

(b) Distribution of shared/solo laughter in topic continuation vs. topic transition moments.

In order to investigate (a) and (b), I refer to previous studies that explore similar distributions in a telephone conversation corpus of English native speakers. Holt [Holt, 2011, 2010] proposes a correlation between shared laughs and topic termination sequences. According to Holt [2011] shared laughs may be part of a topic termination sequence and may introduce to end of the topic. The mutual acceptance of a laugh relates to the common agreement of a completed topic. Hence, I analyze whether, in our corpora, evidence is found of shared laughter occurring closer to the end of the topic than solo laughter. I refine our previous analysis of μ(LT) vs μ(TL), distinguishing between shared (SH) and solo (SO) laughs. Since the interest is only in the topic termination subsection, I focus on the topic boundary left neighborhood μ(LT) and explore the distance between shared laughs (SH) and topic change μ(ShLT) and solo laugh

(SO) and topic change μ(SoLT). As shown in Fig. 4.27, in TableTalk, some evidence is found of shared laughs being closer than solo laughs to topic termination boundaries, but this tendency does not reach significance. In particular, it has to be noticed how the median distance of a SH from topic termination is 7 sec, while the median distance of SO from topic termination is 12 sec.

This result is different from that reported in the initial work of Bonin et al. [2012b], due to the differences in annotation between that work and the present analysis, as described above (Section 4.7.1). In what remains we retain the constraint that only temporally overlapping laughs count as moments of shared laughter.

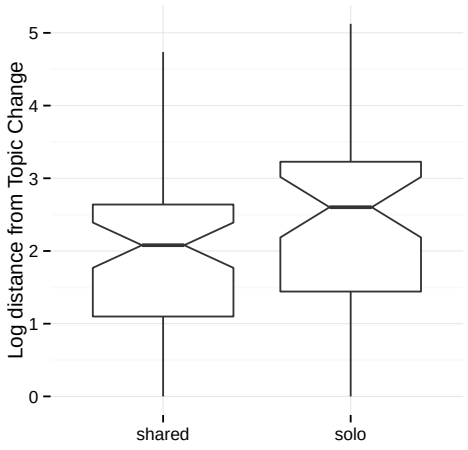


Figure 4.27 – $\mu(ShLT)$ and $\mu(SoLT)$ in TableTalk.

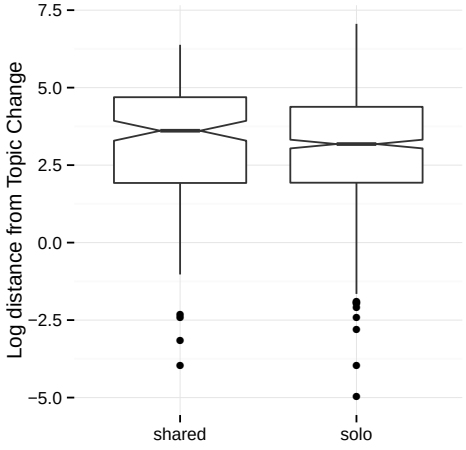


Figure 4.28 – $\mu(ShLT)$ and $\mu(SoLT)$ in AMI.

A similar behavior is found in AMI. I compare μ(ShLT) and μ(SoLT), finding that SH do not tend to occur more in proximity of the end of the topic than SO (no significant difference in the distributions). This is shown in Fig. 4.28.[13] In the AMI corpus, the median distance of SH and SO from topic termination is 28 sec. and 30 sec. respectively. Therefore, differently from previous studies, in these corpora topic termination sequences do not appear to be characterized by shared laughter more than by solo

[13]Fig. 4.28, I report the logarithm of the distribution to emphasize differences visually. One tailed Student's T-Test on the logarithm of the distribution is in line with the Wilcox test on the raw data.

laughter.

4.8 Discussions of the experiments

This Chapter reports the results of the analyses of the dynamics of five signals belonging to the social sphere around a structural event as a topic change. These social signals could represent a first indicator of the general level social activity, called here interactional entropy, in the conversation.

One could imagine three different but equally reasonable distributions depicting the dynamics of social interaction in a conversation.

Distribution [A] - Right skewed distribution: high social activity in the beginning when the topic is new and everybody wants to contribute to it, followed by a continuos decrease until the topic dies and a new one comes in.

Distribution [B] - Normal distribution: low activity in the beginning of a topic, higher activity in the central part of the conversation, where the discussion is more lively, low activity in the end when a general agreement is reached and one person is summarizing it (meeting scenario), or the topic is exhausting (spontaneous chat scenario).

Distribution [C] - Left skewed distribution: low activity in the beginning when someone is introducing the new topic, with an increase of activity until the introduction/change to a new topic.

Figure 4.29, 4.30, 4.31 provide a visualization of the three different scenarios.

Analysis 1 has been deployed to verify **Distribution [B]**, since it provides an understanding of the distribution in the central part of a topic and the edges of the topic (transition vs. continuation segments). **Analyses 2 -3** have been deployed to verify **Distributions [A] and [C]**; in both cases an understanding of the dynamics of social activity during the topic was necessary. It emerges that social activity has the tendency to increase during the topic, but not in a monotonic way. Clearly though,

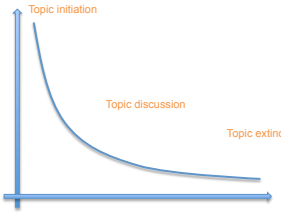


Figure 4.29 – Distribution [A]: high social activity in the beginning of a topic, low at the topic termination

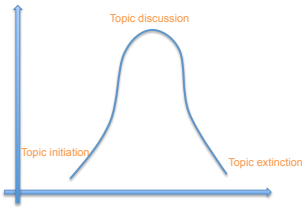


Figure 4.30 – Distribution [B] low social activity in the beginning and termination, high in the topic discussion

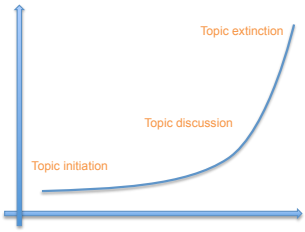


Figure 4.31 – Distribution [C]: low social activity in the beginning of a topic, high at the topic termination

evidence is found of difference in social activity between the end of and the beginning of a topic. **Analysis 4** is intended to verify the difference between the end and the beginning of a topic, by looking at topic termination and topic beginning segments.

Analysis 1, see Section 4.5.1, showed that topic transition vs topic continuation segmentation does not provide deep insight in the distribution of social signals as reported in Table 4.4 and Table 4.3, excluding Distribution [B]. The reason for that could be the fact that topic transition segments cover a variety of socially different moments (the end and the beginning of a topic), flattening possibly interesting patterns. For this reason further analyses have been conducted to identify the distribution within a topic. In Section 4.5.2, the distribution of the signals between first and second halves of conversations, as in Fig. 4.2, is compared. This analysis uncovers an increase in social activity between the first and the second half of a topic. This increase is evident in the case of laughter for both corpora, and it is significant also for all the other signals in AMI data. However, to achieve a more fine grained perspective on the distribution of the signals, Analysis 3 (Section 4.5.3) provides a detailed investigation of the first, second and third thirds of a topic, showing how in AMI a real increase of the considered social signals occurs after the first third of the topic.

This might be interpreted as a change in the social dynamics between the beginning of a topic and the rest of the topic: in other words, *something* is lacking in the beginning of a topic, and is re-appearing when the topic is established and enters in its

full development. In fact, while analyzing the beginning and termination of a topic, an evident decrease in social activity is found in the beginning segment.

Further analyses (Section 4.6.1) give better insight into the causal relationship of these events, showing whether it is the topic change to cause to drop in events or the drop in events to cause the topic change. Since the decrease in events is consistently found immediately after a topic change, one could reasonably suppose that this decrease in social events is a consequence of the topic change.

This has led to the conclusion that Distribution [C] is the one that models the actual dynamics of interactional entropy within a topic.

It appears that the beginning of a topic is characterized by a participant taking the flow, with very few laughs, interruptions and lexical backchannels overlapping his speech, while the center and end of a topic present a higher level of social interaction, laughs and interruptions. In other words, one could say that the beginning of a topic has lower social activity.

To conclude, in this Chapter, I have analyzed the dynamics of five signals belonging to the social sphere around topic change moments in dialogue. Although not exhaustive, these social signals could represent the general level of social activity in the conversation, defined in 4.5 as *interactional entropy*. It has been noticed that the beginnings of a new topic show a lower presence of social activity, but a greater amount of lexical content. In contrast, topic terminations show higher social activity and lower lexical volume. I found that both in AMI and TableTalk there is a drop in interactional entropy when a new topic begins. One could interpret this as a social order introduced by the new topic: from a situation of high social interaction, with a higher number of overlaps, feedback, laughter, the new topic leads to a monological situation, in which one speaker takes the floor, reducing the interactivity among the participants. Although limited to the datasets considered, a pattern has, therefore, been found between the fluctuation of social interaction (given by the amount of social signals) and discourse phenom-

ena such as topic changes. In addition, it has also been noticed how topic changes are the cause of the drop in interactional entropy, and not vice-versa, as the drop occurs immediately after a topic change.

These analyses show that it is possible to derive a particular behavior of social signals with respect to topic changes, leading to the conclusion that, although usually categorized as having a non-linguistic function, social signals do have a correlation with events at the discourse level and therefore a discourse function [RL_1].

4.9 Summary

In this Chapter I explored the relation between social signals and discourse phenomena such as topic changes, investigating whether social signals have a discourse function in addition to their social function. I proposed different analyses that investigated the temporal dynamics of laughter, backchannels, silences and overlaps finding a relationship between topic changes and a decrease in social signals. The results show that immediately after a topic change there is a significant drop in social activity. Defining *interactional entropy* as the amount of social signals in a given segment of a conversation, I conclude that after a topic change a decrease of interactional entropy occurs. This information might be used to better understand the discourse structure via non-linguistic information such as laughter, overlaps backchannels and silence, and shed new light upon the discourse functionality of social signals. While in this Chapter the focus has been on the social context of the conversation, in the next Chapter, I consider the situational context, as the ensemble of situational events in which a conversation takes place, and their relationship with the discourse structure.

CHAPTER 5

Situational Context and Event Segmentation

In this Chapter, I provide the analysis related to *RL2* and investigate whether and how situational signals influence the prediction of events in social interactions. As discussed in Chapter 1, the general purpose of this thesis is exploring the relationship between discourse and context. While in Chapter 4 I have analyzed the relation between discourse and social context, here I analyze the relation between discourse and the situation in which the conversation takes place. In particular I analyze the distribution of relevant events, specifically noteworthy events, providing a definition of noteworthiness, and investigate the relation between situational information and the annotation of noteworthy events in telephone conversations. I finally propose a classification experiment showing how situational information boosts the performance of an automatic system for the detection of relevant events.[1]

[1]The analyses of this Chapter are the result of a collaboration with Telefonica Research, partly published in [Bonin et al., 2014c].

5.1 Definitions of situational context and relevant event

The previous Chapter focuses on the relations between the social context in a conversation, as the ensemble of the social signals exchanged during the conversation, and a structural event such as topic changes. In this Chapter I broaden the perspective and explore the role played by the situation in which the conversation takes place. As described in Chapter 2, many researchers have discussed the concept of context at an extended level, considering context as the situational condition under which the language is spoken [Malinowski, 1947]. Hymes [1972] consider interactions as a series of speech events in which the settings of the conversation play an important role, and Duranti and Goodwin [1992] as a way towards understanding the relationship between language structure, social organization and culture. Within this framework, in this chapter the answer to the following question is sought: *does the situational context influence the way we, as speakers, structure conversation?* I explore whether, in a real life scenario, the situation, such as the fact of being in a car driving while talking on the phone, influences the prioritization of relevant information. In this, two concepts need to be defined: situational context and relevant event. I propose a working definition of situational context of a conversation (as mentioned in Section 2.3.2). I define situational context as: the set of conditions which belong to the situation in which the conversation takes place. This information can be known or can be derived from other known elements of the conversations. Situational signals are then elements of this set.

