TRINITY COLLEGE DUBLIN

SCHOOL OF COMPUTER SCIENCE AND STATISTICS

THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Quantification of Mutual Understanding in Task-Based Human-Human Interactions

Author:
Justine REVERDY

Supervisor:
Dr. Carl VOGEL

Submitted to Trinity College, the University of Dublin
March 15, 2021

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Declaration

I declare that this thesis details entirely my own work, and has not been submitted as an exercise for a degree at any other university. Due acknowledgements and references are given to the work of others where appropriate.

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.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Justine Reverdy

Dated: March 15, 2021

Abstract

This thesis explores the quantification of mutual understanding in task-based interactions by observing the relation between patterns of repetitions and measures of communicative success. Two important characteristics of mutual understanding have to be kept in mind: it cannot be established for certainty and cannot be directly measured. However, signs of understanding can be detected and quantified, based on two elements: (1) the way by which conversational partners achieve understanding is dependent on their communicative behaviour, and (2) dialogues exhibit repetitions, despite the immense number of possibilities to compose sentences with words that are at our disposal.

These repetitions of linguistic choices between conversational partners, a process known as *alignment*, are argued to play an important role in the establishment of a common ground that leads to understanding. The exact dynamic of *alignment* – and related phenomena such as *synchrony* – is still under debate, which has created a large body of research interested in determining its scope. However, fewer studies have been conducted that systematically examine its relation with communicative success, and even fewer studies do it in an automatic way that does not require human annotations.

It is in this perspective that the research presented here compares repetition patterns to different communicative assessment methods, namely task-success scores, presence of high levels of negative/positive cognitive states, and third-party moderator evaluation. Five corpora with a total of 192 dialogues (about 32 hours) are analysed in terms of *other-shared* and *self-shared* repetitions, at different levels of linguistic representations and utterance lengths.

The main contribution of this thesis is the establishment of the extent to which repetitions – categorised as happening outside chance variation – may function as a proxy measure of mutual understanding. Results suggest a higher proportion of *other* than *self*-repetitions happening above chance in task-based interactions. While participants in the position of information givers have a higher volume of speech and use longer utterances, information followers repeat the giver and themselves more. Information givers repeating themselves seem to relate to higher task success, even more so when repeating themselves structurally, in particular for women. Furthermore, familiarity emerged as a decisive factor for success. Participants being familiar with each other unsurprisingly achieved better scores whether they exhibited signs of linguistic alignment or not, however, unfamiliar partners seemed to benefit from alignment, in particular at first attempt of a task. In computer-mediated interactions, both *other* and *self* repetitions happened in high proportions, and a significant drop in *self*-repetitions of long utterances was observed in troublesome dialogues; in interactions monitored by a human facilitator, more encouragements were provided where the method detected less alignment and inversely less encouragement when alignment was present. These two findings highlight the potential of (1) detection of problematic communication, (2) indication of the state of an interaction – mutual understanding taking place or not, of the described method. It was also found that American speakers repeat themselves more than Scottish speakers. However, in both dialects, familiar participants did not need to exhibit alignment to succeed in the task. Finally, *divergence* – taken as the opposite behaviour of *alignment* – was very seldom exhibited in the task-based corpora analysed.

Altogether, the proxy measure of mutual understanding described in this document stress that the research efforts made in this direction have a great potential both for the improvement of dialogue systems and monitoring critical human interactions.

Acknowledgements

Firstly, I would like to thank my supervisor, Carl Vogel, for the time and never-failing attention to detail that he provided me during the years of the PhD. I sincerely thank him for his patient support and inspirational advice. His guidance was always able to keep me on track when I got carried away, and without his precious feedback at all stages of this work, it would not have been possible.

I acknowledge here the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106), co-funded under the European Regional Development Fund, and all the great colleagues I met there.

Also, I would like to thank my previous supervisor, Nick Campbell, for his encouragements and support during my internship in Trinity College, and the early stages of this work.

My earnest thanks also go to Naomi Harte and Tim Fernando, my Confirmation examiners, for their valuable and constructive criticism that helped to improve the research direction of this thesis. I also wish to express my gratitude towards my Viva Voice examiners, Christine Howes and Gaye Stephens, with whom the detailed discussion of this thesis work was deeply enriching, as well as a positive and valuable exercise. I would like to thank my colleagues for their academic help and friendship, Carmen Klaussner, Akira Hayakawa, Erwan Moreau, Milena Lopes, Camille Nadal, Kevin Doherty, Maria Koutsombogera, Fahim Salim, Fasih Haider, Liliana Mamani Sanchez, Loredana Cerrato, Ashjan Alsulaimani, Marguerite Barry, Bérenger Arnaud, Francesca Bonin, Emer Gilmartin, Alfredo Maldonado, Andreas Balaskas, Zaynab Salman, Leysan Nurgalieva and Seamus Ryan.

I would like to particularly thank my co-authors, Akira Hayakawa and Maria Koutsombogera, for their inspirational help in shaping my research with their valuable insights, experience and expertise, as well as their trust.

I also wish to thank my family members and friends in both France and Ireland, that provided support during all steps of this thesis, and notably for checking on my increasingly erratic sleeping patterns and making sure I kept discovering new parts of this beautiful country.

Finally, I would like to express all my gratitude to my parents Joëlle and Denys and my sister Isabelle, for all their support and love, throughout these years of the PhD, but also during all my previous life stages.

Thank you all.

Justine Reverdy

University of Dublin, Trinity College

March 15, 2021

“ Language analysts believe that there are no genuine philosophical problems, or that the problems of philosophy, if any, are problems of linguistic usage, or of the meaning of words. I, however, believe that there is at least one philosophical problem in which all thinking men are interested. It is the problem of cosmology: the problem of understanding the world — including ourselves, and our knowledge, as part of the world.”

The Logic of Scientific Discovery (Preface, 1959)

– Karl Popper

———————————————-

“ Roi Arthur .– Vous n’êtes pas sans savoir que les tentatives d’invasions saxonnes se multiplient dangereusement ces derniers temps.

Élias de Kelliwic’h .–

Roi Arthur .– Vous êtes au courant de ça?

Élias .– Pardon?

Roi Arthur .– Vous êtes au courant de ça?

Élias .– Oui.

Roi Arthur .– Et bah dites-le!

Élias .– Bah je vous écoute! Je ne vais pas dire ‘oui oui’ toutes les cinq minutes! ”

Kaamelott - Livre III (Épisode 21) - Le Renfort Magique (2006)

– Alexandre Astier

Associated Publications

Reverdy, J., Vogel, C. (2017).

Measuring Synchrony in Task-Based Dialogues

In *Proceedings of INTERSPEECH' 2017 : the 18th Annual Conference of the International Speech Communication Association* (pp.1701-1705). Stockholm, Sweden: International Speech Communication Association (ISCA).

Reverdy, J., Vogel, C. (2017).

Linguistic Repetitions, Task-based Experience and A Proxy Measure of Mutual Understanding

In *Proceedings of CogInfoCom 2017 : the 8th IEEE International Conference on Cognitive InfoCommunications* (pp. 395-400). P. Baranyi, A. Esposito, P. Földesi, and T. Mihálydeák, (Eds.), Debrecen, Hungary: IEEE. – Best Paper Award

Reverdy, J., Hayakawa, A., Vogel, C. (2018).

Alignment in a Multimodal Interlingual Computer-Mediated Map Task Corpus

In *Proceedings of LREC 2018 : the 11th edition of the Language Resources and Evaluation Conference.* H. Koiso P. Paggio (Eds.) Paris, France: European Language Resources Association (ELRA), (pp. 55–59), Workshop on Language and body in real life & Multimodal Corpora 2018.

Reverdy, J., Koutsombogera, M., Vogel, C. (2020).

Linguistic Repetition in Three-Party Conversations

In *Neural Approaches to Dynamics of Signal Exchanges.* (pp. 359-370) Springer, Singapore.

Reverdy, J., Hayakawa, A. Vogel, C. (2020).

Map Task Deviation Scores: A Reconstruction

In *Proceedings of CogInfoCom 2020 : the 11th IEEE International Conference on Cognitive InfoCommunications* (In Press). P. Baranyi, A. Esposito, P. Földesi, and T. Mihálydeák, (Eds.), Debrecen, Hungary: IEEE.

Contents

Chapter 1

Introduction

As unscripted spoken interactions unfold, people tend to repeat their interlocutors and themselves. From a computational linguistics and cognitive science perspective, this research examines dialogues, a form of multi-party interaction that can, in most cases, be seen at aiming to reach mutual understanding. How the achievement of understanding can be quantified, and which methods can be used to detect linguistic behaviours displayed by interactants that can signal this understanding, are the main themes I am going to focus on, over the course of this document.

Within the many theories of dialogue used to develop current speech interfaces (interactive voice response systems or conversational agents) in a broad range of applications, the idea that speakers cooperate in a coordinated manner is a fundamental element in the achievement of successful communication (Lester, Branting, & Mott, 2004).

In 1996, Herbert H. Clark, following previous research in the domain of discourse analysis, suggested that unscripted conversations are joint productions of the interlocutors (Clark, 1996). A dialogue is more than the sum of two monologues but rather a coordinated construct in which establishing common ground is essential to communicative success, and where positive evidence of mutual understanding is required at different levels and updates on the current state of knowledge are made in a continuous manner (Clark & Brennan, 1991; Clark, 1996; Fussell & Krauss, 1989; Clark & Krych, 2004).

An unscripted dialogue is a highly dynamic system, in which interactants adapt over time, and is highly influenced by contextual constraints. A phenomenon of communication that is considered as central in this system is *synchrony*, the product of interpersonal coordination

in which interactants coordinate with each other in their verbal and non-verbal behaviours (Bernieri & Rosenthal, 1991; Ramseyer & Tschacher, 2010; Vicaria & Dickens, 2016), and relate to *alignment*, i.e. the repetition of linguistic choices. The two notions of *synchrony* and *alignment* are entangled, and frequently used in the definition of each other in the literature, however, *synchrony* is more often referred to as a general whole body non-verbal behaviour while *alignment* is regularly used in the context of linguistic behaviour. (See § 2.1.2 and § 2.1.3 for more details.)

Among the features of dialogue that structure discourse, repetitions are considered as appropriate elements for quantitative treatment (Schegloff, 1993), particularly, but not exclusively, in the form of other-initiated repair. Also, the repetition of linguistic choices holds multiple communicative functions (Tannen, 2007; Ursi, Oloff, Mondada, & Traverso, 2018), that researchers have suggested establish conversational involvement and may support communicative progress towards mutual understanding. From repetitions in dialogues to the notion of mutual understanding, a number of assumptions have to be considered and validated, hence the necessity to proceed carefully. The purpose of this dissertation is to understand to what extent repetition can be considered a reliable feature to the estimation that understanding has been reached between dialogue interactants.

Several communication models and typologies for repetitions are reviewed to place in context the central usage of repetitions in language, however, this thesis is not concerned with the categorisation of repetitions by their functions in discourse, but rather by the quantification of a specific construction of a repetition. That is the repetition of a linguistic element (Token, Part-Of-Speech, or Lemma), from one interlocutor turn from the preceding turn (his own turn or another speaker's turn). This construction is given with more details in chapter 4.

The method's core shape, designed to examine interlocutors synchrony, was introduced by Lydia Behan and Carl Vogel (2010; 2012). The present work extends, refines and uses this method in multiple contexts while associating the notion of synchrony with measures of communicative success in task-based interactions. This association allows the development of a proxy measure of mutual understanding.

1.1 Research Question

This research lies on two foundations: repetitions constitute a key component in the building of a common ground that leads to mutual understanding, and that component can be used to predict successful communication once interactional patterns are established in context.

Whether mutual understanding is achieved by interlocutors can never be asserted with complete certainty; however, interlocutors can achieve a state in which they lack direct evidence of misunderstanding (Taylor, 1992). I adopted the existence of a null hypothesis as described by Vogel (2013, p. 384): unless a significant (in its statistical meaning) amount of communicative cues, such as repetitions, are evident in dialogue, mutual understanding cannot be reliably asserted. Which leads to the main counter hypothesis: measures of repetition can be used as proxy measures of mutual understanding. It is not expected that repetitions perfectly index understanding; language is a complex phenomenon influenced by a potentially infinite range of factors. The context and social situation will have a critical impact.

This thesis argues that, in task-based interactions in particular, the relation between linguistic adaptation and communicative success can be estimated through the quantification of repetitions between and within interlocutors.

Following these considerations, the main research question that shapes the work presented in this document is:

To what extent is an automatic method focusing on one feature of dialogue structure, repetition as cues of an alignment process, able to capture interactional behaviours and patterns with sufficient accuracy to quantify a degree of mutual understanding?

1.2 Motivations

Determining if alignment exists between interlocutors has been the subject of extensive previous work. However, less work has been interested in coupling the extent to which this alignment may happen in relation to mutual understanding, in the specific context of unscripted task-based interactions. In addition to the work from Behan and Vogel (2012) mentioned above, two research efforts in particular have pushed these boundaries and are part of the motivational basis for this thesis. Reitter and Moore (2007; 2008) investigated the pre-

diction of task success in relation to priming effects at the syntactic level, defining priming as the decay of repetition probability over time. They found a relation between task success and long term priming effects. Colman and Healey (2011; 2012), explored the relation between mutual understanding and patterns of distributions of repair and the phenomena of ellipsis/anaphora. By developing a manual coding protocol, they showed that the distribution of those phenomena were different in task-based and ordinary conversation and that the medium and familiarity also impacted that distribution. A more complete review of those works is given in chapter 2, but these two works laid part of the foundation for this thesis by showing the differences in distribution of alignment cues in task-based conversations as well as the possibilities in their automatic detection.

The capacity to use low-level cues[1] present in dialogue, such as repetitions, as an index for the higher level process of mutual understanding, has multiple implications.

Along with the improved understanding of conversational content and human behaviour that corpus-based conversation analysis provides, the potential uses for an automatic measure of interactional success are diverse; in particular if the method used does not require a human annotation phase, that is often a long and delicate process. Among those possibilities are the improvement of performance of conversational agents (from chatbots to dialogue systems in general), either in the form of design guidelines, or in the assessment of human-human/human-machine interactional performances. In the rapidly growing field of artificial personal assistants such as Siri (Apple), Alexa (Amazon), Bixby (Samsung) or Cortana (Microsoft), how to interpret and assess conversational content, with respect to mutual understanding and common ground building, in an automatic way, is still in need of more insights (Brennan, Galati, & Kuhlen, 2010; Fusaroli & Tylén, 2016).

Many domains could benefit from interactional measures, where the monitoring and assessment of the likelihood of understanding can be essential. To give a few examples of such situations, the evaluation of interactional content remains critical between pilots and air-traffic controllers, medical personnel and patients,[2] call-centre officers and clients, or courtroom/police interrogations.

[1] Also called sometimes *low-hanging fruits*: lexical and syntactic levels; as opposed to *high-hanging fruits*: semantic and pragmatic levels.

[2] Notably in the detection of schizophrenia or depression.

1.3 Contributions

In order to answer the research questions defined in section 1.1, five experiments were undertaken in chapter 5, using five task-based corpora: 1. the HCRC Map Task corpus (Anderson, Bader, et al., 1991), 2. the ILMT-s2s corpus (Hayakawa, Luz, Cerrato, & Campbell, 2016), 3. the MULTISIMO corpus (Koutsombogera & Vogel, 2018), 4. the MIT American English Map Task (AEMT) corpus, and 5. the PARDO 2006 Map Task corpus (Pardo, 2006). In each corpus, the tasks are to be achieved by the collaboration of two participants through the medium of speech. Their degree of success in the task are directly linked to their ability to cooperate linguistically, and it is hypothesised that the better their cooperation is, the better they perform in the task. These are the elements that binds together the corpora that could be seen as a meta-data set for the exploration of communicative success. These corpora each use different methods to measure task success, which I use to estimate mutual understanding: (1), (4), and (5) task scores, (2) the presence of negative/positive cognitive states, and (3) third party assessments. These measures each relate to mutual understanding in different ways, which are described in more details along with the corpora they are associated with in chapter 3.

The HCRC Map Task corpus contains the largest number of dialogues and controlled factors such as role, gender, familiarity between participants, eye-contact, and familiarity with the task. Which is why it constitutes the core dataset, that the other corpora are compared against.[3] While the measurement of task success cannot be assumed to match communicative success perfectly, it can be reasonably hypothesised as a sufficient indicator. This is because the measurements contained in the HCRC Map Task corpus are close to an objective measurement of communicative success, since it does not depend on an annotator's assessment that can contains a certain subjectivity. To ensure reliability in the task success scores, the method was refined and reconstructed by the author and all scores of the 128 maps of the HCRC corpus were counted and compared to the score given by the authors of the corpus. The analysis presented in this document revealed the presence of alignment in the majority of the HCRC Map Task, but no overall correlation between communication success and alignment. However, this link is present in certain conditions, notably for unfamiliar partner

[3]Since the MULTISIMO corpus does not use the map task method, it is not directly compared *per se* with the HCRC Map Task corpus.

at first attempt of the task. This empirical observation highlights the current paradigm that is: alignment, rather than being a universal process occurring in all types of conversation, it is mostly present in task-based interactions, and favour successful communication in only certain conditions, of which I go into details in chapter 5 and chapter 6. To confirm findings from the study of the HCRC Map Task, four subsequent studies were conducted. They are presented in this document in the order that they have been conducted, as each experiment led to further questioning that aimed at answering the main research question, of which each corpus helped answer a partial aspect, within goal-oriented task-based interactions. Analysis of the ILMT-s2s corpus revealed a (1) lack of structural alignment in the dialogues where high levels of frustration were found and an examination of the MULTISIMO corpus showed (2) a match between the behaviour of an interactional facilitator and quantified levels of repetitions relating a lack of alignment with more encouragement and evidence of alignment with less encouragement. An additional experiment (see § 5.5) revealed similar patterns of the relation between repetitions and task-success measures, even if no perfect matching was found, by the analysis of two American English Map Tasks corpora. The fact that even if overall repetitions patterns differed, notably revealing significantly more self-repetitions in American than in Scottish English, the patterns of repetitions in interaction with task success showed high similarities, which constitute another validation of the method's potential. Finally the method was explored to determine if repetitions happening under what can be considered chance (see § 5.6), taken as the sign of divergence, would bring in some cases relevant information. The patterns found in this last experiment did not allow for a conclusive finding, except that the task-based corpora analysed exhibited almost no significant under chance repetitions. All the above described patterns constitute the main contributions of this thesis, some of which were published:

– Reverdy, J., Vogel, C. (2017). Measuring Synchrony in Task-Based Dialogues. In *Proceedings of INTERSPEECH' 2017 : the 18th Annual Conference of the International Speech Communication Association* (pp.1701-1705). Stockholm, Sweden: ISCA.

– Reverdy, J., Vogel, C. (2017). Linguistic Repetitions, Task-based Experience and A Proxy Measure of Mutual Understanding. In *Proceedings of CogInfoCom 2017 : the 8th IEEE International Conference on Cognitive InfoCommunications* (pp. 395-400). P. Baranyi, A.

Esposito, P. Földesi, and T. Mihálydeák, (Eds.), Debrecen, Hungary: IEEE.

– Reverdy, J., Hayakawa, A., Vogel, C. (2018). Alignment in a Multimodal Interlingual Computer-Mediated Map Task Corpus. In *Proceedings of LREC 2018 : the 11th edition of the Language Resources and Evaluation Conference.* H. Koiso P. Paggio (Eds.) Paris, France: European Language Resources Association (ELRA), (pp. 55–59), Workshop on Language and body in real life & Multimodal Corpora 2018.

– Reverdy, J., Koutsombogera, M., Vogel, C. (2020). Linguistic Repetition in Three-Party Conversations. In *Neural Approaches to Dynamics of Signal Exchanges.* (pp. 359-370) Springer, Singapore.

– Reverdy, J., Hayakawa, A. Vogel, C. (2020). Map Task Deviation Scores: A Reconstruction. In *Proceedings of CogInfoCom 2020 : the 11th IEEE International Conference on Cognitive InfoCommunications* (In Press). P. Baranyi, A. Esposito, P. Földesi, and T. Mihálydeák, (Eds.), Debrecen, Hungary: IEEE.

Other contributions in terms of methodology are the development of the analytical method itself for characterising repetition patterns as well as the extension to multiple levels of linguistic representation and the systematic comparison with measures of communicative success. A visual summary of the experiments is given in Figure 1.1. It shows each experiment associated with their main characteristics and research questions, all designed to answer the principal one given in section 1.1: the preliminary experiment (0) given in Appendix A, that was a replication of previous experiment by Vogel and Behan (2012), and tested first the addition of different levels of representation; the Human-Human task-oriented exploratory experiment (1) around which the following experiment radiates to explore Computer-Mediated interactions (2), Three Party interactions (3), Cultural Influence (4), and Divergence (5). All of which are detailed in chapter 5.

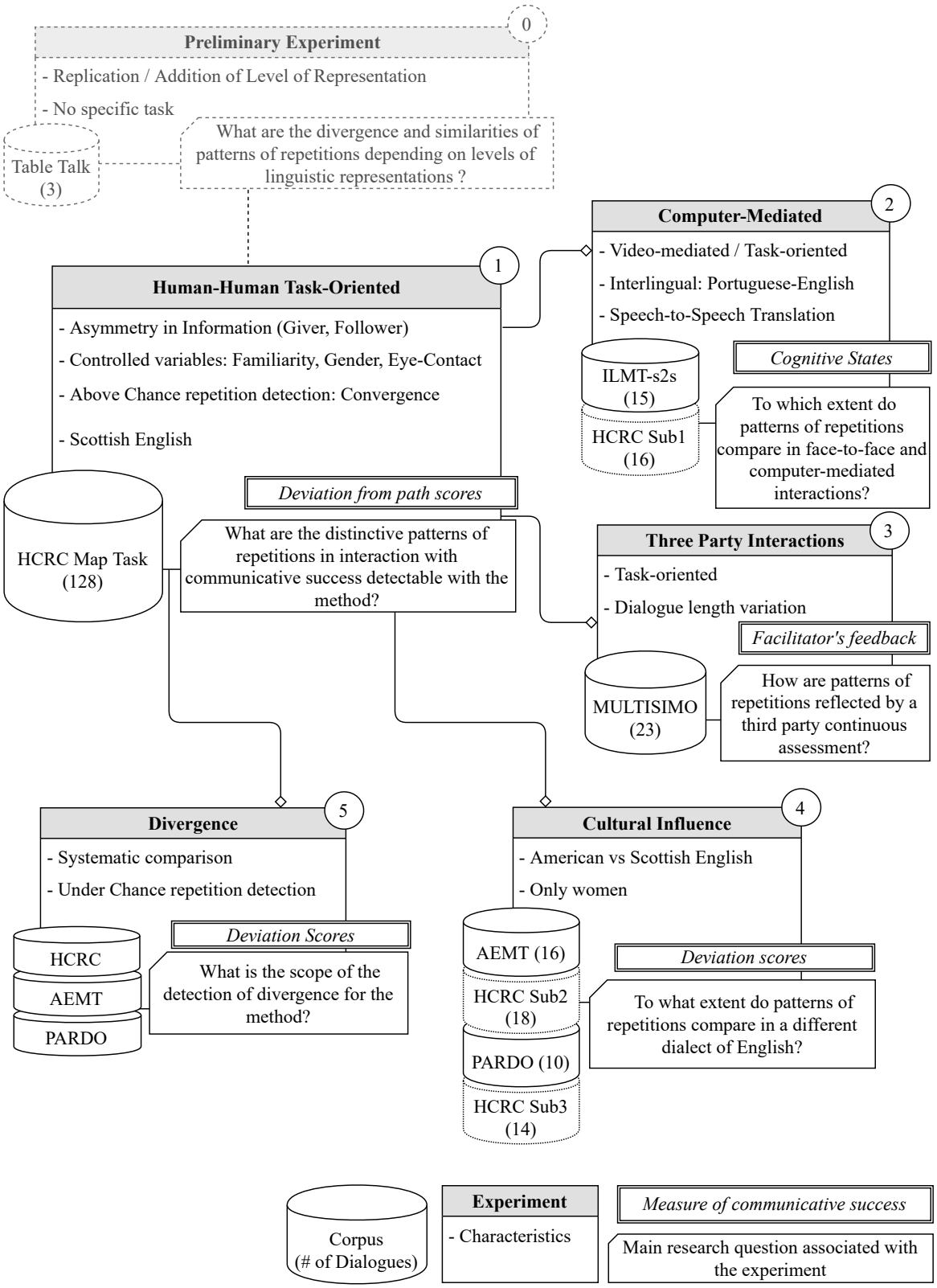


Figure 1.1: Quantifying Mutual Understanding – Experiments Summary

1.4 Thesis Structure

This document is structured as follows: chapter two reviews the literature in the domains corresponding to the research question. Chapter three describes the corpora used as material, presented before the methods as their structure inform the methods designs. Chapter four present the methods with their previous uses, the extension to multiple linguistic levels as well as the reconstruction of task-success scores. Chapter five reports the experiments undertaken along with the results they brought. Finally chapter six gives a general conclusion that emphasize the possibilities of the developed methods to provide a reliable proxy measure of mutual understanding.

Chapter 2

Literature Review

This chapter describes the theoretical framework that the present research is built upon. As mentioned in the introduction chapter, this subject is fundamentally interdisciplinary and associates elements from computational linguistics and cognitive sciences, more generally. The main notions that constitute the components of this thesis' approaches are presented in the context of their respective fields and their relevance to the subject.

2.1 Communication Models

During the last decades, and one could go as far back as the last century, researchers in social sciences have been dissecting human behaviours – describing processes, creating scales of discrete or continuous units, and making observations of large or restricted scopes – in a manner that is not far off the way physicists were preoccupied with the description of matter. Yet, there is no "unified theory" of speech interaction, but rather a series of proposals that describe communication models, with each its definitions, fitted to different perspectives and purposes (Turnbull, 2003). Each model has built on its predecessors, however, the model this thesis takes as a starting point as it is reaching consensus and is considered fundamental in human communication as well as in a wide range of applied dialogue systems research is the model that advocate the concept of *common ground* (Clark, 1996).

2.1.1 The Grounding Phenomenon

In human communication research, the collaborative process in which interlocutors gradually built a *common ground* (Clark & Wilkes-Gibbs, 1986; Clark & Brennan, 1991; Clark, 1996) is a form of *joint action* that combines general beliefs (cultural commonalities) and joint experiences (conversations, events). Robert Stalnaker (1978) first technically described the common ground as the presuppositions of common knowledge, while David Lewis (1979) argued that "presuppositions can be created or destroyed in the course of conversation" (p. 339) which corresponds to a constant incrementation of the state of knowledge. Clark (1996) gives a detailed account on common ground building and insists on the social purpose of language that involve speaker's meaning and addresses understanding. In the establishment of an element as grounded, two phases are distinguished: the presentation phase and the acceptance phase, that allow interlocutors to reach joint closure. An element is presented in the conversation by one speaker, that will consecutively look for positive evidence that this element has been understood by its interlocutor. Once acceptance is made – of which the positive evidence can take multiple signalling forms (words, gestures, noises and so on) and include side sequences of repair in case a misunderstanding is noticed – then the joint action reaches closure. As David Traum (1994, p. 3) notes: "All that is important for communication is that one agent used a particular locution to convey some content to another agent, and that the speaker's intention to convey the content becomes mutually understood (grounded) by both agents, regardless of any objective meaning of the utterance."

In ordinary and task-based conversations it may be the case that both interlocutors are satisfied that mutual understanding has been reached, whether or the not the actual semantic and pragmatic meaning of their utterance has been grounded. This situation occurs when the reference of a word or expression is different in the world representation of the interlocutors. This is the reference problem. Consider the following situation:[1]

Two friends may decide that they will meet in a coffee shop. Both are sure of which location they are referring to, while ignoring the fact that they are mutually referring to different shops of the same franchise, that happened to be located in a relatively small area. In such a case, the satisfaction that meaning has been conveyed and understood is reached

[1] This is a real example that happened to the author of this document.

in the mind of both participants, while in fact a break in the grounding process occurred. In that example, the reference to the place is the element that lacks grounding between the interlocutors, but this is a mere example of the vast number of insufficiently grounded elements that can occur. The element misunderstood was unnoticed and both parties left the conversation convinced that mutual understanding was reached, while the consequences of that misunderstanding only became apparent later on. If, given the local proximity of the two destinations, the two parties met unintentionally on the sidewalk before the appointment, the misunderstanding might not ever have become apparent. This reference problem highlights the fact that assessment from the participants in a conversation is not enough to conclude that mutual understanding has been reached.

Nonetheless, the mechanisms by which the common ground is built that leads to understanding are rooted in a phenomenon that has been described in multiple variations: behavioural adaptation. This thesis is less preoccupied by the distinction between communication models than by the mechanism by which understanding can be quantified: the phenomenon of synchronized interactional patterns, and the hypothesis that this synchrony is at the same time a natural result of the coordinated process of talk and linked to efficient communication.

Corpus-based descriptions of patterns of interaction is a central aspect of the present work, that takes the stance that this phenomenon is detectable at its linguistic level by observing repetitions, and to this aim it is necessary to establish the distinction between two notions that share similarities, but are distinguishable: *synchrony* and *alignment*.

2.1.2 Synchrony

In 1996, Condon and Ogston, defined *interactional synchrony* as the "harmony existing between speaker and listener" and coined *self-synchrony* as the "harmony existing between the speech and body motion of the speaker"(p.342) (Condon & Ogston, 1966). The phenomenon of synchrony is more often associated with non-verbal body/face motions, notably in mother-infant interactions (Condon & Sander, 1974; Feldman, 2007), student-teacher interactions (Bernieri, 1988; LaFrance & Broadbent, 1976), or in psychotherapy (Charny, 1966; Ramseyer & Tschacher, 2006, 2010). However, this thesis is interested by the possible effects the

verbal – linguistic – aspect of this phenomenon has on the outcome of interactions.

In her effort to unravel between the concepts of adaptation phenomenon in the literature of communication and psychology, Toma (2014) notably review Communication Accommodation Theory (CAT), Interaction Adaptation Theory (IAT), Interactional Synchrony, Motor Mimicry, and Linguistic Style Adaptation. She distinguishes those theories in terms of commonality and differences – to find areas of conceptual overlap – according to four criteria: type of behaviours referred to, the mechanism behind the behaviour, reception by the interlocutor, and the effect the behaviour has on the interaction. She notes:

> "Generally speaking, synchrony is postulated to lead to increased positive affect in interactions. However, because it is seen as a natural, built-in tendency, the effects of synchrony on interactions are generally not addressed. Rather, the absence of synchrony is described as distressing or negative, especially for infants." (p.164)

She also points out that much less research have been dedicated to linguistic adaptation than to vocal and non-vocal adaptation; which is still in need for refinements for two of the criteria she inspected: the mechanisms that are behind the behaviour, and the subject that directly concerns this thesis, the effects this linguistic adaptation have on interactants and the outcome of conversation.

A central notion is that if synchrony is a genuine phenomenon, then it should be more pronounced in real interaction than in pseudo-interactions (Ramseyer & Tschacher, 2010). Reidsma, Nijholt, Tschacher, & Ramseyer (2010) automatically calculate synchrony between speakers motion by using a time-lagged cross-correlation technique from Ramseyer & Tschacher (2006). In their method, they randomly shuffled interaction measures and compared them to the actual level of synchrony measure to assess whether the actual levels were higher than what would be considered by chance. An analogous Monte Carlo approach is adopted in this work, while applying this method to linguistic adaptation, using ten times randomised conversational content to compare to actual content (see detailed description in chapter 4). A similar approach of measuring actual versus one random conversational content is used by Healey et al. (2014), in the context of alignment in everyday conversations.

The notions of synchrony and linguistic adaptation are also closely related to a number

of other concepts: accommodation, convergence and entrainment. These terms are sometime used as synonyms as they have overlapping meaning and have been subject to a number of academic treatments and formulations, which require definition. The term "accommodation" is defined as "a multiply organized and contextually complex set of alternatives [...] (and at another level) characterize wholesale realignments of patterns of code or language selection, although again related to constellations of underlying beliefs, attitudes, and sociostructural conditions" (Giles, Coupland, & Coupland, 1991, p. 2), a description that encompasses surface and underlying representations of the world. "Convergence" is defined as "a strategy whereby individuals adapt to each other's communicative behaviors in terms of a wide range of linguistic-prosodic-nonverbal features" (Giles et al., 1991, p. 7), which distinguishes the notion as an adaptation between interactants by the use of multiple features. The term "entrainment" is defined as the "adjustment or moderation of behavior to coordinate or synchronize with another" (Bernieri, Reznick, & Rosenthal, 1988, p. 243). The extent to which those terms differ is still an object of discussion, however, one could notice the usage in these definitions of words from previously defined notions, such as "alignment" in the definition of "accommodation" and "synchrony" in "entrainment". The terms "synchrony" and "entrainment" have originally been characterized in the context of body movement and non-verbal behaviour and later applied to speech analysis, while the terms "alignment", "accommodation" and "convergence" appear more closely related to a linguistic characterization. The present work being interested in the linguistic aspect of the phenomenon, I will principally use the term *alignment* in future descriptions. I might also occasionally use the term *convergence*, to describe the opposite phenomenon to *divergence*. *Divergence* refers to the situations when interlocutors show explicit signs that they do not, intentionally or not, linguistically align with each other.

2.1.3 Alignment

Repetitions are viewed by a dense literature on human communication as the cues of a phenomenon of *alignment*, defined as a coordination occurring between interlocutors, to merge their representation of the world at multiple levels (Garrod & Anderson, 1987; H. P. Branigan, Pickering, & Cleland, 2000; Brennan & Hanna, 2009).

The *Interactive Alignment Model* (IAM) (Pickering & Garrod, 2004), postulates that the reason that makes dialogue one of the easiest forms of communication, is that its production and comprehension are coupled in a largely automatic and unconscious process that leads to the alignment of linguistic representations and situation model. According to the IAM, alignment is achieved through priming mechanisms, referring to the tendency to repeat a previously "primed" word or structure, that would not require extensive cognitive loads and is supposed rather automatic and resource-free. Within the alignment theory, the idea of *structural priming* (the term "structural" is chosen to detach the notion from specific structural theories that are attached to syntax) is central and suggested to ease cognitive processing, see Pickering & Ferreira (2008) for a comprehensive review. Alignment between partners would happen by "percolation" from higher levels of representation (pragmatic and semantic) to structural then lexical levels and vice versa. The IAM model relies on the principle of output/input coordination given by Garrod and Anderson (1987), where they observed that the players of a maze-game were inclined to repeat the semantic and pragmatic choices of the previous utterance given by their dialogue partners.

The tendency toward alignment has been observed at different levels, such as syntactic and lexical (H. P. Branigan et al., 2000; Reitter & Moore, 2007; Garrod & Anderson, 1987); however, it is still an open debate whether syntactic representation is independent from, partially dependent on, or dependent on, lexical representation. Other evidence is given for the existence of alignment in prosody, speech rate or phonetic realisations (Giles et al., 1991; Curl, 2005; Pardo, 2006; Xia, Levitan, & Hirschberg, 2014). Alignment is also viewed as an indicator of conversational engagement in interaction, a notion that is receiving considerable attention in the past decades and of which definition also varies depending on the fields (Glas & Pelachaud, 2015). In 1966, Goffman gave a definition for face-to-face engagement that is: "engagements comprise all those instances of two or more participants in a situation joining each other openly in maintaining a single focus of cognitive and visual attention – what is sensed as a single mutual activity, entailing preferential communication rights" (Goffman, 1996, p. 89). Engagement can be characterized as being at the crossing of *alignment* and *synchrony* as interlocutors engage using different linguistic features but also use non-verbal strategies involving social signals such as mimicry, gaze, facial expressions or gestures (Sun

& Nijholt, 2011; Jokinen, 2009) to construct mutual understanding throughout the conversation (Turnbull, 2003). This interactional engagement is argued as being at the root of linguistic mutual understanding (Gumperz, 1982).

Nonetheless, repetitions of linguistic choices, taken as the sign that alignment or linguistic synchronisation is occurring, lead with a number of other factors to engagement, provide on their own a basis for the quantification of mutual understanding in conversations. To understand how the apparently simple feature of repetition can be used to reach the more abstract concept of understanding, they must be placed within conversation.

2.2 Repetitions in Conversation Structure

Repetition plays a central role from infancy (Kadar, 1993) and remains pervasive through all language use (Ursi et al., 2018). As Deborah Tannen notes:

> "In summary, then, repetition is at the heart of language: in Hymes's (1982) terms, language structure; in Bolinger's (1961), language production; in Becker's, all languaging." (Tannen, 2007, p. 56)

This section is concerned with the functions, the types, and the structural aspects of repetitions that are considered of interest for quantification.

2.2.1 Functions of Repetitions

Tannen (2007) divides the functions of repetitions, that she also coins as "patterning", in conversation under four categories, namely, (1) **production**, (2) **comprehension**, (3) **connection** and (4) **interaction**:

1. Repetitions help to produce language with ease and fluency, they can signal high-involvement, and enable automatic behaviours. Having a set of prepared situation utterances allows more cognitive load to be allocated to meaning and variations.

2. Being the other side of the same coin, repetitions help to comprehend language as a repeated item carries less semantic information than an entirely new one. The effort needed to understand content decreases as there are more repetitions, the listener is

able to focus on the meaningful variations rather than thinking of the utterance wording.

"Turn to the left, then turn to the right"

vs.

"Turn to the left then take a right"

These two formulations carry essentially the same semantic content, yet the first will be more easily processed as the structure is repeated.

3. Repetitions' connective function allows a speaker to tie ideas together, and make reference of them easily accessible. Repeating the same sentence while changing one word in it interestingly both emphasizes the part that is repeated and the part that is different, an aspect that can be illustrated while revealing a number of functions. Consider the following example, extracted from an actual informal conversation of the author:

   " – I study repetition patterns.

   – You study repetition patterns?"

   Here, the repetition of the token[2] sequence *"study repetition patterns"* can be interpreted as having three functions: (1) it signals participatory listening by an almost exact repetition of the whole sentence, (2) the modification of pronoun and addition of the interrogative mark while keeping the same main clause is a request for a clarification, and (3) an effect of style is added as the content comically matches the meaning of the sentence.

4. And finally the interactional function of repetition combines the first three together as well as including: the act of getting and holding the floor, stalling, showing appreciation to an utterance or indicating to a third party who just joined the conversation what has been discussed.

I note here that while the notion that a repetition has a specific function for each given utterance is acknowledged, qualifying those occurrences is not the object of the present work.

[2] A *token* is a representation of a word as it is pronounced by a locutor, which does not necessarily correspond to its orthographic form. This designation can for example include truncated or elided forms (Ursi et al., 2018); and is more generally a string of characters between two spaces (or punctuation marks) that allows for its automatic treatment as a unit.

Those functions are here summarized as an indication of the tie between the situated events that are repetitions to the structure of conversation.

Repetitions have also been emphasized to **avoid miscommunication** (Cushing, 1994), in particular in task-based interaction where the certainty that an information is grounded is critical. The last positively viewed function is one that has been largely explored, in particular in Conversation Analysis, repetition can signal a misunderstanding and **induce repair** (Hutchby & Wooffitt, 2008; Heritage, 2009; Colman & Healey, 2011), notably through feedback mechanisms. Feedback can takes multiple forms and is a phenomenon that is central in communication (Sundberg Cerrato, 2007) and one of its form is verbal repetition (Loewen, 2012). The process of repair connects two feedback functions that seem at first contradictory: signaling understanding in certain cases, for example when used in association positive visual feedbacks such as head-nods, and signaling misunderstanding in other cases, for example when used in association with negative visual feedbacks such as frowns (Healey, Mills, Eshghi, & Howes, 2018). These two functions of feedback leads to repairs that help create local coordination that may leads ultimately to mutual understanding (Healey et al., 2018). Schegloff also links intersubjectivity, a notion close to mutual understanding, to the organisation of repair in conversation.

> "The achievement and maintenance of this sort of intersubjectivity[3] is not treated in a theoretically satisfactory manner by invoking socialization as a mechanism, for intersubjectivity is achieved for a virtually inexhaustable range of types of events always contextually specified [...]. The solution surely is provided for by a resource that is itself built into the fabric of social conduct, into the procedural infrastructure of interaction. [...] this involves a self-righting mechanism built in as an integral part of the organization of talk-in-interaction – what has been termed the organization of repair." (Schegloff, 1992, p. 1299)

This statement indicates the central importance given by one of the leading figures of Conversation Analysis, of the use of repair – which translates into repetition in certain cases – in the establishment of mutual understanding between interlocutors.

[3] Schegloff here restricts intersubjectivity to "particular bits of conduct that compose the warp and weft of ordinary social life" (p.1299) rather than intersected knowledges and beliefs between individuals.

Repetitions can also be viewed negatively, and thus for two reasons. They might be perceived as the sign of an immature (i.e. in the case of language acquisition) or unimaginative (i.e. inability to produce original clause) mind. Some even take the capacity of divergence, producing entirely new and different linguistic content, as the sign that an immature speaker (i.e. a child) is gaining mental maturity. The current method has no means to distinguish for this type of negative repetition from the positive functions described above. They can also be viewed as markers of disfluencies, and are categorised as such in many studies (Shriberg, 1996; Colman & Healey, 2011). In consequence of that second perspective, repetitions of this type should be eliminated by Automatic Speech Recognition systems to allow a better processing by Language Understanding components (notably improve labelling of Part Of Speech or other higher level of linguistic representations).[4] This last type, usually described as "self-repair", is often found within one speaker single contribution, and is therefore not relevant to the present work which is concerned with repetitions from one turn to the next, either by two different speakers or from the same speaker distinct contributions.

2.2.2 Repeating The Other or Repeating Oneself

In addition and in interaction with the functions described in the previous section, it is important to distinguish two types: **self-repetitions** and **other-repetitions**, repetition of oneself or of utterances of the interlocutor. Self-repetition can be the sign of individual communication patterns, such as personal preference for certain structures or lexicons over others, yet are still involved in a more general communication process. For example, self-repetitions can be an indication to the other speaker that they should change the focus of conversation (Stivers, 2004), or a strategy to hold the floor and gain planning time (Rieger, 2003).

Self-repetitions also reveal another aspect of human communication that has been shown to ease understanding and reduce cognitive load (Brennan and Clark, 1996): behavioural consistency. This notion of individual self-consistency was also highlighted by Fusaroli & Tylén (2016): they monitored individual patterns and took them as a control baseline to make sure that interpersonal coordination was not reducible to individual behaviour only. The mechanisms behind self-repetitions are multiple but this type of works show that they

[4]For example in the sentence: *"I think that that is not true."* the second *that* could be interpreted as a disfluency and therefore removed, no matter what were the intentions of the speaker.

are separated to a certain extent from the mechanisms behind other-repetitions while still being an important feature to take into account when observing interpersonal coordination.

The main repetition type that concerns alignment theory remains however other-repetitions, as they signal the most strikingly communicative behaviours influenced by conversational partners. Swerts, Koiso, Shimojima, & Katagiri (1998) distinguish two pragmatic functions of other-repetitions: "integration" i.e. a repetition to signal understanding, and "non-integration" i.e. a repetition to induce repair. Those functions may be considered through the scope of who holds information in an interaction. In many informal interactions, in particular the ones that do not have a specific aim else that creating social link, it could be considered that participants hold information by turns, while in specific tasks with pre-defined roles (such as map task interactions) this distinction can be maintained throughout the course of the dialogue.

In relation to other-repetitions, Reitter et al. proposed in 2010 a cognitive model of syntactic priming. As mentioned in section 2.1.3, priming effects are presumed to be at the root of the alignment phenomenon. This model postulates that when producing an utterance, the speaker is influenced by *recent* linguistic experience. The next section is interested in what *recent* entails.

2.2.3 Local Routine and Long Term Adaptation

One point of particular concern is defining the span of alignment under study. It is possible to consider two angles, in one case long-term accommodation (alignment between partners within a long time window, from the course of a whole dialogue, for which defining the length remains a challenge, to accommodation over days), and in another case short-term accommodation ("local" level alignment, repetitions of elements contained in previous utterances or within a short time window).

Over the past decade, Reitter, Moore, & Keller, while also scrutinising repetition patterns, obtained results that highlighted the effects of long-term accommodation (Reitter & Moore, 2007; Reitter et al., 2006; Reitter & Moore, 2014). They undertook studies relating a task success measure given in the HCRC Map Task corpus to the proportion of repetitions, with an emphasis on phrase-structure analysis, which is considered close to lexical indepen-

dence, in specific time windows. With their method, they did not find direct evidence that short-term priming effects, while being present, correlated overall with task success. However, they established a link between repetitions (long-term priming effects) and task success, and their observation of repetitions at both lexical (same token) and structural levels inspired the preliminary experiment conducted at the beginning of the chapter 5, an extension of previous work by Vogel & Behan (2012). This experiment assured the reproducibility of earlier studies that used only lexical level as a feature, and confirmed the existence of variations in results when taking into account a number of different linguistic level of representation (see Appendix A). However, the present thesis followed a different analytic approach than Reitter et al., testing to see whether repetitions happened above chance or not, as explained in detail in chapter 4. These differences in the method used, as well as the differences in results they produced, constitute a part of the contribution of this work.

Another perspective on dialogue structure that inspires this research comes from the concept of interpersonal synergy applied to conversation analysis (Fusaroli, Rączaszek-Leonardi, & Tylén, 2014), borrowing the idea from its previous characterisation in movement coordination (Riley, Richardson, Shockley, & Ramenzoni, 2011). This notion states that speakers become interdependent by relying on "local routines" that structure dialogues, in particular in the case of pre-determined complementary roles for each participant. This concept relates to the notion described by Pickering & Garrod (2004) in the Interactive Alignment Model that states that interlocutors also use "local routines" created and maintained during a dialogue to ease their speech production and comprehension.

In a study that compares the two approaches of interactive alignment and interpersonal strategies, in terms of their impact on task-success, Fusaroli & Tylén (2016) found that if the aspects they determined as relevant from interpersonal synergy seem to provide the best predictors of collective performance, the local structural organisation of task-oriented dialogue was crucial in their success. The "local" aspect of alignment is also given evidence of influence among social signals associated to transcribed speech, amidst which Beňuš, Levitan, & Hirschberg (2012) explored entrainment of acoustic features. They examined the repetitions of filled pauses (such as uh, ah, eh, and um) between lawyers in Supreme Court hearings and Justices, in relation with the favorability of the Justice Vote. They found that the occurrence

of above chance adjacent filled pauses (local alignment) between a lawyer and a Justice related to more favourable decision of the Justice for the case being discussed. Whereas when observing over-all dialogue filled pauses amounts, not at a local level, no relation with favorability was found, a finding that supports the hypothesis that short-term accommodation have a positive effect on communication.

There are therefore, in the described above studies, two different views present in the literature of task-oriented interactions that are difficult to conciliate a priori. One view gives evidence of alignment existing at a local level (short-term accommodation) and having an impact on communicative success, while the other, concerned with lexical and structural priming effects, argue that alignment detected at local level does not influence success. Retaining parts of both views, the distinctive approach developed here investigates whether that patterns of repetitions can be observed at local level that relate to communicative success in specific tasks. More specifically, the present work observes if an overall correlation between local alignment and success can be found or not, but also propose an extended investigation that takes into account a number of extra-linguistic features and their possible impact to local alignment on communicative success.

2.3 Quantification of Communication Phenomena

To identify which elements in conversations could be taken as reliable indexes of mutual understanding it is necessary to first assess if a phenomenon is quantifiable and by which features, the context in which it is relevant, and the areas where it might be useful and desirable to do so.

2.3.1 Quantification in Conversation Analysis

Conversation Analysis traditionally builds the description of interactional behaviours on collections of data fragments in specific settings (Albert, 2017), and practitioners overwhelmingly tend to avoid quantification (Turnbull, 2003, p. 215).

In that regard, a series of cautionary remarks were issued by Schegloff (1993) on the validity of quantification in conversation study in an essay to clarify challenges that he found

were left inexplicit. According to him, quantitative analysis is first built on single instance analysis. His first remark is on *significance* that he reminds not being the "only way of establishing relevance" (p.101). In the matters of speech interactions in sociology, the study of single instances is a rather common practice that permitted to highlight a number of phenomena. If one still wants to quantify in *talk-in-interaction*, three criteria that needs to be met (p.103):

1. "environments of possible relevant occurrence"

2. "set of types of occurrences whose presence should count as events"

3. "an analytically defensible notion of the domain or universe being characterized"

He argues that some studies that, for example, try to assess a degree of sociability by counting 'laughter per minutes', cannot fulfil the above enunciated conditions, as the *positioning* of laughter in the conversation will have a great deal to do with the appropriateness of the behaviour. Schegloff insists that "what is to be counted [needs to be] analytically relevant because it is organizationally related to it in the conduct of interaction" (p.104). He points out the wild variety a laughter may have, and the failure to occur in "environments of possible relevant occurrence" may be as relevant as its occurrence. This point is evocative of the duality that repetitions may have, that they might signal understanding just as much as misunderstanding. To him, it is furthermore the aspect 'per minute' that breaks the link the feature could have with the analytical notion of laughter. The presence or absence is meaningful in conversation as much as where the event happens.

The present study counts repetition from the previous turn. The relevance of repetition to the structure of conversation is kept, and it is precisely this relevance that is scrutinised in the randomisation of turns that is applied and subsequently quantified. There is indeed an exception to the conversational features he considers inappropriate for quantification: "Unlike the earlier discussed practices of reference to persons, it appears that this domain of practices of *talk-in-interaction* – other-initiation of repair and its sequelae – can be 'qualified' for quantitative treatment." (p.115) Once the elements deemed fit for quantification are identified, their specific contexts of analysis and the measures against which they are to be compared must also be selected; aspects on which the next section focuses.

2.3.2 Communicative Success Measures

Patterns of repetitions are currently said to differ substantially in casual everyday conversations and task-based dialogues. Healey and colleagues argue against the generalisation of the theory of alignment to any type of communication and in particular of structural priming effects existing in dissociation with lexical priming effects, and provide cases where people repeat each other less than expected in ordinary conversations (Healey, Purver, & Howes, 2014; Howes, Healey, & Purver, 2010). Evidence that supports the alignment theory mainly comes from interactions where participants perform a task, often in laboratories settings. Closed-class words (typically function words, such as pronouns, determiners, conjunctions and prepositions), that are used in the expression of specific syntactic structures, are argued to be the actual prime in the repetition of syntactic structures. Within these task-based interactions, even if their results suggest the existence of the abstract phrase structures, confirmation of the existence of structural priming are mostly given for specific constructions, such as active/passive forms, prepositional object/double objects constructions, or dative alternations (Pickering & Ferreira, 2008; Bock, 1989; Tree & Meijer, 1999). There are exceptions that I review below.

I chose to focus on task-based dialogues for the development of this proxy measure of mutual understanding, mainly because such dialogues have an independent notion of success available to them, which enables the operationalisation of the notion of mutual understanding, although I do not focus on a specific grammatical structure. Measuring interactional success on its own is not trivial. Cappella (1991) distinguishes four approaches: coding, rating, participant judgement, and observer judgement. Coding refers to the assignation of a value to an interactional segment, rating is similar but the values are assigned on a scale for each segment; both coding and rating can be carried out automatically or by trained humans. Judgement approaches do not necessarily require training but have to be done by humans, acting as judges, either judging their own behaviours or someone else behaviours. Cappella notes that each provides different frames of reference, but that instead of searching for a privileged frame of reference, research should be focused on "transforming the results from one frame of reference to another by developing mappings from the more objective measurement frames to those represented by participants and observer judgements" (Cappella,

1991, p. 111). From this statement, one could consider that the verification of the existence of matching patterns from one communication setting to another is desirable, while taking into consideration how the changes in settings could affect the patterns. How much was understood by each participant of a casual conversation remains dependent on subjective interpretation. However when a specific task that requires coordination is given, the accuracy of completion of the task gives a starting point to assess the quality of the interaction at hand in a more objective manner.

In his approach to quantifying mutual understanding, which he defines as "the process whereby interlocutors satisfy themselves that the intended meaning is being conveyed and understood." (Colman, 2012, p. 19), Colman assumes that problematic dialogues are the ones in which participants signal issues and attempt to correct them, and therefore uses repair as a negative index of mutual understanding. His approach however, relies on the full manual annotation of the corpus. This allows him to distinguish different types of repetitions, with great refinement for each category, but also requires substantial human training, time, and efforts, for each corpus analysed.

In a wide range of domains using statistics, metrics are used as tools of comparison, between different variables of interest. In the Natural Language Processing field a great number of specific custom-made measures have been created, however some typical evaluation metrics can be quoted such as Precision, Recall and F-measure for systems output, word-error rate or BLEU scores that are more often associated with Machine Translation. Some well-defined measures of similarity can be cited when trying to quantify interactional measures: Euclidean distance, Cosine Distance, or Jaccard Distance. Concerning more specifically mutual understanding, some also mention the use of length of instalments, the use of pronouns, or ellipsis and anaphora (Colman, Eshghi, & Healey, 2008).

Aware of the multiplicity of forms that linguistic behaviours can adopt depending on context and social factors, this thesis is constrained to the analysis of task-based interactions, that are described for each corpus used in section 4.5. Despite being considered a limitation by some researchers (Healey et al., 2014; Duran, Dale, & Galati, 2016), this type of interaction is still not fully understood and does not provide uniform patterns across the literature. In addition, task-based conversations are a type of interaction in which the development of

tools for the detection of mutual understanding is pertinent for possible applications, notably dialogue systems, that remain mostly task-based in their usage.

2.3.3 Spoken Dialogue Systems

Although dialogue systems are not in focus in the present work, one of the aims is to participate in the deciphering of task-based human-human interactions to ultimately inform dialogue systems builders, either for dialogue modelling or dialogue management (D. Traum, 2017); notably as some models already partly include the management of repetitions inspired by grounding and alignment theories. This section situates very briefly the current state of dialogue systems and the challenges they face that makes the study of interactions from a computational linguistic perspective relevant to their development.

Traditional spoken dialogue systems are built in modules; notably employing Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialogue Management (DM), Natural Language Generation (NLG) and Text-to-Speech synthesis (TTS), to cite a few. To model a progression in dialogue viewed as "natural", many approaches have been explored, starting with simple scripts following sequences of possible actions. Following those, local structures systems, that may include more appropriate feedback, and global structures systems that allow the monitoring of high-level dialogue elements, have been progressively implemented. However those were limited by a lack of flexibility if the user deviates too far from the possible options developers thought of. Unexpected answers from a user immediately bring up the limits of the system. Plan-based approaches, inspired by Speech Act Theory (Austin, 1962; Searle, 1969) brought more dynamic systems, oriented to users' intentions. Recent advances in statistical approaches (machine learning models), such as sequence-to-sequence (seq2seq) or partially observable Markov decision process (POMDP), are starting to produce results that bring meaningful responses from conversational agents and allow more flexibility in users' behaviours (Young, Gašić, Thomson, & Williams, 2013; Serban, Lowe, Henderson, Charlin, & Pineau, 2015; Serban, Lowe, Charlin, & Pineau, 2016; Zhang et al., 2018; Han & Gmytrasiewicz, 2019).

A model to monitor real human-human dialogues that was inspired by the grounding theory described in section 2.1.1, is the Degree of Grounding Model proposed by Roque &

Traum (2008). They established a set of evidence of understanding that defines a degree of grounding of each utterance. An important component of the model is repetition, even if the feature is used inside a higher level evidence set (namely: Submit, Repeat back, Resubmit, Acknowledge, Request Repair, Move On, Use, Lack of Response). It was applied on a task-based corpus of radio military training. They interestingly note a double function of repetition in relation to mutual understanding, that: "[...] in this domain [military] Resubmit evidence generally indicates lack of understanding; in general conversation, it is not true that the repeated mention of material indicates that it is not understood [...]" (Roque & Traum, 2008, p. 61). In 2006, Varges reported the implementation of a spoken dialogue system following an 'overgeneration and ranking' approach that took into account alignment and variation phenomena (Varges, 2006). In that system, the final candidate for generation followed alignment and variations scores. *Alignment* was defined as the repetition of "bag-of-words" unigrams and bigrams, while *variation* refereed to alternative sets of realisations (for continuation queries). Dialogue systems should be modelled with alternatives, to avoid a 'boredom effect' that could lead to a disengagement form the user (Cushing, 1994). However, the formula chosen for this system gave more weight to alignment than to variations, while acknowledging that deriving the exact weight to give should come from empirical corpus data, which was the direction to take for future work.

From Eliza (Davis, 2001) and ALICE (Wallace, 2009), to the ones made accessible by technology giants such as Apple *Siri* or Amazon *Alexa*, considerable improvements have been made. However, achieving naturalness in dialogue is still considered a long term goal and is far from the staple of science fiction of an agent displaying human-like abilities, either only by a voice as seen in the movie "Her" (Jonze, 2013) or by corporal humanoid artificial intelligence as in the episode of the British series *Black Mirror* "Be Right Back" (Brooker & Harris, 2013).

Dialogue systems are improved by the integration of human data analysis (D. Traum, 2017), to which the present work aspires. According to Ward & Devault (2015), one of ten challenges that highly-interactive dialogue systems currently face is that developers and social science researchers are still not communicating enough. "The behaviors in today's dialog systems are seldom based on the findings of social scientists, and conversely, the

results of dialog system research are rarely noticed by them" (Ward & Devault, 2015, p. 106). A way to achieve this would be by a increase on behavioural descriptions "specific enough to use by dialog systems". This supports the stance that research at the crossing of social and computer science is needed to achieve more robust and adaptable systems.

2.4 Factors Influencing Conversation Structure

Discriminating between the factors that influence conversation structure, and subsequently its success, is not mundane. The HCRC Map Task (Anderson, Bader, et al., 1991), from the inception of the map-task method (Brown, Anderson, Shillcock, & Yule, 1985), provides a context in which task accomplishment is not possible without some levels of linguistic success in communication. In addition to a task score provided by the authors, this corpus contains controlled non-linguistic features, such as defined task roles, gender, eye-contact, familiarity and task practice, that might influence communication and therefore patterns of repetitions.

2.4.1 Non-Linguistic Sociological Features

H. Branigan, Lickley, & McKelvieDavid (1999) examined the influence of non-linguistic factors on disfluencies in the HCRC Map Task. They hand-labelled respectively: repetitions, deletions, insertions and substitutions (disfluencies) and reparandum words (discard); and calculated rates for each of these categories by counting the number of words labelled as disfluent per 100 words intended. Their method for coding was concerned with within-contribution disfluencies, also termed self-repair, which differ from the present work (see chapter 4). With their measures, they found repetitions accounting for 30 to 60 percent of all disfluencies and described a series of detailed patterns:

- Female participants are overall less disfluent than males in eye-contact conditions, without a significant influence of the gender of the addressee.
- Overall, task roles have a considerable influence: Information Givers are more disfluent than Followers, which is not only considered attributable to their tendency to produce longer utterances, but rather to a more complex conceptual processing.

- A higher discard rate for Information Giver in familiar pairs, while no significance was found for familiarity on disfluency rate.
- The no eye-contact condition induces significantly more within-turn self-repetitions.
- A higher rate of within-turn self-repetitions at first attempt for Information Givers – interpreted as a lower conceptual planning for subsequent attempts.

They concluded that non-linguistic factors of disfluencies are not uniform in their influence on the interaction. The deduction of the impact that one factor can have on disfluencies might lead to an over-simplification, and should rather be interpreted as a complex interaction between factors. Therefore, interaction between factors can also be expected when observing turn to turn repetitions in interaction with task success, which is the aim of this thesis.

Colman & Healey (2011) showed that repair mechanisms are approximately double in task-oriented dialogues than in everyday conversations and pointed out that "two different task roles are associated with divergent patterns of repair" (Colman & Healey, 2011, p. 1567). A corpus study using the HCRC Map Task, where the topic remains constant (at a level of granularity associated with dialogue task as opposed to unstructured conversation), also found gender-based differences in the distribution of specific linguistic items, in particular back-channels signalling involvement that men would be less likely to produce than women (H.-J. Schmid, 2015).

In his formulation of mutual adaptation, a notion closely related to synchrony and mutual understanding, while focusing on interpersonal communication, Cappella states that an understanding of communication patterns can only be made by taking into account the "association between patterns of message interchange between partners and the partners' experienced state of the relationship" (Cappella, 1991, p.103). From this stance, whether the relationship between partners is long established or two interactants are meeting for the first time, in other terms their degree of familiarity is going to have a strong impact on a given exchange, and one can expect to observe a high degree of alignment between unfamiliar partners as they try to engage with each other. That familiarity between participants is a high influencing factor of task successful completion is nothing new. Indeed, this point can be illustrated by the NASA Apollo missions procedure that was that in case a member of the prime crew would become unable to fly for some reason, the entire back-up crew would

replace them rather than just that individual.[5] Even if both crews were highly trained to do the same precise procedures (Phinney, 2015), is was theoretically considered that crew familiarity with each other and knowledge of personality and reactions was a factor crucial to mission success. This principle was famously by-passed by the replacement of the Lieutenant Commander Thomas K. Mattingly by Jack Swigert as Command Module Pilot in the Apollo 13 mission. Indeed, as the Mission Operations Report mentions, "A vigorous simulation program was successfully completed prior to launch to ensure that Lovell, Swigert, and Haise could function with unquestioned teamwork through even the most arduous and time-critical simulated emergency conditions" (p. 4),[6] even though they could only have 2 or 3 days of training together, according to Charles M. Duke.[7] This example only shows the idea that familiarity within a crew or team might in some cases be considered above task familiarity.

In Clark's communication model, it is the amount of personal common ground that distinguish friends from strangers. He illustrates (Clark, 1996, p. 115) acquaintedness with four degrees:

- strangers (no personal common ground)
- acquaintances (limited personal common ground)
- friends (extensive personal common ground)
- intimates (extensive personal common ground, including private information).

As people get acquainted, they build an interpersonal lexicon that eases communication, a process that unfamiliar participant have to add to the task itself when collaborating to achieve the said task. In the subsequently described studies, two degrees of acquaintance are present and can be distinguished in terms of repetition patterns: strangers and friends (it

[5]From: Apollo 13 Prelaunch Mission Operations Report, 1970-03-31 (p.41) https://history.nasa.gov/afj/ap13fj/pdf/a13-prelaunch-mission-ops-report-19700331.pdf — Last Accessed 10.05.2020

[6]From: Apollo 13 Postlaunch Mission Operations Report, 1970-04-28. https://history.nasa.gov/afj/ap13fj/pdf/a13-postlaunch-mission-ops-report-19700428.pdf — Last Accessed 10.05.2020

[7]From: NASA Johnson Space Center Oral History Project, Edited Oral History Transcript. Charles M. Duke, Jr. Interviewed by Doug Ward Houston, Texas – 12 March 1999 https://historycollection.jsc.nasa.gov/JSCHistoryPortal/history/oral_histories/DukeCM/DukeCM_3-12-99.htm — Last Accessed 10.05.2020

is not possible to determine from the information found in the corpora if the speakers are intimates).

2.4.2 Computer-mediated Interactions

The validation of methodological choices in other contexts and using different measures of interactional success is of primary concern in order to confirm that patterns of communication are not unique to one particular interactional setting. O'Malley, Langton, Anderson, Doherty-Sneddon, & Bruce (1996, p. 177) for instance, stated that "when speakers are not physically co-present, they are less confident in general that they have mutual understanding [...], and therefore over-compensate by increasing the level of both verbal and non-verbal information". This finding gives another aspect to explore, a different setting where alignment, and therefore repetition patterns, might change: computer-mediated interactions.

Previous experiments have also found that communication style used by the subjects during task-oriented computer-mediated communication differs substantially from direct face-to-face communication. For example, dialogue acts distribution differs, with backchannel utterances (acknowledging understanding) reduced significantly in computer-mediated interlingual communication (Hayakawa, Luz, & Campbell, 2016), and participants use a more concise, specified style of communication not echoed in human face-to-face interactions (Newlands, Anderson, & Mullin, 2003). Another study examined alignment in machine-translated communication, yet in a de-contextualized setting (not two humans trying to achieve a joint task), a Wizard-of-Oz experiment where participants were asked to answer supposedly machine-translated questions (Schneider & Luz, 2011). Half of the questions contained translation mistakes resembling ones a Machine Translation system would produce. The authors interpret their results by arguing that people align their answer and reproduce the obvious errors (translation mistakes), assuming that a speech-to-speech machine translation system would understand them better.

This behaviour echoes the results found in studies about the alignment process with a virtual agent that reported evidence of exaggerated alignment when the speakers thought they were talking to a machine (H. P. Branigan, Pickering, Pearson, & McLean, 2010; Dubuisson Duplessis, Clavel, & Landragin, 2017). How talking to robots of various size and shapes,

one robot or multiple robots (Saito et al., 2018), impact communication and alignment is currently extensively studied alongside with the development of companions or learning robots (such as the well-known Pepper from SoftBank Robotics). Those results suggest an increase in alignment with agents and robots are for now related to beliefs about their linguistic abilities (H. P. Branigan, Pickering, Pearson, McLean, & Brown, 2011).

2.4.3 Human-mediated Interactions

Another dimension of successful communication is joint idea formulation and consensus building, aspects that were possible to investigate in the present work through third party assessment. The presence of an interaction facilitator might introduce a different dynamic to patterns of repetition between a dyad of interactants. Mediation could have no impact, or bring novel patterns impacting successful communication.

The construction of common ground might not require the same amount of feedback or communicative cues signalling understanding, at the beginning of the interaction, while as the conversation becomes longer, confirmation that communication had been effective might become more important. One could make the hypothesis that the amount of cues that is necessary to suppose mutual understanding might not be the same at the beginning of a conversation as it is at the end, which gives yet another aspect to explore.

2.4.4 Across Language Variations

What is considered appropriate to say in a particular situation is deeply rooted in the socio-cultural context. Time and location, even within the same language, in our case English, can have a determinant impact on appropriateness. One could imagine a conversation having in appearance the same topic, being discussed vastly differently in a 17th-century literature salon in Paris and in a 2019 “Meet up” event organised in Dublin, Ireland. Yet, both will use English, and both will contain repeated expressions from each participants background. Expatriates have all noticed the phenomenon when going back to their home country after an extended period of time: some of the everyday expressions in fashion have changed. Some collocations are less used, if one uses them she/he might seem out-dated, and some new expressions have appeared. By repeating the expressions new to her/him, the expatriate will

at the same time show adaptation and contribute to the enforcement of the expressions as appropriate language items. The local aspect of language is well-known (Pennycook, 2010), as well as the within communities language.