These include, for example: the gender of the participants (if known); the location in which the conversation takes place (outdoor/indoor or the exact geo-spatial location if known); the number of participants, the time of the call, the day of the call, the relation among the participants in a telephone conversation belong to the situational context.

Differently from social signals, situational signals do not change during the unfolding of the conversation (the relation among participants will be the same from the

beginning to the end, the fact that a participant is *at work* or *at home* does not change in analyzed conversations); therefore the timing of situational signals does not belong to the same scale of social signals timing: while the latter follows the evolution of the conversation, the situation signals timing follow the evolution of the situation.

By relevant event, in this thesis, I refer to an event that the speaker would not like to forget, therefore, an event in the conversation the speaker would like to take note of.

My hypotheses is that what is considered relevant and, most of all, when participants decide to express during a conversation relevant chunks of information may depend on the characteristics of the participants' relation and on the situation in which the call takes place. In any conversation there are chunks that participants do not consider worthy to remember (hence taking note of) and chunks that are considered worthy to take note of. The former situation may arise for two reasons: *1)* the event is not an important point of the conversation or *2)* the event belongs to the background knowledge of the participants, and, as a consequence, there is no possibility that it will be forgotten.

To clarify this concept I present two scenarios.

[Scenario 1] A conversation is taking place between Mary and a doctor's assistant, and, the end of the call, Mary needs to take an appointment with the doctor;

[Scenario 2] A conversation is taking place between Mary and her friend, and they agree to meet to go to the cinema on Saturday.

In scenario one, it is plausible to imagine that Mary will feel the need to remember the name of the doctor she was talking to and the date/time of the appointment after the call, because all these are *new* information gathered from the call. In scenario two, Mary might feel the need to remember the date/time fixed with her friend, but it is quite unlikely (though not impossible) that she will feel the need to annotate the name of the friend. Whereas while taking an appointment with the doctor, it is plausible that

the need exists to annotate the name of the doctor, in a social call with a friend or a relative, the name of the interlocutor is part of the background knowledge of the user. At the linguistic level the name of the doctor and the name of the friend lie at the same level and belong to the same category (named entity of Person); therefore, while from a content (and an NLP) point of view both names are Person named entities and carry the same amount of information, from the point of view of the user they might have different weights (*no need of taking note* opposed to *need of taking note*).[2]

So the question that arises is how can a system distinguish *names* that are important to remember, and names that are not important to remember. Situational context, such as knowing among who is the conversation taking place, provides in this case an unambiguous information to detect whether some piece of information belongs to the background knowledge of a user or not.

In this work I take in consideration a subset of situational conditions and analyze their interaction with the discourse; namely: information about the participant, the location and the time.

5.1.1 Participants of the Conversation

As mentioned, characteristics of the participants of the conversation may play a role in the distribution of events. Some NLP tasks have exploited user information for different purpose. Topic change detection, for example, have used the role of the participants to infer topic changes in news datasets [Arguello and Rosé, 2006], as well as meeting summarization have exploited the role of participants to gain information about who is leading, hence most likely summarizing, the meeting [Oya et al., 2014]. However, in this case, I am interested in understanding whether some of the participants' features

[2]Referring to the information structure theory one could talk about the rheme, and the new in the conversation, and the theme as the given in the conversation Halliday [1967]. However while the rheme is what is new with respect to what is known within the conversation, in the concept of relevance given here, an information is new with respect to the background knowledge of the speaker. For example, in the following example:

It is Paul, who went to the cinema.

Paul, is the new information, independently from the fact that Paul is well know by the speaker of not.

(e.g. gender, age) are correlated with the distribution of relevant events.

5.1.2 Location of the Conversation

The location of a conversation refers to *where*, as well as in which situation, or performing which action the conversation takes place. For example, if a conversation takes place in an outdoor noisy environment, the participants might prioritize the relevant information that needs to be delivered according to the loudness of the environment; similarly, if a conversation takes place over a mobilephone while one of the participants is driving, the actual dynamics of the call might depend on the stress level of the situation that the participant is experiencing. Since conversations have become ubiquitous, there is a need to take the location element into account.

5.1.3 Time of the Conversation

Similarly to the conversation's location, the conversation time can be a factor influencing the dynamics of the speech exchange. In particular this relates to the purpose of the conversation. If the conversation takes place during working hours, one can reasonably assume it is a *working* conversation, with a particular structure, while in the evenings or weekends the conversation is more likely to be a spontaneous chat. This might have an influence on the length of the conversation as well as on the prioritization of the items to be discussed.

In particular, I analyze when within a conversation a noteworthy event takes place, with respect to the time of the day, and categorical distinctions of working vs non-working hours.

5.2 Definition of relevant events

Following the definition provided in Bonin et al. [2014c], I consider relevant an event that a user would like to remember at a later time: an event that the user would like

to take note of (i.e noteworthy). As described in Chapter 2, noteworthiness detection in conversations is to be considered a particular form of summarization: the aim is to summarize the conversation maintaining only the chunks that the participants wish to recall. However, the concept of *relevance* in noteworthiness detection does not include informative fragments of content, which would be part of a summary, unless those fragments are also worth remembering for future recall. To the best of my knowledge not many scholars have investigated the possibility as well as the necessary knowledge for automatically detecting of what is worth taking notes of in a conversation. See Section 2.4.4.

5.3 Dataset and Annotation of relevant events: noteworthiness

The dataset used for the analyses in this Chapter is the Callnotes corpus [Carrascal et al., 2012], described in details in section 3.4. The corpus consists of 796 spontaneous telephone conversations among Spanish speakers, recorded during a study carried on by Telefonica Research. The Callnotes process of data acquisition lasted 64 days and its objective was to collect a set of mobile calls, their transcriptions, and information related to their context. Participants were recruited through popular Web portals in Spain and they were asked a pre-study questionnaire to obtain demographic information, calling habits and annotation habits. Participants installed a VoIP application on their mobile phones and whenever they made a call with their mobile, it was routed through specific servers and recorded. The use of the application was transparent for the user, since it was installed as the default phone application and they could choose at any time whether to use the callnotes application to make the call or not. As compensation for participating in the study, all calls made through the system to national landlines or mobile lines were free of charge. The calls were transcribed and made available to participants by means of a Web application. There, they could see a list of

Class	Annotations
I	*We are in front of the fruit shop*
RoA	*Tomorrow we go to look for the swimsuit*
RI	*Are you coming to eat? At what time*
O	*Sure, it's normal*

Table 5.1 – Examples of annotations.

their recorded calls, and for each call, its audio recording and its transcription. Participants had the possibility to delete any call they considered to have sensitive content within the 24 hours after each call was made available through the Web application. Otherwise, the calls were considered to be willingly contributed to the study by the participants, as stated and accepted by participants in the terms of consent of the study. A very extended annotation process was conducted. Participants were in fact asked to annotate in the calls the chunks of information (either turns or constituents) which they retained relevant to take note of. In the case that the participants did not find any important information, they were asked to explicitly say so. In addition to this participants were asked to provide information about the situation they were at the time of the call: the time of the day, the day of the week, the action that they were performing, etc. This resulted in a unique dataset with extended annotations of noteworthy events made by the actual participant of the call and situational information, features which make this a perfect dataset for analyzing the influence of situational signals over the call.

As described by Bonin et al. [2014c], a qualitative analysis was conducted on the corpus to understand the nature of the annotations entered by the participants in the study.

Four types of annotations were distinguished: *Giving Information (I)*, *Requesting Information (RI)*, *Reporting on an Action (RoA)* and *Other (O)*. Examples of these 4 types of annotation are presented in Table 5.1. Three annotators labelled a total of 54 randomly selected turns from the dataset (IAA, Fleiss Kappa=0.54, Fleiss [1971]). 47% of the turns

were classified as belonging to the *Giving Information* category, 22% of the turns to the *Request Information* category, 26% to the *Other* category, and only 3% were classified as *Reporting on an Action*. Intuitively, one could expect the *Giving Information* category to be the most common in the annotated turns. However, the results obtained show that the other types of annotations are also well represented in the data.

Two primary features of interest emerge. First, while the vast majority of annotations correspond to turns where a piece of information is given (*e.g. We meet at 3pm*), turns where information is requested are also well represented in the sample. Second, more than 25% of this manually annotated dataset was marked under the *Other* category, which includes turns with very diverse functionalities (*e.g.* greetings, statements of agreement). This reveals that participants tend to annotate as noteworthy, turns with very diverse functional aspects.

5.4 Methodology

To answer RL_2, and, therefore, verify the hypothesis that the situational context influences the structure of the conversation, two analyses are conducted on the Callnotes dataset.

The first (***Analysis 1***) consists of a preliminary investigation of the correlation between the temporal dynamics of the distribution of relevant events in a conversation and situational factors such as: where the conversation is taking place and when the conversation is made (which day of the week and which time of the day). I investigate whether speakers tend to distribute relevant chunks of information (aka noteworthy) according to situational factors.

The second (***Analysis 2***) consists of an automatic classification of noteworthy events using situational information. If a relationship exists between the distribution of noteworthy information and the situational factors, then those factors, used as features describing a conversation, should be discriminative in the detection of such informa-

tion. In the next two sections, I present the methodology and the results of these two analyses.

5.5 Analysis 1: Preliminary feature analysis

Participants of a call may decide whether to condense all the relevant information in the beginning of a call as opposed to spread it throughout the evolution of the conversation. Knowing this distribution represents a key piece of information for automatically detecting relevant chunks, but, unfortunately this distribution varies between conversations. In this section, an analysis is conducted to detect whether some correlation is to be found between this distribution and the situational factors of a call. It is reasonable to assume that the distribution of relevant events may vary according to the situation in which the call is made. If a user is driving, he will tend to deliver all the information immediately in order to not be distracted from his/her main activity, in contrast to a situation in which he is relaxing on a sofa with his interlocutor. In order to explore this, I investigate the distribution of relevant events, calculating the cumulative distribution of the events over time.

Figure **??** shows the cumulative distribution of the relevant events in conversations factorized by hour of the day (over a 24 hours representation). The x axis represents the timeline of the conversation, while the y axis represents the cumulative distribution of the relevant events for that conversation. Since both the number of relevant events and the length of the conversation may vary, both values are normalized over the total number of relevant events (y axis) and the total length of the conversation (x axis). The plot is to be interpreted examining the skewness of the curve. Conversations with a cumulative distribution skewed to the left, have all the relevant information expressed at the beginning, while conversations whose distributions are right skewed have the relevant events towards the end. In order to visualize the correlation with situational factors, one should compare the distributions for conversation in different situational context.

Fig. 5.1 represents the distribution factorized by time of the day and the faceting is used to divide the calls made in the weekend (on the left -0) from the call of the weekdays (on the right - 1).

As one can see during late in the day (i.e. 20h, 21h, 22h) the cumulative distribution is not entirely expressed in the first part of the conversation; on the contrary it is relatively spread throughout the conversation: half way into the conversation, 70% of relevant information is expressed, attaining the 100% between halfway point and the end of the conversation. In contrast, in the middle of the day (this is particularly visible at 2pm, 3pm and 4pm of working days), there are calls in which the cumulative distribution of relevant events reaches 100% within the first 5% to 15% of the conversation.