Each new learner of a second language start by learning lists of vocabulary but also collocations, and the ability to repeat appropriately expressions are taken as the sign of a higher command of a language as a social act. What determines that a particular set of collocations and expressions is being part of a certain community has been extensively studied in linguistics and remains a central aspect of Human Computer Interactions (Cowan et al., 2019). The set of expressions that are perceived as part of "American English" or "Irish English" is not fully clear-cut, but those sets exist as part of the local aspect of language (Pennycook, 2010), without any judgement of value among those sets. It is only part of the descriptive effort that is at the root of understanding. Another aspect is how much repetition must be used depending on the group under study. Tannen points out that the frequency of usage of patterned expressions also varies among cultures: " 'Among the Ibo[8] the art of conversation is regarded very highly, and proverbs are the palm-oil with which words are eaten.' (*Things fall apart*, 1958) [...] Americans, in contrast, are inclined to regard relatively fixed expressions with suspicion and are likely to speak with scorn of cliches, assuming that sincerity is associated with novelty of expression and fixity with insincerity" (Tannen, 2007, p. 51). From this quote, we may expect a variation in the relative frequency of repetitions that is considered appropriate depending on the population observed. Other studies also suggest the existence of priming between two different languages, which suggest that the existence of purely lexical priming is less likely, but rather the existence of a higher level of structural representation in the language production (Hartsuiker, Pickering, & Veltkamp, 2004; Loebell & Bock, 2003).

2.5 Conclusion

To conclude this review, the varieties of aspects that still remain unknown within the distribution of repetition counts in various settings as well as the disparities of the patterns described in the literature in both methodological aspects and unity of findings opens up a range of pos-

[8]The Ibo (also spelled Igbo) are a meta-ethnicity native to present-day Nigeria.

sibilities to investigate in the quantification of mutual understanding. The aspects possible to study in that framework which can be controlled by experimental settings and available for research shaped in part the smaller research questions presented in Figure 1.1 and first introduced in section 1.1 in the choice of materials that the next section describes. These factors, namely role, gender, eye-contact, familiarity, task experience, computer-mediated settings, human-mediated settings, and dialects, are the reasons why the following presented materials were chosen.

Chapter 3

Materials

3.1 Introduction

This chapter presents the materials used to carry out the analyses presented in this thesis. Five corpora are used, the HCRC Map Task (Anderson, Bader, et al., 1991), the ILMT-s2s corpus (Hayakawa, Luz, Cerrato, & Campbell, 2016), the MULTISIMO corpus (Koutsombogera & Vogel, 2018), the MIT American English Map Task (AEMT) corpus, and the PARDO 2006 Map Task corpus (Pardo, 2006). The motivation for choosing these materials is that all corpora were based upon task-based dyadic interactions, with each having a measure of communicative success. The measure was either available by the nature of the task, or the corpus had an element that permitted creating one. Namely, the three face-to-face Map Task corpora, the HCRC, the AEMT and the PARDO 2006, have a measure called a *deviation score* (described in more detail below), the ILMT-s2s corpus was labelled for cognitive states of participants, and labels describing the feedback given by an interaction facilitator were created for the MULTISIMO corpus. Table 3.1[1] show a summary of the used corpora by the number of tokens, turns, duration,[2] number of dialogues and language. All these corpora use variations of English except for the ILMT-s2s that had half of its participants speaking Portuguese. For the ILMT-s2s, the AEMT and the PARDO 2006, as they were created with the intent to follow the HCRC Map Task technique, subsets of the HCRC Map Task were

[1]The HCRC (Sub1) was extracted for using the same maps as the ILMT-s2s corpus, the HCRC (Sub2) was extracted to correspond to the eye-contact, female, familiar conditions of the AEMT, and the HCRC (Sub3) was extracted to correspond to the no eye-contact, female, unfamiliar conditions of the PARDO.

[2]The duration refers to the size of the recordings.

extracted following similar conditions, to allow a direct comparison. I chose to present in this document the material used before the methods as they are in some cases informed by features of the corpora used in the experiments.

Table 3.1: Summary of all used Corpora. The term "Mixed" refers to a mix of native and non-native English speakers from various backgrounds. The grey sections are not kept in the total as they are subsets of the HCRC Map Task used as comparison.

Corpus	**Tokens**	**Turns**	**Duration**	**Dialogues**	**Language**
HCRC (Full)	160,697	27,069	14:24:19	128	English (Scottish)
ILMT (Eng)	13,761	2,310	09:38:57	15	English (Mixed)
ILMT (Por)	12,671	2,236	09:38:57	15	Portuguese (Brazil)
HCRC (Sub1)	22,106	3,790	02:02:22	16	English (Scottish)
MULTISIMO	35,928	11,511	03:45:45	23	English (Irish/Mixed)
AEMT	24,878	4,750	02:17:49	16	English (US)
HCRC (Sub2)	27,700	5,041	02:12:56	18	English (Scottish)
PARDO	17,767	2,490	01:43:56	10	English (US)
HCRC (Sub3)	15,220	2,551	01:29:18	14	English (Scottish)
Total	**265,702**	**50,366**	**31:54:02**	**192**	

3.2 The HCRC Map Task Corpus

The Human Communication Research Centre (HCRC) Map Task corpus consists of 128 dialogues released in 1992 (Anderson, Bader, et al., 1991). This corpus uses the map task technique (described below) to elicit spontaneous communicative behaviours in the frame of Human-to-Human task-based interactions. Two subjects per dialogue, with either the role of Information Giver (IG) or Information Follower (IF), were each given A3 maps containing landmarks. Almost all participants were native Scottish speakers of English. The IG had a route drawn on the map with a START and a FINISH, and was tasked with guiding the IF through a map containing only landmarks. To add to the difficulty of the task, landmarks from the two maps and their placement differed a little. For example, some landmarks were present on the IG maps but not on the IF maps, or present in both but labelled with different names.

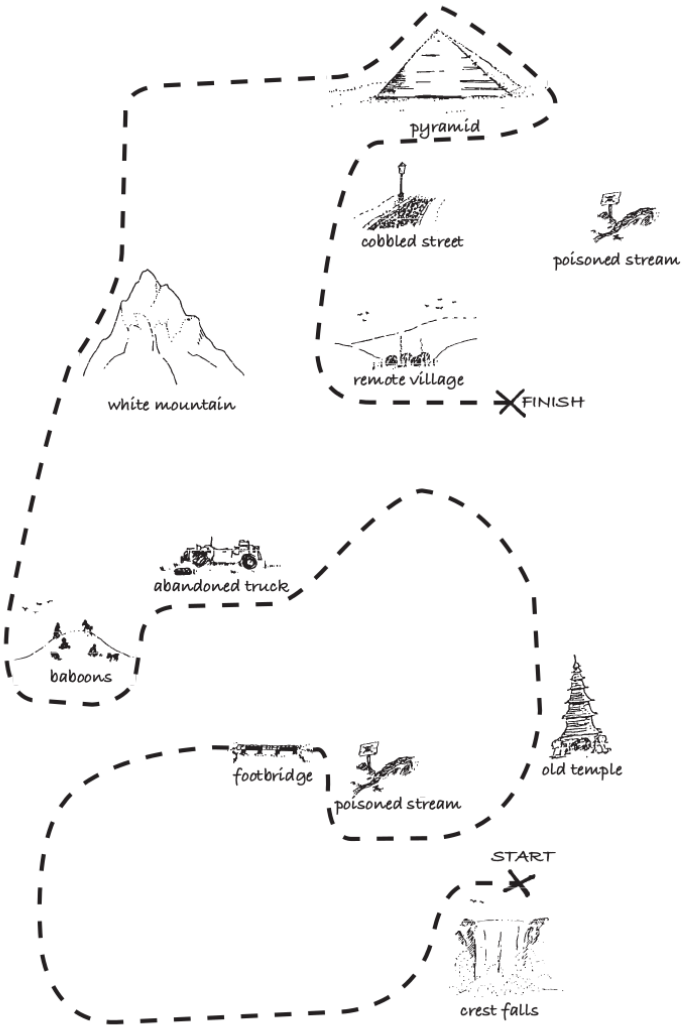


Figure 3.1: Information Giver's Map number 1 of the HCRC Map Task

The 128 recordings have been divided into “quads” of two pairs of participants, in which each participant did the task four times, twice as IG and twice as IF. Participants were recruited by pairs of Familiar partners that knew each other well and matched with another pair of participants they did not know prior to the recording. Figure 3.2 represents those quad divisions (16 quads in total, with half having eye-contact while the others did not, variable that was used to do 8 grouping) in which each dialogue is given a five symbol code: "q" for “quad” followed by the quad number, then "ec" or "nc" respectively meaning "eye-contact" and "no eye-contact" followed by the dialogue numbers identification within each quad.

These divisions allowed the control of three variables susceptible to impact speech variation: eye-contact, gender and familiarity between participants. Figure 3.3 divides the corpus according to these variables to give a visual representation of the number of dialogues involved in each category. This representation highlights that some categories, such as Eye-contact/Male only/Familiar or No eye-contact/Female only/Familiar contain a high number of dialogues (18 in each) while others contain many fewer (only four dialogues in No eye-contact/Male only/Unfamiliar and Eye-contact/Female Only/Unfamiliar). I mention here this situation naturally deriving from the condition divisions as it will subsequently impact the possibilities for meaningful comparisons made in the following experiment chapters.

In the division concerned with the visual cues accessible, half the subjects were able to see their interlocutor’s face (i.e., Eye-contact), while the other half had opaque screens placed between them (i.e., No eye-contact). The subjects could not see their interlocutor’s map at any point. A summary of the number of tokens and number of turns is given in Table 3.2. The IF used on average 393.31 tokens per dialogue and the IG 858.10. The participants were 64 in total, with 32 females and 32 males [Reported Gender], that would participate in the task four times, twice as IG and twice as IF, and in each role once with a familiar partner and once with an unfamiliar one.

Table 3.2: HCRC Map Task Summary of tokens and turns per conditions

	IG	**IF**	**Fam**	**UnFam**	**Female**	**Male**	**Eye**	**NoEye**
Tokens	110,075	50,622	89,047	73,647	66,519	66,043	74,034	88,660
Turns	15,203	11,866	15,357	12,077	11,792	10,475	12,412	15,022

The task consists in verbal guidance and the participants were told not to use gestures. The IG has to guide the IF along a predefined route, and any *deviations* from that route were assumed to be the result of less successful communication between the two participants, as the subjects were precisely told not to stray from the route.

The *deviation scores* are described in section 4.5.1. The precomputed HCRC Map Task corpus *deviation score*, ranges from 4 (best) to 227 (worst). The reconstructed *deviation score* ranges from 7 to 329 for the method retained for usage (see § 4.5.3, which describes the scores reconstruction by the author of this document from descriptions given in the HCRC corpus documentation). The higher the score, the more the route deviates from the original route, which is taken as an indication of less successful communication. The relative objectivity behind this reasoning makes the *deviation score* a particularly good fit to measure success in the frame of this research's methodological development, mostly because it shapes communicative success into a numerical scale. This aspect is discussed in more details in section 4.5.

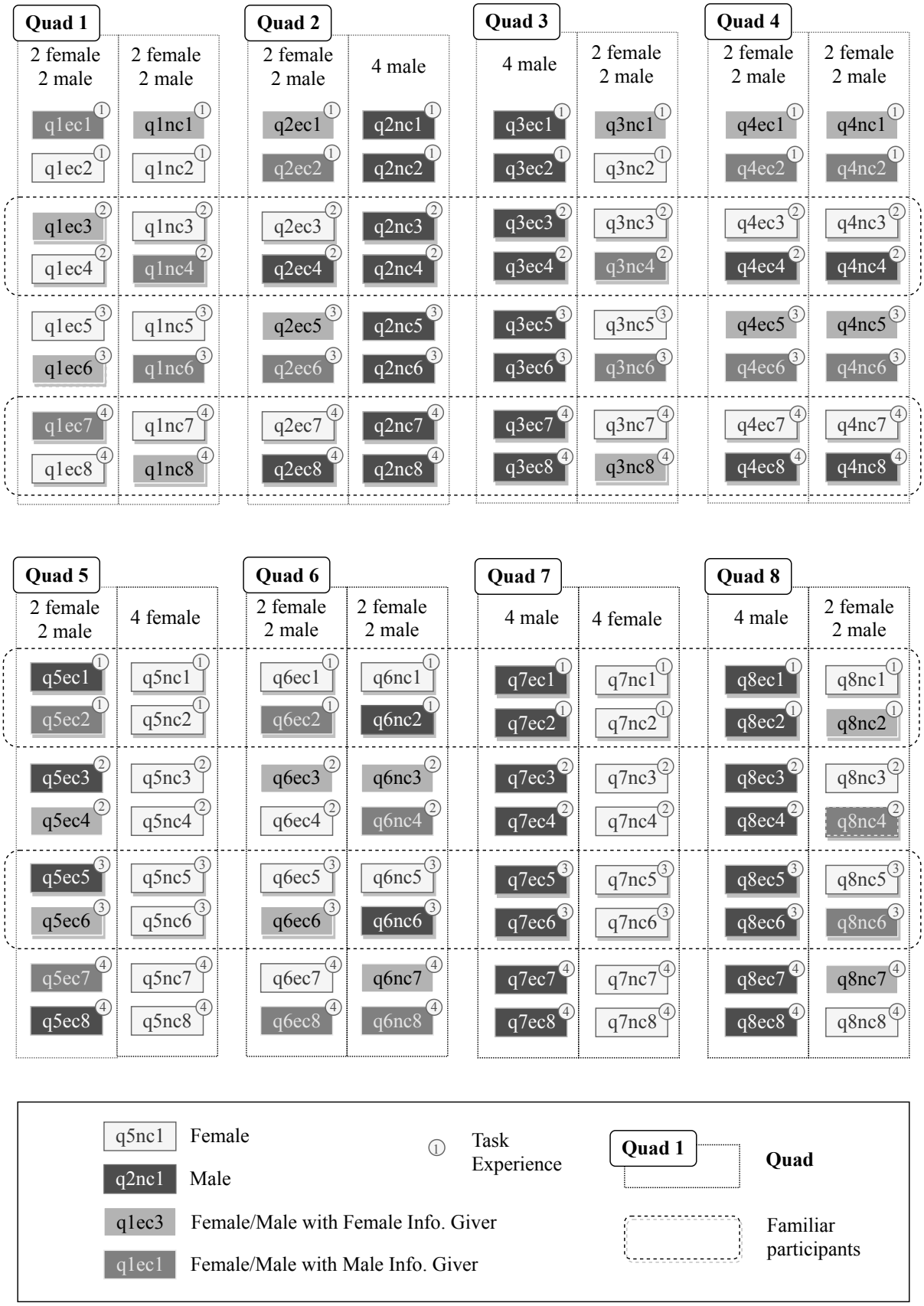


Figure 3.2: Summary of the HCRC Map Task dialogues by Quads, each dialogue is represented within its quad and the conditions are given: Eye-contact, Gender and Familiarity (UnFam : Unfamiliar ; Fam : Familiar), and Task Experience.

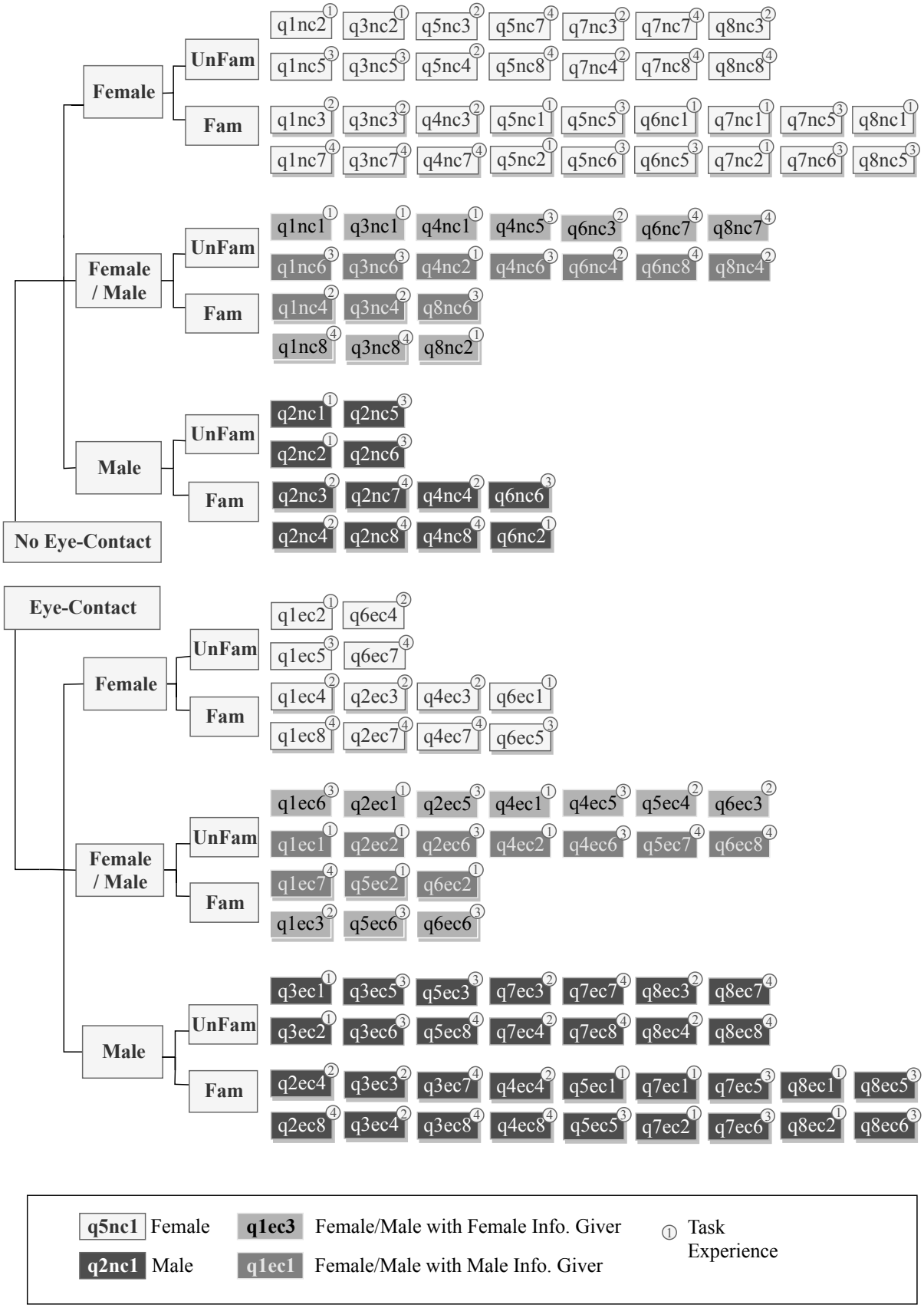


Figure 3.3: Summary of the HCRC Map Task dialogues by Conditions, each dialogue is represented within the dividing categories: Eye-contact, Gender and Familiarity (UnFam : Unfamiliar ; Fam : Familiar).

3.3 The ILMT-s2s Corpus

The Interlingual Machine Translation Speech-to-Speech (ILMT-s2s) corpus (Hayakawa, Luz, Cerrato, & Campbell, 2016) consists of fifteen dialogues between fifteen English and fifteen Portuguese subjects speaking to each other as pairs in their native language via a Speech-to-Speech Machine Translation system (S2S-MT) – the ILMT-s2s System. As with the HCRC Map Task corpus, the dialogues use the map task technique to elicit spontaneous communication, but with a difference that the subjects are located in different rooms, speak different languages to each other and communicate via a Speech-to-Speech Machine Translation system (S2S-MT) system. The maps that are used are two maps taken from the HCRC Map Task corpus, in their original version for the English speakers, and translated for the Portuguese speaking subjects.

The ILMT-s2s System is a system that uses off-the-shelf components — Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech synthesis (TTS) — to perform S2S-MT. It is activated by a "Push-to-talk" button that the subject will click-and-hold for the duration of the utterance and release once the subject has finished. Neither subject can hear the other's voice, since the output of the ASR/MT is provided by a synthetic voice. The subjects (aged 18 – 45) were recruited from the Trinity College Dublin digital noticeboard and personal connections of the authors of the corpus. Fifteen recordings of fifteen native English speakers (5 female, 10 male), and fifteen native Portuguese speakers (11 female, 4 male), were collected. The corpus was annotated by two trained students for cognitive states, namely: Frustration, Amusement, and Surprise, for each speaker in all the dialogues. The inter-coder agreement for the labels was calculated[3] and the results are well above 0.6. It is to be noted that Amusement and Surprise are considered negative here along with Frustration as these two perceived cognitive states were a reaction to high word error rate utterances output, and high amounts of the three states grouped also matched dissatisfaction with the translation system according to the user's survey (Hayakawa, Vogel, Luz, & Campbell, 2017).

In this section the choice was made to relate mutual understanding, and more precisely its possible lack, with the presence of high amounts of negative cognitive state for each speaker.

[3]Using the modified kappa feature of ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) version 4.9.0's "Inter-Annotator Reliability…" function.

The hypothesis is that if there is high amounts of negative cognitive state in the task setting, this correlates negatively with mutual misunderstanding, and therefore lack of repetitions.

In the direct human-to-human dialogues of the HCRC Map Task corpus of the previous study, the count was carried out between the utterances of the two human subjects. However, for the interlingual computer mediated dialogues of the ILMT-s2s corpus, the count was carried out within the same language — the utterances from the English speakers are coupled with the English translations (human transcriptions) of the Portuguese speakers utterances and vice-versa, which created two fully monolingual dialogues. To compare the face-to-face and computer-mediated settings, the HCRC Subset 1 was created, corresponding to the same map used to create the ILMT-s2s corpus. The HCRC Subset 1 contain 16 dialogues. A summary of the number of tokens and the number of turns is given in Table 3.3.

Data of the English dialogues of the ILMT-s2s corpus were labelled with the TreeTagger English training set (H. Schmid, 1994), while the Portuguese dialogues of the ILMT-s2s corpus were labelled using the TreeTagger tag-set proposed by Pablo Gamallo (Gamallo & Garcia, 2013).

Table 3.3: HCRC Map Task Subset 1 and s2s-ILMT Corpus Summary of tokens and turns.

	HCRC Subset 1			ILMT-s2s					
Language	English			English			Portuguese		
	IG	**IF**	**Total**	**IG**	**IF**	**Total**	**IG**	**IF**	**Total**
Tokens	14,903	6,666	21,569	16,042	11,480	27,522	13,690	11,652	25,342
Turns	2,117	1,540	3,675	2,626	1,994	27,522	2,476	1,996	4,472

3.4 The MULTISIMO Corpus

The newly created multimodal MULTISIMO corpus (Koutsombogera & Vogel, 2018), consist of 23 dialogues of collaborative group interactions, where two players work together to provide answers to a quiz. The two participants are guided by a facilitator, who monitors their progress and provides feedback when needed. Although being the only corpus of this body of work not being a map task, this corpus is still derived from task-based interactions and it has features that can be used as indicators of mutual understanding and task-based success. This corpus was also chosen to explore the potential of the method for communicative success detection outside of map task interactions.

The task of the players was to converse with each other with the aim of estimating and agreeing on the 3 most popular answers to each of 3 questions, and rank their answers from the most to the least popular. An independent survey of a sample of 100 people was used as reference for correctness of the answer and ranking. The overall task was to guess the popularity of answer rather than exactitude, as in the television show *Family Feud*. The corpus consists of synchronised audio and video recordings, and its overall duration is approximately 4 hours, with an average session duration of 10 minutes. The average age of the participants is 30 years old and their gender is balanced (25 female, 24 male). The language used is English, with one third of the participants being native speakers of English, with thirteen Irish, two British and one American. The other two thirds represent fifteen nationalities of fluent speakers of English [4] all living in the region of Dublin, Ireland.

Each group consists of 3 members, who collaborate with each other in a quiz: 2 players and 1 facilitator. Out of the 49 corpus participants, 3 were designated as facilitators, and 46 were assigned the role of players and were randomly paired in 23 groups. Facilitators coordinated those discussions, i.e. provided the instructions of the game and confirmed participants' answers, but also assisted participants throughout the session and encouraged them to collaborate. Facilitators were briefly trained before the session recordings, i.e. they were given the quiz questions and answers and they were instructed to monitor the flow of the discussion.

The facilitator role is critical in the setup design, considering that it is a role to be mod-

[4] Greek:13, French:4, Brazilian:2, Indian:2, Pakistani:2, Chinese:1, Croatian:1, Egyptian:1, Italian:1, German:1, Kazakh:1, Mexican:1, Romanian:1, Slovenian:1, Thai:1

elled for an embodied conversational agent that would coordinate group interaction and would help participants achieve their goals. In this respect, the facilitator role was designed to enable the extraction of behavioural cues for the development of an agent responsible for managing the interaction and choosing actions that maximise the collaboration effort and the performance of the group participants.

All corpus sessions were fully transcribed by 2 annotators using `Transcriber`.[5]

The transcription consists of the segmentation of the audio signal into speaker turns, the transcription of speech, and the segmentation of the dialogue into 11 sections, i.e. introduction, question 1, 2 and 3, categorisation of each question in 2 parts (namely answering phase and ranking phase), and closing. Transcripts were then imported into the `ELAN` annotation editor,[6] so that all the information recorded in the transcript was visible and further editable.

For the purpose of the present study, the introductory and closing parts of each session were disregarded, and I focus on the following section types:

- *Full*: consisting of the 3 questions of the quiz as a whole (23 sections in total)
- *Question*: each of the 3 questions, cutting *Full* into 3 parts (69 sections in total)
- *Answer*: the answering phase within each *Question* (69 sections in total)
- *Ranking*: the ranking phase within each *Question* (69 sections in total)

The *Full* section embeds the 3 *Questions*; the *Question* embeds the *Answer* and *Ranking* phases, while those last two are mutually exclusive. All the facilitator's turns occurring during the question-answer sequences were further annotated for their feedback type. To this end, 2 annotation layers were introduced: the first annotation layer includes the values of *positive*, *neutral* and *negative* feedback. Table 3.4 presents the mean and median duration of each section type as well as the number of turns per feedback type, the turn mean and median duration, and the mean and median number of tokens encountered in each *feedback type*. The term *feedback* in this study refers to attitudinal, behavioural, and linguistic reactions of the interaction facilitators to the participant's contributions and behaviours in the game.

[5] `http://trans.sourceforge.net/` — Last Accessed 11.05.2020
[6] `https://archive.mpi.nl/tla/elan` — Last Accessed 11.05.2020

Table 3.5 goes further into details and presents the mean and median number of turns and tokens of each section type, as well as the maximum and minimum values for the number of turns and tokens of each section type. Feedback values are further refined at a secondary level, that is, feedback subtypes. The annotation values of feedback type and subtype are listed in Table 3.6, together with a brief description and an example for each value from the corpus.

Table 3.4: Section type mean (μ) and median (M) duration (in minutes); and number (n) of turns, turn mean (μ) and median (M) duration (in minutes) and mean (μ) and median (M) number of tokens per feedback type

Sections			**Feedback**					
Section type	**Duration (μ)**	**Duration (M)**	**Feedback Type**	**Turns (n)**	**Turn Length (μ)**	**Turn Length (M)**	**Words (μ)**	**Words (M)**
Full	8.50	8.51	Positive	1062	1.01	0.48	141	108
Questions	2.59	2.45	Negative	360	1.24	1.06	69	70
Answer	1.58	1.31	Neutral	1154	1.40	1.16	391	298
Ranking	1.20	0.47						

Table 3.5: Section type mean (μ) and median (M) number of turns per section, mean (μ) and median (M) number of tokens per section, and Maximum (Max.) and minimum (Min.) number of turns and tokens per section.

Section type	**Turns number (μ)**	**Turns number (M)**	**Tokens number (μ)**	**Tokens number (M)**	**Turns Max.**	**Turns Min.**	**Tokens Max.**	**Tokens Min.**
Full	357.26	340	1054.9	990	758	193	2313	523
Questions	119.04	102	351.29	307	289	43	898	86
Answer	76.61	58	213.01	154	258	24	838	41
Ranking	42.46	34	138.72	94	119	11	499	16

The facilitator's feedback was coded by one annotator, and annotations were edited for validity and consistency issues by a second annotator. The annotation task resulted in 2576 annotations, and their distribution per feedback type is presented in Table 3.7. The distribution is detailed per Question section (Q1, 2, 3), per answer [A] and per ranking [R] phase.

The most frequent value is that of *neutral* feedback, indicating that the facilitator often intervenes to help the participants by providing hints and examples. Almost equal to the number of the *neutral* values are the occurrences of *positive* feedback, implying that the facilitator not only confirms the correct answers, but also has a positive disposition towards participants, aiming at their successful results. This positive disposition created delicate

Table 3.6: Annotation values for the Facilitator's Feedback Type and Subtype.

Type	Subtype	Subtype Description	Subtype Example
Positive	General	Positive feedback that the participants are doing well	*Great job, well done!*
	Confirmation	Confirms the correctness of the answers	*That was the right ranking.*
Negative	General	Negative feedback while they discuss possible answers	*It doesn't have to do with food.*
	Disconfirmation	Disconfirms replies that are not correct	*Unfortunately you didn't get this one.*
Neutral	Elaboration	Provides helping cues	*It is related to food, but think of a different category of food.*
	Feedback elicitation	Poses direct or indirect questions	*Is that your final decision?*
	Topic change	Manages the sequence of questions	*Now let's move on to the second question.*

Table 3.7: Distribution of feedback type values in section types Full, Question (Q), Answer, Ranking

	Positive		**Negative**		**Neutral**	
	n	%	n	%	n	%
Full	1062	41	360	14	1154	45
Q1	380	36	89	25	402	35
Q2	345	32	123	34	377	33
Q3	337	32	148	41	375	32
Answer	766	72	205	57	838	73
Ranking	296	28	155	43	316	27

cases to annotate, and the annotators often exploited multimodal information to disambiguate certain instances, that is, the speech prosody and facial expressions of the facilitator. For example, cases such as "They are all very good answers but they're not the popular answers." were considered as negative feedback, even if the facilitator's words are positive in the first clause, because the audio and visual information indicated otherwise. Moreover, there is no significant difference in the quantity of expressed feedback among the three questions. However, it seems that the majority of feedback responses for all positive, negative and neutral types is occurring in the answering phase, where players need to identify the 3 most popular answers. The annotations were performed by the author of the present document and verified by the author of the corpus.

Table 3.8: MULTISIMO Summary per Conditions. The Facilitator speech is taken out of the count for gender. There are 10 female/ male, 6 female only and 7 male only dialogues.

	Participants	**Facilitator**	**Female**	**Male**
Tokens	24,244	11,684	12,488	11,756
Turns	8,217	3,294	4,323	3,889

Table 3.9: MULTISIMO Summary per Dialogue Sections.

	Full	**Q1**	**Q2**	**Q3**	**Answer**	**Ranking**
Tokens	24,244	9,341	7,244	7,654	14,698	9,539
Turns	8,217	3,013	2,425	2,776	5,286	2,930

3.5 The MIT American English Map Task Corpus

The Massachusetts Institute of Technology (MIT) American English Map Task (AEMT) was recorded in 1999, in Speech Communication Group of the Research Laboratory of Electronics, by Olga Goubanova.

Following the HCRC Map Task method (Anderson, Bader, et al., 1991), eight female native American English speakers participated in the recordings, seven from the Boston area and one from California, all young recent college graduates. The participants were highly familiar to each others, described as close friends (LaVoie, 2002). The dialogues were transcribed by a graduate Computer Science and Language student, Eunice Oreoluwa Fasan, and verified by the author of the present thesis. The corpus consist of 16 task-based conversations, two group of four participants recorded each eight map task dialogues.

As in the HCRC Map Task, an Information Giver (IG) was tasked to give instruction to an Information Follower (IF) to draw a route on a map containing landmarks. The task was made more difficult by slight mismatches in the set of landmarks and the names given to them, between the the IG maps and the IF maps. An opaque screen prevented the participants from seeing each other, emulating the no eye-contact condition of the HCRC Map Task. Participants were encouraged to speak freely and cooperate.

1 of the 8 sets of maps created for the HCRC Map Task was used. The recordings are available under the Creative Commons license.[7] A *deviation score* between the IG and IF routes was calculated for each map following the counting methods described in section 4.5, ranging from 2 (best) and 322 (worst). To compare the American English and Scottish English dialects repetition patterns, the HCRC Subset 2 was created, corresponding to the no eye-contact, female, and familiar conditions. The HCRC Subset 2 contain 18 dialogues. A summary of the number of tokens and turns for the two corpora is given in Table 3.10.

Table 3.10: American English Map Task Summary.

	AEMT			HCRC Subset 2		
	IG	**IF**	**Total**	**IG**	**IF**	**Total**
Tokens	17,208	7,670	24,878	18,415	9,285	27,700
Turns	2,812	1,938	4,750	2,830	2,211	5,014

[7] `https://dspace.mit.edu/handle/1721.1/32533` — last accessed 01.08.2019

3.6 The PARDO 2006 Map Task Corpus

The PARDO 2006 Map Task corpus (PARDO) was recorded in 1998, at Yale University, New Haven, United States, by Jennifer Pardo (Pardo, 2006).[8]

This corpus originally consisted of 30 dialogues between unfamiliar to each other participants (half female, half male) in no eye-contact conditions using the HCRC Map Task technique (Anderson, Bader, et al., 1991). The participants were native American English speakers, undergraduate students at the Yale University. None exhibited a strong regional accent, even if originating from various regions of the United States.

Each participant did the task five times, the Information Giver and Information Follower's role staying the same for each conversation. The original landmarks drawn on the HCRC Map Task's maps were used on 8.5 by 11-inches sheets with adjustments made in the landmarks labels to correspond to American rather than Scottish English pronunciation. Spoken samples of the landmarks were originally collected before and after the map task session to observe the degree of phonetic convergence.

In order to facilitate the comparison with the female-only AEMT corpus, the 10 female-only dialogues were transcribed by the author of this document from audio files to be used in the present study. [9]

To also compare the American English and Scottish English dialect repetition patterns, the HCRC Subset 3 was created, corresponding the no eye-contact, female, and unfamiliar conditions. The HCRC Subset 3 contain 14 dialogues. A summary of the number of tokens and turns for the two corpora is given in Table 3.11.

Table 3.11: PARDO 2006 Map Task Corpus Summary.

	PARDO			HCRC Subset 3		
	IG	**IF**	**Total**	**IG**	**IF**	**Total**
Tokens	11,451	6,316	17,767	11,149	4,071	15,220
Turns	1,289	1,201	2,490	1,438	1,113	2,551

[8] I am grateful to Professor Jennifer Pardo for the granting me access to this data.

[9] A score of performance, similar to the *deviation score*, was calculated by the author of the corpus, for each map by superimposing a 1 by 1 centimetre grid over the Information Giver's map and the Information Follower's corresponding map. This score is expressed as the proportion of total centimetre squares of IF's path drawing that overlap with original path. The scores range from 0.96 (best) to 0.65 (worst) in the transcribed dialogues. These scores were eventually not used in the study, they are mentioned as they were originally intended to be used.

3.7 Conclusion

This chapter described five task-based corpora used for discourse analysis independently recorded by different institutions. Each corpus allows for the exploration of a different dimension of repetition patterns in relation with communicative success measures. The reference corpus is in almost all the following experiments the HCRC Map Task that gives a point of comparison. The exception is the three-party interaction experiment involving the MULTISIMO corpus, that lets us explore a differently structured type of task-based interaction.

None of the material presented here have been originally created to specifically measure mutual understanding, which helps to avoid the bias associated with the experimenter effect: that is the influence the experimenters might have on the participants as they unconsciously know what their desired outcome is (Rosenthal, 1963; Harris & Rosenthal, 1985). The only annotation process that was applied directly to conduct one of the below described experiments concerns the MULTISIMO corpus' feedback annotations; the corpus in itself was recorded prior to its usage in the present thesis, without the possibility to foresee this usage. However, as designed by the author of this document and the corpus author, the focus of the annotations are on the positive, negative or neutral valency of the feedback given by the facilitator rather than directly coding for understanding, which is aimed to avoid that possible confound, that yet must be mentioned here.

Chapter 4

The Methods

4.1 Introduction

This chapter presents the methods that are applied to analyse the dialogue contents in chapter 5. The assessment of a degree of alignment and its relation to communicative success implies a series of constructs and assumptions about the nature of dialogue and interactions. The methods presented evolved with usage, and adjustments necessary to fit corpora specificity's were also made, while the source materials are transcripts from dialogues. The base method was first described by Vogel & Behan (2012), and also used in (Vogel, 2013). This method is based on word frequency and designed to inspect interactional content by measuring repetitions in actual dialogues in contrast with randomised versions of those, arguing that repetitions may be randomly distributed in discourse, and if not, be the expression of communicative patterns linked to mutual understanding. In this thesis, this method is extended to multiple levels of linguistic representations and applied in task-based interactions where a crucial element is present: a measure of task success. It is hypothesised that the combination of the classification of repetitions as happening *above* or *under chance* when between interlocutors (defined in this work as phenomenon of alignment/convergence and divergence) and task success measures of those interactions allow for the quantification of a degree of mutual understanding. The methods are applied on the five distinct datasets: the HCRC Map Task that is the main corpus studied, both for its fitness to the subject and its volume, and four other corpora: the ILMT-s2s, the MULTISIMO corpus, the AEMT corpus and the PARDO corpus, described in chapter 3. A preliminary experiment using the Table

Talk corpus is reported in Appendix A, mainly performed as a comparison with previous uses and indications of usefulness of the methods' extension to multiple linguistic levels of representation. The five datasets are used to vary both the context of usage and the type of assessment of successful communication. This chapter starts by describing the base method followed by its previous uses, then describes the extension which represents its current applicable form. A summary step-by-step guide of how the method is applied in practice can be found in Appendix B.

4.2 Base Method

The base method (Vogel & Behan, 2012; Vogel, 2013) consists of counting repetitions in dialogue transcripts. A *repetition* is here defined as a token of a contribution repeated in the immediately preceding contribution of each participants. The *contributions* of each speakers, also referred to as *turns*, are following the segmentation that was applied by the authors of each corpora described in chapter 3. The original method explored temporal overlap, which was chosen not to be developed here. A count is made of each repetition of a token uttered by another participant (other-repetitions or OTHERSHARED), and for each repetition of a token uttered by the same participant (self-repetitions or SELFSHARED). Sequences of tokens, known as *n*-grams, are recorded as counts at lengths $n = 1$ to $n = 5$. For example, a repetition of a sequence of five tokens would be recorded as five repetitions at the *n*-gram level $n = 1$ and one repetition at the *n*-gram level $n = 5$. This definition of repetition analytically differs from studies concerned with priming effects (Healey et al., 2014; Reitter et al., 2006), and does not includes within contribution repetitions. Once the count is made in the actual dialogue, each contribution is indexed and randomly shuffled within each dialogue, and a count is made again in the randomly re-ordered dialogues. Those re-orderings and countings are made ten times, to observe if a significant contrast emerges between the actual dialogues and the shuffled ones in repetition counts. For each dialogue a first data-preparing step aimed at normalizing the pronouns (i, me, you, us, we, your, my, our, mine, yours, ours) to the token "IY", was applied. This step aims at capturing repetitions in dialogue contributions in which the structural dynamic of complementary first-person and second-person personal pronouns is occurring.

(1) A: You got that?

B: I got that.

(2) A: IY got that?

B: IY got that.

In example (1) the sequence of two tokens *got that* is recorded as two repetitions at the *n*-gram level $n = 1$ and one repetition at the *n*-gram level $n = 2$, and in example (2) as three repetitions at the *n*-gram level $n = 1$, two repetition at the *n*-gram level $n = 2$ and one repetition at the *n*-gram level $n = 3$. Punctuation was disregarded. The focus is on the proportion of the total number of *n*-grams that could have been shared but were not (NON-OTHERSHARED, NON-SELFSHARED) and the ones that were shared (OTHERSHARED, SELFSHARED), both in actual and randomised dialogues.

4.3 Previous Uses

4.3.1 Casual Talks and Air Traffic Crisis

The basis of the method's relation with mutual understanding may appear counter-intuitive: in a given interaction, language use does not lead to mutual understanding unless proven otherwise. This statement may seem odd at first. After all, the ability to express complex idea and concepts through language is argued to be strongly interwoven with our evolution as a species, as it is seen as a major adaptive advantage (Malle, 2002). However, this same complexity that forms its strength and richness also create a major potential for misunderstandings: the countless possible ambiguities.

Results from the previous analysis by Vogel & Behan (2012), that are the foundation of this work, of the Table Talk corpus (Campbell, 2009) – unscripted casual conversations among five participants in English over three days – showed a significantly higher proportion of repetitions on *Actual* than on *Randomised* dialogues. This constitutes the first step in the potential of the method to distinguish meaningful repetitions from randomly occurring repetitions.

This study also examined the Air Traffic communication corpus, a transcript of the dialogue between the US Airways Flight 1549 captain, his first officer and other members of

the crew that were all assimilated as one participant, during the landing of January 15, 2009 on the Hudson River.[1] From this corpus which was chosen because of the expectation of extensive repetitions, results showed more OTHERSHARED repetitions in the *Actual* than in *Randomised* dialogues, although not reaching statistical significance, while less SELF-SHARED, that were however significantly higher in *Actual* for the Captain and First officer. In the discussion, the results observed lead to the remark that the importance of social role in the conversation matter just as much as individual personality in task-based interactions, while noting that communication during a crisis event might differ from air traffic communication in general.

4.3.2 Forensic Interrogations: Negative Evidence

Evidence toward the robustness of the method was explored for another type of interactions: courtroom interrogations – a legal context in which the assessment of understanding necessitates great caution – and a political television interview (Vogel, 2013). It was necessary to test the method on dialogues that are known to display naturally occurring failure in communication. Three dialogues were examined: Jeremy Paxman's 1997 interview of Michael Howard, the former UK Home Secretary, and two courtroom transcripts: People v. Herrero and State v. Cunningham. This study was focused on the idea that if no evidence of synchrony is detected, mutual understanding remains undetermined, while the higher the level of synchrony that can be established, the higher the chance that mutual understanding is taking place, without claiming certainty. In the political interview, both participants were native English speakers, therefore the apparent lack of understanding was supposed to be somewhat intentional avoidance. In People v. Herrero, the specific short "yes/no" answers given by the defendant for whom English was not the mother tongue, even though coherent within the dialogue, questioned both his involvement and understanding of the trial content. In State v. Cunningham, it was the ability of a jury member to understand the concept of presumed innocence and predisposition toward a particular verdict that was questioned. The analysis of these three dialogues concluded in the failure to reject the null hypothesis tested: the proportion of shared vs. non-shared repetitions contained in the actual dialogues did not

[1]The plane had lost all engine power after hitting a flock of Canadian geese, and the crew successfully managed the landing with no human loss.

exceed or were found equal to the proportion of shared vs. non-shared repetitions in the randomised dialogues. In other terms, there was no evidence that mutual understanding – significant level of linguistic engagement – was reached in those interactions.

These two studies exhibited the potential usefulness of determining if a significant amount of repetitions happened in actual conversation over their randomised counterparts, and suggested its validity to be used as a proxy for mutual understanding, while using a subjective assessment of communication.

4.4 Extended Method

The method was extended with the aim of exploring more levels of linguistic representation than the tokens transcribed in the dialogue (lemmas and part-of-speech labels were considered alongside tokens, and sequences thereof) and doing so in the context of task-based interaction corpora in order to relate repetition effects to successful communication that could be a sign of mutual understanding. It was extended for two reasons: to discern the scope to which different linguistic levels of representation provide information reliably as indicators of synchrony within the frame of the method, and to observe to what extent success in communication, through definite measures, is associated with repetitions patterns.

The first characteristic of the extension consists of a pre-processing labelling designed to measure five linguistic types of repetitions (referred to as 'Levels'): Token (which was the unit previously analysed), Lemma, Part-Of-Speech (POS), and a combination of Token with POS and Lemma with POS. I labelled the corpora transcripts with the decision-tree based parser TreeTagger as trained for English (H. Schmid, 1994), which was used to keep consistency in labelling for each corpora.[2]

For each dialogue, proportions of repetitions were extracted, per Dialogue type (*Actual* versus *Randomised*), per speakers (depending on the corpus, having a specific role), per *n*-grams (All *n*-grams [up to length 5]; N1: $n = 1$ [length 1]; N2+: $n > 1$ [length 2 to 5]), per type of sharing (OTHERSHARED and SELFSHARED),[3] and per Level:[4] TOKEN (Level

[2]It is acknowledged that the tagger was not trained on speech data specifically. The complete lists of tags can be found at `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/` — last accessed 01.02.2021

[3]The extraction by type (dialogue, *n*-grams and sharing) was already present in previous uses.

[4]The ordering of the levels is arbitrary.

1), LEMMA (Level 2), LEMMA+POS (Level 3), POS (Level 4), TOKEN+POS (Level 5). An example of three dialogue contributions is given in Table 4.1, at each linguistic level, with a highlighted OTHERSHARED repetition of same token for a length of n-grams $n=2$ (*straight up*), and a highlighted SELFSHARED repetition of lemma for a length of n-grams $n=1$ (*curve*). Other repetitions are present on this example but the highlight is made to emphasise the difference between OTHERSHARED and SELFSHARED repetitions and between token and lemma repetitions.

Table 4.1: Extract from dialogue q4ec6 of the HCRC Map Task (IF: Information Follower, IG: Information Giver).

Level 1	(Token)
IG	then you go **straight up** and **curve** over the top of the disused monastery
IF	**straight up** right
IG	have you over [...] **curved** over the monastery
Level 2	(Lemma)
IG	then you go **straight up** and **curve** over the top of the disused monastery
IF	**straight up** right
IG	have you over [...] **curve** over the monastery
Level 3	(Lemma+Part-Of-Speech)
IG	then+RB you+PP go+VV **straight+RB up+RP** and+CC **curve+NN** over+RP the+DT top+JJ of+IN the+DT disused+JJ monastery+NN
IF	**straight+RB up+RP** right+RB
IG	have+VH you+PP over+RP [...] **curve+NN** over+RP the+DT monastery+NN
Level 4	(Part-Of-Speech)
IG	RB PP VV **RB RP** CC **NN** RP DT JJ IN DT JJ NN
IF	**RB RP** RB
IG	VH PP RP [...] **VVN** RP DT NN
Level 5	(Token+Part-Of-Speech)
IG	then+RB you+PP go+VV **straight+RB up+RP** and+CC **curve+NN** over+RP the+DT top+JJ of+IN the+DT disused+JJ monastery+NN
IF	**straight+RB up+RP** right+RB
IG	have+VH you+PP over+RP [...] **curved+VVN** over+RP the+DT monastery+NN

Comparing the proportions of repetitions of Tokens vs. Lemmas is of particular interest, as one might expect a different distribution of repetitions for this conventional representation of lexemes. Indeed, as the lemma represent the canonical form – uninflected form – of a set of semantically related words,[5] it is possible that the repetition effect captured at this level might be related to a repetition of meaning, but this remains at a surface level. Tannen (2007) distinguishes instances of repetitions along a *continuum*, a scale of fixed forms. This scale goes from exact word repetition (same word used) to paraphrasing an idea (equivalent meaning with different words). Repetition of lemma can be viewed as midway on that scale, which allows capturing variations in repetition types. While it might not be considered as a method designed to look at syntactic repetitions per se, the POS labelling allows us to observe two different form of repetitions; lexical repetitions for N1: $n = 1$, and structural repetitions for N2+: $n > 1$ in combination with Level 4 (POS).

[5]For example, "curve" is the lemma of the inflected form "curved" in Table 4.1

4.4.1 Statistical Modelling

The model, originally described by Vogel (2013), is built to determine whether a significantly higher proportion of repetition appears in the actual dialogues than in their randomised counterpart, or in other terms, the aim is to observe possible contrasts between the two types of dialogue ordering. The proportion of interest is the number of linguistic items that were shared compared to the ones that were not shared. This section describes the generalized linear model with a binomial error family computed using R (Version 3.5.1) (R Core Team, 2018), followed by a single-step Tukey Honest Significant Difference (HSD) multiple comparison test using the package ***multcomp*** (Bretz, Hothorn, & Westfall, 2016). This model allows us to observe the contrasts between the Actual and Randomised dialogues as well as taking into account the speakers and level of linguistic representations described in section 4.4. The following hypothesis was tested for each dialogue:

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

$$H_1 : Random.Speaker.Level - Actual.Speaker.Level < 0$$

This H_0 null hypothesis states that the proportion of shared repetitions in the randomised dialogues should equal or exceed the proportion of shared of repetitions in the actual dialogues if repetitions are simply due to chance. The alternative (H_1) hypothesis states that the actual repetitions exceed the proportion of shared random repetitions, which is interpreted as the repetitions having a role in communication. This hypothesis was tested at three levels of n-grams granularity: N: n-gram= all ($1 \leq n \leq 5$), N1: $n = 1$ (lexical level), and N2+: $n > 1$ (phrasal level).

Logistic Regression Model

Generalized linear models (GLM) are a unification of both linear and nonlinear regression models that can be used for response variables that do not follow the Normal Gaussian distribution (Montgomery, Peck, & Vining, 2012), but only need to be a member of the exponential distribution family (normal, Poisson, binomial, exponential and gamma). Logistic regression models are used to investigate count data expressed as proportions (Crawley, 2005).

In this case a response variable y is modelled as function of x in the simple linear model:

$$y = \beta_0 + \beta_1 x \tag{4.1}$$

Where β_0 is the intercept and β_1 is the slope. The response variable y is assigned two possible values: either no linguistic items were repeated from the previous utterance (absence) or one or more repetitions were present (presence). This binary variable has a binomial distribution where the probability of one realisation is given by: $p^x(1-p)^{n-x}$. Here p is the parameter that describes the probability of x successes out of n attempts. The logistic model for p as a function of x is:

$$p = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}} \tag{4.2}$$

Taking q as the probability of failure out of n attempts, the way of linearizing this logistic function is by applying a simple transformation, that is substituting the probability p by the **odds** p/q, which is:

$$\frac{p}{q} = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}} \left[1 - \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}\right]^{-1} = e^{(\beta_0+\beta_1 x)} \tag{4.3}$$

As $ln(e^x) = x$, this equation simplifies to give a linear predictor for the logit transformation of p:

$$ln\left(\frac{p}{q}\right) = \beta_0 + \beta_1 x \tag{4.4}$$

The logit $ln\left(\frac{p}{q}\right)$ is the link function that relates the value of p to the linear predictor. The explanatory variable x is created by three two-factor categorical variables in interaction, *DialogType* (*Actual* or *Randomised*), *Speaker*, and *Level* of linguistic representation:

$$x = DialogType^{*}Speaker^{*}Level$$

We are here only interested in the interaction of those variables, in order to establish contrasts, and the individual effects of each variable are disregarded in the final model.

Hypothesis Testing – Tukey Contrasts

The regression model $y = \beta_0 + \beta_1 DialogType^* Speaker^* Level$ allows us to perform a pairwise comparison Tukey's Honest Significant Difference (HSD) test.[6] The Tukey's HSD test is used to make all pairwise comparisons. The test is one-sided to the left, as the contrasts of interest are when *Actual* exceeds *Randomised*.[7] For each test the different factors create a large number of possible combinations that need to be tested independently, which makes a single-step procedure appropriate, as it assures this independence. The widely used alpha threshold of $\alpha = 0.05$ was adopted for the rejection of the null hypothesis. At this stage, it is crucial to be aware that statistical significance does not equal practical importance and of the intense debate among statisticians for the last decades surrounding the links between scientific and statistical inferences (Krantz, 1999; Gelman & Stern, 2006). The controversy surrounding the misuse of *p*-values among scientists and the risk of making overconfident claims when the results of a test are incorrectly interpreted is also taken into consideration (Amrhein, Greenland, & McShane, 2019; Briggs, 2012).

4.4.2 Meta-Analysis

The procedure that consists in summarising the results of independently tested hypothesis has been called a meta-analysis for more than four decades (Glass, 1976), even if the topic was already among the interests of the statisticians Ronald Fisher and Karl Pearson in the 1930s. The original aim of this method, that are combined tests, is to integrate the findings of different studies having the same hypothesis to obtain an overall summary, and therefore have a stronger claim in drawing general conclusions. This type of procedure is all the more important in domains of behavioural sciences, in which this work aspires to be included as it is at the crossing of computer sciences and linguistics. The common definition of a meta-analysis is to group studies from diverse origins testing a similar research question. The procedure that I use in the experiments below relate to that definition as all results of a series of Post-Hoc single-step Tukey tests are combined, even if not coming from various origins with different experimental designs, as is common practice. Nonetheless, the procedure is

[6] The appropriateness of the use of this frequentist method is justified by modelling the sequences of speech production as sequence of repeatable event (occurrence of linguistic item repetition).

[7] *p*-value adjustments are incorporated in the Tukey test present in the ***multcomp*** package.

comparable, in the sense that each dialogue analysed individually at different levels of granularities, then the results of those tests are combined, which result in an analysis of analyses (Glass, 1976), that synthesizes the results of independent tests of the same hypothesis (Wolf, 1986). A different combination of methods was used. Initially the number of rejections of the null hypothesis in each category compared to the number of tests made, or rate of rejection, is observed.[8] Then I characterise the given interaction (dialogue or dialogue section) as ABOVE CHANCE or NOT ABOVE CHANCE, (ABOVE CHANCE: $p \leq 0.05$, the null hypothesis is rejected; NOT ABOVE CHANCE: $p > 0.05$, the null hypothesis was not rejected). Mann-Whitney-Wilcoxon tests for population distribution and a Hedge'*g* test for effect size, were subsequently used to distinguish task success measures distributions in dialogues categorized as ABOVE CHANCE or NOT ABOVE CHANCE. The linguistic levels were first tested in isolation with the different factors and further in groups, where appropriate. The following section addresses the degree to which a measure of task success can be associated with communicative success.

4.5 Measuring Communicative Success in a Task

To establish if interlocutors may or may not have understood each other in a dialogue, a means to evaluate their communication has to be agreed upon. All types of face-to-face interactions have their own type of appropriate communicative behaviour, an informal meeting with a friend and a political debate, for example, will have vastly different underlying motives and expectations that will change the communicative style of the participants, influence turn-taking and topic-change. An informal discussion may even have no other motivations than creating social links and the information exchanged have no conscious or exact purposes. In task-based interactions on the other hand, one of the purposes of the exchange, that is to accomplish the task, is explicit, and a measurement of how well a task has been accomplished is available, which gives a means to evaluate the communication. When carrying out this evaluation, the type of the task and the medium will influence communicative style in various ways. The outcome of a task that involves persuasion might be more influenced

[8] I qualify this relation as a *rate* as it is the ratio of the number of rejections to the number of possible rejections: $\left(\frac{rejections}{tests}\right)$.

by the presence of the medium of video, which conveys facial expressions, while problem solving tasks might not be influenced by the presence or absence of a video channel, and if anything its presence could even be a disruption (Whittaker, 2003).

This section describes in depth one type of task success assessment methods that is used in this thesis to evaluate communicative success: The *deviation scores* of the map tasks. The *deviation scores* are described as the centimetre square difference between the route on the map of the Information Giver and the Information Follower, with the map divided into a grid of one centimetre squares. Those scores are here reconstructed, evaluated, and results from its replication suggest a good reliability of the measure against other possible measure of task success, such as completion time, for the particular task at hand. As discussed in chapter 3, two other measures are used in the following experiments: negative cognitive states and positive/negative interaction's facilitator's feedback, for two of the corpora. However, this section is only concerned with the *deviation scores* created from the map tasks, that measure by how much a route drawn on a map by an Information Follower deviates from the original route described verbally by an Information Giver.

4.5.1 Deviation Scores in the Map Task Technique

Since its first description (Brown et al., 1985), the map task technique has been used in a wide range of experiments testing for different conditions for which many corpora have been created. Among them the HCRC Map Task (Anderson, Bader, et al., 1991) is noticeable by its design and accessibility to researchers. The technique was used to create corpora in a variety of English dialects (Scottish, American, Australian), other languages (French Occitan, Italian, Japanese, Portuguese, Dutch, Swedish), in medical contexts (Bard, Sotillo, Anderson, Thompson, & Taylor, 1996), in computer-mediated interactions (Hayakawa, Luz, Cerrato, & Campbell, 2016), with the use of avatars (Clayes & Anderson, 2007), and so on. Many studies use the HCRC map task, and the particular success measure provided by the authors as a variable indicating successful task management. The design of the map task is made so that success in the task is subordinated to successful verbal communication, at least to a certain degree. Half the corpus is in no eye-contact condition with a screen blocking the entire view of the other participant. In the other half – eye-contact conditions

– a low-height screen was placed to block the view of participants' maps, they could only see each other's face, and they were instructed not to use gestures to communicate (the low-height screens helping with the enforcement of this instruction). In both conditions, the privileged medium is verbal, and the extent of non-verbal communication is limited to facial expressions but seldom body language. The controlled design of the HCRC Map Task, makes it an exceptional object of analysis for the study of human verbal communicative behaviour, even 30 years after its release. Even taking into account the evolution of language during that time, it is reasonable to assume that the underlying communicative processes involved in speech dynamics remain fairly similar, if not exactly the same, as they are now.

Following those considerations, much scientific research from fields such as cognitive science, computational linguistics, natural language processing and so on, legitimately used this material in various studies (Carletta et al., 1997; Sotillo, 1997; Davies, 1997; Pardo, 2006; Truong & Heylen, 2012), and also made a use of the measure of success available: the deviation scores (Reitter & Moore, 2007; Colman & Healey, 2011; Rothwell, 2018). An interesting aspect of this task success measure, is that its evaluation gives theoretically the least space to subjective interpretation from the evaluator.

4.5.2 Original Pre-Computed Deviation Scores

My work uses the material given by the authors of the HCRC Map Task corpus (The Information Givers' and Followers' maps) to recreate the precomputed score given. Despite its apparent straight-forwardness, there were a number of issues, even when interested in overall tendencies, that appeared to make the scores worth being replicated to ensure their trustworthiness, in particular to researchers interested in using the deviation scores as their primary measurement of successful communication. Far from disparaging the measure, I am here rather raising some issues that require resolution and that might be of interest for future researchers in their usage of this method of task success measurement. One of the main concerns to assure sound scientific research is the replicability of experiments. To build solid foundations for hypothesis testing that will in turn lead to the construction of the theories that constitute scientific knowledge, the capacity to replicate experimental results is crucial in all domains. In particular in domains closely related to psychology such as computational

linguistics and cognitive sciences, where the difficulty of replicating results have been more than often pointed out (Asendorpf et al., 2013).

4.5.3 Reconstruction of the Deviation Scores

The reason for the redesign of path deviation counting methods was mainly that the same results were not found after few attempts at replicating the scores of different maps; even when carefully reverse engineering the A3 to A4 specifications and the 1 cm grid overlay, that should have led to the same scores. Similar, but not exact scores were found, which led to the question of replicabilty and most importantly reliability of the scores that were meant to be used as a starting point to evaluate communication in human behaviour in subsequent experiments. Those scores needed to be accurate and possible to replicate with precision in order to constitute a sound object of study. A number of issues appeared, such as the poor quality of the scanned Information Follower (IF) maps and their apparent distortion during the scanning process of the completed maps. It was indeed difficult to match the given IF original maps that were the templates for the maps and the actual IF maps that had the results routes drawn on them. The software Adobe Illustrator was used to respect precisely the instructions given (A3 grids over A3 IF/IG maps). Following the instructions resulted in the definition of each possible cases in which a square can be in 4 different methods of counting, from the most restricted possibility to the broadest interpretation.

A number of descriptions of the method created to count deviation scores can be found in the literature, the most pertinent are listed in the Table C.1, Table C.2, Table C.3, of Appendix C, and were used to reconstruct those scores from the original maps given with the HCRC Map Task corpus. In a widely cited work as originating the deviation scores (Anderson, Clark, & Mullin, 1991), no mention of a scoring method is found. This work uses the map task technique to create 170 dialogues of children between 5 and 13 (in three groups), and mention that this technique was created by Brown, Anderson, Yule & Shillcock (1985). No mention of deviation scores is made in that last publication either. The results in that publication are concerned with the proportions of definite and indefinite article usage among the different categories of speakers — as it is the declared aim of the research article

— but not in relation with the participants success in the task.[9] The book cited by Anderson et al., *Teaching Talk: Strategies for Production and Assessment* (Brown et al. 1985), however, does mention an assessment method, even if it is not a deviation score per se. The map task was originally designed to assess pupil's information transferring skills, and was tested among other tasks such as wiring-broad task or story summarising. As the authors that created the corpus highlight, despite the requirement that the route is to be drawn as precisely as possible in order to avoid potential danger, "it is not draughtsmanship which is being assessed, but the ability to recognise that the other person needs to be told whether to go right or left, or up or down at crucial parts of the map" (Brown et al., 1985, p. 111). However, the relative precision of the reproduction of the route remain key to the reliability of the measure. As mentioned earlier, the *deviation scores* are described as the centimetre square difference between the route on the map of the Information Giver and the Information Follower, with the map divided into a grid of one centimetre squares.

The description given on the HCRC Map Task website[10] was used to define four methods of counting, depending on the area between the route drawn by the Information Follower and the original route, each described in Table 4.2.

[9] No mention in the study of a task score or any variable linked to task success.

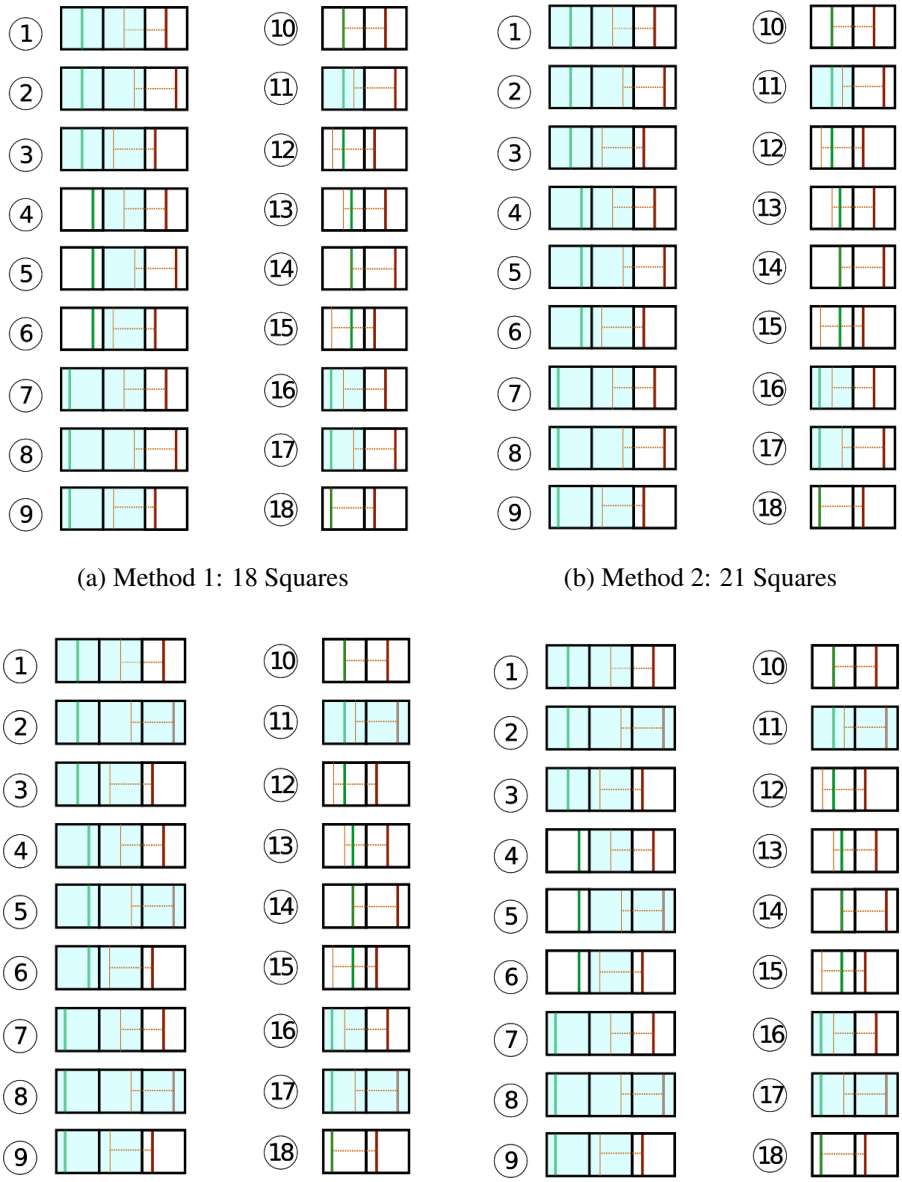
[10] `http://groups.inf.ed.ac.uk/maptask/maptask-description.html` (Last consulted: 20/03/2017) See Appendix C, Table C.1 for full text.

Table 4.2: Description of Deviation Scores Counting Methods.

	Description
Method 1	Counting squares of which the area is more than 50% between the 2 routes. Refer to Figure 4.1 (a).
Method 2	In addition to Method 1, include the squares that cover the IF path but of which the area is not more than 50% between the two routes. Refer to Figure 4.1 (b).
Method 3	In addition to Method 2, count the squares including the IG route when more than 50% of the area is between the two routes. Refer to Figure 4.1 (c).
Method 4	In addition to Method 1, count the squares including the IG route when more than 50% of the area is between the two routes. Refer to Figure 4.1 (d).

Each grid of 1 cm squares has been placed to overlay the IF maps from the top left. The grids and the Information Follower maps are subsequently placed on top of the Information Giver maps by trying to match the starting cross first, then the different landmarks. In some cases, this operation required slight distortions for a better match between the maps, as it seemed the IF maps provided suffered distortions in the scanning process. How to estimate the distance between the two lines and decide if a square of the grid that overlay the two lines should be included in the score, or in other terms: 'is the area between the two routes large enough to be counted as a deviation?'. In an attempt of precision and after an observation of the real resulting maps, I identified eighteen possible situations (involving forty-five squares) in which the lines and overlays could be.[11] This is illustrated in Figure 4.1.

[11]This number of eighteen situations comes from the three possible positions in which a line can be in terms of division of the space inside a square; either the area of the square is divided into 50% of the space on each side of the line, or the percentage is higher/lower on each side. (As there are two lines that can be in three positions in each square, with possibly an entire square in between this, results in: 2 lines times three positions times three situations equals eighteen.)