To better investigate this, Fig 5.2 provides a different visualization of the same phenomenon. Each box summarizes the distribution of the temporal span elicited when all the relevant information for that call has been exchanged; more formally, being t_0 the beginning of the conversation, t_e represents the time when all the relevant events have happened (in this case all the noteworthy turns for that conversation have been uttered); D_e will be the difference between t_e and t_0, hence the time elapsed since the beginning when all the relevant events have been seen. Each box represents the distribution of D_e in conversations per hour of the day.

Therefore, lower distributions represent sets of conversations in which the totality of relevant information is expressed very early within the conversation, higher distributions represent sets of conversations where the totality of the relevant information is reached towards the end of the conversation. Once again, the faceting divides weekends (0) from weekdays (1), and the x axis represents the time of the day. One can see how, in line with the previous plot, the conversations at 2pm on the weekdays show a median value of around 15%, meaning that in median the conversations of that time of the day in our dataset reach the 100% of the information within the first 15% of the conversation. These figures are normalized to the length of the conversation which is

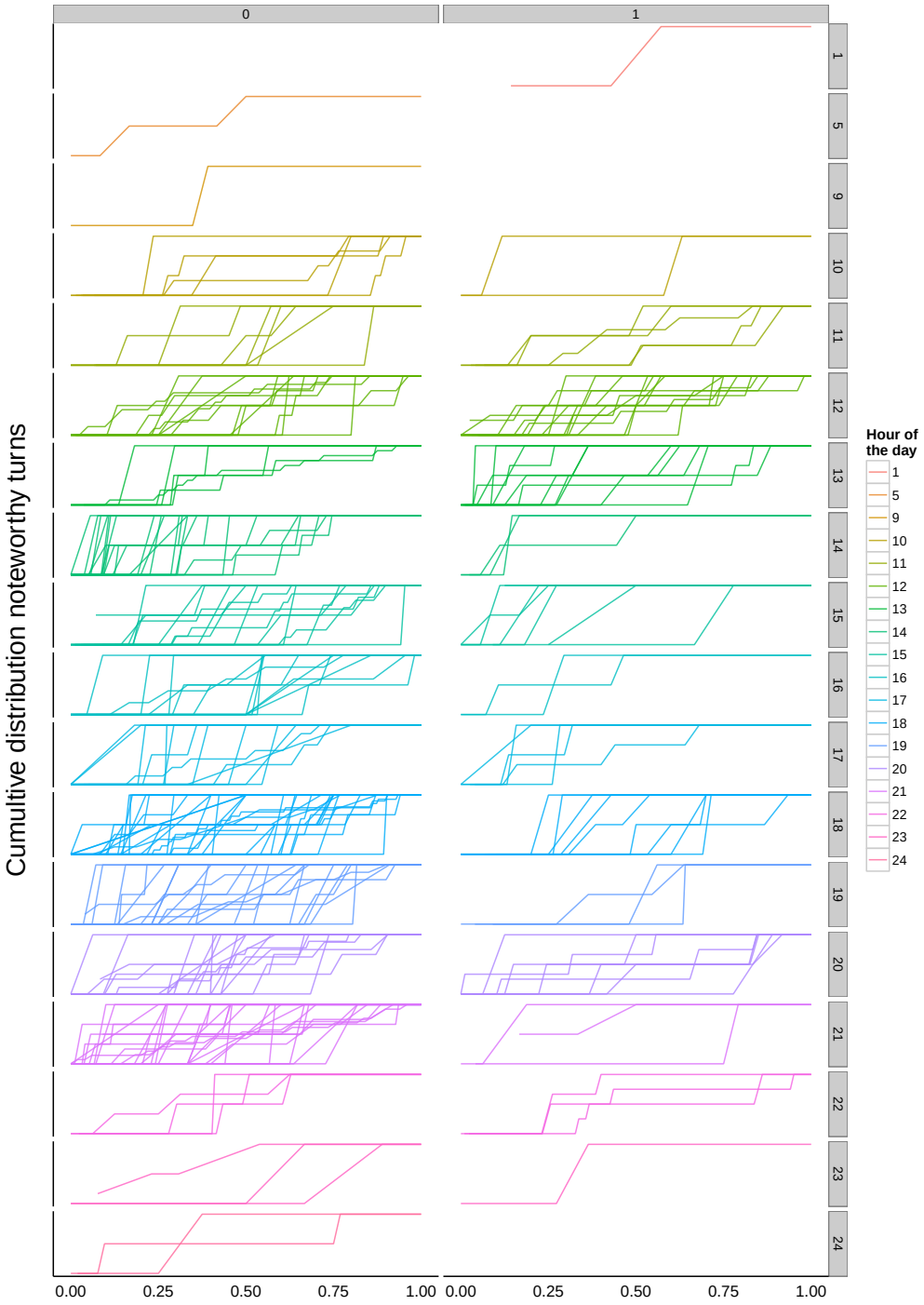


Figure 5.1 – Cumulative distribution of relevant events in the conversations of the Callnotes corpus divided by hours of the day. Each line in each subplot represents a phone call, the x axis reports the normalized length of the call and the y axis the cumulative distribution of noteworthy events. A distinction is made also between calls at the weekend (left part) and calls during the week (right part). The colors are provided for helping distinguishing between early hours (greenish), middays (bluish), and evenings (pinkish).

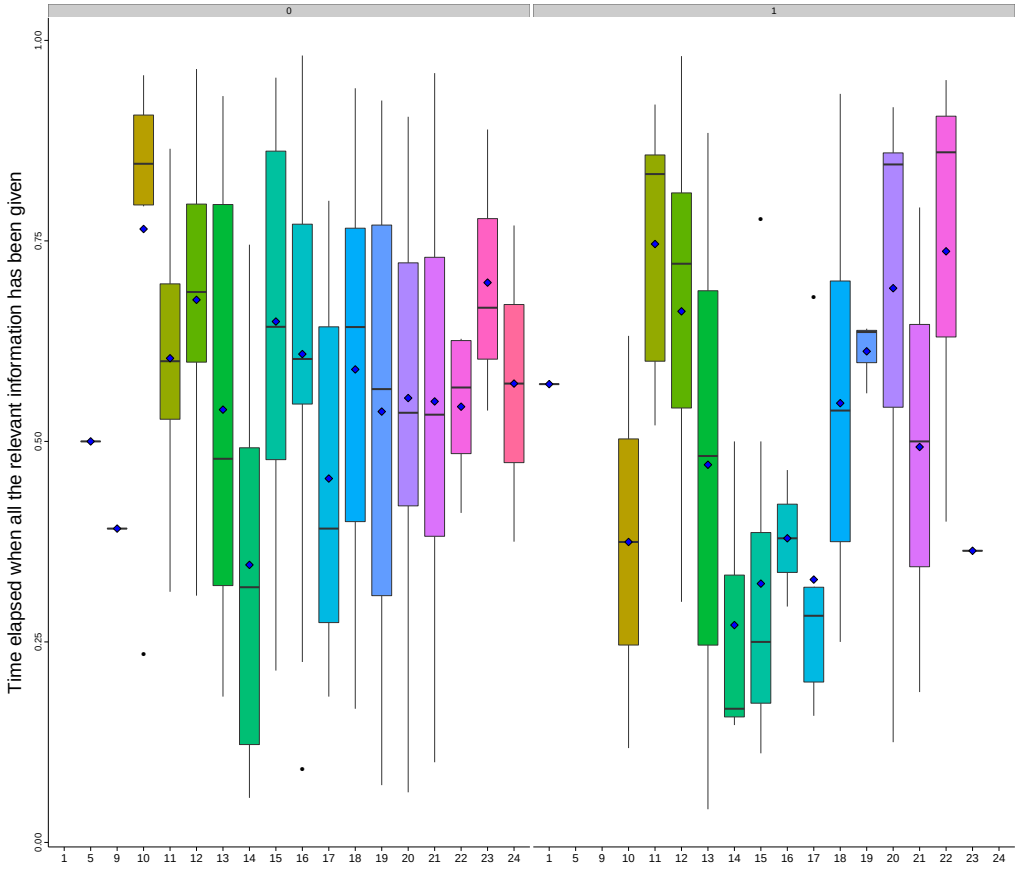


Figure 5.2 – Distribution of the time elapsed from the beginning of the conversation when all the relevant events have been expressed. The x axis shows the time of the day. A distinction is made also between calls at the weekend (left part) and calls during the week (right part). As one can see the weekend conversations do not show a clear patters, while during the weekdays, conversations in the middle of the day tend to deliver the relevant information in the beginning of the conversation. The colors are provided only to help the visualization and distinction between early hours (greenish), middays (bluish), and evenings (pinkish).

not a factor in this analysis.[3] As one can see, the right side of Fig 5.2 shows a clear pattern, where the central hours of the day have conversations which reach the totality of relevant information exchanged very early; one could say that in these hours, speakers *go straight to the point.*[4]. As similar conclusion can be drawn in Fig. 5.3, where the

[3] I am interested in analyzing when the relevant information is provided, independently of the duration of the conversation, which might continue without any further relevant chunks of information being uttered.

[4] A possible interpretation of this stands in the distinction between working hour (in the central part of the conversation) and non working hour. However, this might depend on the professions' distribution

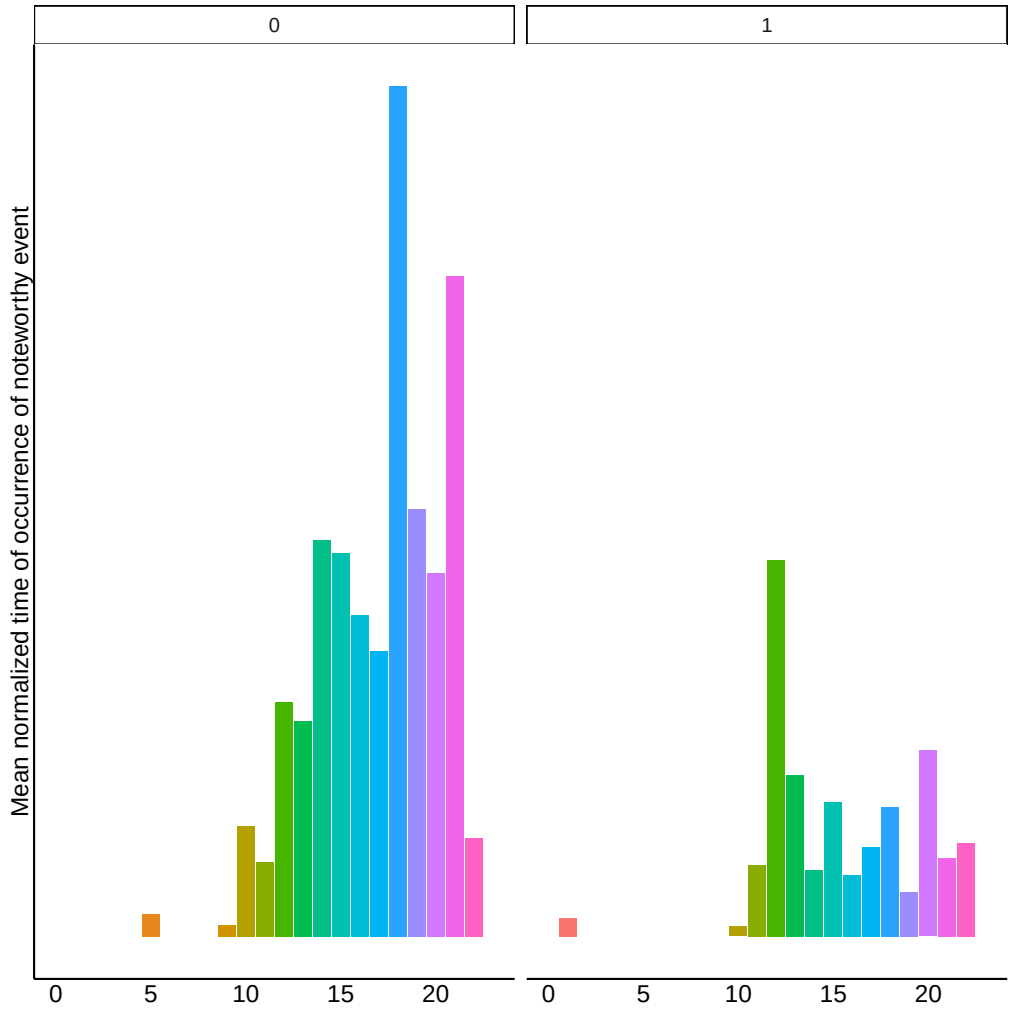


Figure 5.3 – Mean normalized time of occurrence of noteworthy event per hour. The x axis shows the hour of the day. A distinction is made also between calls at the weekend (left part) and calls during the week (right part). As one can see in the weekday conversations, the noteworthy events tend to occur on average in the beginning of the conversation. The colors are provided only to help the visualization and distinction between early hours (greenish), middays (bluish), and evenings (pinkish).

mean normalised time of occurrence of noteworthy events $\frac{\sum_{e \in Events} time_e}{Events}$ per hour is reported.