(a) Method 1: 18 Squares

(b) Method 2: 21 Squares

(c) Method 3: 26 Squares

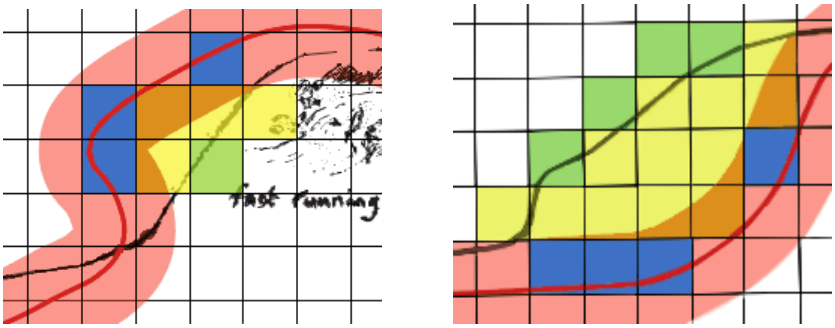
(d) Method 4: 23 Squares

Figure 4.1: Possible cases for each square in the Methods of counting (18 possibilities); the red line (right) represents the original route, the yellow line (center) represents the distance of one centimeter and the green line (left) represents the route drawn by the IF.

In the eighteen possible configurations in which the drawn route can be positioned with respect to the original route that have been isolated, each method kept a certain number of squares. Respectively 18 squares for Method 1, 21 for Method 2, 26 for Method 3 and 23 for Method 4. in which the red line represents the original route (right), the yellow line the distance of one centimetre (centre), and the green line the route drawn by the IF (left). This can be interpreted as the distance from which the path starts to be considered deviant, Method 1 being the least inclusive (the strictest interpretation of the described method) and

Method 3 the most inclusive (the broadest interpretation).

The methods of counting have been applied to the 128 maps of the HCRC Map Task and the 16 maps of the AEMT corpus. The results of these counts can be found in Table C.4 in the Appendix C. An example of counting on the real maps, to exemplify the possibilities of different interpretations if no clear definition of the counting method is given, can been seen in Figure 4.2 and Table 4.3. The red line represents the original route, and the black line the route drawn by the IF, the light red area represents one centimetre distance from the original route, the squares in yellow are counted using method 1, the squares in green are the addition created by following counting method 2, the squares in blue are added for method number 3, finally method number 4 retains the squares in blue and yellow only.



(a) Example 1 (b) Example 2

Figure 4.2: Counting Methods real example from HCRC maps

Table 4.3: Squares counted by Methods in the real example given in Figure 4.2

Method	(a)	(b)
1	4	12
2	5	16
3	8	20
4	7	16

To assess which method has the minimum distance with the HCRC Deviation Score, a series of Correlation Tests have been applied on the methods (see Table 4.4). A plot of the deviation scores counted on the HCRC Map Task maps, by methods of counting, indexed on the original pre-computed deviation score is given in Figure 4.3.

Table 4.4: Pearson Correlation Coefficients for Pre-Computed Deviation score and given Counting Methods

Scores compared	Correlation coef. R
HCRCDevS,M1	0.944509
HCRCDevS,M2	0.9426531
HCRCDevS,M3	0.9492633
HCRCDevS,M4	0.9501321
M1,M2	0.9938681
M1,M3	0.9922469
M1,M4	0.9956481
M2,M3	0.9966103
M2,M4	0.9886506
M3,M4	0.9946619

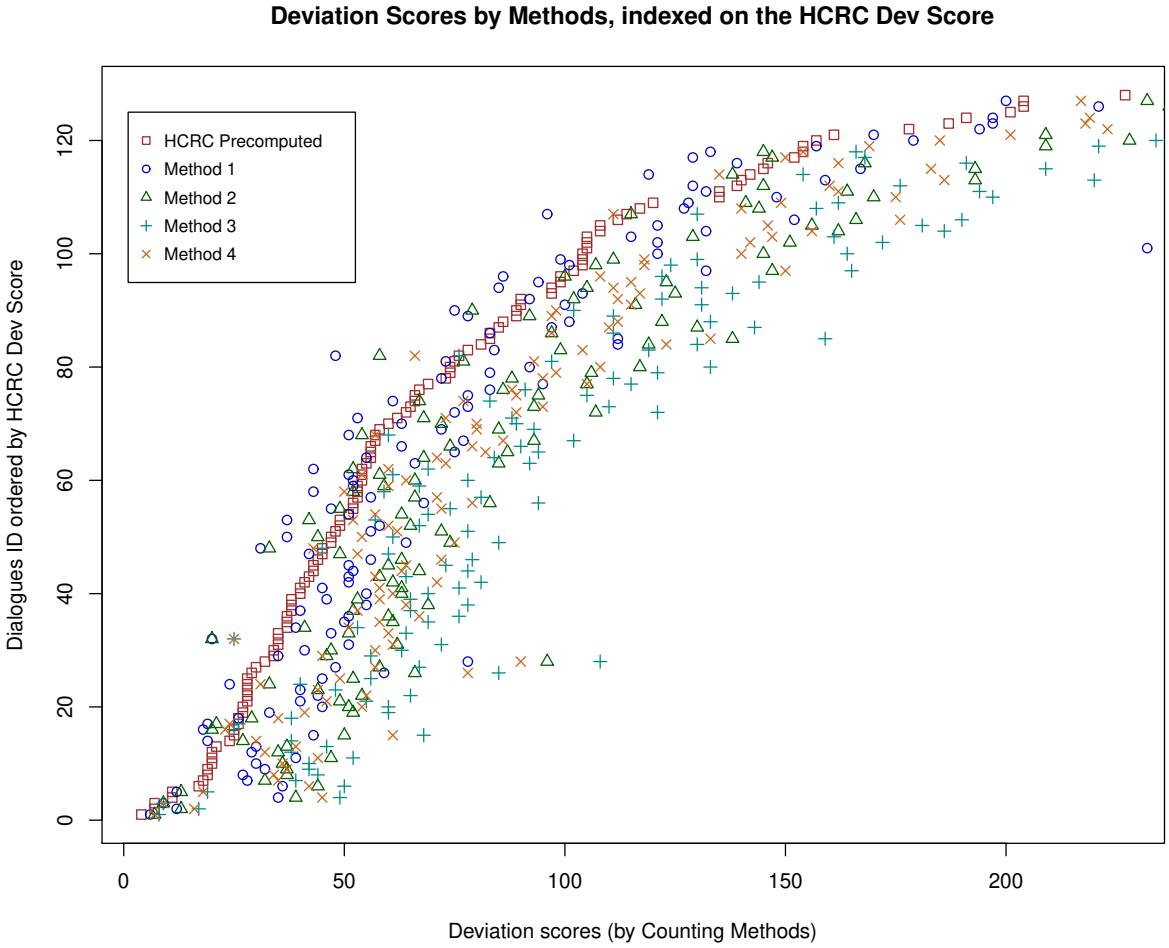


Figure 4.3: Plot of Deviation Scores by Methods, indexed on the HCRC Precomputed Deviation Score

Method 4 is adopted as the closest to the HCRC pre-computed method, as it is the method showing the highest correlation with the original pre-computed scores. Method 4 is the one used in the assessment of similarities and divergence in repetition patterns between the AEMT

corpus and the HCRC Map Task, labelled as M4, in the experiment described below in section 5.5. The results show a very high correlation between the pre-computed deviation scores and the recounted ones, which still indicates the relative precision of the original scores, even if it was needed to verify their accuracy after noticing small differences. The original HCRC deviation scores are used in the experiments reported in section 5.2, section 5.3, and given as comparison with the Method 4 in section 5.5.

4.5.4 Other Map Task Scoring Systems and Limitations

Time of completion is also used as measure of success for map tasks (Rothwell, 2018), however I did not find a correlation between deviation scores and time in the HCRC Map Task (Pearson correlation coefficient, $r = -0.08$) and note that participants in the map task were told to be accurate (Brown et al., 1985) and not to finish as fast as possible, which makes time of completion a less appropriate measure for this particular task. In her studies examining the structuring principles of task-oriented dialogues, Bethan L. Davies uses the HCRC Map Task for which she has created an alternative scoring system (Davies, 2006, 1997): the Incorrect Entity score. She identifies two disadvantages of the deviation scores method. Her first point is that (p.102) "[...] estimating portions of grid squares is not straightforward: it is inevitable that the section being calculated will not always contain whole grid squares." This first point is addressed with the above described methods of counting, which define all possible situations in which a square can be – and the problem of subjectivity that arises when including a square in the counting or not – is greatly reduced. Her second point remains however valid despite the counting method proposed in this section: as it is not draughtsmanship that is measured, Information Followers that negotiate correctly the landmarks but do not accurately follow the original route are penalised by the method. They were told to be accurate, as explorer that had to follow the route described precisely as it is the only "safe" route, but not to the centimetre precision that the deviation scores suggests. Despite this shortcoming, the deviation scores remain pertinent measures in the frame of this study by its objectiveness and accuracy, above time completion. Those limitations in the deviation scores lead to consider the limitations of the method described in this chapter as a whole in the following section.

4.6 Limitations of the Methods

The methods used contain a certain number of points of concern, that can be considered limitations but also may simply require awareness in interpretation of results, that are addressed in this section. Firstly, in expanding to multiple levels of representations, such as lemma and Part-Of-Speech, it necessarily relies on previously constructed Western language oriented linguistic theories of syntactic rules and word cutting representations (Ansaldo, Don, & Pfau, 2010). While this fact is not exactly a limitation, awareness is necessary when comparing different languages and using speech from non-native speakers, for whom the usage of these particular constructs in their native languages might be vastly different and influence their ultimate use of English. There are also profound controversies in the usage of statistical hypothesis testing, the validity of *p*-values, in particular in the domains concerned with psychology and human behaviour analysis (Asendorpf et al., 2013; Krantz, 1999). Many studies on speech that can to a certain degree be called empirical, including most of the ones presented in the preceding literature review, are based on relatively small samples. Moreover small samples that behavioural scientists have the least difficulties to recruit: Western, Educated, Industrialized, Rich and Democratic (WEIRD) people – often university students – and the corpora used here correspond to this description, which *de facto* reduces the possibilities to generalise findings to other human populations (Henrich, Heine, & Norenzayan, 2010). One of the current issues is the size of the samples available for speech linguistic analysis, as well as their quality. Even with the remarkable improvements made in the domain of Automatic Speech Recognition (ASR), most transcriptions are still human-made, or if an ASR tool is used, manual checking is required to ensure quality. The fact that the chance to detect an effect increases with a larger data set (increasing the chance of Type I error), also cannot be ignored. The decision to keep, in the first instance, all the levels of representation in testing for differences among groups in terms of communicative success scores, may risk combining five data points from the same dialogue at its five different levels of representation, to test against only one or two levels in other dialogues. This is only a concern if we consider that syntactic priming effects are not independent from lexical priming effects, but that lexical effect "boost" syntactic effects. For now the full dependence of the syntactic level is controversial but evidence points towards no full independence (Pickering & Ferreira,

2008). This issue is addressed here by performing a grouped test in the first instance then subsequently testing levels in isolation to curb that possible confound. The problem posed by the risk of detecting a faint phenomenon, or in other words inflating artificially effects that would have been undetectable with fewer layers of testing might in that case allow us to highlight a quantitative pitfall common in linguistics studies: "Phenomena of theoretical interest may be so sparsely represented in naturally occurring speech that huge corpora may still fail to supply sufficient instances to support robust conclusions"(Anderson, Bader, et al., 1991, p. 351). A drawback that is less frequent in non-speech text analysis where massive amounts of data are often available. Using only transcripts and without associating speech features in the analysis is also a limitation that needs to be addressed in future work. That the study of speech should include sound and even non-verbal features – with an increase in the available methods interested in multi-modality – to capture as much as possible the dynamics of spoken interaction is widely agreed upon. However the extension to speech (i.e. prosodic) and non-verbal features should be the next step in the evolution of the method. The large body of research that still mostly uses transcripts, also highlights the richness and complexity that can be extracted from this material.

4.7 Conclusion

This chapter presented the methods used to examine the qualification of mutual understanding within the frame of task-based interactions, from its foundations in previous use, extension and current state. This chapter also presented a method for reconstructing deviation scores, an important element in the assessment of successful communication in map task dialogues, as well as the limitations of the methods used to analyse textual conversational content.

Chapter 5

Experiments

5.1 Introduction

The experiments reported in this chapter are designed to answer the main research question repeated here:

> *— To what extent is an automatic method focusing on one feature of dialogue structure, repetition as cues of an alignment process, able to capture interactional behaviours and patterns with sufficient accuracy to quantify a degree of mutual understanding?*

Each experiment carried out is concerned with different dimensions in which patterns of repetitions are reported and compared in a series of contexts along with their measure of communicative success. Figure 5.1 gives an outline of the main research questions associated with each experiment and summarises the characteristics of the corpora used.

Section 5.2 reports the work undertaken with the HCRC Map Task, grouping a preliminary experiment, and two in-depth studies exploring the Map Task controlled conditions: task role, gender, the possibility of eye-contact, the familiarity between interactants and with the task they are committed to accomplish. Effects between these conditions, task success and repetition patterns, are expected according to the literature given in chapter 2:

> (1) The distinction between above chance and not above chance proportions of repetitions potentially have an impact on task-success.

(2) Women seem overall less disfluent than men (however a pattern found in same turn repetitions), could be found in turn to turn repetitions patterns.

(3) Information givers seem overall more disfluent than followers (however a pattern found in same turn repetitions), could be found in turn to turn repetitions patterns.

(4) The absence of eye-contact induces more repetitions, in particular at first attempt of a task (H. Branigan et al., 1999).

The following sections (§ 5.3, § 5.4, § 5.5, and § 5.6), present the four subsequent studies, investigating the ILMT-s2s, the MULTISIMO, the AEMT, and PARDO 2006 corpora, with adjustments and improvements added gradually as the methods evolved. These studies are represented in Figure 5.1 (as presented in the Contributions) as radiating around the first section as results obtained from the HCRC Map Task informed them and the HCRC Map Task corpus is also used for comparison in three of the four experiments.

Each section gives partial answers to the main research question, repeated above, by exploring the relationship between task success using a score and patterns of repetitions depending on different features, then between patterns of repetitions and cognitive states and types of feedback given by an interactional facilitator. Parts of the analyses and results presented in this chapter have been published (Reverdy & Vogel, 2017b, 2017a; Reverdy, Hayakawa, & Vogel, 2018; Reverdy, Koutsombogera, & Vogel, 2020) and are reported here with co-authors' permissions.

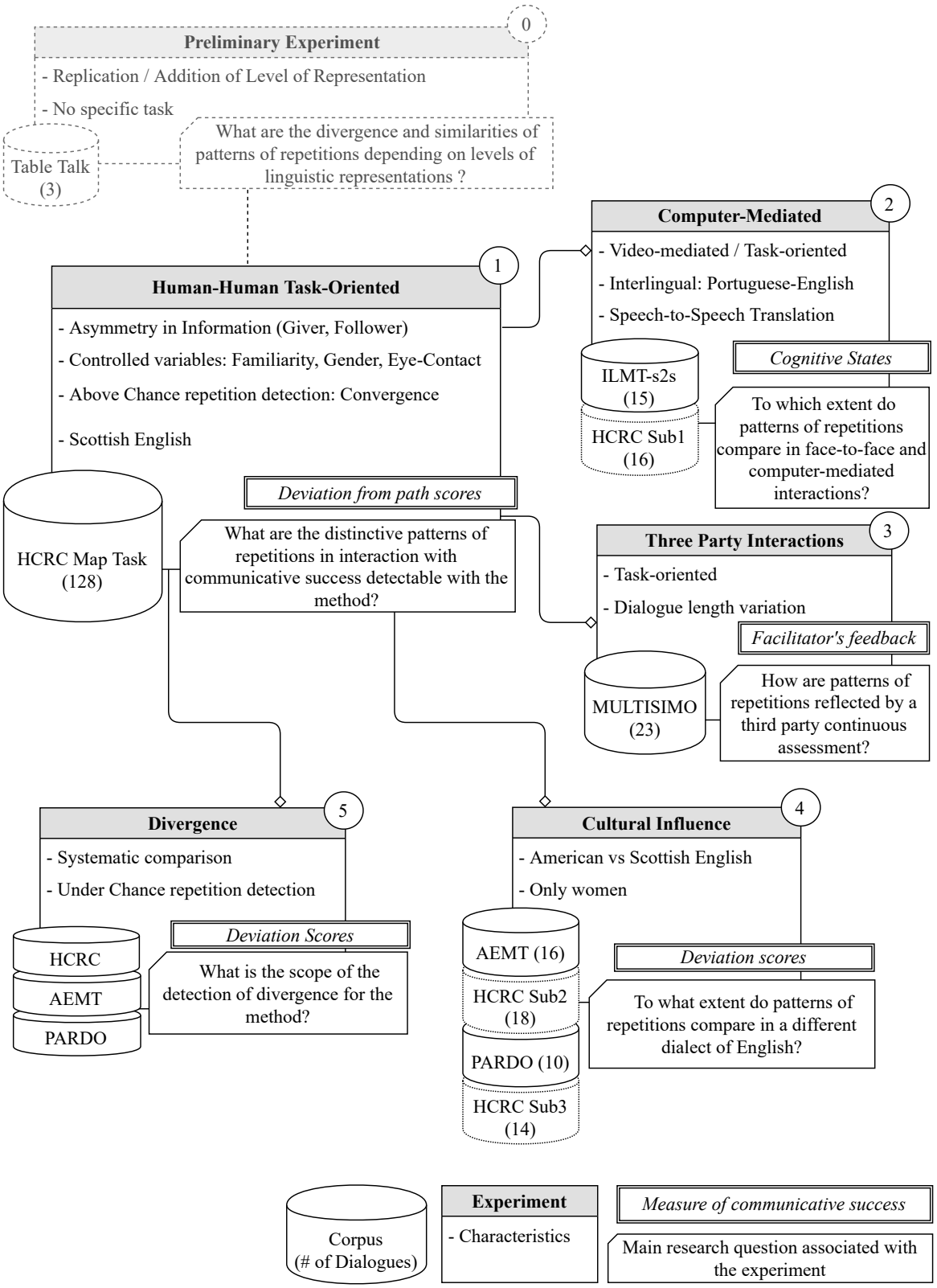


Figure 5.1: Experiments Summary

5.2 Human-Human Task-Oriented Interactions

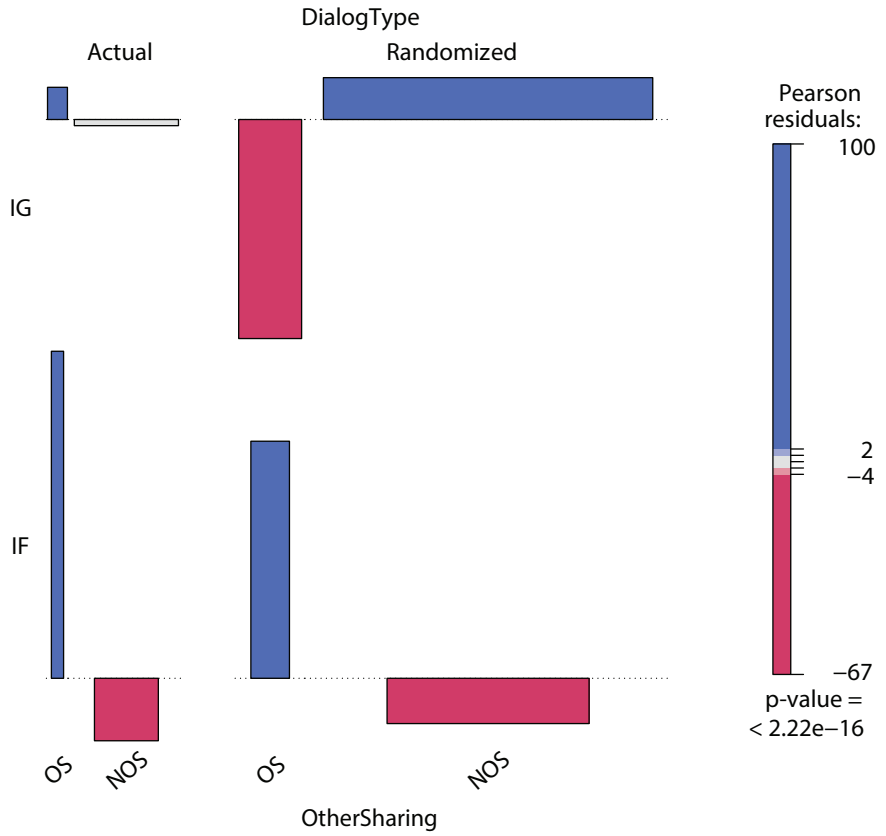
This section reports the preliminary experiment I did with a subsection of the corpus, then I present two published studies that explore the method using the full HCRC corpus in conjunction with the controlled features that constitute, from the perspective of assessing mutual understanding, one of its essential interests. The ensemble formed by the preliminary experiment and the two following sections forms an *exploratory study* of the HCRC corpus.

5.2.1 Preliminary Experiment within the HCRC Map Task

A subset of the HCRC Map Task, that I call HCRC Subset 1, comprised of 16 dialogues, was used in this first approach. As noted by the authors of the corpus at its release (1991), there is a significant asymmetry between the different roles in terms of speech volume, the Information Giver (IG) producing on average more than twice the number of words uttered by the Information Follower (IF). To explore if and how this asymmetry impacted repetition patterns, our analysis starts by observing the proportion of repetitions between actual and randomised dialogue by speakers role.

As it can be seen in Figure 5.2 and Figure 5.3,[1] the speakers in IF positions show on average a higher proportion of SELFSHARED and OTHERSHARED repetitions than would be expected by chance in the actual dialogues, according to the Pearson residuals. Figure 5.2 and Figure 5.3 also show that for the IG the proportion of repetitions are above expected levels for OTHERSHARED but not for SELFSHARED repetitions. Those observations seem to indicate that the Information Follower repeats the Information Giver much more than the contrary. As mentioned in section 3.2, the IG has to guide the IF along a predefined route, any changes in that route were assumed being the result of less successful communication between the two participants and the *deviation scores* were then computed with that assumption. To observe a possible link between ABOVE CHANCE repetitions in dialogue and task success, the significance test results for each dialogue were compared to the *deviation scores*

[1] Figure 5.2 and Figure 5.3 present two association plots of residuals, determined by the difference between observed and expected values, using a loglinear model (Meyer, Zeileis, & Hornik, 2006): the magnitude of a box corresponds to the magnitude of residuals; shading intensity encodes significance (residuals between 2 and 4 are significant at the $p < 0.05$ level); boxes projecting up from the horizontal line correspond to divergences in excess of expectations and boxed projecting down from the horizontal convey the extent to which observations are fewer than expected, where expectations are those of the null hypothesis, which is that there is no interaction among the categories examined.



(a) OTHERSHARED (OS) vs. Non-OTHERSHARED (NOS)

Figure 5.2: Proportions of OTHERSHARED repetition units at All Levels, for 16 HCRC Map Task dialogues, per Speaker Role in Actual and Randomised dialogues

in a preliminary naive way: the mean score (71.82), is used to divide the dialogues into two categories:

– Problematic: The score is above average

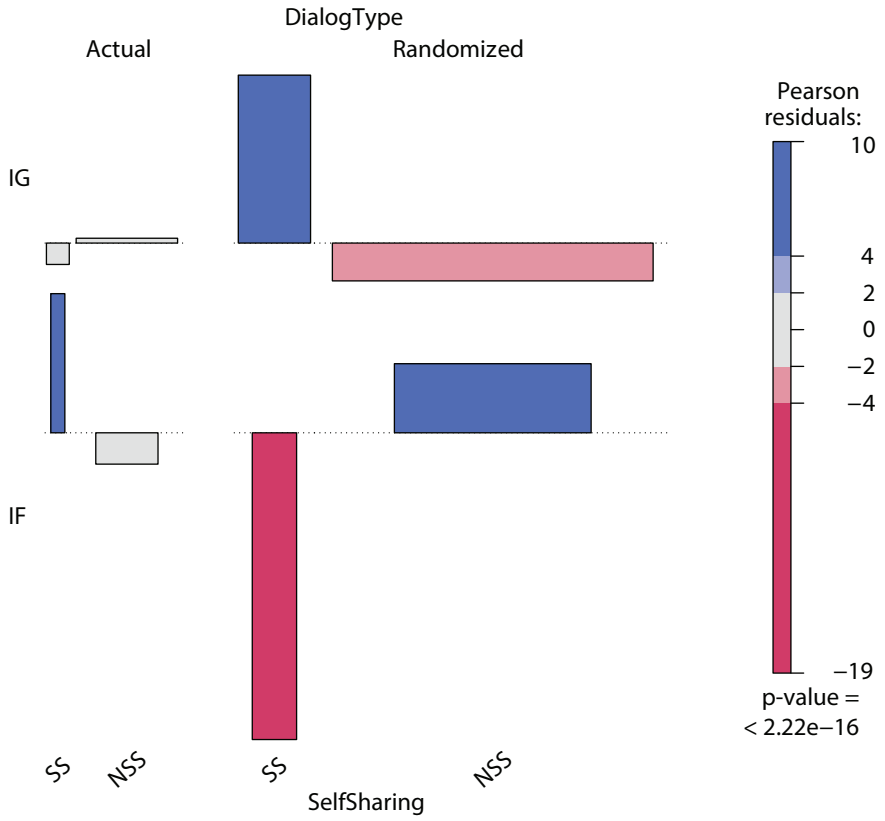
– Successful: The score in below average

Then dividing the dialogues based on the rejection of the null hypothesis according to different selection in a binary manner:

– Negative = 0: No rejection for more than 2 linguistic levels

– Positive = 1: Rejection for more than 2 linguistics levels

The results given by this first comparison are shown in Figure 5.4. A negative correlation for the Information Follower between low score and successful response emerges, while the failure to reject the null hypothesis (indicating a non-significant difference in the proportion of repetitions between *Actual* and *Randomised*) relates with higher *deviation scores*. Despite the fact that the IG had on average a greater quantity of speech than the IF, the latter repeated



(a) SELFSHARED (SS) vs. Non-SELFSHARED (NSS)

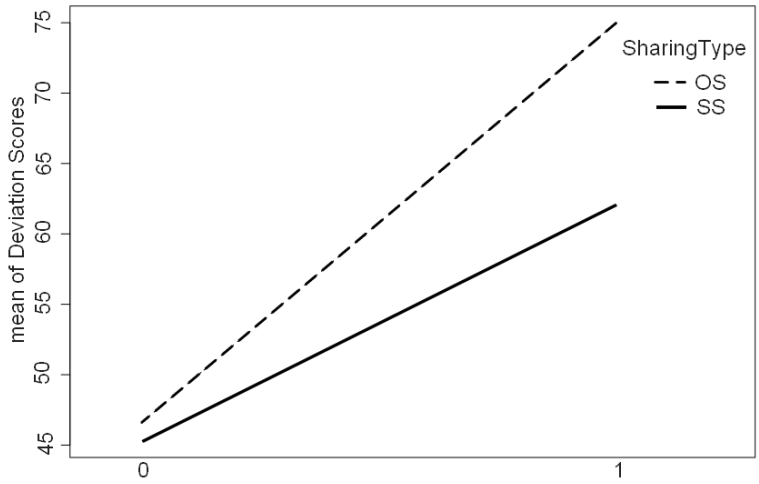
Figure 5.3: Proportions of SELFSHARED repetition units at All Levels, for 16 HCRC Map Task dialogues, per Speaker Role in Actual and Randomised dialogues

the IG at above expectations level and when ABOVE CHANCE repetition occurred it related to higher task-success (low *deviation score*). The relation found between the *deviation score* and ABOVE CHANCE repetition level are encouraging signs that this method could be effective as a possible predictor of success in task-based communication. In summary, this first approach shows a clear role distinction, however more refinements are needed.

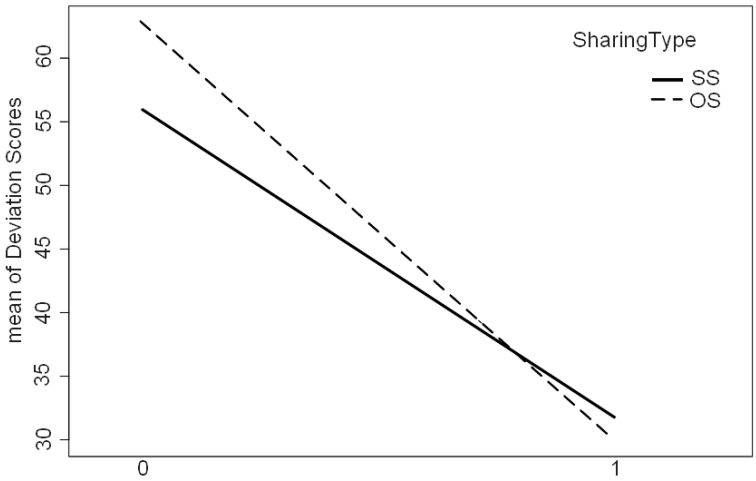
5.2.2 Task Success and Non-linguistic features

The following section describes the experiment conducted using the entire HCRC Map Task Corpus and explores the potential differences held by the controlled features it contains. Table 5.1 gives a summary of the number of repetitions per conditions. As described in section 4.4.1, Tukey's tests were performed on all dialogues, resulting in 1280 tests of the four variables against the two repetition types (OTHERSHARED, SELFSHARED), first including all n-grams, then for n-grams with $n = 1$, and finally for n-grams with $n > 1$.

Following a threshold of ($p \leq 0.05$), the Null Hypothesis was rejected 902 times for OTH-



(a) Information Giver (IG)



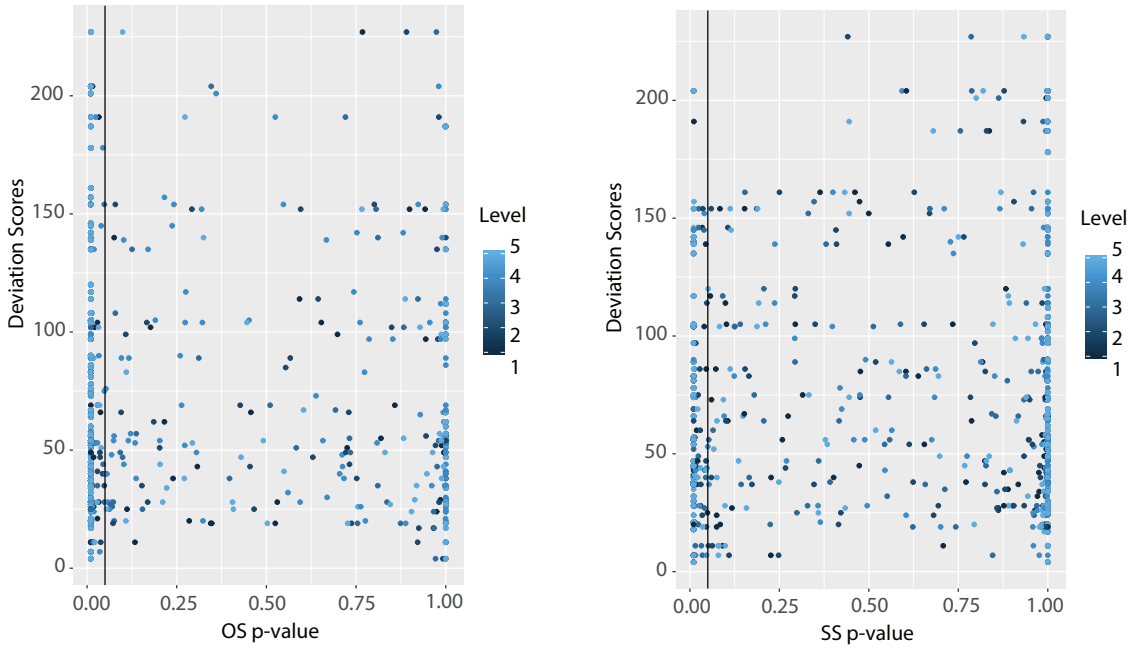
(b) Information Follower (IF)

Figure 5.4: Interaction between binary division of null hypothesis H_0 and *deviation scores*

ERSHARED and 281 for SELFSHARED respectively, for all *n*-grams. Showing that across all variables, there is a high proportion of OTHERSHARED repetitions in this task-based corpus. Figure 5.5 shows the distribution of *p*-values resulting for the Tukey's tests, from 0 to 1 on the *x*-axis (the 0.05 threshold adopted in the method is represented by a black vertical line). Figure 5.6 shows this distribution for Token only. These scatter-plots interestingly expose the clear divide between 0 and 1, with relatively few values in-between.