Table 5.2 shows the correlations between D_e and the situational features taken into consideration.

in the sample. An investigation of this is left to future studies.

5.6 Analysis 2: Automatic Classification of Relevant Events

Analysis 2 consists in a automatic classification of relevant events using situational information.[5] If a relationship exists between the distribution of noteworthy information and the situational factors then those factors, used as features describing a conversation, should be discriminative in the detection of such information.

Noteworthiness detection in conversations could be considered a particular form of topic summarization: the aim is to summarize a topic maintaining only the chunks which the participants wish to recall. Previous work has been conducted in noteworthy detection in meetings. Banerjee *et al.* investigate the feasibility of discovering noteworthy chunks in meetings, exploring if a notes-suggestion task can be accomplished by a human being. A Wizard of Oz experiment was then conducted over nine meetings, and the wizard reported a precision of 35% and recall of 41%. Banerjee and Rudnicky [2008] apply techniques developed in extractive meeting summarization for automatically identifying noteworthy information from meetings, creating a baseline for noteworthiness detection in meetings of 0.14 F-score. Section 2.4.4 provides an in-depth overview of the state of the art in noteworthiness detection in previous studies. These results give an idea of the challenges of the task, which deals with an extremely

[5]The work described in this section has been published in Bonin et al. [2014c], and is reported with authors' consensus.

Class	Pearson correlation ρ
Time - categorical	0.005**
Time of the day (when3)	0.013
Location (where)	-0.09*
Age	0.72
Gender	0.24
Education	0.033
Occupation	0.04

Table 5.2 – Linear regression between D_e (distribution of time elapsed from the beginning of the conversation to the moment in which all the relevant events for that conversation occurred) and all the variables.

subjective concept like noteworthiness, in a noisy scenario such as conversations. The work of Banerjee and Rudnicky [2008] on meetings exploits the content information of meeting transcriptions in a lab simulated static situation. In the case of telephone conversations, the complexity of the situation and the errors due to automatic transcription in noisy outdoors create the basis for an even more challenging scenario. However, telephone conversations can benefit from situational information, which might give hints about the discourse structure (e.g. where the noteworthy information is more likely to be located).

A supervised machine learning approach is used here to automatically detect noteworthy turns in conversations, where the conversations are the documents represented as a collection of values in a feature space and the ground-truth are the noteworthiness annotations provided in the dataset collection stage (See Chapter 3.4). The goal of the classification is to automatically identify information annotated by users in terms of its relevance for future needs; however, I am particularly interested in understanding what contributes to detecting relevant turns and whether information on the situation helps improve the results. For the purposes of this study, the turn is considered as the basic unit of analysis. A preliminary analysis showed that often users tend to highlight complete turns as relevant, instead of parts of the turns. On average, 66.57% (s = 35.87) of the words within an annotated turn were highlighted, with a median value of 80%. Hence, the turn is chosen as the basic unit of analysis considering relevant a turn with at least one annotated word.

5.6.1 Features description

In this section I describe the features computed to represent conversations and which have been engineered to capture information relevant to the problem at hand at the turn level. Every turn is described as a vector of features modeling the characteristics of the turn. Specifically, two sets of features can be distinguished: **C**ontent features,

denoted hereafter with the letter **C**, and situational conte**X**t features, denoted hereafter with the letter **X**.

Not only information at the situational level is considered, due to the coarse grained nature of this information. Situational signals are constant during the conversation (the location of a call does not change within the timeframe of a conversation), and, therefore, the only use of situational information does not provide any turn-level information, necessary to detect relevant turns.

5.6.1.1 Content Features

Content features are computed by analyzing the content of the conversations; while some of them are derivations of the features used in meeting summarization, others are novel features engineered for the task at hand. The input of the C-features is the textual information resulting from the semi-automatic transcription of the calls.[6] In order to extract features from the transcript, the datasets are first pre-processed (split in turns, lemmatized, PoS tagged). Finally, Named Entities (NEs) are recognized and classified using Freeling Language Processing tools [Padro et al., 2010]. As noted in Chapter 2, the detection of relevant information in conversations shares similarities with meeting summarization. For this reason I rely on previous literature in meeting summarization in the feature engineering process. The final set of 42 content-based features includes both variations of features previously used in the meeting summarization literature and novel features particularly adapted to the task at hand. A main deviation from meeting summarization is the fact that while a summary will report information on the main topic of the meeting, a noteworthy chunk of information in a call is not necessarily related to the main topic of the conversation; for this reason, in contrast to related work on meeting summarization, no content features are based on lexical similarity to the entire call or to the main topic of the call. In addition, due to the poor quality of

[6] While the analysis of the acoustic signal may reveal additional cues useful for noteworthiness detection, it lies out of the scope of this work.

the transcription, no long distance dependency information (*e.g.* argument predicate relations) or deep syntactical parsing is involved.

Table 5.3 provides a summary of all the content-based features extracted from the dataset. Where applicable, two vector representations are used: binary and frequency-based. I will refer to these two different encoding schemes as **Bin** for the binary case, and **Freq** for the frequency case.

CONTENT FEATURES	
C-BoW (Bag of Words)	
BoW	BoW for all words (except hapax)
C-T (Turn-based)	
NE-SP	presence (or frequency) of Person
NE-G	presence (or frequency) of Location
NE-O	presence (or frequency) of Organization
NE-V	presence (or frequency) of misc. NE
Number	presence (or frequency) of Numbers
Date	presence (or frequency) of Dates
TLN	Turn length in # words normalized
PoS	PoS distribution
TF	Max and Mean term frequency
IDF	Maximum and Mean inverse document frequency
C-D (Dynamic)	
Rep	Repetition between t and t-1,t+1,t-2,t+2
Int	Presence (or total amount) of Int. pro./adj. in t-1
Q	Presence (or total amount) of question in t-1
C-C (Conversational)	
Dur	Duration of the call (# turns and # words)
Cent	Conversation centrality
Spk	Speaker
Dom	Speaker dominance

Table 5.3 – Final set of content features extracted from the conversation

Turn-Based content features (C-T) Turn-based content features take into account information related to individual turns. We distinguish between lexical and non-lexical C-T:

Lexical content features: Lexical C-T features capture the lexical properties of

a turn and include NEs, *Locations, Organizations, Persons*, *Miscs*, *Numbers*, *Dates*, and temporal expressions. Information regarding the presence of NEs has been widely exploited in text summarization [Gupta and Lehal, 2010; Maskey and Hirschberg, 2005]. Maskey and Hirschberg [2005] reports that the total amount of NE is among the most discriminative lexical feature in predicting sentences considered to be relevant for a summary. However, in the case of phone conversations, attention has also been given to the presence of temporal expressions under the intuition that temporal cues are good indicators of upcoming pieces of information (*e.g. The meeting is tomorrow*).

For each turn t, the presence of any NE and temporal expression, as well as the presence of individual classes of NEs is detected and for each of these classes of entities, a binary and a frequency feature vector is extracted.

Non-lexical content features: capture characteristics of the turn which do not involve lexical information, namely: turn length, part-of-speech (PoS) distributions and Tf-idf descriptive statistics at the turn level. In the area of meeting summarization research, the average length of a turn has been found to be a valid feature [Xie et al., 2008]. In the Callnotes dataset, preliminary analyses revealed that annotated turns tend to be longer than average, suggesting that the turn length, calculated as the number of tokens per turn normalized over the average turn length (punctuation excluded), could be useful. To further gauge discourse characteristics, the distribution of PoS at the turn level is detected: *i.e.* for each turn, the frequency of nouns, pronouns, adjectives, adverbs, interjections, verbs, prepositions and conjunctions is calculated. Finally, I extract the maximum and mean term frequency (Tf) and inverse document frequency (idf) measures. In Xie et al. [2008], authors report that idf is among the most discriminative features in sentence selection for text summarization.

Dynamic content features (C-D) Dynamic content features are designed to capture the semantic relationships between each turn and its precedent and subsequent turns. In this context, by semantic relationship I refer to possible semantic links that

each turn may have with its neighboring turns, such as lexical and topical cohesion, question-answer relationship, and the appearance of general cues that may anticipate relevant pieces of information in the subsequent turn.

Repetitions: words repeated by different speakers in consecutive turns are used. Participants of a conversation tend to align at several linguistic and paralinguistic levels in order to ease communication and increase mutual understanding [Pickering and Ferreira, 2008]. This phenomenon has been investigated in terms of prosody, lexicon and syntax [Levitan and Hirschberg, 2011; Brennan, 1996; Bonin et al., 2013; Branigan et al., 2010]. From a lexical point of view, the alignment mechanism, often referred to as priming, is realized by means of word repetition among speakers. In this context the priming phenomenon is exploited to detect concepts in the conversation that are considered important by both participants, relying on the fact that repeated words convey concepts that participants want to make sure have been successfully communicated to their interlocutor. Given a dataset D, a turn in D, $t \in D$, and $t-i$ and $t+i$ turns in the context of t, I calculate the amount of repeated lemmas between t and $t-i$, and t and $t+i$ for $1 \leq i \leq 2$. In order to consider semantically meaningful repetitions, only content words (nouns, adjectives, adverbs, verbs) are taken into account.

For sake of clarity I report an example of consecutive turns with repetitions:[7]

```
Turn Utterance
t-1: Starting at half past four.
t:   Starting at half past four, yes.
```

Interestingly a relationship is found between the presence of repetitions among consecutive turns and the relevance of a turn; being A the set of annotated turns, I compare the frequency of repeated terms among t and $t-i$ for $t \in A$ and $t \notin A$. The pairs where $t \in A$ showed a significant higher frequency of repeated terms with respect to the pairs where $t \notin A$ (Student T-Test on normal distribution,$p<0.005^{**}$).

[7] I report the English translation of the Spanish original.

Interrogative pronouns and questions: As shown in Sec. 5.3, 47% of the annotations of relevant chunk of information of the call were marked as *giving information*, that following the natural flow of the conversation might have been triggered by a *request of information*.

Therefore, for each turn t, I extract information about the presence of interrogative pronouns/adjectives in $t-1$, and the presence of a question in $t-1$ to capture all the *giving information* turns, which might have been triggered by a *request of information* in the preceding turn.

Conversational flow features (C-C) These features are designed to model information about the conversation's flow and speakers' interactions.

Centrality of the turn: captures the position of the turn within the complete dialogue, computed as the distance of the turn from the middle of the conversation. This feature is inspired by the sentence location features used in text summarization Chen et al. [2002]. Chen *et al.* assign different weights to sentences in the first, middle and final part of a paragraph, in order to favor sentences which lie in the central part of the paragraph as they are considered to be more informative for a summary. This distance is measured in terms of number of words, excluding punctuation.