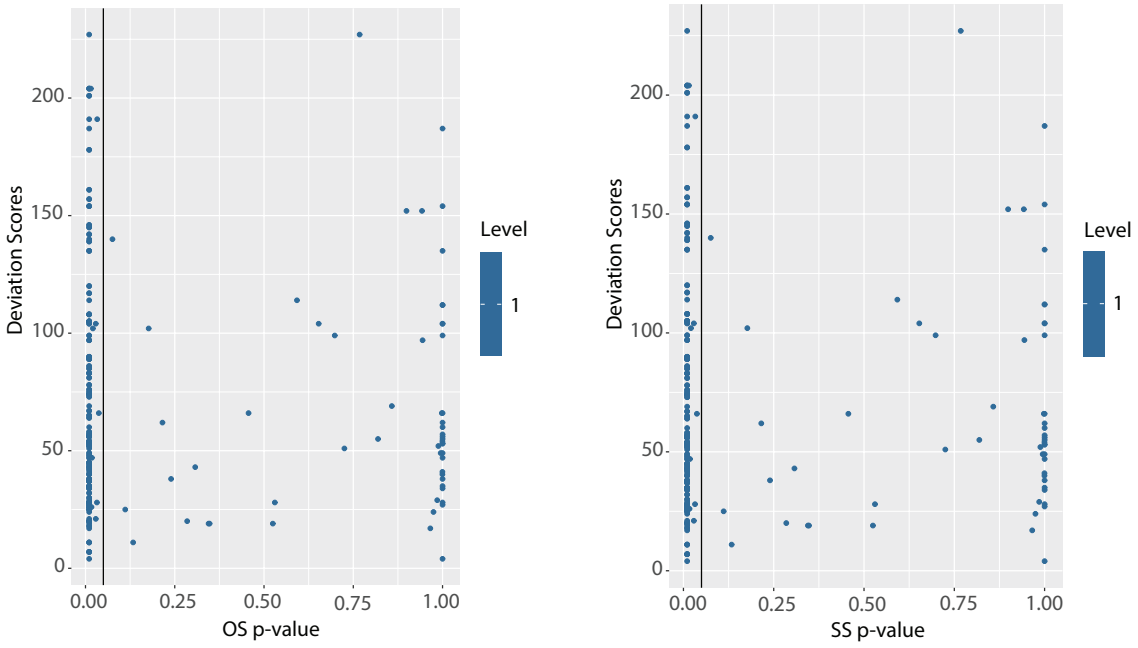
Table 5.1: HCRC Map Task Summary of repetitions per conditions, SELF REP and OTHER REP (see § 4 for the definition) are given for the linguistic representation level token only.

	IG	**IF**	**Fam**	**UnFam**	**Female**	**Male**	**Eye**	**NoEye**
OTHER REP	13,492	5,834	11,886	9,232	8,125	9,020	9,990	11,128
SELF REP	11,281	9,781	10,503	9,073	7,575	8,287	9,057	10,519



(a) Scatterplot OTHERSHARED (b) Scatterplot SELFSHARED

Figure 5.5: Distribution of *p*-value resulting from Tukey's tests in interaction with Deviation scores from the HCRC Map Task (See § 3.2), and Level (1: Token, 2: Lemma, 3: POS+Lemma, 4: POS, 5: Token+POS) for OTHERSHARED and SELFSHARED



(a) Scatterplot OTHERSHARED L1 (Token) (b) Scatterplot SELFSHARED L1 (Token)

Figure 5.6: Distribution of *p*-value resulting from Tukey's tests in interaction with Deviation scores from the HCRC Map Task (See § 3.2) at Level 1: Token Only for OTHERSHARED and SELFSHARED

Table 5.2: Rejections of H_0 for OTHERSHARED and SELFSHARED in the HCRC, in relation to roles (IF: Information Follower; IG: Information Giver), in each case (each cell) the Null Hypothesis can potentially be rejected 128 times

HCRC Map Task

	OTHERSHARED								**SELFSHARED**							
All n-grams | H_0 : $Rand.Speaker.Level - Actual.Speaker.Level \geq 0$																
	Tok	Lem	LemP	POS	TokP	Mean	%		Tok	Lem	LemP	POS	TokP	Mean	%
IF	112	109	109	82	107	103.8	81	IF	36	35	37	19	38	33	25.7
IG	88	87	80	47	81	76.6	59.8	IG	27	26	30	5	28	23.2	18.1
N1: n-gram=1 | H_0 : $Rand.Speaker.Level.N1 - Actual.Speaker.Level.N1 \geq 0$																
IF	78	78	74	46	75	70.2	54.8	IF	8	10	11	4	11	8.8	6.8
IG	49	47	51	18	54	43	34.2	IG	4	4	4	0	5	3.4	2.6
N2+: n-gram>1 | H_0 : $Rand.Speaker.Level.N2+ - Actual.Speaker.Level.N2+ \geq 0$																
IF	108	104	105	81	107	101	78.9	IF	38	38	39	26	37	35.6	27.8
IG	90	91	88	58	89	83.2	65	IG	44	49	43	16	46	39.6	30.9

A closer look at the number of times the Null Hypothesis (H_0) was rejected depending on the Speakers and Level is given in Table 5.2. In this table, a higher number of rejections is observed for OTHERSHARED than SELFSHARED, for both Information Giver (IG) and Information Follower (IF). Nonetheless, a significant asymmetry between the different roles arise, with the IF repeating herself/himself and the IG, more in the Actual dialogues than in the Randomised ones. A low number of rejections is noticed for SELFSHARED for the IG. The IG only repeated herself/himself significantly in five dialogues for the Level 4 (POS) in particular. Yet, the IG overall number of rejections for N2+: n-grams $n>1$ is slightly higher than the IF, which signals that when the IG self-repeats, it tends to be longer utterances. Mann-Whitney-Wilcoxons tests[2] for population distribution and Hedge tests for effect-size showed overall non-significant differences and negligible effect-sizes of the distribution of *deviation score* between male and female (W= 9049.5, $p = 0.13$, and between Eye-contact and No eye-contact (W= 8278, $p = 0.88$). The only significant difference found for the non-linguistic factors was (W= 6572, $p = 0.006$) between Familiar ($\bar{x} = 64.37$) and Unfamiliar ($\bar{x} = 79.28$) participants. Tests showed all non-significant differences for IF (Gender: $p = 0.10$; Eye-Contact: $p = 0.92$; Familiarity: $p = 0.053$) and IG, even if a small effect size appeared between Gender for the IF ($g = 0.30$), with pairs having a male IF obtaining better scores ($\bar{x} = 64.4$) than pairs with a female IF ($\bar{x} = 79.2$).

[2]The results reported in this section used the precomputed *deviation scores* given by the authors of the HCRC Map Task. The tests have been done a second time using the Method 4 for deviation score calculation given in section 4.5.3, for which similar results have been observed, significance and non-significance found between the same tested variables.

Table 5.3: Sums of *deviation score* per Conditions, at all linguistic levels of representation (5), along with the number of dialogue involved in each division, in the HCRC Map Task: significant OTHERSHARED *p*-values (Above Chance | Not Above Chance), Speakers (IG: Information Giver | IF: Information Follower), Eye-contact, Familiarity, and Gender. The sum of dialogues for Unfamiliar participants amount for 640 (five times 128), and it is also the case for Familiar partners.

Above Chance

	No EyeContact				EyeContact			
	Female		Male		Female		Male	
	IF	IG	IF	IG	IF	IG	IF	IG
UnFam	7034	5048	3860	2989	4574	2666	4987	4360
Num of dial	85	53	50	35	47	29	72	52
Fam	5744	4955	1834	1293	3380	1858	5746	4360
Num of dial	82	67	40	33	47	34	96	80

Not Above Chance

	No EyeContact				EyeContact			
	Female		Male		Female		Male	
	IF	IG	IF	IG	IF	IG	IF	IG
UnFam	1325	3092	235	1326	1071	2289	2283	3600
Num of dial	20	52	5	20	8	26	33	53
Fam	1146	2250	1081	1307	1090	1472	579	2473
Num of dial	23	38	15	22	8	21	9	25

Table 5.4: Sums of *deviation score* per Conditions, at all linguistic levels of representation (5), along with the number of dialogue involved in each division, in the HCRC Map Task: significant SELFSHARED *p*-values (Above Chance | Not Above Chance), Speakers (IG: Information Giver | IF: Information Follower), Eye-contact, Familiarity, and Gender. The sum of dialogues for Unfamiliar participants amount for 640 (five times 128), and it is also the case for Familiar partners.

Above Chance

	No EyeContact				EyeContact			
	Female		Male		Female		Male	
	IF	IG	IF	IG	IF	IG	IF	IG
UnFam	2287	807	916	607	941	105	1435	1436
Num of dial	19	13	15	11	9	1	24	16
Fam	1482	576	1065	696	898	739	2265	2005
Num of dial	26	15	18	15	12	16	42	29

Not Above Chance

	No EyeContact				EyeContact			
	Female		Male		Female		Male	
	IF	IG	IF	IG	IF	IG	IF	IG
UnFam	6073	7333	3179	3708	4704	4850	5835	6524
Num of dial	86	92	40	44	46	54	81	89
Fam	5408	6626	1850	1904	3572	2591	4060	5460
Num of dial	79	90	37	40	43	39	63	76

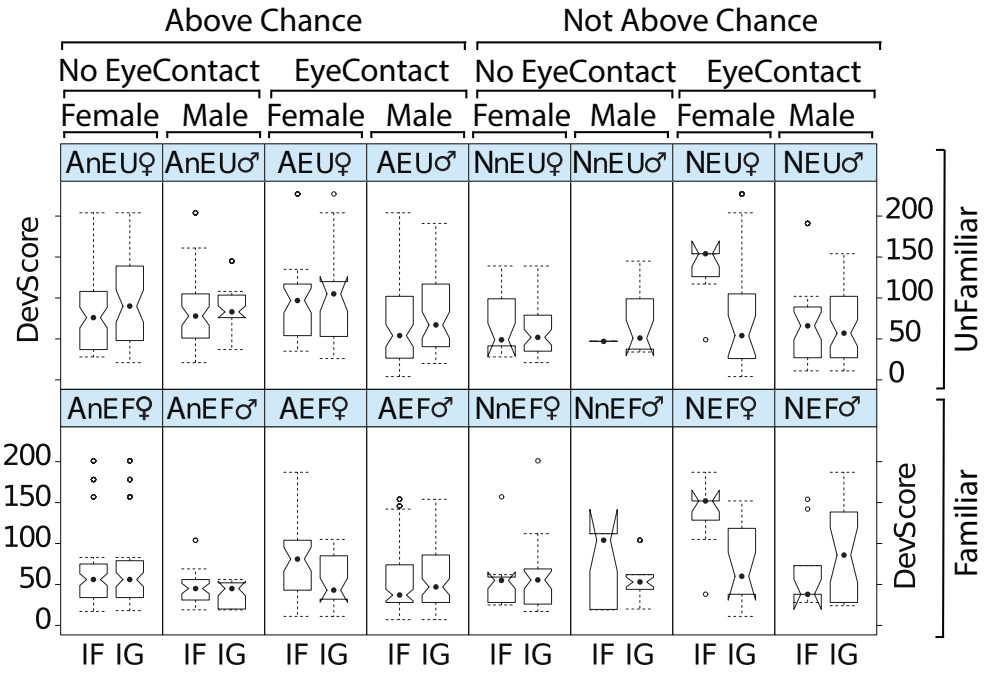


Figure 5.7: (Distribution of Dialogues in interaction with *deviation score*, Speakers (IG: Information Giver | IF: Information Follower), significant OTHERSHARED *p*-values (A: Above Chance | N: Not Above Chance), Eye-contact (nE: No eye-contact | E: Eye-contact), Familiarity (U: Unfamiliar | F: Familiar), and Gender (♀: Female | ♂: Male)

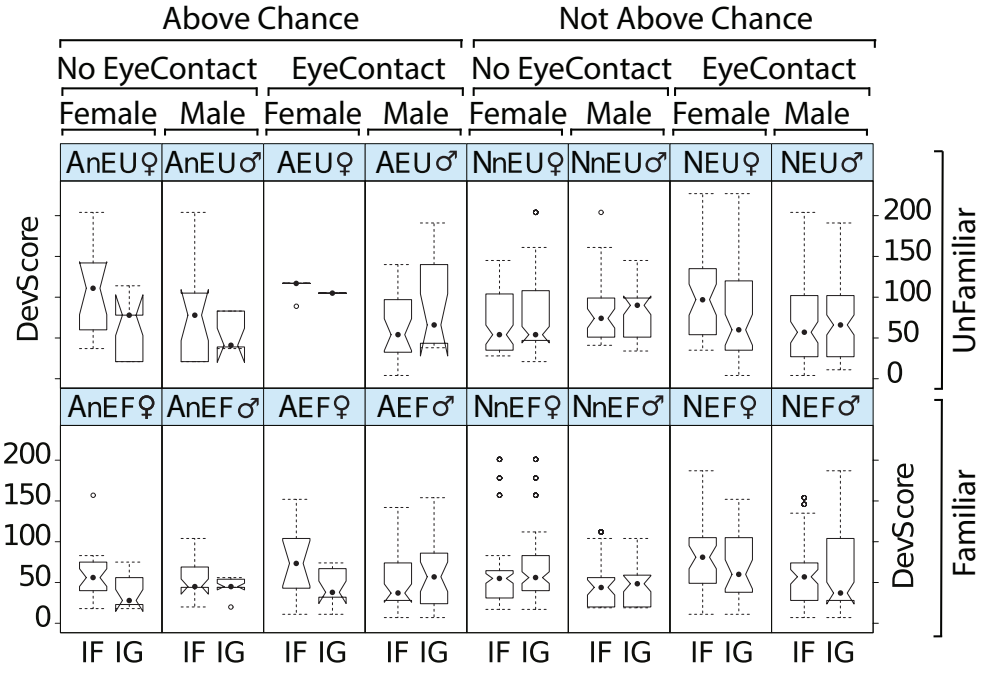


Figure 5.8: Distribution of Dialogues in interaction with *deviation score*, Speakers (IG: Information Giver | IF: Information Follower), significant SELFSHARED *p*-values (A: Above Chance | N: Not Above Chance), Eye-contact (nE: No eye-contact | E: Eye-contact), Familiarity (U: Unfamiliar | F: Familiar), and Gender (♀: Female | ♂: Male)

However, significant differences appeared at the introduction of the categorization as ABOVE CHANCE and NOT ABOVE CHANCE. Table 5.4 and Table 5.3 gives the sum of deviation scores per Conditions, at all linguistic levels of representation (5), along with the number of dialogues involved in each division. The difference observable in these two tables are reflected in the visual representation of the spread of the deviation scores along the 0 to 200 scale in each of the division in Figure 5.7 and Figure 5.8. Figure 5.7[3] and Figure 5.8,[4] show the distribution of the dialogues along the *deviation scores* depending on Role, in interaction with significant OTHERSHARED and SELFSHARED p-values, Eye-contact, Familiarity and Gender for All n-grams and all linguistics Levels. A large effect size ($g = -0.92$) was found between female and male IF OTHERSHARED NOT ABOVE CHANCE p-values. In Figure 5.7 and Figure 5.8, the combination No eye-contact and Familiar subjects is related to the lowest *deviation score*, without strong effects from Gender, Speaker or Shared type, except for the combination OTHERSHARED NnEF♂, where the male IF not repeating significantly the IG have an average *deviation score* ($\bar{x} = 72$) higher than any other Familiar No eye-contact combination. However, in Eye-Contact situation for NOT ABOVE CHANCE OTHERSHARED repetitions, female IF have an average *deviation score* much higher ($\bar{x} = 135$), than male in the same conditions ($\bar{x} = 55.5$). (See NEF♀ and NEU♀ vs. NEF♂ and NEU♂, in Figure 5.7)

In addition to observe the distribution for All n-grams, a closer look at the notion introduced in section 4.4 is given: structural repetitions. Indeed, another large effect-size ($g = 0.89$, W= 44.5, $p = 0.18$) was detected for ABOVE CHANCE self-repetitions of the IG at Level 4 (POS) for $N2+$ n-grams $n > 1$, with pairs having a female IG obtaining better scores ($\bar{x} = 43.2$) than pairs with a male IG ($\bar{x} = 80.14$). It can be seen in Figure 5.8 that the ABOVE CHANCE SELFSHARED combination relates to lower scores for the IG, except for male participants in Eye-Contact settings. This tendency seems to indicate that if the participants cannot see each-other, self-repetition from the IG plays a role toward a higher task success.

[3] DevScore vs. Speakers | (OSPValue<=0.05)+ (EyeContact)+ (Familiarity)+ (Gender)
[4] DevScore vs. Speakers | (SSPValue<=0.05)+ (EyeContact)+ (Familiarity)+ (Gender)

5.2.3 Task-based Experience: The Influence of Familiarity

In the previous section (§ 5.2.2) the focus was on the repetition differences within gender, with or without eye-contact, and familiar with the interactant or not. From that study, a strong influence of familiarity between participants was noticed. The study in this section focuses on another aspect of the HCRC Map Task corpus, the influence that repeating the task may have on repetition patterns, with a closer examination at how the familiarity factor might impact this relation. As it is established in the previous section that familiar pairs obtained a higher success than unfamiliar pairs, and one may expect that experience will positively increase success over task attempt, one can wonder the way repetitions behave according to those expectations. If short-term repetition plays a role in communication in relation to task-success, then distinctive patterns of interacting linguistic features of repetitions and non-linguistics features of Experience and Familiarity should appear. In particular, where interlocutors are not familiar with each other, it is expected that the presence of significant above chance repetition will relate with task-based success. Similarly, where interlocutors are not familiar with their task, it is expected that significant amounts of repetition will relate to task-based success. This section first examines the influence Experience and Familiarity have on task success, then observes how repetitions impact these relations to determine which repetition patterns are more likely to relate to successful or unsuccessful communication, in the context of the map task.

Figure 5.9 allows us to observe the importance of Experience on the *deviation scores*. The first attempt having the highest average *deviation score* ($\bar{x} = 109.4$) by far in comparison to the next three attempts (Second: ($\bar{x} = 69$), Third: ($\bar{x} = 54.2$), Fourth: ($\bar{x} = 54.5$)). Figure 5.9 also shows us the difference in *deviation scores* between familiar and unfamiliar pairs of participants, that a Mann-Whitney-Wilcoxons test for population distribution found significantly different ($W = 6572$, $p = 0.00625$). This phenomenon is also clearly visible in Figure 5.10, with the first attempt displayed in the darkest shade of grey. The observation that the *deviation score* lowers as experience increase is also an indication of its suitability as task success indicator.

For Level 1 (Token only), no significant difference between dialogues with an above chance amount of repetitions (ABOVE CHANCE) and non-significant amount of repetition

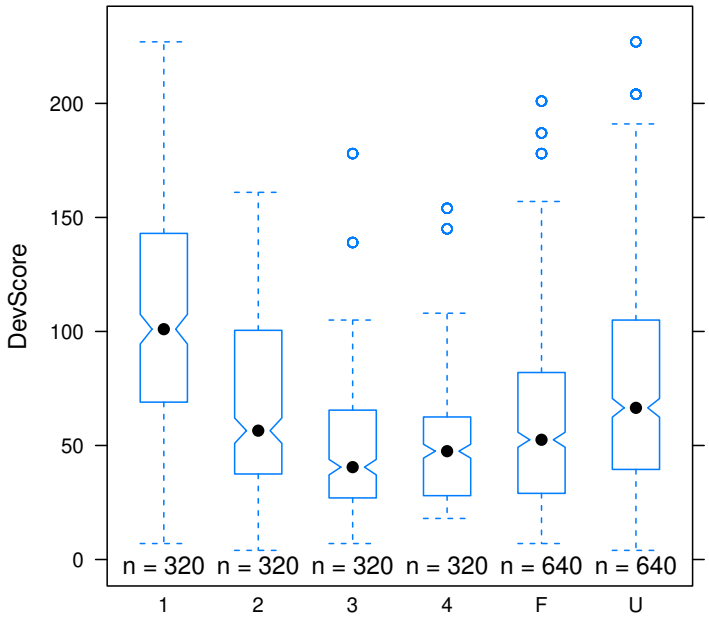


Figure 5.9: Distribution of Deviation Score by Experience (Attempt 1, 2, 3, 4), along with Familiarity (U: Unfamiliar | F: Familiar)

(NOT ABOVE CHANCE) during the first attempt was found, for both OTHERSHARED (W= 78.5, $p = 0.45$) and SELFSHARED (W= 73.5, $p = 0.14$). Even if a medium effect-size (Hedge test) was found for SELFSHARED (g estimate= -0.54). Whether significant differences between ABOVE CHANCE and NOT ABOVE CHANCE appeared for each linguistic level in isolation was tested and none was detected. However, when testing with all linguistics Levels (TOKEN (Level 1), LEMMA (Level 2), LEMMA+POS (Level 3), POS (Level 4), TOKEN+POS (Level 5) in combination, a significant difference was found (W= 7015.5, $p = 0.03$), relating ABOVE CHANCE to a lower deviation score ($\bar{x} = 105$) than NOT ABOVE CHANCE ($\bar{x} = 122.59$) and thus higher success.

The association plots in Figure 5.11 are displaying the relation between the sum of the *deviation scores*, Familiarity and the amount of repetitions (above chance or not) detected at all Levels. The Pearson's standardized residuals point out that for Unfamiliar pairs that repeat each other ABOVE CHANCE, the observed value is below expectations, indicating lower scores (therefore higher communicative success). For Familiar pairs, the observed value is above expectations. In those figures, as they display the sum of the *deviation scores*, results under the baseline for independence could relate to higher success as a low deviation score suggests so. For both cases chi-square tests indicate the association present between the

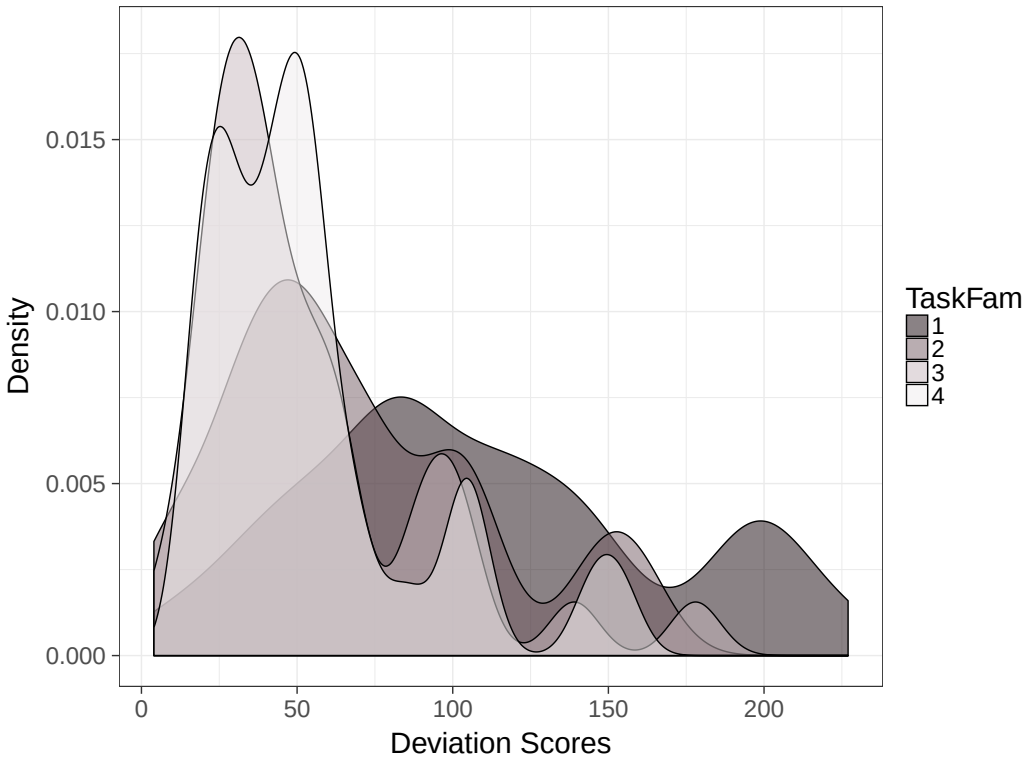
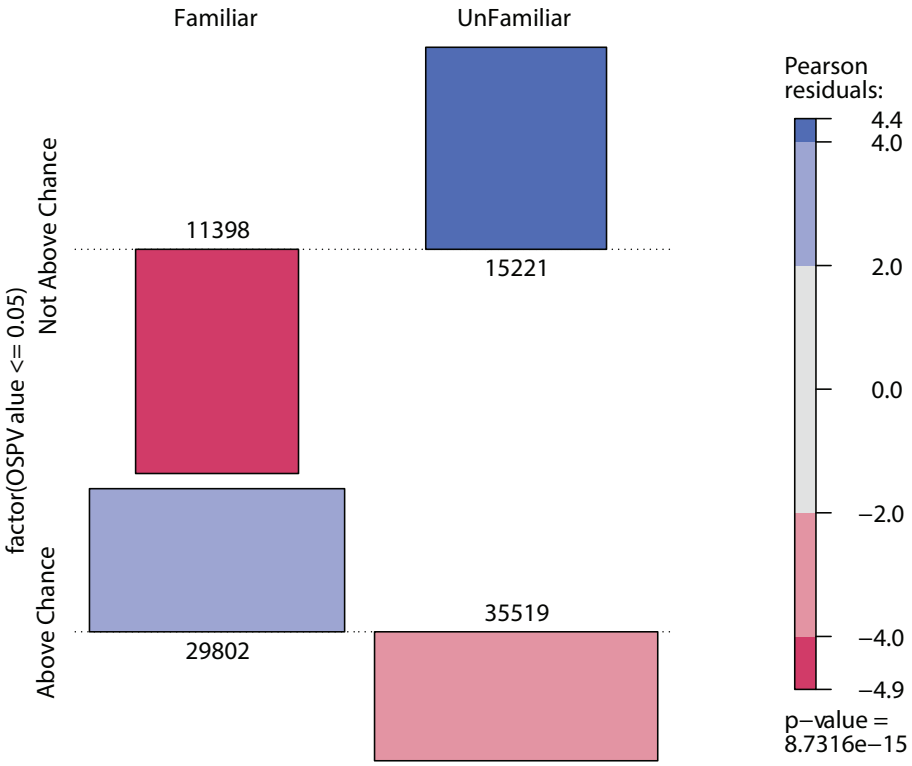


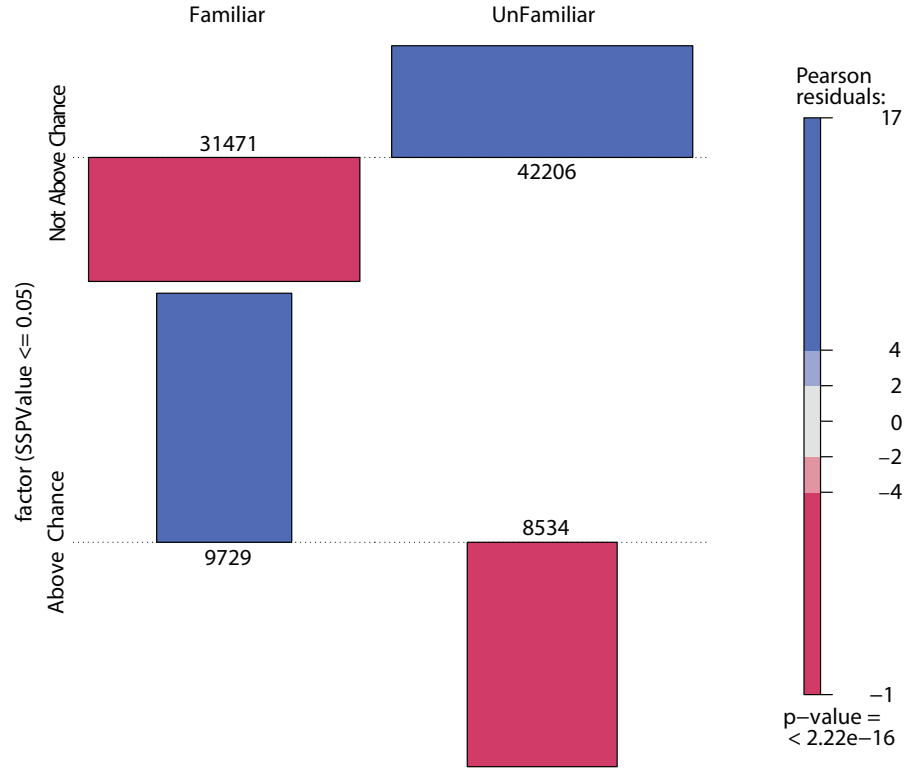
Figure 5.10: Density plot of Deviation Score per Experience (By grey shading, First Attempt: Dark grey to Fourth Attempt: Light grey). For each distribution $n = 32$.

variables (OTHERSHARED: $p = 8.7316e-15$; SELFSHARED: $p =< 2.22e-16$). This indicates the effect repetitions detected by the method are having a higher impact on unfamiliar pairs task success than on familiar pairs.

Among all levels and participants, a significant difference in *deviation score* distribution was seen between ABOVE CHANCE and NOT ABOVE CHANCE for both OTHERSHARED and SELFSHARED, as well as First Attempt in isolation but not for attempt 2 to 4. Figure 5.12 displays the distribution of the *deviation scores* between Familiar and Unfamiliar pairs. If Unfamiliar pairs seem to have on average a deviation score always higher in all conditions, a clear distinction between ABOVE CHANCE and NOT ABOVE CHANCE is observable at First Attempt. For OTHERSHARED no significant difference was found between Familiar pairs at First Attempt ($p = 0.106$), however, a difference was found for Unfamiliar pairs ($p = 0.039$), with ABOVE CHANCE having a lower mean ($\bar{x} = 123.41$) than NOT ABOVE CHANCE ($\bar{x} = 141.29$). A significant difference was found at second attempt for Familiar pairs ($p = 0.004$), with ABOVE CHANCE having a lower mean ($\bar{x} = 68.19$) than NOT ABOVE CHANCE ($\bar{x} = 93.74$), but not for unfamiliar pairs ($p = 0.106$). A combination of the *deviation score* distribution from

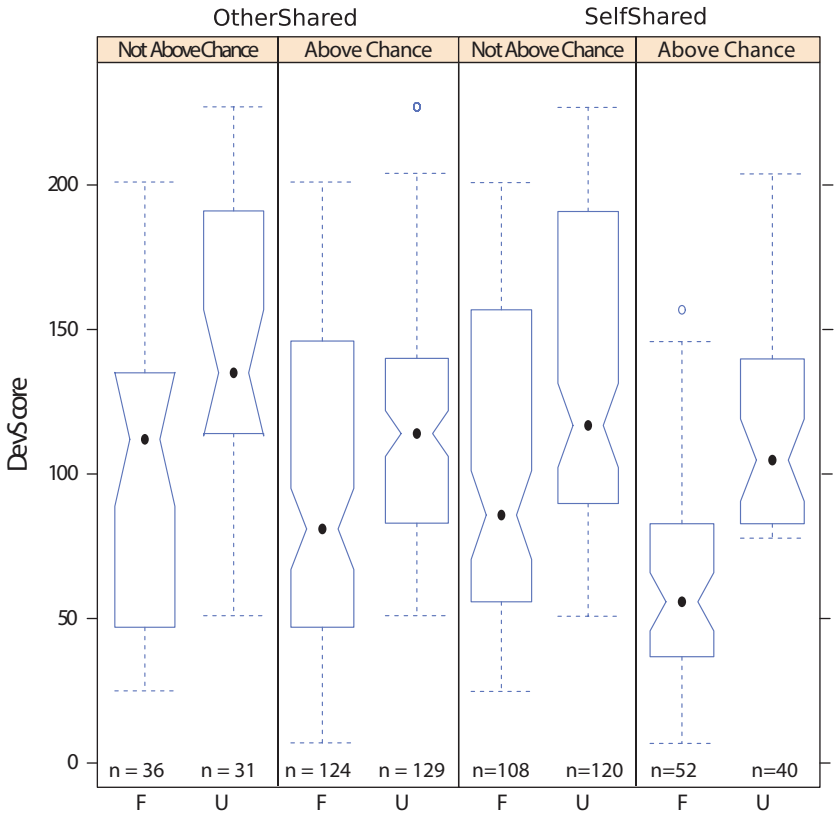


(a) OTHERSHARED

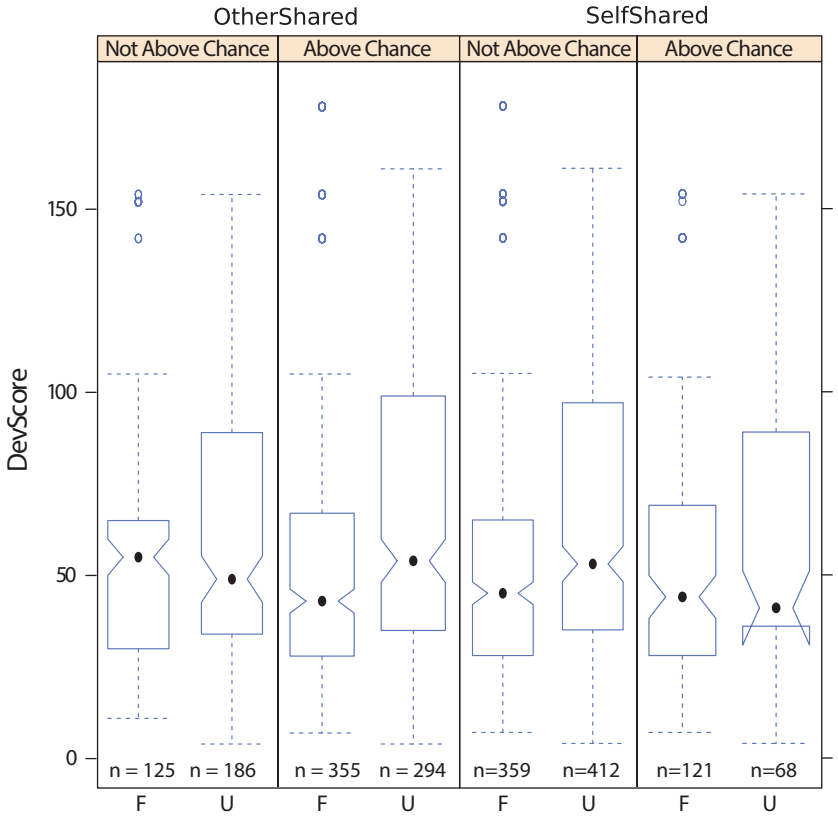


(b) SELFSHARED

Figure 5.11: Association Plot of significant p-values (Above Chance | Not Above Chance) and Familiarity



(a) First Attempt



(b) Attempt 2 to 4

Figure 5.12: Distribution of Deviation Score in interaction with Familiarity (U: Unfamiliar | F: Familiar) for OTHERSHARED and SELFSHARED

attempt 2 to 4 was found significant for both Familiar and Unfamiliar pairs, with ABOVE CHANCE having a lower mean ($\bar{x} = 53.42$) than NOT ABOVE CHANCE ($\bar{x} = 60.03$) for Familiar pairs, and with NOT ABOVE CHANCE having a lower mean ($\bar{x} = 58.43$) than ABOVE CHANCE ($\bar{x} = 66.97$). For SELFSHARED, no significant difference was found between ABOVE CHANCE and NOT ABOVE CHANCE, except for first attempt of familiar pairs ($p = 1.393e-05$), with ABOVE CHANCE having a higher mean ($\bar{x} = 105.85$) than NOT ABOVE CHANCE ($\bar{x} = 63.03$).