Speaker: Who is uttering the turn (caller vs callee).

Conversation duration: Length of the conversation in number of turns and in number of words. The number of turns captures the dynamics of a dialogue (shorter turns equals a more dynamic exchange), while the number of words captures the overall duration.

Speaker dominance: The dominance, in terms of amount of productions during the call, is extracted. This is calculated by comparing the number of turns of speaker *a* vs speaker *b*, normalized over the total amount of turns per call.

Bag-of-Words (BoW) Finally, the performance of a naive bag-of-words scheme to represent the content at the turn level is explored . Given the large vocabulary size of our corpus ($10,144$ tokens) and the sparsity organic to bag-of-word representations, a trivial dimensionality reduction strategy is used, filtering out the terms that appear only once in the corpus. No stop-list of function words is applied, as they appeared to have a high discriminative value.

SITUATIONAL CONTEXT FEATURES	
X-C (Call-based)	
X-C-T	Time of the call
X-C-Loc	Location of the call
X-C-Day	Day of the call
X-C-Obj	Objective of the call
X-U (User-based)	
X-U-G	Gender
X-U-A	Age
X-U -I	Income
X-U-E	Education
X-U-Ms	Marital Status

Table 5.4 – Final set of situational context features

5.6.1.2 Situational Context Features

Situational Context features (X) are introduced under the hypothesis that relevant information may depend on the characteristics of the user and on the situation in which the call takes place. An initial analysis of this hypothesis was presented in Section 5.5, where it was shown that a relationship exists between the time of the day in which the conversation takes place and the position within the call of relevant information. Now, I will explore whether this information influences the classification performances in the detection of relevant turns.

In fact, as noted in Section 2.4.4, pure NLP approaches applied to automatically detect relevant information in meetings are able to achieve an F-score of only 0.14, a low F-score which underlines the complexity of the task and the limitations of a purely

content-based approach. Situational cues may be used to increase the discriminative power of the classification model. A schematic overview of these features is provided in Table 5.4, where the two main classes of situational context features are visible: call-based (X-C) and user-based (X-U) situational context features.

Call-Based Features (X-C) Call-based features are meant to capture contextual information at the call level and include information about *where*, *when* and *why* a call is made, under the intuition that calls made, for example, during working hours may specify relevant information with a different timing with respect to calls made during the weekend.

Where: In the dataset it is possible to identify six *location* categories: home, work place, while commuting, while exercising, while shopping and other, locations that was provided by participants through the post-call questionnaire. However location information is typically available from the mobile network.

When: In terms of *temporal* features, the actual time of the call (over 24 hours) is considered and categorized into two classes: working vs non working hours, and the day in also two classes: weekday vs weekend.

Why: the *objective* is finally considered. Specifically the users were asked to indicate whether the call had been made for *giving/receiving information*, *discussing a topic*, *taking an appointment*, *asking for a favor*, *for having a chat*.

User-Based Features (X-U) In addition to the information on the situation, an interesting question concerns the influence of the participants information. A set of features that feed the model with information about the user is introduced, capturing age, gender, educational level, income and marital status. Gender is represented as a binary feature, while age is categorized in 5 groups: below 20 years old, between 20 and 30, between 30 and 40, between 40 and 50 and above 50. The education status is represented by the following categories: Primary education, Secondary education, Bachelor

degree or Postgraduate education (Master or PhD). Yearly income is categorized by: up to 10k, 20k, 30k, 40k and more than 40k. Finally, marital status is categorized as: single, in a couple (married, with a stable partner), other.

5.6.2 Experiments

In this section I present the results of the classification task. The goal of the experiment is to automatically identify information annotated by the caller of a telephone conversation in terms of its potential relevance for future recall. As mentioned previously, this classification task presents two main challenges. First, the restrictive nature of annotation leads to a very unbalanced dataset, where less than 3% of the corpus has been labeled as relevant by the participants. Second, the subjectivity of the task leads to a high variability of annotation behaviors, as described in Section 3.4.

For sake of simplicity I label the dataset as $\mathcal{G}$ dataset, and all the characteristics are given in Chapter 3.6.[8]

A standard Support Vector Machine (SVM) with RBF kernel is used, as this classification approach provided the most consistent results throughout all the evaluated configurations compared to the rest of methods tried (SVMs with polynomial kernels, logistic regression, and naive Bayes). The dataset has been divided with a constant split of training and test sets for all the experiments, accounting for 70% and 30% of the dataset respectively. A grid-search approach to tune the hyperparameters of the SVM model using F-score as the quality metric to optimize is used.

[8] In a derived work, published in Bonin et al. [2014c], we have investigated the performance over a subset of $\mathcal{G}$, called $\mathcal{A}$ dataset where only conversations having at least one relevant turns are considered. While the $\mathcal{G}$ dataset represents a situation where all conversations are stored and processed without any human intervention, the $\mathcal{A}$ subset represents a semi-supervised scenario where users would label conversations (but not the turn) as noteworthy (or not) right after finishing their phone conversation. This second scenario, while useful in the context of a real word application, is not related to the question posed at the beginning of this chapter, being whether situational features helps detecting relevant information.

5.6.3 Classification Results

This section presents the results obtained for the prediction of individual relevant noteworthy turns within a conversation. Although the aim of this section is to understand whether situational signals improve the performance, in order to provide a more complete overview over the results, I also report the results of the different sets of content features. It will be shown that the content features described in Section 5.6.1.1 and traditionally used in meeting summarizations, fail to provide a valid classification in such a challenging task.[9] Therefore I report here the prediction results given by the following feature sets: C-T only, C-D only, the combination of C-T and C-D (C-TD), and the combination of C-T, C-D and C-C (C-TDC). The results of these feature sets are shown in Table 5.5.

Features	Precision		Recall		F-score	
Encoding	Bin	Freq	Bin	Freq	Bin	Freq
BoW	0.081	0.083	0.730	0.720	0.150	0.150
C-T	0.087	0.088	0.53	0.32	0.15	0.139
C-D	0.03	0.26	0.26	0.12	0.15	0.05
C-TD	0.087	0.09	**0.754**	0.33	0.1505	0.1419
C-TDC	0.09	0.093	0.58	0.37	0.158	0.149
C-TDC+BoW	**0.11**	**0.11**	0.52	0.51	**0.18**	**0.18**

Table 5.5 – Classification performance using different configurations of features.

As shown in Table 5.5, the maximum F-score is achieved by the combination of all content features including the BoW. The low score ($F = 0.18$) is a direct consequence of the low precision obtained ($p = 0.11$). Interestingly the C-TDC feature set outperforms the pure BoW approach ($F = 0.15$), using a fraction (about 1%) of the number of BoW features, which leads to a considerably simpler model.

These results serve as a framework for better understanding the classification effectiveness of different combinations of content and context features that is now presented. Specifically I consider **C**ontent features (corresponding to the C-TDC set de-

[9] One should consider the unbalance nature of the dataset and the subjectivity of the concept of relevant turn.

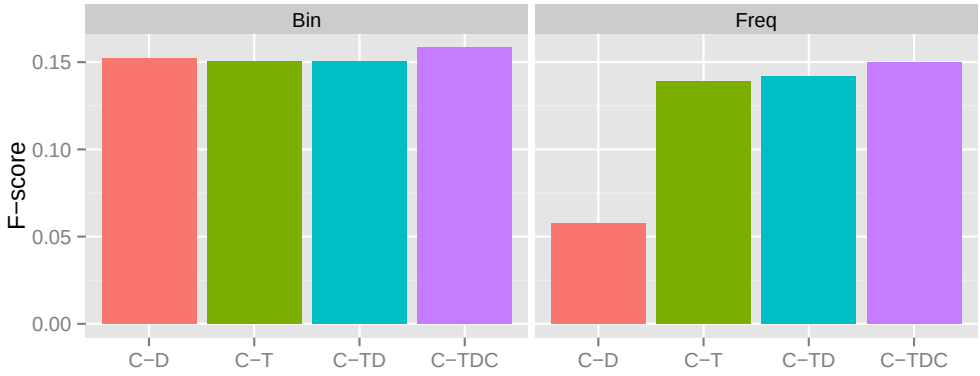


Figure 5.4 – Classification performance using different configurations of features.

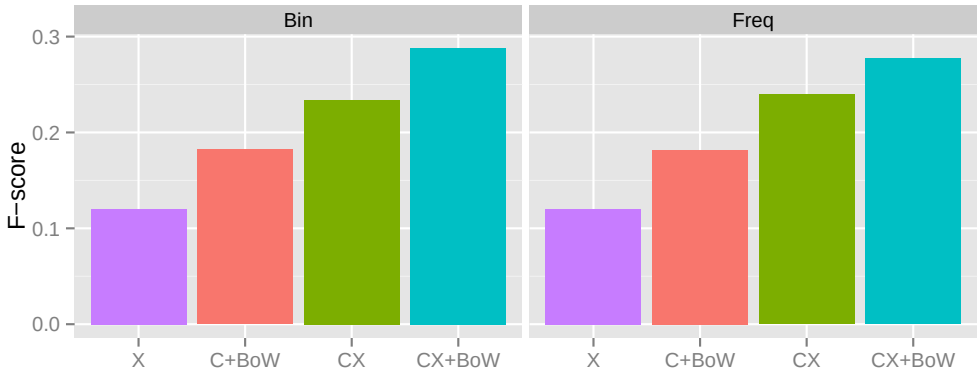


Figure 5.5 – Classification performance using Content, Context features and their combination

Features	Precision		R		F-score	
Rep.	B	F	B	F	B	F
C+BoW	0.11	0.11	0.52	0.51	0.18	0.18
CX	0.169	**0.20**	0.38	0.286	0.2354	0.2394
CX+BoW	0.189	0.1919	0.524	0.5022	**0.288**	0.277

Table 5.6 – Classification performance using the combination of context and context-based features

scribed above), situational conte**X**t features, bag of words, and the combination of content and context features (**CX**). For simplicity, in the remainder of this section I refer to the entire set of content features as C, to the entire set of context features as X, and

to their combination as CX. When the test is done with the addition of BoW features the *+BoW* convention is applied. The results for F-score are shown in Table 5.6 and Figure 5.5. One can observe that the fusion of content and context features (CX and CX+BoW) provides a noticeable overall increase in the F-score by almost a factor of 2, from $F = 0.18$ to $F = 0.28$, with a precision rising from 0.11 to 0.19.

This result gives empirical evidence that these two sets of features convey complementary information that is relevant to the task at hand: the same words can carry different relevance depending on the contextual information of the conversation.

5.7 Discussion

Analysis 1 and **Analysis 2** focus on exploring the relationship between situational signals and discourse events, such as the distribution of noteworthy events. With **Analysis 1** a correlation has emerged between some situations such as the time of the call, the day of the week and the distribution of noteworthy events in a conversation. Specifically, it has been noted how during working hours of weekdays the callers tend to anticipate noteworthy information at the beginning of the conversation. A possible interpretation of this behavior is that users are aware of the situation while structuring the information flow in their conversation. If the call is made in the central part of the day (working hours), the conversation will generally tend to a specific task having all the relevant information within the first part of the call. This is evident in Fig 5.2 where boxplots indicate the distribution of the time elapses from the beginning of the conversation to the moment in which all the relevant information has been expressed. During working hours of working days boxes show a regular pattern, different from the rest. On the contrary at the weekends, as shown in the left side of Fig 5.2, no pattern emerges. In order to better investigate this, **Analysis 2** shows how the use of such information (time of the day, day of the week, etc) improves the results of a classification task aimed at predicting noteworthy information in telephone conversations.

Situational information is therefore shown to influence the way in which speakers organize the information distribution over a conversations to the point that exploiting situational information helps in detecting noteworthy events in the conversation.

5.8 Summary

This Chapter provides the analyses related to RL_2 and investigates whether and how situational signals influence the prediction of events in social interactions. The focus is on the relation between noteworthy events and situational signals, analyzedin two ways: a correlation analysis and a classification analysis. The correlation analysis has shown a relationship between some situational signals and the distribution of noteworthy events in the conversation of the considered dataset. The classification analysis has shown that the use of situational features improves the performances of an automatic classification task for noteworthy information. In conclusion, a relationship emerged between situational context and noteworthy event distribution.

CHAPTER 6

Conclusions

6.1 Wrapping up

This thesis proposes a novel view of the social context in communication, exploring the relations of social and situational aspects with the discourse structure, rather then with the affective and emotional sphere of communication. Social aspects such as laughter, overlaps, silences and backchannels have been here explored as the knowledge of their dynamics can be very useful in the field of human machine interaction.

I have explored the influence on human-human communication of the social context, divided into the social (the ensemble of the social signals exchanged by the participants) and the situational aspects (the situation in which the conversation takes place).

The original hypothesis was that both, social and situational aspects, play a role at the discourse level and, therefore, influence the way speakers organize their speech productions in spontaneous conversations. I have considered two research lines (RL)

exploring the contribution of:

RL_1: social context to topic segmentation;

RL_2: situational context to event segmentation;

and I have investigated RL_1 in Chapter 4 and RL_2 in Chapter 5. As discussed in Chapter 4, social signals play a function at the discourse level, in addition to the emotional level, in relation to topic changes. In fact, a drop in social signals appears to occur immediately after a topic change when the interactional entropy, defined as the amount of social activity, is reduced. Participants show the tendency to limit the interaction immediately after a topic change, probably to leave the floor to the speaker who has introduced the new topic. Chapter 5 shows how the situation in which the conversation takes place influences the discourse structure, particularly in the production of noteworthy events. Speakers adjust the timing of their production according to the situation in which the conversation takes place. For example, relevant information is concentrated at the beginning of the conversation when the exchange takes place in working hours.

Chapter 4 has addressed RL_1, concluding that the social context has a relation to topic segmentation, and that, particularly, this can be seen in the fact that variations in social activity are influenced by discourse events such as topic changes.

Chapter 5 has addressed RL_2 showing that there is a relation between noteworthy events and situational factors, and that situational factors might help in segmenting the conversation into noteworthy and non-noteworthy information.

Overall, this thesis provides an overview on how content, social context and situational context are intertwined, influence the discourse structure and in which way they contribute to the co-construction of speech exchanges.

6.2 Future work

This work analyzes a new conversational model where linguistic, social and situational signals are intertwined and contribute to the co-construction of the conversation. I believe this introduces a novel research framework, where both linguistic, social and situational elements (the content and the context of a conversation) are analyzed in their continuous interaction with each other towards a better understanding of the conversation. To this extent, there is room for novel investigations on the correlation between discourse events and social events.

I foresee two potential and intertwined approaches of research that natural language processing and social signal processing fields could follow to improve our understanding of social dynamics and their relation with discourse events: an *analytical approach* and *synthesis approach*. From an *analytical* point of view, more explorative study could be conducted on natural and spontaneous conversations in order to better understand the nature of human-human interactions and the varied ensemble of social rules behind them. Such new knowledge could be soon used, from a *synthesis* point of view, to improve the naturalness of conversational agents as described later in section 6.3.

Possible future novel investigations of these phenomena could involve other speech exchange systems (intended as type of interacitons), by collecting and exploring corpora of different nature. Even though in this work, I try to explore a varied range of datasets, covering different types of conversation ranging from meeting conversation to spontaneous calls. However there are still many other speech exchange systems that need to be explored. Situations like job interviews, political debates or lecturing could, for example, stress the existence of different phenomena, or a different relation between content and social context. Nonetheless, cultural and linguistic framework different from English should be explored.

In addition, an interesting line of research would be to explore the influence of

the medium on the relation between content and social context, where by medium I refer to the channel of communication (e.g. telephone, face-to-face conversation, digital presence such as remote conferences). While in this thesis, the considered datasets provide interactions with different media (face-to-face conversations such as AMI and TableTalk versus telephonic interactions such as Callnotes), the analysis of the effects due to the medium was outside the scope of this work. Future works could further focus on this aspect.

Other signals could be explored; for example, hesitations, false starts, gestures, gazes, postures, proximity. These could reveal interactions with the discourse and help defining the discourse structure. Finally, a deeper investigation of the concept of interactional entropy could be conducted, in order to take into account the distribution over the all set of social signals and all the other linguistic vocalizations.

6.3 Potential benefits for existing and evolving applications

A better understanding of the relation and the dynamics between content, social and situational context in conversations can be a valuable resource in many current applications. In this last section, I would like to conclude by providing an overview of state of the art technologies that could benefit from the finding of this thesis.

Natural-language based interfaces, such as dialogue systems, have become a feasible and attractive trend as they offer the most natural model of interaction: natural language. In this, a key role is played by multimodality: a range of input and output modalities which people normally employ in spontaneous communication, such as speech, gesture, gaze direction and facial expressions is used for engaging the user in a natural conversation [Jokinen, 2009c]. The design of dialogue systems having an interactive behavior which is natural to its users and exploiting the full potential of spoken and multimodal interaction would benefit from a deep understanding of human dialogue characteristics and from the incorporation of social intelligence. In the following

sections, I describe how the know-how emerged from this work could bring an added value to existing technologies.

6.3.1 Conversational assistants - Virtual Agents

In recent years, many companies have followed the trend of integrating Virtual Agents (also referred to as: Mobile Virtual Assistant, Virtual Personal Assistant, Intelligent Software Assistant[1]), in their technologies [Riccardi, 2014]. Apple's Siri has been the first Virtual Agent (VA) to give the users the impression that smarthphones could speak, be funny, answer a limited set of questions and execute simple tasks. After Siri, Google, Nuance, and AT&T [Johnston et al., 2014] have followed with their own implementation of VAs. The common functionality that characterizes VAs is their ability to interpret Natural Language via spoken interaction and return responses either in the form of a software program execution (e.g. opening the contacts folder) or a spoken dialogue (Question Answering). While some VAs can hold limited conversations, others tend to work in the background providing information when and where needed. However, the rate of success of these strategies is limited with respect to the level of context understanding, and input's noise. State-of-the-art VAs, like Siri and Cortana, are similar to Question Answering or command-and-control systems and they are not able to maintain a complex dialogue flow. They can remember context across a few (mostly one or two) turns, and tend to treat most utterances as individual queries and commands. When they are not able to understand the context, their fall back strategy is to present the user with the results of a web-search of the query. Of course, in real world scenario, situations in which the input is noisy, and the linguistic context difficult to understand, are very common, and a repeated use of the fall back strategy can be quite frustrating for the user.

[1] Additionally, they are referred to as Knowledge Navigators, for their navigation capabilities

Potential contribution of this work In this context, the contribution that this study could bring is twofold.

a) From a conversational point of view, a better understanding of the conversation dynamics and of the natural timing of linguistic and social events in a conversation is a key information for improving the dialogue management techniques with a VA.

b) From a natural language understanding perspective, in case of noisy input, the knowledge of the discourse value of non-linguistic signals, such as laughter, hesitations, silences, backchannels[2] can facilitate the understanding of the discourse structure without a full understanding of the content. For example, if Siri detects a moment of high social activity (the speaker laughing) followed by a moment of low social activity and high lexical volume, it can reasonably assume that the speaker has passed to a new topic (or a new way to express the same concept[3]): the VA can therefore reasonably ignore the input with high social activity, and focus its natural language understanding strategies on the rest. This would produce a sort of coarse grain automatic text cleaning.

6.3.2 E-Health Conversational agents

In recent years, socially assistive robotics for domestic use, also called Companion Technologies, have been a rapidly increasing field of research and development. By socially assistive technology it is to be intended a range of technologies aiming at providing a social, other than a physical interaction with an elderly person. The variety of assistive functionalities that can be provided by a robot companion is manifold and ranges from situation-specific, intelligent reminding (e.g. taking medication or drinking) and cognitive stimulation by means of tailored training exercises, up to the detection and evaluation of dangerous situations [Gross et al., 2011]. Although, those systems have made substantial progress in the last years, possible development could

[2] Overlaps are not mentioned here, as the situation presented is a one human/one machine interaction.
[3] On the basis of what discussed in Chapter 4

include the integration of context awareness; as reported in [Riccardi, 2014], a companion system should be always aware of the user's current context. It should be aware of both physical and emotional state of the user by monitoring and interpreting the personal and world signals of the users, detecting his/her physical condition, possible danger, falls, but also requests for help.

Potential contribution of this work As mentioned above, future developments of companion systems are likely to be in the area of context aware technology. In other words, companion technologies could well benefit from a deep understanding of the context in which the user (the elderly person in this case) is living, moving and interacting. This also includes the sphere of social interaction of the user. Therefore, if the user is engaged in a conversation with someone in the room, or on the phone, the system will need to be able to understand the most appropriate way and time to enter into the conversation and maintaing a good timing interaction. Knowing the social activity variability and dynamics of conversation can help the system participate in the most natural fashion.

6.3.3 Commercial chat-bot

A chat-bot, referred to as talkbot, Bot, chatter-box, Artificial Conversational Entity, is a program able to carry on a conversation usually in a textual form. Today, chat-bots are also used in dialog systems for various practical purposes including customer service or information acquisition on the websites of many private companies. Some chat-bots use sophisticated natural language processing systems, but many simply scan for keywords within the input and pull a reply with the most matching keywords, or the most similar wording pattern, from a textual database.

Potential contribution of this work Textual chats can be considered spontaneous conversations. Although lacking of the acoustic channel, social signals can, and usually

are expressed, by other means where laughs can become series of *ahahaha*, or codes like '*LOL*' or emoticons. Overlaps and silences can be detected by looking at the timestamps of the chats and backchannels are usually translated in lexicalized versions like *ehm, mm, uhm, etc.* Situational signals on the other hand can be detected in human/chat-bot conversations, using external sensors (GPS for example if the user is typing from a smartphone). Similarly to what anticipate for VA, this information can be used to monitor the social dynamics of the conversation for detecting the system engagement with the user, and also the user intentions (whether the user is just having a social chat, or is determined to complete a task).

6.3.4 Advertising in chats

Today, many conversations take place through chats services like GTalk, Skype, and similar context. Many of these are free services that gain their profit from targeted advertisements displayed on the user's page during the interaction (this is the case of Gmail, where links to services related to the user email or chats content are displayed).

Potential contribution of this work Detecting social signals like overlaps, silences, laughter and backchannels in users' chat would grant these services powerful cues on the dynamics of the conversation and on the structure of the discourse. For example, if, during a chat, a socially dynamic moment is followed by a moment of low social activity, one could reasonably hypothesize (on the basis of what discussed in Chapter 4), that a topic change has occurred. Furthermore, this could possibly be detected without scanning the entire content of the conversation, but only the social signals exchanged. The system could then perform topic modeling only on the first exchanges after the hypothetical topic change and adjust the advertisement accordingly to the new topic. This would allow the system to conduct continuous high level screening of the conversation, and to focus upon the content only when a relevant event (like a topic change) occurs.

6.3.5 Smartphone Applications

The range of smartphone applications is varied and constantly growing. However, an application capable of automatically detecting noteworthy chunks of a conversation has not yet been developed. A common use case scenario depicts a user having a conversation on his/her smartphone while performing other activities, such as driving. Having both hands busy, the user is not in the position of taking a physical note on a piece of paper or on the phone itself. As mentioned in Chapter 5, Carrascal et al. [2012] found, in a large user study, that this situation is common among smartphone users and that they would appreciate an application able to automatically retrieve noteworthy pieces of information from an ongoing conversation.