5.2.4 Conclusion

These three studies revealed a number of patterns in the presence or absence of repetitions above what could be considered as chance in the speech of partners carrying out a task in cooperation, that appear to be influenced by sociological factors. Firstly, it confirmed the importance of the task roles in the distribution of repetitions, and secondly, highlighted the differences that gender, eye-contact and mostly familiarity, bring to the outcome of a task as a function of the presence or absence of repetitions above chance. The extent to which these repetitions taken as cues of alignment can index mutual understanding are further discussed in detail in the chapter 6 and summarized in chapter 7. The following section is interested in another aspect found in the literature that can possibly influence the alignment between partners trying to perform a task: having an interaction mediated by a computer.

5.3 Interlingual Computer-Mediated Interactions

This section describes the use of the method in the context of computer-mediated interactions, and its comparison with the previously used corpus of Human-to-Human dialogues. To standardise the data with the ILMT-s2s corpus (§ 3.3), only dialogues that used the same maps (maps 1 & 7) were kept in this study, resulting in 16 dialogues from the HCRC Map Task corpus (Subset 1). The literature (see section 2.4) leads us to expect that :

(1) A greater alignment will be found in computer-mediated interactions than in Human-to-Human interactions.

(2) To see variations in the amount of repetitions happening above chance depending on the amount of negative cognitive states observed in the participants.

The ILMT-s2s corpus is therefore compared in terms of OTHERSHARED and SELFSHARED with the HCRC Subset 1, and the amounts of negative cognitive states present in the ILMT-s2s corpus are examined in relation to the presence or absence of OTHERSHARED and SELFSHARED repetitions at the different levels of granularity that the method allows (chapter 4). Table 5.5 gives a summary of the number of repetitions per conditions.

Table 5.5: HCRC Map Task Subset 1 and s2s-ILMT Corpus Summary; SELF REP and OTHER REP (see § 4 for the definition) are given for the linguistic representation level token only.

	HCRC Subset 1			ILMT-s2s					
Language	English			English			Portuguese		
	IG	**IF**	**Total**	**IG**	**IF**	**Total**	**IG**	**IF**	**Total**
OTHER REP	1,383	1,252	2,635	2,660	2,154	4,814	1,216	998	2,214
SELF REP	1,663	782	2,445	4,864	2,890	7,754	2,582	2,030	4,612

5.3.1 Human-to-Human vs. Computer-Mediated

The null hypothesis (H_0), with the threshold of $p \geq 0.05$, was rejected 233 times out of 300 for OTHERSHARED and 273 times out of 300 for SELFSHARED in the ILMT-s2s corpus across all linguistic levels while in the data from the HCRC Map Task, OTHERSHARED was rejected 111 times out of 160 and SELFSHARED was rejected 25 times out of 160 (Table 5.6).

This reveals a considerable difference in the rejection rate for SELFSHARED repetitions between the direct human-to-human dialogues of the HCRC Map Task corpus (25/160 = 15%) and those of the ILMT-s2s corpus (273/300 = 91%), with SELFSHARED repetitions happening above chance more often in the computer-mediated corpus. A Mann-Whitney-Wilcoxon test found that across all linguistic levels, the number of SELFSHARED repetitions is significantly different ($p = 2.686e-06$) between the HCRC Map Task (with an average rejection of $\bar{x} = 2.5$) and the ILMT-s2s corpus (with an average rejection of $\bar{x} = 13.65$). However, no significant difference ($p = 0.9636$) was found between the two corpora concerning OTHERSHARED repetitions at level *n*-grams = All, both corpora showing a high rate of rejection of H_0. No significant difference was found between the two corpora in terms of speaker role, language spoken, and eye-contact modality at level *n*-grams = All.

Table 5.6: Rejection count of H_0 for levels L1 to L5 and mean (*M*) values in the ILMT-s2s corpus and HCRC Map Task corpus for all *n*-grams. For each dialogue at each level, the number of possible H_0 rejection is 15 in the ILMT-s2s corpus, and 16 in the HCRC Map Task corpus.

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

5.3.2 Within Computer-Mediated Interactions

No impact of above chance repetitions in relation to the cognitive states of the participants was found at *n*-grams length *n* = All (count listed in Table 5.7). However, differences appeared for OTHERSHARED repetitions of Portuguese (IF) at *n*-gram length *n* >1 (N2+) in

Eye-contact conditions (Table 5.8). While in all other settings the rate of rejections of H_0 remains high, the Portuguese IF speakers did not repeat the English speakers' words in the same proportion in the Eye-contact conditions.

Table 5.7: Number of Cognitive States per Subject Role (Information Follower, Information Giver), Spoken Languages (English, Portuguese) and Cognitive State Type (Frustrated, Surprised, Amused) in the ILMT-s2s corpus

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

Table 5.8: Rejection count of H_0 for levels L1 to L5 and mean (M) values. In each case the number of possible H_0 rejection is 8 (modality: eye-contact).

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

This relation is highlighted with Pearson's standardized residuals from log-linear models in Figure 5.13. For long sequences of n-gram repetitions (N2+), when there is Eye-contact, the Portuguese speakers show higher levels of negative cognitive states than expected when they are at the same time not repeating the English speaker. Meanwhile they show less frustration than expected if they repeat the English speaker for long sequences (N2+). A closer look to the location of the rejection made us exclude the possibility that it was only a dialogue or two that where responsible for the differences, but on the contrary the rejections were scattered over different dialogues. The distributions of negative cognitive states was found significantly different between ABOVE CHANCE and NOT ABOVE CHANCE OTHERSHARED repetitions for the Portuguese IF speakers at n-gram>1 level (W = 883, p-value = 0.027). The low rate of N2+ repetitions detected is echoed in the user survey conducted in the ILMT-s2s corpus. The Portuguese speakers (IF) in Eye-contact conditions showed the lowest appreciation of the system (Median score = 3.0; Overall Median score = 5.0), which coincide

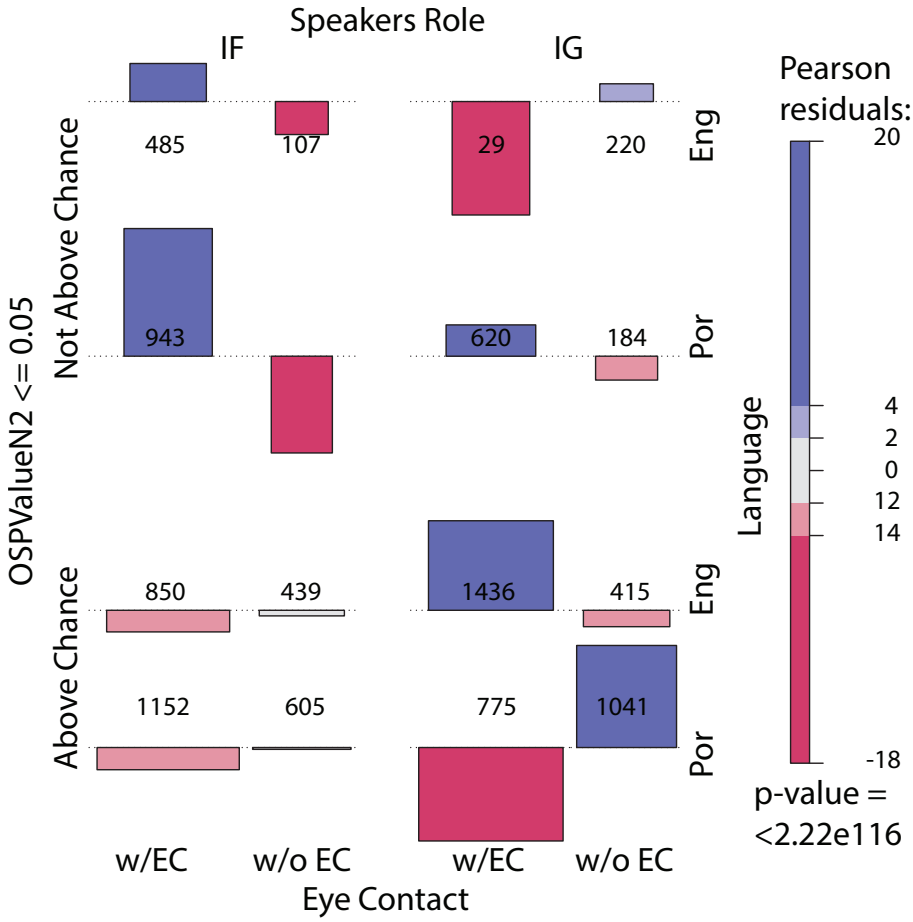


Figure 5.13: Association Plot of significant OTHERSHARED residuals (Above Chance: $p <= 0.05$ | Not Above Chance: $p > 0.05$) for n-gram>1 (N2+), Subject Role (IG: Information Giver | IF: Information Follower), Eye-Contact (w/ EC: with Eye-contact | w/o EC: no eye-contact), and Language Spoken (En: English | Pt: Portuguese)

with a high amount of negative cognitive states for those speakers.

5.3.3 Conclusion

This section presented the comparison in terms of repetitions happening above chance between an human-to-human corpus and an interlingual computer-mediated corpus. Similar cues of alignment were found in both corpora, however, computer-mediated interactions held more above chance self-shared repetitions than human-to-human interactions. Another pattern, that indicates the potential usefulness of the method for the monitoring of computer-mediated interactions, was a low presence of other-shared repetitions in the group of speakers who were also observed as having difficulties with the system. These patterns are discussed further in the next chapter. The following section investigates the possible differences in patterns of repetitions if the interaction is mediated not by a computer, but by a human that plays the role of a facilitator.

5.4 Third Party Assessment Interactions

This section describes the fourth study conducted in the exploration of the method. Three-party game-based interactions, where two players participate in a quiz, while supervised by a facilitator, are analysed and the relation between repetition patterns and the type of the facilitator's feedback is investigated. Table 5.9 and Table 5.10 gives a summary of the number of repetitions per conditions and dialogue sections.

Table 5.9: MULTISIMO Summary of repetitions per conditions: SELF REP and OTHER REP (see § 4 for the definition) are given for the linguistic representation level token only. The Facilitator speech is taken out of the count for gender. There are 10 female/ male, 6 female only and 7 male only dialogues.

	Participants	**Facilitator**	**Female**	**Male**
OTHER REP	2,974	958	1,493	1,481
SELF REP	2,038	858	1,099	939

Table 5.10: MULTISIMO Summary of repetitions per dialogue sections, for the participants only; SELF REP and OTHER REP (see § 4 for the definition) are given for the linguistic representation level token only.

	Full	**Q1**	**Q2**	**Q3**	**Answer**	**Ranking**
OTHER REP	2,974	1,093	907	961	1,600	1,055
SELF REP	2,038	755	592	685	1,210	815

5.4.1 Dialogue Length Variations

An important interest in the exploration of this method is the time length over which the potential alignment is occurring. Despite not being a map task, which is the privileged type of corpus used in this thesis (which allows for a good comparison between variables), the corpus used in this section also contains tasks that participants are required to solve, but clearly divided in parts. An aspect that is difficult to model in the map tasks is how to divide the dialogues into shorter segments of an individual task. One could think of dividing the dialogues by "landmark passed". However, after a closer examination of the dialogues, it becomes apparent that such segments have blur limits. Indeed, a wide range of situations can happen that leads them to interrupt smaller individual tasks, for example, participants

often go back to previous landmarks if they notice that they made an error, or cite landmarks simply to discard them. The map task makes the contours of possible dialogue segments (by each smaller task) blur. The MULTISIMO corpus offers an interesting feature to this respect: interlocutors have to accomplish the same task (divided into two distinct parts) three times. This configuration allows for delimited cuts of small, clearly divided tasks, within a longer interaction.

Given the playful nature of the task which is modelled on that of a television game show, *Family Feud*, and the approach to facilitation provided (see section 3.4), substantial quantities of patently negative feedback from facilitators are not to be expected. Rather, in the context of the task, contributions from the facilitator might either tend towards introducing participants to discrete phases of the interaction (and therefore be deemed neutral) or will be positively encouraging. However, encouragement for a task is a natural response to a perception of communication difficulties. Therefore, for this data set, expectations are fewer positive and more neutral facilitator contributions in contexts where interlocutors experience success, and more (encouraging) positive contributions in contexts where interlocutors experience difficulty.

Obtaining objective measures of success regarding human communication is delicate. We believe that the feedback given by the facilitator represents a continuous assessment of the ongoing success of the interaction and the success in the task the two players were given to achieve. Whether repetitions happened ABOVE CHANCE or not within each dialogue section, or if they perform a meaningful role, signalling alignment among players; and whether the degree of alignment among the two players is reflected in the facilitator's feedback was also investigated. Interactional facilitators' style has an impact on dialogue outcomes (e.g. "supportive" vs. "oppositional" on qualities of reflection (Cacciamani, Cesareni, Martini, Ferrini, & Fujita, 2012), "task oriented" vs "socially oriented" on perceptions of efficacy (van Dolen, de Ruyter, & Carman, 2006)). The amounts and types of facilitators' feedback to the players is chosen here as a measure of interactional success to be related to cues of alignment in the players' speech.

For this experiment, the method described in chapter 4 was applied to the dialogues cut by sections: *Full*, *Question*, *Answer*, and *Ranking*, as mentioned before, to observe if the section

type, by their nature and length, show variations in amount of repetitions happening above chance or not. Here again for each section, whether significantly more repetition appears in the actual dialogue sections than in the randomised dialogue sections is determined. Our hypothesis are that:

(1) The divisions in dialogue sections should expose different patterns of above chance repetitions at different dialogue lengths, if they are present. We expect that longer sections of dialogues will exhibit larger amounts of above chance repetitions.

(2) The alignment detected by the method should be reflected in the facilitator's feedback. On one hand, lack of mutual understanding signalled by a lack of alignment between the players would to correspond to a high amount of positive feedback. On the other hand, a scarcity of positive feedback (with the presence of negative or neutral feedback) is expected where there is substantial evidence of participant's alignment.

Although three categories of feedback were annotated, two were grouped together (negative and neutral) to form a binary opposition: positive vs non-positive feedback. This approach toward neutral feedback (which consists of the facilitator's elaboration or elicitation of players' contribution to the dialogue) is consistent with a view of the content of neutral feedback as simple game guidance. Neutral questions, such as *"Is that your final decision?"*, without indications that the participants are going in the correct direction or not, are here considered likely when no great difficulties in communication are perceived by the facilitator.

5.4.2 Above Chance Repetitions and Facilitators' Feedback

This section focuses on the repetition behaviours of the two participants and the possible interaction with the facilitators' feedback.

For the *Full* dialogues, at the Level Token, following a threshold of ($p \leq 0.05$), the Null Hypothesis was rejected 30 times over 46 for OTHERSHARED and 27 over 46 for SELF-SHARED, for all *n*-grams (1 to 5) for the two players, which shows that there was a slightly higher proportion of significant OTHERSHARED repetitions in the corpus. The detail of the rejections of H_0, per dialogue section, linguistic representation level and repetition types can

Table 5.11: Rejections of H_0 for OTHERSHARED and SELFSHARED, at All n-grams and N2+, the Total is the sum of of rejections of H_0 across the five linguistic levels, Possible Rej. is the number of Possible Rejections per cell in each level, see § 5.4.2.

	All n-grams					Total	**N2+ (n-grams, $n>1$)**					Total	Possible
Level	Tok	Lem	LemPPOS		TokP		Tok	Lem	LemPPOS		TokP		Rej.
OTHERSHARED													
Full	30	31	29	14	30	134	19	23	19	14	21	96	46
Question	37	42	47	20	51	197	20	17	20	7	20	84	138
Answer	19	21	22	5	24	91	8	9	7	4	6	34	138
Ranking	10	6	13	3	13	45	3	2	1	0	1	7	138
SELFSHARED													
Full	27	25	28	12	29	121	27	27	28	12	29	123	46
Question	20	17	20	7	20	84	15	19	15	15	11	75	138
Answer	6	4	4	2	4	20	6	4	4	2	4	20	138
Ranking	34	29	31	13	31	138	3	2	1	0	1	7	138

be found in Table 5.11. For the *Full* dialogues, in each case the Null Hypothesis can potentially be rejected 46 times, as there are 2 speakers in 23 dialogues. For the other dialogue sections, the Null Hypothesis can potentially be rejected 138 times, as each section is repeated 3 times. For OTHERSHARED repetitions, the rate of rejection of the null hypothesis is the highest in the *Full* dialogues, and decreases as the dialogue sections shorten (see Table 5.11). For SELFSHARED repetitions, the rate of rejection is also the highest for the *Full* dialogues; however, the section *Ranking* contains a higher rate of rejections despite being the shortest dialogue section. Since this pattern is not present in longer sequences of n-grams (N2+), one can conclude a high rate of lexical unigram repetitions in those sections.

A binary classification of the facilitators' feedback was adopted: positive and non-positive (negative and neutral), as described in chapter 4. Figure 5.14, shows that when there are ABOVE CHANCE OTHERSHARED repetitions, the amount of positive feedback is less than one would expect and non-positive feedback is in greater amount than one would expect if there were no interaction between the categories of facilitator feedback and the degree of repetition in the dialogue. Conversely, where OTHERSHARED repetitions are at a level that is NOT ABOVE CHANCE, there is more positive feedback than one would expect and less non-positive feedback than one would expect if there were no interaction between feedback type and degree of repetition.

Using Mann-Whitney-Wilcoxon tests, the following pattern was observed for the *Full* dialogues: the amount of positive feedback found in the dialogues categorized as ABOVE

CHANCE OTHERSHARED repetitions was significantly different from the amount found in NOT ABOVE CHANCE (W= 4092, $p = 2.487e-06$), with ABOVE CHANCE accompanied by less positive feedback ($\bar{x} = 43.11$) and NOT ABOVE CHANCE accompanied by more positive feedback ($\bar{x} = 50.44$). The Mann-Whitney-Wilcoxon test applied to the amount of non-positive feedback between ABOVE CHANCE and NOT ABOVE CHANCE *Full* dialogues did not return a significant result. The same observations were made for the *Question* sections: the amount of positive feedback found in the dialogues categorized as ABOVE CHANCE OTHERSHARED repetitions was significantly different from the amount found in NOT ABOVE CHANCE (W= 39046, $p = 5.33e-05$), relating ABOVE CHANCE to on average less positive feedback ($\bar{x} = 14.21$) and NOT ABOVE CHANCE to on average more positive feedback ($\bar{x} = 15.86$). No significant difference was found for the amount of non-positive feedback between ABOVE CHANCE and NOT ABOVE CHANCE. With respect to the *Ranking* sections, the amount of positive feedback found in the dialogues categorized as ABOVE CHANCE OTHERSHARED repetitions was not significantly different from the amount found in NOT ABOVE CHANCE (W= 13602, $p = *0.4768*$). No significant difference was found for the amount of non-positive feedback between ABOVE CHANCE and NOT ABOVE CHANCE.

In the *Answer* section type, the amount of positive feedback was not found to be significantly different depending on above chance repetitions, while the amount of non-positive feedback was (W= 32320, $p = 0.004$). The answer sections with ABOVE CHANCE levels of repetition have more non-positive feedback ($\bar{x} = 17.62$) and the sections with NOT ABOVE CHANCE repetition have less non-positive feedback ($\bar{x} = 14.56$).

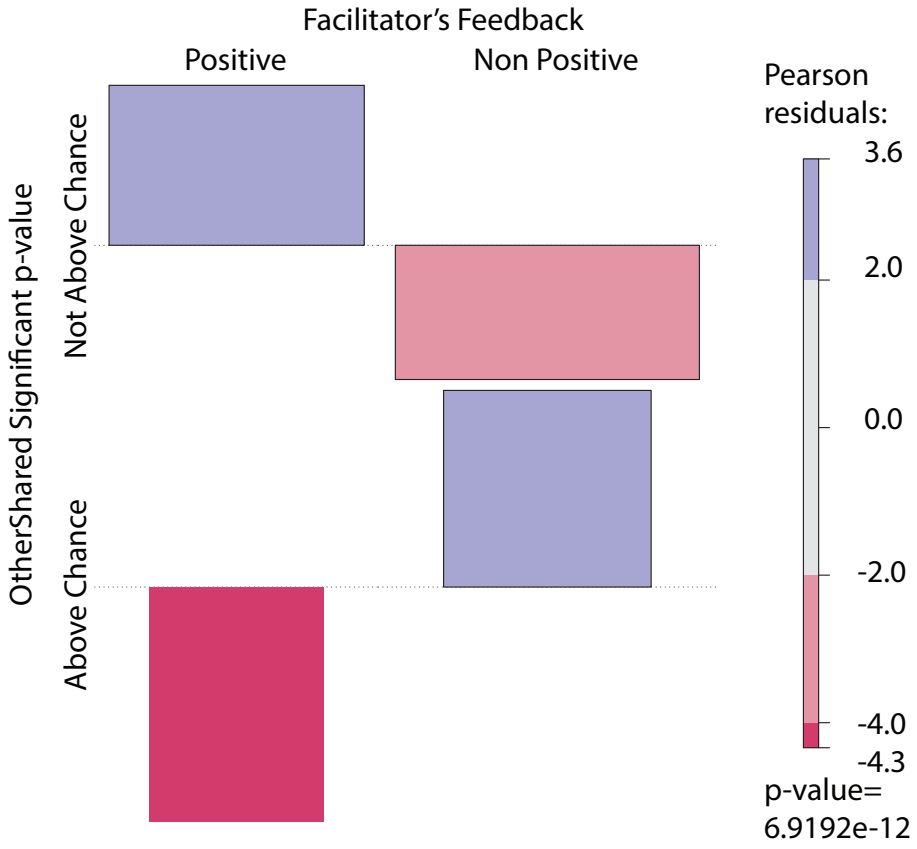


Figure 5.14: Association Plot of significant OTHERSHARED p-values (Above Chance | Not Above Chance) and Facilitator's feedback (All Positive | All Non-Positive) across the *Full* dialogues

5.4.3 Conclusion

The above study presented the application of the method on a corpus of human-mediated interactions with specifically cut sub-tasks that allowed the investigation of different lengths of task-based interactions. The amounts of repetitions happening above chance were reflected in the behaviour of the dialogue facilitator: where alignment was detected by the method, less than expected positive feedback was provided, where a lack of alignment cues was found, more positive feedback than would be expected by chance was found. These patterns are discussed in greater extent in the next chapter (see chapter 6). This study's participants were a mix of native and non-native speakers, from different geographical regions and therefore potentially used different dialects of English, without the possibility in the corpus design to control for the potential impact these different dialects may have on the mutual understanding of the participants. The section below explores yet another variable that could influence the alignment between dialogue partners: the English dialects.

5.5 Variations Across Dialects

This section describes the fifth study conducted in the exploration of the method. It investigates the extent to which different dialects of the English language might impact repetition patterns, in relation to communicative success. Two American map task corpora and two subsets of matching conditions extracted from the HCRC Map Task (which is in Scottish English) are used. As mentioned in chapter 2, the socio-cultural context of an interaction has an important impact on conversational style. Do the variations induced by the use of different dialects of English have a significant impact on repetition patterns? Are repetition patterns related to successful communication interacting in a similar manner?

Two corpora that are directly inspired and emulate comparable conditions (medium and task-type) of the HCRC Map Task are analysed: The MIT American English Map Task corpus (16 dialogues), recorded in 1999, between familiar female participants and the PARDO 2006 Map Task corpus (10 dialogues), recorded in 1998, between unfamiliar female participants.[5] These materials are both recorded in No eye-contact conditions. Two subsets of the HCRC corpus were extracted to match these conditions, the HCRC Subset 2 (Familiar, Female, No eye-contact) which amounts to 18 dialogues, and the HCRC Subset 3 (Unfamiliar, Female, No eye-contact) which amounts to 14 dialogues. This represents a quarter of the full HCRC Map Task. The first subsection analyses the corpora in which the pairs are Familiar (AEMT & HCRC SUB2) and the second subsection analyses the corpora in which the pairs are Unfamiliar (PARDO & HCRC SUB3). Table 5.12 and gives a summary of repetitions in the two corpora and the corresponding subsets HCRC of the per speaker's roles.

[5] See chapter 2 for more complete descriptions.

Table 5.12: American English Map Tasks Summary of repetitions per speaker's roles.

	AEMT			HCRC Subset 2		
	IG	**IF**	**Total**	**IG**	**IF**	**Total**
OTHER REP	1,292	1,144	2,436	1,906	1,554	3,460
SELF REP	2,643	1,202	3,845	2,068	1,136	3,204

Table 5.13: PARDO 2006 Map Task Corpus Summary of repetitions per speaker's roles.; SELF REP and OTHER REP (see § 4 for the definition) are given for the linguistic representation level token only.

	PARDO			HCRC Subset 3		
	IG	**IF**	**Total**	**IG**	**IF**	**Total**
OTHER REP	1,176	990	2,166	802	724	1,526
SELF REP	2,352	967	3,319	1,234	378	1,612

5.5.1 American vs. Scottish English: Familiar Partners

The *deviation scores* computed from the Method 4, described in section 4.5.3, of which the higher the score the more unsuccessful the communication is assumed, ranged from 2 to 322 in the AEMT and 24 to 293 in the HCRC Subset 2.[6] For each dialogue, the method described in chapter 4 is applied to establish a proxy measure of mutual understanding (whether repetition levels exceeded chance, leading to H_0 rejection, $p \geq 0.05$: this yields the meta-analysis categorization: ABOVE CHANCE or NOT ABOVE CHANCE. According to patterns found previous studies, given that the only main differences are the English dialect and a 7 years gap between the recordings, the following results are expected:

(1) The previously found patterns of ABOVE CHANCE repetitions are going to be similar in the AEMT corpus than in the HCRC Subset 2.

(2) As the AEMT is a corpus of familiar speakers, OTHERSHARED ABOVE CHANCE repetitions should not relate to higher success in the interaction.

Overview of the Results

Contrary to expectations, differences in terms of rate of rejection to H_0 appeared between the AEMT and the HCRC Subset 2 for both OTHERSHARED and SELFSHARED, as can be seen in the Table 5.14. The Information Giver repeated ABOVE CHANCE the Information Follower on average less in AEMT ($\bar{x} = 5.2$) than in HCRC Subset 2 ($\bar{x} = 12.2$). The Information Follower appears to have repeated the Information Giver a similarly high proportion in both corpora. The low proportion of self-repetitions found in the HCRC Subset 2 is not echoed in the AEMT, where the rate of rejection of H_0 is high for both roles. Structural repetitions are more frequent than lexical repetitions, in both corpora, for both roles and in OTHERSHARED and SELFSHARED. The next two subsections detail these results in regard to task success and task experience.

[6] To keep the comparison possible with previous experiments using the pre-computed *deviation scores* by the authors of the HCRC Map Task, they are also reported next to the M4 scores.

Table 5.14: Rejections of H_0 for OTHERSHARED and SELFSHARED in the AEMT and the HCRC Sub2, by to roles (IF: Information Follower; IG: Information Giver), in each case (each cell) the Null Hypothesis can potentially be rejected 16 times in the AEMT and 18 times in the HCRC Subset 2.

OTHERSHARED

AEMT							**HCRC Subset 2**						
All n-grams | H_0 : $Rand.Speaker.Level - Actual.Speaker.Level \geq 0$													
Level	Tok	Lem	LemP	POS	TokP	Mean	Level	Tok	Lem	LemP	POS	TokP	Mean
IF	13	13	11	12	12	12.2	IF	16	16	15	11	15	15
IG	**6**	**5**	**8**	**2**	**5**	**5.2**	**IG**	**13**	**14**	**14**	**8**	**12**	**12.2**
N1: n-gram=1 | H_0 : $Rand.Speaker.Level.N1 - Actual.Speaker.Level.N1 \geq 0$													
IF	7	8	4	4	6	5.8	IF	12	11	11	7	11	10
IG	3	3	5	0	4	3	IG	10	9	9	2	9	7.8
N2+: n-gram>1 | H_0 : $Rand.Speaker.Level.N2+ - Actual.Speaker.Level.N2+ \geq 0$													
IF	14	13	10	11	12	12	IF	17	17	15	12	17	16
IG	10	11	12	4	11	9.6	IG	16	16	16	11	15	15

SELFSHARED

AEMT							**HCRC Subset 2**						
All n-grams | H_0 : $Rand.Speaker.Level - Actual.Speaker.Level \geq 0$													
Level	Tok	Lem	LemP	POS	TokP	Mean	Level	Tok	Lem	LemP	POS	TokP	Mean
IF	**13**	**11**	**13**	**10**	**12**	**11.8**	IF	5	6	5	3	7	5.2
IG	**14**	**10**	**12**	**6**	**11**	**11**	IG	3	4	4	0	4	3
N1: n-gram=1 | H_0 : $Rand.Speaker.Level.N1 - Actual.Speaker.Level.N1 \geq 0$													
IF	9	7	9	4	9	7.6	IF	1	3	3	0	3	2
IG	6	5	8	2	5	5.2	IG	1	0	2	0	1	0.8
N2+: n-gram>1 | H_0 : $Rand.Speaker.Level.N2+ - Actual.Speaker.Level.N2+ \geq 0$													
IF	12	11	12	11	11	11.4	IF	5	6	6	4	5	5.2
IG	12	10	11	10	11	11	IG	6	8	7	2	7	6

AEMT vs HCRC Subset 2: Task Success

When compared to the task success measure, the divergences found in terms of rate of rejection are interpretable in the light of previously found patterns. The results of Mann-Whitney-Wilcoxon tests can be seen in Table 5.15, Table 5.16 and Table 5.17. For OTHERSHARED repetitions, both corpora show similar patterns in regards to task success: a significant difference between ABOVE CHANCE and NOT ABOVE CHANCE at all levels for all n-grams. In the AEMT ($p = 0.004$), ABOVE CHANCE related to lower task success ($\bar{x}$ =104.66) , and NOT ABOVE CHANCE to higher task success ($\bar{x}$ =61.15), and similarly in the HCRC Subset 2 (p-value=0.001), ABOVE CHANCE is related to lower task success ($\bar{x}$ =101.31) , and NOT ABOVE CHANCE to higher task success ($\bar{x}$ =70.52). Those similarities were echoed for both IG and IF. No significant difference was found at individual linguistic levels of representation (Token only, Lexical nor Phrasal) for both corpora. For SELFSHARED repetitions, no significant difference was found in the AEMT corpus. While in the HCRC Subset 2, a sig-

nificant difference between ABOVE CHANCE and NOT ABOVE CHANCE at all levels for all n-grams was found for the IG ($p = 0.014$), relating ABOVE CHANCE to higher task success ($\bar{x}$ =57.66), and NOT ABOVE CHANCE to lower task success ($\bar{x}$ =100.6).