Potential contribution of this work In Chapter 5, I have shown that the main challenges of automatically detecting noteworthy information from conversations is the subjectivity of the concept of noteworthiness, and the small number of noteworthy turns within a conversation. However, despite these challenges, the use of contextual situational information, allows reasonable performances to be reached. One could imagine a real system retrieving all the possible noteworthy candidate turns and providing a list of candidate noteworthy concepts to the users. The user could than select from this list only the ones that really wants to store as a memo-note. In addition, in order to refine the results of the systems, one could imagine a semi-automatic system, asking the user to input, at the end of a call, whether the call contains information worth remembering or not. This would provide the system with a much cleaner training dataset. An experiment with this kind of scenario is reported in [Bonin et al., 2014c], and shows an improvement in the results of the 4%.

6.4 Final remarks

To conclude, as technology has become more and more present in our daily life, research centers as well as industry are evolving and improving current technologies with natural language interfaces, and transparent technologies capable of being present and augmenting our daily lives, making them safer and easier. In this scenario, the use of social and situational contextual information could be exploited within these existing technologies to provide more socially-aware machines able not only to use the information of where, to whom and in which condition users are talking, but also behave according to the social rules of human-human conversations.

APPENDIX A

Appendix A

Below is an excerpt, from the AMI corpus, where share laughter anticipates a topic change from the AMI corpus:

```
FEE005 Yeah, so uh
 [disfmarker]
MEE008 Probably when
 he was
 little he got lots
 of attention for doing
 it and has forever
 been conditioned .
FEE005 Yeah , maybe .
FEE005 Maybe.
[vocalsound-laugh] Right,
 um where did you find this?
 Just down here ? Yeah .
```

MEE008 [vocalsound-laugh]

MEE006 [vocalsound-other]

FEE005 Okay .

TOPIC - CHANGE

FEE005 [vocalsound-other]

Um what are we doing next ? Uh um .

FEE005 Okey , uh we now

need to discuss the project finance. Um

FEE005 so according to the

brief um we’re gonna be selling

this remote control for twenty five

Euro, um and we’re aiming to make

fifty million Euro . [...]

Below an excerpt, from the AMI corpus, in which laughter does not anticipate a topic change:

MEE008 A beagle .

FEE005 [vocalsound-laugh]

MEE008 Um charac favourite

characteristics of it ?

Is that right?

Uh , right , well basically

um high priority for any

animal for me is that they be

willing to take a lot of physical

affection from their family .

And , yeah that they have lots

of personality and

uh be fit and in robust good health .

So this is blue.

Blue beagle.

My family’s beagle .

FEE005 Yeah . Yeah .

[MEE006 [vocalsound-laugh]

FEE005 Right . Lovely .

[vocalsound-laugh]

MEE008 [vocalsound-laugh]

MEE007 [gap]

MEE007 Well , my

favourite animal

would be a monkey .

FEE005 [vocalsound-laugh]

MEE006 [vocalsound-laugh]

Bibliography

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on knowledge and data engineering*, 17(6):734–749, 2005.

J. Allwood, E. Ahlse‘n, J. Nivre, and S. Larsson. Own communication management: Kodningsmanual. *Computer*, 1997.

J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4):273–287, 2007.

J. Arguello and C. Rosé. Topic segmentation of dialogue. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech, Association for Computational Linguistics*, pages 42–49, 2006.

M. Argyle. *The psychology of interpersonal behaviour.* Penguin Press, IV edition, 1983.

D. Balota and E. Marsh. *Cognitive Psychology: Key Readings.* Key readings in cognition. Psychology Press, 2004.

S. Banerjee and A. Rudnicky. Detecting the noteworthiness of utterances in human meetings. In *Proceedings of the SIGDIAL 2009 Conference, Association for Computational Linguistics*, pages 71–78, 2009.

S. Banerjee and A. I. Rudnicky. Smartnotes: Implicit labeling of meeting data through user note-taking and browsing. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 261–264, 2006.

S. Banerjee and A. I. Rudnicky. An extractive-summarization baseline for the automatic detection of noteworthy utterances in multi-party human-human dialog. In *Proceeding of the IEEE Spoken Language Technology Workshop (SLT)*, pages 177–180, 2008.

J. Barwise and J. Perry. Semantic innocence and uncompromising situations. *Midwest studies in philosophy*, 6(1):387–404, 1981.

D. Bauer. Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, 67(339):687–690, 1972.

M. Bazire and P. Brézillon. Understanding context before using it. In *Proceedings of CONTEXT 2005, Modeling and Using Context, 5th International and Interdisciplinary Conference*, pages 29–40, 2005.

J. A. Bea and P. C. Marijuán. The informational patterns of laughter. *Entropy*, 5(2): 205–213, 2003.

F. Bonin, R. Böck, and N. Campbell. How do we react to context? annotation of individual and group engagement in a video corpus. In *Proceedings of SocialCom/PASSAT*, pages 899–903, 2012a.

F. Bonin, N. Campbell, and C. Vogel. Laughter and topic changes: Temporal distribution and information flow. In *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pages 53–58, 2012b.

F. Bonin, N. Campbell, and C. Vogel. Laughter and topic changes: Temporal distribution and information flow. In *Proceedings of CogInfoCom, 2012 IEEE 3rd International Conference on Cognitive Infocommunications*, pages 53–58, Dec 2012c. doi: 10.1109/CogInfoCom.2012.6422056.

F. Bonin, C. D. Looze, S. Ghosh, E. Gilmartin, C. Vogel, A. Polychroniou, H. Salamin, A. Vinciarelli, and N. Campbell. Investigating fine temporal dynamics of prosodic and lexical accommodation. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 539–543, 2013.

F. Bonin, N. Campbell, and C. Vogel. Time for laughter. *Knowledge-Based Systems*, 71 (0):15 – 24, 2014a.

F. Bonin, E. Gilmartin, C. Vogel, and N. Campbell. Topics for the future: Genre differentiation, annotation, and linguistic content integration in interaction analysis. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges, RFMIR@ICMI 2014*, pages 5–8, 2014b.

F. Bonin, J. San Pedro, and N. Oliver. A context-aware NLP approach for noteworthiness detection in cellphone conversations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 25–36, Dublin, Ireland, August 2014c.

F. Bonin, C. Vogel, and N. Campbell. Social sequence analysis: temporal sequences in interactional conversations. In *Cognitive Infocommunications (CogInfoCom), 2014 5th IEEE Conference on*, pages 403–406, Nov 2014d.

F. Bonin, N. Campbell, and C. Vogel. The discourse value of social signals at topic change moments. In *INTERSPEECH, Accepted*, 2015.

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

S. E. Brennan. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, pages 41–44, 1996.

G. Brown and G. Yule. *Discourse analysis*. Cambridge University Press, 1983.

H. Bunt. Dynamic interpretation and dialogue theory. In *Taylor, M., Neel, F., and D., B., editors, The structure of multimodal dialogue II*, pages 139–166. John Benjamins, Amsterdam, 1999.

H. Bunt. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2):222 – 245, 2011.

N. Campbell. An audio-visual approach to measuring discourse synchrony in m ultimodal conversation data. In *Proceedings of Interspeech 2009*, 2009.

J. P. Carrascal, R. de Oliveira, and M. Cherubini. A note paper on note-taking: understanding annotations of mobile phone calls. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, MobileHCI '12, pages 21–24, New York, NY, USA, 2012.

L. Cerrato. *Investigating Communicative Feedback Phenomena across Languages and Modalities*. PhD thesis, KTH Computer Science and Communication, Stockholm, Sweden, 2007.

W. Chafe. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li, editor, *Subject and Topic*, pages 25–55. Academic Press, New York, 1976.

W. Chafe. Polyphonic topic development. *Conversation: Cognitive, communicative and social perspectives.*, VIII:41–54, 1997.

E. D. Chapple. Quantitative analysis of the interaction of individuals. *Proceedings of the National Academy of Sciences of the United States of America, 25(2)*, pages 58–67, 1939.

F. Chen, K. Han, and G. Chen. An approach to sentence-selection-based text summarization. In *Proceedings of TENCON'02, IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, volume 1, pages 489–493. IEEE, 2002.

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition) 2nd Edition.* New Jersey: Lawrence Erlbaum, 2nd edition, 1988.

C. L. Crown and S. Feldstein. The perception of speech rate from the sound-silence patterns of monologues. *Journal of Psycholinguistic Research*, 20(3):47–63, 1991.

A. K. Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1): 4–7, 2001.

A. Dielmann and S. Renals. Automatic meeting segmentation using dynamic bayesian networks. *Multimedia, IEEE Transactions on*, 9(1):25–36, 2007.

P. Dourish. What we talk about when we talk about context. *Personal Ubiquitous Comput.*, 8(1):19–30, February 2004. ISSN 1617-4909.

A. Duranti. Language matters in anthropology: A lexicon for the new millennium. In *A Special Issue of the Journal of Linguistic Anthropology*, volume 9, pages 1–2, 1999.

A. Duranti and C. Goodwin. *Rethinking the Context: Language as an interactive phenomenon.* Cambridge, England: Cambridge University Press, 1992.

J. Eisenstein, R. Barzilay, and R. Davis. Gestural cohesion for topic segmentation. In *Proceedings of ACL-08: HLT*, pages 852–860, Columbus, Ohio, June 2008.

A. Fetzer. *Recontextualizing Context: Grammaticality Meets Appropriateness*. Pragmatics & beyond. John Benjamins Pub., 2004.

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 364–372, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

M. Galley, K. R. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In E. Hinrichs and D. Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–03)*, pages 562–569, Sapporo, Japan, July 2003.

H. Garfinkel. *Studies in Ethnomethodology*. Eglenwood Cliff, NJ. Prentice Hall, 1967.

S. Ghosh, R. Johansson, G. Riccardi, and S. Tonelli. Improving the recall of a discourse parser by constraint-based postprocessing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 2791–2794, 2012.

E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell. Laughter and topic transition in multiparty conversation. In *Proceedings of SIGDIAL 2013 Conference,Metz, France, 22-24 August*, pages 304–308, 2013a.

E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell. Exploring the role of laughter in multiparty conversation. In *Proceedings of SEMDIAL 2013, (DialDam), Amsterdam, Netherlands, December 2013*, pages 191–193, 2013b.

J. Ginzburg. Resolving questions i. *Linguistics and Philosophy*, 18-5:459–527, 1995a.

J. Ginzburg. Resolving questions ii. *Linguistics and Philosophy*, 18-6:567–609, 1995b.

T. Givón. *Topic continuity in discourse: A quantitative cross-language study*, volume 3. John Benjamins Publishing Company, 1983.

E. Goffman. *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.

E. Goffman. Replies and responses. *Language in Society*, 5:257–313, 12 1976. ISSN 1469-8013.

E. Goffman. *Behavior in public places.* Simon and Schuster, 2008.

E. Goffman. On face work. *Interaction Ritual*, pages 5–46, 1967.

F. Goldman-Eisler. The predictability of words in context and the length of pauses in speech. *Language and Speech*, 1(3):226–231, 1958.

C. Goodwin. *Conversational organization: Interaction between speakers and hearers.* Academic Press New York, 1981.

C. Goodwin. Embedded context. *Research on Language & Social Interaction*, 36(4):323–350, 2003.

A. Grobet. *L'identification des topiques dans les dialogues.* Champs Linguistiques: Recherches. De Boeck Supérieur, 2002.

H. Gross, C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, A. Bley, C. Martin, T. Langner, and M. Merten. Progress in developing a socially assistive mobile home robot companion for the elderly with mild cognitive impairment. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 2430–2437, Sept 2011.

B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

A. Gruenstein, J. Niekrasz, and M. Purver. Meeting structure annotation: Data and tools. In *6th SIGdial Workshop on Discourse and Dialogue*, 2005.

J. Gumperz and D. Hymes. *Directions in sociolinguistics: the ethnography of communication.* Holt, Rinehart and Winston, 1972.

J. Gumperz. Interactional sociolinguistic a personal perspective. In H. E. H. edited by Deborah Schiffrin, Deborah Tannen, editor, *The Handbook of Discourse Analysis*, pages 215–228. John Wiley & Sons„ 2008.

V. Gupta and G. S. Lehal. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3):258–268, 2010.

M. Halliday and R. Hasan. *Cohesion in English.* English language series. Longman, 1976.

M. A. Halliday. Notes on transitivity and theme in english: Part 2. *Journal of linguistics*, 3(02):199–244, 1967.

D. Harrah. Message theory and the semantics of dialogue. In L. Vaina and J. Hintikka, editors, *Cognitive Constraints on Communication*, volume 18 of *Synthese Language Library*, pages 267–276. Springer Netherlands, 1984.

P. Healey and C. Vogel. Dressing dialog for success. In *Proceedings of the Munich Workshop on Formal Semantics and Pragmatics of Dialog.*, pages 82–99, 1997.

M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pages 33–64, 1997.

M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

J. Heritage. *Garfinkel and Ethnomethodology.* Social and political theory from Polity Press. Polity Press, 1984.

J. Heritage. The new blackwell companion to social theory. pages 300–320. Oxford University Press, 2008.

J. Hirschberg and B. Grosz. Intonational features of local and global discourse structure. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, pages 441–446, 1992.

J. R. Hobbs. *On the coherence and structure of discourse.* CSLI, 1985.

C. Hockett. Two models of grammatical description. In *Readings in Linguistics*. University of Chicago Press,Chicago, 1958.

E. Holt. The last laugh: Shared laughter and topic termination. *Journal of Pragmatics*, 42(6):1513–1525, June 2010.

E. Holt. On the nature of 'laughables' : laughter as a response to overdone figurative phrases. *Pragmatics*, 21(3):393–410, September 2011.

C. Hovland. *The order of presentation in persuasion.* Bibliographie - pp. 190. Published for the Instituted of Human Relations [by] Yale University Press, 1966.

P.-Y. Hsueh and J. D. Moore. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, page 1016–1023, Prague, Czech Republic, June 2007. Association for Computational Linguistics, Association for Computational Linguistics.

D. Hymes. Models of interaction of language and social life. *Directions in sociolinguistics: the ethnography of communication*, pages 35–71, 1972.

D. Hymes. Introduction: Toward ethnographies of communication1. *American Anthropologist*, 66:1–34, 1964.

R. Jakobson. Structure of language and its mathematical aspects. Proceedings of symposia in applied mathematics. American Mathematical Society, 1961.

V. J. Jensen. Communicative functions of silence. *ETCA Review of General Semantics 30,*, pages 249–257, 1973.

M. Johnston, J. Chen, P. Ehlen, H. Jung, J. Lieske, A. Reddy, E. Selfridge, S. Stoyanchev, B. Vasilieff, and J. Wilpon. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Associatio for Computational Linguistics*, chapter MVA: The Multimodal Virtual Assistant, pages 257–259. 2014.

K. Jokinen. Gaze and Gesture Activity in Communication 1. *Communication*, pages 1–11, 2009a.

K. Jokinen. Gaze and gesture activity in communication. In C. Stephanidis, editor, *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, volume 5615 of *Lecture Notes in Computer Science*, pages 537–546. Springer Berlin / Heidelberg, 2009b.

K. Jokinen. Nonverbal feedback in interactions. In J. Tao and T. Tan, editors, *Affective Information Processing*, pages 227–240. Springer London, 2009c.

K. Jokinen. Pointing gestures and synchronous communication management. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, and A. Nijholt, editors, *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967 of *Lecture Notes in Computer Science*, pages 33–49. Springer Berlin Heidelberg, 2010.

K. Jokinen and G. Wilcock. Contextual inferences in intercultural communication. *SKY Journal of Linguistics*, 19:291–300, 2006.

H. Kamp, J. Van Genabith, and U. Reyle. Discourse representation theory. In *Handbook of philosophical logic*, pages 125–394. Springer, 2011.

E. Keenan. Reference restricting operators in universal grammar. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, pages 227–237, 1976.

A. Kendon. Some context for context analysis: a view of the origin of the structural studies of face to face interaction.the problem of meaning in primitive language. *Conducting interaction: Patterns of behavior in focus encounters*, pages 15–50, 1990.

A. Kratzer. Situations in natural language semantics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014.

K. Lambrecht. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge Studies in Linguistics. Cambridge University Press, 1994.

S. Levinson. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1983.

R. Levitan and J. Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

S. Luz. Locating case discussion segments in recorded medical team meetings. In *Proceedings of the Third Workshop on Searching Spontaneous Conversational Speech*, SSCS '09, pages 21–30, 2009. ISBN 978-1-60558-762-2.

S. Luz. The nonverbal structure of patient case discussions in multidisciplinary medical team meetings. *ACM Trans. Inf. Syst.*, 30(3):17:1–17:24, 2012.

N. A. Madzlan, J. G. Han, F. Bonin, and N. Campbell. Towards automatic recognition of attitudes: Prosodic analysis of video blogs authors. In *Speech Prosody, Dublin 2014*, pages 91–94, 2014a.

N. A. Madzlan, J. Han, F. Bonin, and N. Campbell. Automatic recognition of attitudes in video blogs - prosodic and visual feature analysis. pages 1826–1830, 2014b.

B. Malinowski. The problem of meaning in primitive language. *The meaning of Meaning. First edition 1923*, pages 296–336, 1947.

P. Maranda. *Mythology; selected readings.* Penguin modern sociology readings. Penguin Books, 1972.

S. Maskey and J. Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH*, pages 621–624, 2005.

I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 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*, 2005.

J. L. McKechnie. *Webster s New Twentieth Century Dictionary of the English Language.* Number 9780671418199 in Dictionary Simon and Schuster; 2nd edition, 1983.

D. McNeill. *Gesture and thought.* University of Chicago Press, 2008.

J. Mey. *Pragmatics: An Introduction.* Wiley, 2001.

J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

E. Ochs and B. Schieffelin. Language acquisition and socialization: Three developmental stories and their implications. *Linguistic anthropology: A reader*, pages 263–301, 2001.

T. Oya, Y. Mehdad, G. Carenini, and R. Ng. *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, chapter A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships, pages 45–53. Association for Computational Linguistics, 2014.

L. Padro, S. Reese, E. Agirre, and A. Soroa. Semantic services in freeling 2.1: Wordnet and ukb. In P. Bhattacharyya, C. Fellbaum, and P. Vossen, editors, *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, Mumbai, India, February 2010. Global Wordnet Conference 2010, Narosa Publishing House.

J. Pallant. *SPSS Survival Manual*. Open University Press, Milton Keynes, UK, USA, 3rd edition, 2007.

M. Pantic, R. Cowie, F. D' Errico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, and A. Vinciarelli. Social signal processing: the research agenda. In *Visual analysis of humans*, pages 511–538. Springer, 2011.

R. J. Passonneau and D. J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, March 1997.

A. Pentland. Social signal processing. *IEEE Signal Processing Magazine*, 24(4):108, 2007.

V. Petukhova and H. Bunt. Towards a multidimensional semantics of discourse markers in spoken dialogue. In *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 09, pages 157–168, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

M. J. Pickering and V. S. Ferreira. Structural priming: a critical review. *Psychological bulletin*, 134(3):427, 2008.

M. Poesio and D. Traum. Conversational actions and discourse situations. *Computational intelligence*, 13(3):309–347, 1997.

M. Purver. Topic segmentation. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317, 2011.

J. C. Reynar. An automatic method of finding topic boundaries. In *ACL*, pages 331–333, 1994.

G. Riccardi. Towards healthcare personal agents. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges, RFMIR@ICMI 2014*, pages 53–56, 2014.

V. P. Richmond, J. C. McCroskey, and S. K. Payne. *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ, 1991.

E. Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

E. Rosch. Principles of categorization. *Concepts: core readings*, pages 189–206, 1999.

H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversationdynamic interpretation and dialogue theory. *Language*, pages 696–735, 1974.

H. Sacks. *Lecture on Conversation*. Wiley-Blackwell, 1995.

E. Schegloff. *Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis*. Sequence Organization in Interaction: A Primer in Conversation Analysis. Cambridge University Press, 2007.

E. A. Schegloff. Sequencing in conversational openings1. *American Anthropologist*, 70 (6):1075–1095, 1968.

E. A. Schegloff and H. Sacks. Opening up Closings. *Semiotica*, 8:289–327, 1973.

D. Schiffrin. *Approaches to Discourse: Language as Social Interaction*. Blackwell Textbooks in Linguistics. Wiley, 1994. ISBN 9780631166238.

B. N. Schilit and M. M. Theimer. Disseminating active map information to mobile hosts. *Network, IEEE*, 8(5):22–32, 1994.

B. Schuller and A. Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, 2013.

A. J. Sellen. Remote conversations: The effects of mediating talk with technology. *Human-computer interaction*, 10(4):401–444, 1995.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. *The ICSI meeting recorder dialog act (MRDA) corpus.* Defense Technical Information Center, 2004.

L. Sidner and Grosz. Attention, Intentions, and the Structure of Discourse. *Discourse*, 12(3), 1986.

L. Smith-Lovin and C. Brody. Interruptions in group discussions: The effects of gender and group composition. *American Sociological Review*, pages 424–435, 1989.

R. Sornicola. Topic and comment. In *Encyclopaedia of Language and Linguistics*, volume 12, pages 766–773. Keith Brwon ed., 2006.

P. Strawson. *Subject and Predicate in Logic and Grammar.* University Paperbacks. Methuen ; [New York], 1974.

D. Tannen. *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse.* Studies in interactional sociolinguistics. Cambridge University Press, 1989.

K. Tracy. Analyzing context: Framing the discussion. *Research on Language and Social interaction*, 31(1):1–28, 1998.

K. P. Truong and J. Trouvain. Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis. *Proceedings of the 4th International Worskhop on Corpora for Research on Emotion Sentiment and Social Signals*, pages 20–24, 2012.

A. Van Dijk. *Sentence Topic versus Discourse Topic*, pages 177–194. Mouton, 1981.

T. Van Dijk. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse.* Longman Linguistics Library. Addison-Wesley Longman Limited, 1977.

A. Vinciarelli. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, November 2009.

A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signals, their function and automatic analysis: A survey. In *Proceedings of ACM Int'l Conf. Multimodal Interfaces (ICMI'08)*, pages 61–68, Chania, Greece, October 2008.

C. Vogel and L. Behan. Measuring synchrony in dialog transcripts. In A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, editors, *Behavioural Cognitive Systems*, pages 73–88. Springer, LNCS 7403, 2012.

E. Weinstein. *Search Problems for Speech and Audio Sequences by.* PhD thesis, 2009.

J. K. Weiqun Xu, Jean Carletta and V. Karaiskos. Coding instructions for topic segmentation of the ami. version 1.1, 2005.

M. William and S. Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281, 1988.

S. Xie, Y. Liu, and H. Lin. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 157–160. IEEE, 2008.

B. S. Yandell. *Practical Data Analysis for Designed Experiments.* Chapman & Hall, 1997.

B. Zellner. Pauses and the temporal structure of speech. pages 41–62. John Wiley, 1994.