Table 5.15: Summary of Wilcoxons Tests AEMT for OTHERSHARED and SELFSHARED, Deviation Method = M4 *(Note: The tests which involved less than 5 dialogues on either side of the tests are not considered for the results, as it makes the comparison unreliable)*

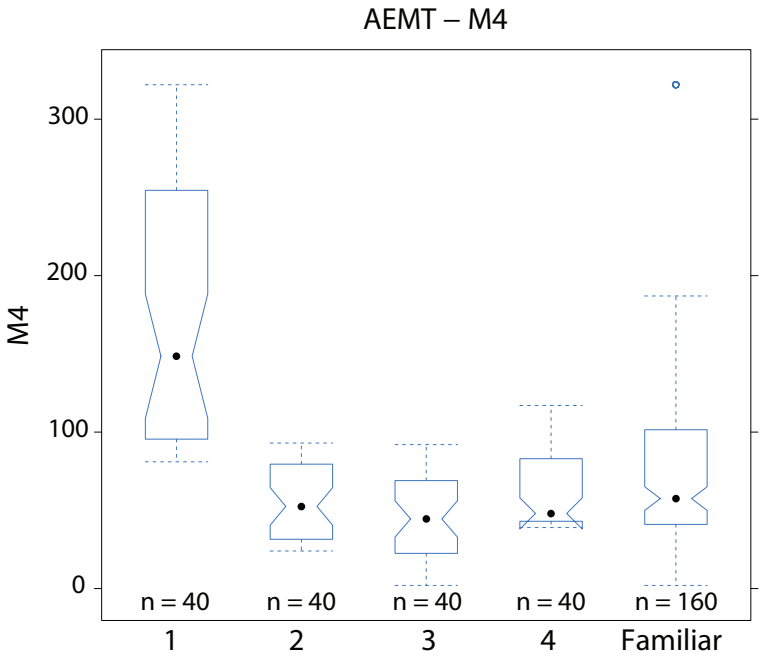
AEMT

	OTHERSHARED						SELFSHARED					
	All Level – All n-grams											
Speaker	IF + IG		**IG**		IF		IF + IG		IG		**IF**	
Wilcox. W	4005.5		914.5		782		2863		780.5		642	
p-value	**0.004**		**0.029**		**0.022**		0.514		0.510		0.809	
Hedges's g	0.6 (med.)		0.84 (large)		0.63 (med.)		0.32 (small)		0.31 (small)		0.33 (small)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	87	73	26	54	61	19	112	48	53	27	59	21
Mean Dev.	104.66	61.15	125.11	65.4	95.95	49.05	92.13	67.72	92.94	68.85	91.40	66.28
	L1 – All n-grams											
Wilcox. W	145.5		34.5		26		60.5		14		16	
p-value	0.408		0.664		0.419		0.735		1		0.686	
Hedges's g	0.54 (med.)		0.68 (med.)		0.58 (med.)		0.21 (small)		0.15 (negl.)		0.15 (negl.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	19	13	6	10	13	3	27	5	14	2	13	3
Mean Dev.	101.63	60.23	118.66	64.5	93.76	46	87.51	70.2	87.78	64	87.23	74.33
	L1 – n-grams n=1 (Lexical)											
Wilcox. W	155		25.5		48		123.5		25		34.5	
p-value	0.069		0.459		0.09		0.894		0.625		0.791	
Hedges's g	0.82 (large)		1.07 (large)		0.7 (med.)		0.3 (small)		0.13 (negl.)		0.44 (small)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	10	22	3	13	7	9	15	17	6	10	9	7
Mean Dev.	126.7	65.77	151.33	69.46	116.14	60.44	97.4	73.70	92.16	80.40	100.88	64.14
	L4 – n-grams n >1 N2+ (Phrasal)											
Wilcox. W	165.5		36.5		34		125.5		31		31.5	
p-value	0.15		0.14		0.49		0.7		0.95		0.69	
Hedges's g	0.78 (med.)		1.46 (large)		0.56 (med.)		0.39 (small)		0.34 (small)		0.4 (small)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	15	17	4	12	11	5	21	11	10	6	11	5
Mean Dev.	115.46	57.76	160.25	59.66	99.18	53.20	95.47	64.45	95.6	66.83	95.36	61.60

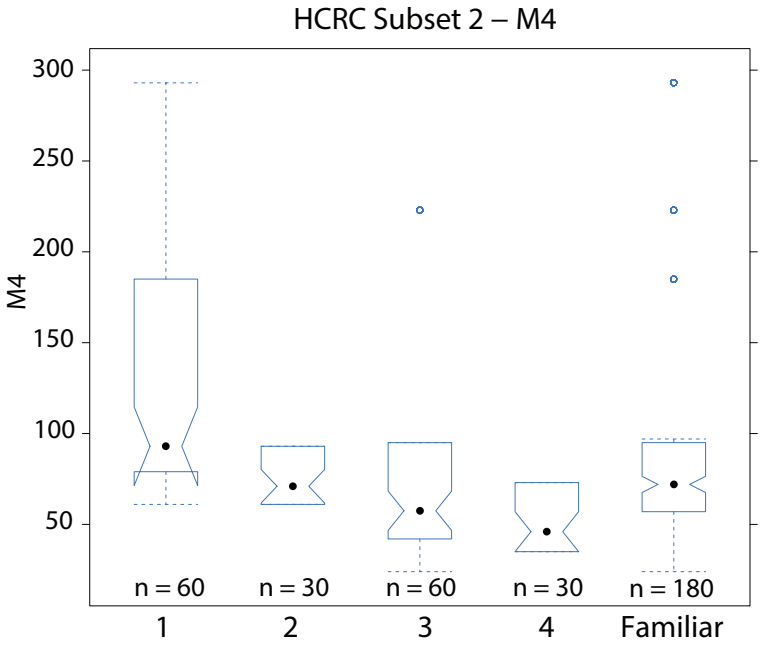
Table 5.16: Summary of Wilcoxons Tests HCRC Sub2 for OTHERSHARED, Deviation Method = M4 and HCRC precomputed Dev scores *(Note: The tests which involved less than 5 dialogues on either side of the tests are not considered for the results, as it makes the comparison unreliable)*

HCRC Sub2 – OTHERSHARED

	All Level – All *n*-grams											
	M4						HCRC DevScore					
Speaker	**IF + IG**		IG		IF		**IF + IG**		**IG**		**IF**	
Wilcox. W	4047		1072		773		4047		1162		825.5	
p-value	**0.001**		0.105		0.116		**0.001**		**0.016**		**0.034**	
Hedges's g	0.45 (small)		0.46 (small)		0.46 (small)		0.5 (med.)		0.52 (med.)		0.49 (small)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	134	46	61	29	73	17	134	46	61	29	73	17
Mean Dev.	101.3	70.52	103.5	72.1	99.41	67.82	74.5	47.76	76.65	48.75	72.69	46.05
	L1 – All *n*-grams											
Wilcox. W	135.5		44.5		21		144.5		49		21	
p-value	0.179		0.256		0.527		0.088		0.114		0.52	
Hedges's g	0.58 (med.)		0.617 (med.)		0.5 (med.)		0.64 (med.)		0.71 (med.)		0.5 (medium)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	29	7	13	5	16	2	29	7	13	5	16	2
Mean Dev.	101.3	60.57	105.9	61	97.68	59.5	74.48	39.42	78.84	38.6	70.93	41.5
	L1 – *n*-grams *n*=1 (Lexical)											
Wilcox. W	185		42.5		49		195		46		50.5	
p-value	0.321		0.858		0.241		0.187		0.624		0.189	
Hedges's g	0.276 (small)		-0.066 (negl.)		0.663 (med.)		0.296 (small)		-0.078 (negl.)		0.727 (med.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	22	14	10	8	12	6	22	14	10	8	12	6
Mean Dev.	101	81.42	91.2	96.25	109.3	61.66	74.09	57.57	65.6	70.25	81.16	40.66
	L4 – *n*-grams $n > 1$ N2+ (Phrasal)											
Wilcox. W	162.5		35.5		45.5		168.5		43		41	
p-value	0.68		0.82		0.399		0.541		0.716		0.673	
Hedges's g	0.404 (small)		0.429 (small)		0.335 (small)		0.348 (small)		0.472 (small)		0.185 (negl.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	23	13	11	7	12	6	23	13	11	7	12	6
Mean Dev.	103.7	75.23	105.8	74.00	101.8	76.67	74.65	55.31	78.27	51.00	71.33	60.33



(a) AEMT



(b) HCRC Subset 2

Figure 5.15: Distribution of Deviation Score (M4: Method 4) by Experience (Attempt 1, 2, 3, 4), along with the average of the four experiences, Familiar (F) being the only condition in the AEMT and the HCRC Subset 2.

Table 5.17: Summary of Wilcoxons Tests HCRC Subset 2 for SELFSHARED, Deviation Method = M4 and HCRC precomputed Dev scores *(Note: The tests which involved less than 5 dialogues on either side of the tests are not considered for the results, as it makes the comparison unreliable)*

HCRC Sub2 – SELFSHARED

All Level – All *n*-grams												
	M4						HCRC DevScore					
Speaker	IF + IG		**IG**		IF		IF + IG		IG		**IF**	
Wilcox. W	2434.5		337.5		849.5		2524.5		360		872	
p-value	0.1566		**0.014**		0.879		0.267		**0.028**		0.7244	
Hedges's g	0.45 (small)		-0.63 (med.)		0.46 (small)		0.5 (small)		0.52 (med.)		0.497 (small)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	41	139	15	75	26	64	41	139	15	75	26	64
Mean Dev.	69.41	100.5	57.66	100.6	76.19	100.4	50.26	72.79	38.6	73.48	57	72
	L1 – All *n*-grams											
Wilcox. W	135.5		44.5		21		144.5		49		21	
p-value	0.179		0.256		0.527		0.088		0.114		0.526	
Hedges's g	0.584 (med.)		0.617 (med.)		0.509 (med.)		0.644 (med.)		0.717 (med.)		0.501 (medium)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	29	7	13	5	16	2	29	7	13	5	16	2
Mean Dev.	101.3	60.57	105.9	61	97.68	59.5	74.48	39.42	78.84	38.6	70.93	41.5
	L1 – *n*-grams *n*=1 (Lexical)											
Wilcox. W	185		42.5		49		195		46		50.5	
p-value	0.321		0.858		0.241		0.187		0.624		0.189	
Hedges's g	0.276 (small)		-0.066 (negl.)		0.663 (med.)		0.296 (small)		-0.078 (negl.)		0.727 (med.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	22	14	10	8	12	6	22	14	10	8	12	6
Mean Dev.	101	81.42	91.2	96.25	109.3	61.66	74.09	57.57	65.6	70.25	81.16	40.66
	L4 – *n*-grams $n > 1$ N2+ (Phrasal)											
Wilcox. W	162.5		35.5		45.5		168.5		43		41	
P-value	0.68		0.82		0.399		0.541		0.716		0.673	
Hedges's g	0.404 (small)		0.429 (small)		0.335 (small)		0.348 (small)		0.472 (small)		0.185 (negl.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	23	13	11	7	12	6	23	13	11	7	12	6
Mean Dev.	103.7	75.23	105.8	74.00	101.8	76.67	74.65	55.31	78.27	51.00	71.33	60.33

AEMT vs HCRC Sub 2: Task Familiarity

When comparing the two corpora with regard to Task Familiarity, as can be seen in Table 5.18 and Table 5.19, one consideration to keep in mind is that the AEMT consists of 8 speakers doing the task four times, switching roles. However, in the HCRC Subset 2 which matches the conditions of the AEMT, this configuration is not the same, meaning that the pairs that are making their first attempt are different pairs than the ones making the subsequent attempts. This situation might be the source of the unexpected differences in the patterns of repetitions.

Table 5.18: Summary of Wilcoxons Tests AEMT for OTHERSHARED and SELFSHARED, at First Attempts and Attempts 2 to 4, Deviation Method = M4, at all levels and all *n*-grams *(Note: The tests which involved less than 5 values on either side of H_0 rejections are not considered for the results, as it makes the comparison unreliable)*

AEMT – Task Familiarity						
OTHERSHARED – Task Familiarity = 1						
Speakers	IF + IG		**IG**		IF	
Wilcoxon W	358.5		90		7	
p-value	0.176		**0.002**		0.72	
Hedges's g	-0.210 (small)		**1.454 (large)**		NA	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial	29	11	10	10	19	1
Mean Dev.	104.67	123.09	233.09	116.7	174.3	187
OTHERSHARED – Task Familiarity = 2 to 4						
Wilcoxon W	1923		359.5		433	
p-value	0.511		0.906		0.3769	
Hedges's g	0.219 (small)		0.152 (negl.)		0.152 (negl.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial	58	62	16	44	42	18
Mean Dev.	59.65	50.16	57.5	53.75	60.47	41.38
SELFSHARED – Task Familiarity = 1						
Wilcoxon W	254.5		77		50	
p-value	**0.003**		**0.003**		0.279	
Hedges's g	**1.138 (large)**		**1.257 (large)**		0.921 (large)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial	29	11	14	6	15	5
Mean Dev.	202	103.8	207.71	98.66	196.66	110
SELFSHARED – Task Familiarity = 2 to 4						
Wilcoxon W	1445.5		379.5		337	
p-value	0.609		0.645		0.807	
Hedges's g	-0.104 (negl.)		-0.273 (small)		0.091 (negl.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial	83	37	39	21	44	16
Mean Dev.	53.74	57	51.74	60.33	55.52	52.62

The two corpora indicate similar relations between OTHERSHARED repetitions and task success. For both the first attempts relate the Information Giver ABOVE CHANCE OTHER-

SHARED repetitions to lower task success. No significant difference was found at further attempts. The main difference between the two corpora in this subsection is in SELFSHARED repetitions. In the AEMT, first attempt significantly relates the Information Giver SELFSHARED repetitions to lower task success (ABOVE CHANCE: $\bar{x} = 207.71$; NOT ABOVE CHANCE: $\bar{x} = 98.66$), while no significance is found at first attempt in the HCRC Sub2. For attempts 2 to 4, a significance is found in the HCRC ($p = 0.03$) for the IG, relating SELFSHARED repetitions to higher task success (ABOVE CHANCE: $\bar{x} = 49.9$; NOT ABOVE CHANCE: $\bar{x} = 78.38$).

Table 5.19: Summary of Wilcoxons Tests HCRC Sub2 for OTHERSHARED and SELFSHARED, at First Attempts and Attempts 2 to 4, Deviation Method = M4 and HCRC precomputed Dev scores, at all levels and all *n*-grams *(Note: The tests which involved less than 5 dialogues on either side of the Wilcoxon tests are not considered for the results, as it makes the comparison unreliable)*

HCRC Subset 2 – Task Familiarity

	OTHERSHARED – Task Familiarity = 1											
	M4						HCRC DevScore					
Speaker	**IF + IG**		IG		IF		**IF + IG**		IG		IF	
Wilcox. W	428		128		82		428		128		82	
p-value	**0.0089**		0.06		0.06		**0.008**		0.06		0.06	
Hedges's g	0.548 (med.)		0.532 (med.)		0.57 (med.)		0.716 (med.)		0.71 (med.)		0.497 (med.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	48	12	22	8	26	4	48	12	22	8	26	4
Mean Dev.	142.9	98.16	145.9	101.2	140.4	92	106.3	62.91	109.4	65.3	103.7	58
	OTHERSHARED – Task Familiarity = 2 to 4											
Wilcox. W	1452		412		298		1562		447		318	
p-value	0.95		0.56		0.89		0.56		0.56		0.82	
Hedges's g	0.34 (small)		0.37 (small)		0.32 (small)		0.34 (small)		0.37 (small)		0.3 (small)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	86	34	39	21	47	13	86	34	39	21	47	13
Mean Dev.	78.06	60.76	79.71	61	76.70	60.38	56.72	42.41	58.17	42.42	55.51	42.38
	SELFSHARED – Task Familiarity = 1)											
Wilcox. W	232		22		85		232		22		85	
p-value	0.11		0.06		0.51		0.11		0.06		0.51	
Hedges's g	-0.67 (med.)		-0.76 (med.)		-0.66 (med.)		-0.55 (med.)		-0.76 (med.)		-0.47 (med.)	
H_0 Rej.	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above	Above	N.Above
# of Dial.	14	46	4	26	10	20	14	46	4	26	10	20
Mean Dev.	92.42	146.6	79	142.4	97.8	152.1	71.5	105.6	56	104	77.7	107.6
	SELFSHARED – Task Familiarity = 2 to 4)											
Wilcox. W	1000.5		157		337		1045.5		164.5		352	
p-value	0.108		**0.031**		0.8		0.18		**0.04**		1	
Hedges's g	-0.4 (small)		-0.59 (med.)		-0.28 (small)		-0.41 (small)		-0.59 (med.)		-0.27 (small)	
# of Dial.	27	93	11	49	16	44	27	93	11	49	16	44
Mean Dev.	57.48	77.72	49.9	78.38	62.68	76.97	39.25	56.55	32.27	57.24	44.06	55.76

5.5.2 American vs. Scottish English: Unfamiliar Partners

As the PARDO corpus is particularly small, 10 dialogues with 4 participants (5 per pair), the tests made in the exploration of task success did not involve enough dialogues to make the possible comparisons meaningful. They are therefore not reported. The comparison in this subsection is limited to rate of Null hypothesis rejections with the AEMT and HCRC Subsets. In the results that can be seen in Table 5.20, the PARDO corpus alone shows comparable rates of rejections of H_0 in OTHERSHARED and SELFSHARED, independently of roles, which is what is also observed in the AEMT (see section 5.5.1). A higher rate of rejection at structural (N2+) rather than at lexical (N1) level, in particular for OTHERSHARED repetitions, is also observed (see Table 5.20), a pattern that is present in both the HCRC Subsets and the AEMT corpus. In the HCRC Subset 3, a difference in OTHERSHARED repetition between roles is visible, with the IG repeating the IF less, which was not the case in the HCRC Subset 2.

Table 5.20: Rejections of H_0 for OTHERSHARED and SELFSHARED in the PARDO and the HCRC Sub3, in relation to roles (IF: Information Follower; IG: Information Giver), in each case (each cell) the Null Hypothesis can potentially be rejected 10 times in the PARDO and 14 times in the HCRC Subset 3.

OTHERSHARED

PARDO							**HCRC Subset 3**						
All n-grams | H_0 : $Rand.Speaker.Level - Actual.Speaker.Level \geq 0$													
Level	Tok	Lem	LemP	POS	TokP	Mean	Level	Tok	Lem	LemP	POS	TokP	Mean
IF	8	5	6	1	8	5.6	IF	12	12	13	7	13	11
IG	7	8	7	2	7	6.2	IG	7	7	5	2	6	5.4
N1: n-gram=1 | H_0 : $Rand.Speaker.Level.N1 - Actual.Speaker.Level.N1 \geq 0$													
IF	3	3	3	1	2	2.4	IF	7	7	8	3	8	6.6
IG	6	6	4	0	5	4.2	IG	3	4	4	1	4	3.2
N2+: n-gram>1 | H_0 : $Rand.Speaker.Level.N2+ - Actual.Speaker.Level.N2+ \geq 0$													
IF	10	10	9	3	9	8.2	IF	11	11	12	6	12	10.4
IG	10	9	9	7	8	8.6	IG	7	6	8	3	7	6.2

SELFSHARED

PARDO							**HCRC Subset 3**						
All n-grams | H_0 : $Rand.Speaker.Level - Actual.Speaker.Level \geq 0$													
Level	Tok	Lem	LemP	POS	TokP	Mean	Level	Tok	Lem	LemP	POS	TokP	Mean
IF	6	5	6	5	6	5.6	IF	2	2	2	0	2	1.6
IG	9	8	9	6	8	8	IG	1	1	0	1	1	0.8
N1: n-gram=1 | H_0 : $Rand.Speaker.Level.N1 - Actual.Speaker.Level.N1 \geq 0$													
IF	4	4	6	4	4	4.4	IF	1	1	1	0	1	0.8
IG	7	8	8	5	8	7.2	IG	0	0	0	0	0	0
N2+: n-gram>1 | H_0 : $Rand.Speaker.Level.N2+ - Actual.Speaker.Level.N2+ \geq 0$													
IF	8	7	7	6	7	7	IF	2	2	2	2	2	2
IG	9	9	7	7	7	7.8	IG	2	2	1	1	1	1.4

5.5.3 Conclusion

This section presented the analysis of three map task corpora, one in Scottish English and two in American English settings. As the two American English corpora had different familiarity conditions, and the importance of this factor on patterns of repetitions was highlighted by previous studies (section 5.2), two matching condition subsets were extracted from the HCRC Scottish English corpus. This analysis revealed different patterns of repetitions between the two dialects, with a higher presence of self-shared repetitions in American than Scottish and a low amount of other-shared above chance repetitions for American Information Givers than the Scottish ones. However, for the familiar participants, both dialects users exhibited similar patterns in relation to task success: low task success when the other-shared repetitions were found above chance. That finding triggered the study described below, that investigates the possibility for the method to detect repetitions happening under chance, as the hint that familiar partners might reach mutual understanding without exhibiting alignment is observed.

5.6 Under Chance Repetitions

This section explores another aspect of communication patterns that results from the previous experiments focusing on familiarity brought to light: the possibility that for familiar partners, divergence could actually indicate understanding. Indeed, section 5.2.3 showed that ABOVE CHANCE repetitions seemed to correspond to Familiar pairs having worse task performance in the HCRC corpus. That aspect can be operationalised through the presented method by exploring UNDER CHANCE repetitions. In previous experiments, it was determined that for familiar people, not repeating the partner had no statistically significant impact on deviation scores at first attempt (See § 5.2.3, p=0.106). Another interesting observation is that if the following attempts are all more successful (for both familiarity types), familiar pairs that align seem to have a better success (even if in this case the effect size is much lower). From these results, this section explores the question: can divergence (actual repetition being significantly less present than random repetitions) for familiar people be indexing understanding, even within the frame of task-based interactions? For the purpose of this experiment, the label ABOVE CHANCE is associated with the notion of "convergence" of which the meaning partly overlaps with *alignment*, but mostly allows the use of the opposite phenomenon that is "divergence", here associated with the label UNDER CHANCE. The speaker for which the proportion of repetitions happened UNDER CHANCE are considered "divergent". The method described in chapter 4 was followed with one modification in the tested hypothesis:

$$H_0 : Random.Speaker.Level - Actual.Speaker.Level \leq 0$$

This H_0 null hypothesis designed to test UNDER CHANCE states that the proportion of shared repetitions in the randomised dialogues should equal or be inferior to the proportion of shared of repetitions in the actual dialogues if repetitions are not happening under chance.[7] If rejected, repetitions in the actual dialogues are happening under what is considered chance. This hypothesis was tested at three levels of n-grams granularity: N: n-gram= all ($1 \leq n \leq 5$), N1: $n = 1$ (lexical level), and N2+: $n > 1$ (phrasal level). This last experiment uses all the previously analysed corpus using the map task technique in non-mediated settings: the HCRC Map Task, the AEMT, and the PARDO corpus.

[7] This hypothesis is not equivalent to H_1 described in chapter 4. The categorisation NOT ABOVE CHANCE does not equal UNDER CHANCE, and the categorisation ABOVE CHANCE does not equal NOT UNDER CHANCE.

5.6.1 Overview of Under Chance

This section gives an overview of the divergence categorization described in the previous section in terms of rejections of the null hypothesis H_0 UNDER CHANCE in the three corpora: the HCRC, the AEMT and the PARDO. Table 5.22 shows that the rate of UNDER CHANCE H_0 rejections of the null hypothesis in the three corpora, is very low with a total of 153 out of 9240 across all linguistic levels in both OTHERSHARED and SELFSHARED repetition types. There was no UNDER CHANCE H_0 rejections for Information Follower. The number of H_0 rejections being extremely low in the AEMT and the PARDO 2006 corpora, they are discarded in the following tests. The dialogues categorized as NOT ABOVE CHANCE do not correspond to the dialogue categorized as UNDER CHANCE. However, Table 5.22 shows that the Information Giver self-repetitions at the level 1, n-grams: $n = 1$ is where the rate of rejection of the null hypothesis is the highest for UNDER CHANCE and the lowest for ABOVE CHANCE.

5.6.2 Under Chance and Task Success

This section explores the relation between the categorization UNDER CHANCE and task success. A Mann-Whitney-Wilcoxon test for population distribution showed a significant (W= 14390, $p = 0.0039$) difference between the IG repeating themselves UNDER CHANCE than ($n = 64, \bar{x} = 53.2$) the IG not repeating themselves UNDER CHANCE ($n = 576, \bar{x} = 73.88$), at linguistic level n-gram $= 1$. This difference is associated with a small negative effect size (g estimate=-0.42). The average Deviation Score is smaller for the Information Giver repeating themselves UNDER CHANCE, however these results must be taken carefully given the large difference in population size (n) in this test. The number of UNDER CHANCE H_0 rejections of the null hypothesis for Information Giver within Familiar pairs is 37 and 27 for Information Giver within Unfamiliar pairs.

Table 5.21 show that the Information Giver both in Familiar and Unfamiliar pairs repeat themselves UNDER CHANCE more as they perform the task again. No significant difference (W= 477.5, $p = 0.7697$) was found between Familiar and Unfamiliar Information Giver repeating themselves UNDER CHANCE, at linguistic level n-gram $= 1$.

Table 5.21: Number of UNDER CHANCE H_0 rejections for Information Giver per Task Attempt (1 to 4) and Familiarity with the Information Follower at linguistic level n-gram $= 1$, in the HCRC Map Task.

Task Attempt	**1**	**2**	**3**	**4**
Familiar	3	6	16	12
Unfamiliar	2	8	5	12

Table 5.22: Rejection count of H_0 for levels L1 to L5 values in the HCRC Map Task corpus, the AEMT corpus and the PARDO corpus. For each dialogue at each level (each cell), the number of possible H_0 rejections is 128 in the HCRC Map Task corpus, 16 in the AEMT and 10 in the PARDO.

		HCRC					AEMT					PARDO				
		L1	L2	L3	L4	L5	L1	L2	L3	L4	L5	L1	L2	L3	L4	L5
ABOVE CHANCE — n-grams = All																
OTHER	IG	88	87	80	47	81	6	5	8	2	5	7	8	7	2	7
	IF	112	109	109	82	107	13	13	11	12	12	8	5	6	1	8
SELF	IG	27	26	30	5	28	14	10	12	6	11	9	8	9	6	8
	IF	36	35	37	19	38	13	11	13	10	12	6	5	6	5	6
Total H_0 rej. HCRC: 1183 / 2560, AEMT: 199 / 320, PARDO: 127 / 200																
ABOVE CHANCE — N1: n-gram=1																
OTHER	IG	49	47	51	18	54	3	3	5	0	4	6	6	4	0	5
	IF	78	78	74	46	75	7	8	4	4	6	3	3	3	1	2
SELF	IG	**4**	**4**	**4**	**0**	**5**	6	5	8	2	5	7	8	8	5	8
	IF	8	10	11	4	11	9	7	9	4	9	4	4	6	4	4
Total H_0 rej. HCRC: 631 / 2560, AEMT: 108 / 320, PARDO: 91 / 200																
ABOVE CHANCE — N2+: n-gram>1																
OTHER	IG	90	91	88	58	89	10	11	12	4	11	10	9	9	7	8
	IF	108	104	105	81	107	14	13	10	11	12	10	10	9	3	9
SELF	IG	44	49	43	16	46	12	10	11	10	11	9	9	7	7	7
	IF	38	38	39	26	37	12	11	12	11	11	8	7	7	6	7
Total H_0 rej. HCRC: 1297 / 2560, AEMT: 219 / 320, PARDO: 158 / 200																
Under Chance — n-grams = All																
OTHER	IG	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	IF	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
SELF	IG	**10**	**10**	**9**	**31**	**7**	0	0	0	0	0	0	0	0	0	0
	IF	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
Total H_0 rej. HCRC: 71 / 2560, AEMT: 4 / 320, PARDO: 0 / 200																
Under Chance — N1: n-gram=1																
OTHER	IG	0	0	0	2	0	0	0	0	3	0	0	0	0	1	0
	IF	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0
SELF	IG	**13**	**10**	**7**	**29**	**5**	0	0	0	0	0	0	0	0	0	0
	IF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total H_0 rej. HCRC: 66 / 2560, AEMT: 4 / 320, PARDO: 3 / 200																
Under Chance — N2+: n-gram>1																
OTHER	IG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	IF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SELF	IG	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
	IF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total H_0 rej. HCRC: 5 / 2560, AEMT: 0 / 320, PARDO: 0 / 200																

5.7 Conclusion

This chapter presented five studies that explored different possibilities given by the method to detect repetition patterns and their relation to mutual understanding. First by highlighting the importance of taking into consideration non-linguistic features such as task roles, gender, eye-contact, familiarity and task experience, as all these features show different patterns of repetition in interaction with task success. Taking into consideration these features can give indications on the likelihood that understanding is reached between conversational partners engaged in a task. For example, a lack of alignment (no statistically significant amount of other-shared repetitions happening above chance) between unfamiliar partners in their first interaction, might indicate a problem in their communication. Secondly by confirming the possible usage of the method in the case of computer-mediated interactions, where a high level of negative cognitive states coincides with a lack of structural alignment, and by observing that the method can also match an interactional facilitator's behaviours and cues of successful communication through alignment. These results show the possibility of usage the method as an interactional indicator in different conditions. However, less clear results can be drawn from the comparison of dialects of English, except that familiar partner repeating each other above chance seem to be related to less success, and from the examination of repetition happening under chance.

Chapter 6

Discussion

This section discusses the results of the experiments described in the previous chapter. The possible interpretations of the results are discussed along with the aspects of the experiments that provided evidence for or against the ability of repetitions to function as a proxy measure of mutual understanding.

6.1 The Exploratory study of the HCRC Map Task

6.1.1 Non-Linguistic Features Exploration

The Information Giver had a much higher volume of speech than the Information Follower and tended to produce longer utterances,[1] and the Information Follower (IF), while talking less, repeated herself/himself and the IG more often significantly in almost all the tested conditions. The results show consistency with previous findings in the sense that in task-based interaction, different social roles leads to a different repetition patterns (Colman & Healey, 2011), and significant OTHERSHARED and SELFSHARED repetitions have an impact on task-success (Reitter & Moore, 2007). It seems that overall, the IF repeats the IG more often than the opposite, and the IG is, on the other hand, repeating herself/himself more (which could be interpreted as keeping the same structure in providing information): both situations that tend to relate to higher communicative success. Differentiation of gender highlighted differences in communication strategies. The results suggested that for the IF,

[1] 110,075 tokens produced in total for IG against 50,622 tokens for the IF, see § 3.2

non-significant OTHERSHARED repetitions mattered less for male than female, for which a small portion of the last meant less successful communication.

Counting repetitions for other linguistic Levels than token only showed some additional information that can be used in the interpretation of the variations of communicative behaviours. In particular in the case of SELFSHARED repetitions, Token only did not always display significant differences to allow the rejection of H_0, but other Levels did, hence indicating the additional information those Level divisions are bringing. "Structural" self-repetitions of the female IGs were related to lower deviation score than male IGs. No eye-contact and Familiarity were related to lower deviation scores, which seem the best combination for task-success, even if men IFs not repeating the IGs were the ones performing least well in those conditions. However, in Eye-contact situations, female IFs not repeating themselves performed on average less well than male IFs. The fact that these repetition patterns showed consistency with known communication strategies is crucial. It can be however noted that the count of Lemma+POS and Token+POS often show little variation, and that no statistically significant difference appeared between Lemma and Token. This can be explained in two ways: the nature of the task did not allow an important variety of inflexions to appear as participants used a simple vocabulary; it is also possible that there would be more inflexions in a different language than modern English, seen more as analytic than synthetic (Haspelmath & Michaelis, 2017). Finally, as mentioned at the beginning of the previous section: among the factors influencing task success, familiarity appears to have the most impact, which is why the next section explores this factor in more detail.

6.1.2 Familiarity & Experience

The first attempt, when participants discover the task, represents the closest observation of an untrained pair of participants in real task-solving conditions, it is therefore the most interesting in the context of this thesis. Both familiar and unfamiliar partners display a high level of ABOVE CHANCE OTHERSHARED repetition during the first attempt, however it impacts the relation with task-success differently. In the first attempt, unfamiliar partners who repeat each other to a significant degree (summing across levels of linguistic representations), and thus align to their partner, have greater levels of task success than unfamiliar partners with-

out a significant degree of repetition. Alignment does not correlate with task-success at the first attempt for familiar pairs, in contrast to unfamiliar pairs. However, familiar pairs with significant self-repetition in the first attempt, compared to familiar pairs without significant self-repetition, achieved greater task-success. While keeping in mind the results given in the HCRC, the following section discusses the repetitions patterns in different task conversation types.

6.2 Mediated Conversations

6.2.1 Computer-Mediated Interactions: The ILMT-s2s

The high rate of ABOVE CHANCE OTHERSHARED repetition in the computer mediated dialogues of the ILMT-s2s corpus indicates that alignment occurs in at least the same proportion as in direct human-to-human communication. No evidence of alignment exaggeration was found with the method, as it detected equally high alignment cues in direct human-to-human communication. However, ABOVE CHANCE repetitions occurred at all linguistic levels at a high rate in the ILMT-s2s corpus, for both OTHERSHARED and SELFSHARED. This is different from the direct human-to-human dialogues from the HCRC where self-repetitions occurred rarely above chance. This high rate of SELFSHARED repetition for both roles could be attributed to the perceived difficulty for the speakers to have their utterance properly recognized by the ASR system and correctly translated to their interlocutor, hence their tendencies to repeat themselves in consecutive turns more. The high rate of repetitions, in both types (OTHER and SELF), in the interlingual computer-mediated corpus, follows past findings (H. P. Branigan, Pickering, Pearson, McLean, & Nass, 2003) that suggest strong alignment in human-computer interaction. Previous work also suggests that exaggerated alignment toward a system is detrimental to the interaction since the subjects also repeated translation errors (Schneider & Luz, 2011). This pattern was not found in the current work, which found similar levels of alignment in human-to-human and computer-mediated interactions. However, a relation emerged within the computer-mediated dialogues, between negative cognitive states and low ABOVE CHANCE repetitions of long sequences. Portuguese speakers in eye-contact conditions displayed higher than expected negative cognitive states

which also related to their low appreciation of the system.

Even if the small size of the two corpora prevents us from making too broad a statement, the repetition patterns detected by the automatic method present S2S-MT software design cues that constitute another step toward aiding human-to-human communication when interacting through machine translation. The results show that the lack of alignment of long token sequences in video conditions indicates problematic interactions. One might wonder if the reason that differences appeared between English and Portuguese speakers could be interpreted as a cultural difference. This could be examined in the future by comparing other language pairs and/or larger data sets.

6.2.2 Human-mediated Interactions: The MULTISIMO

A null hypothesis expects no interaction between facilitator's feedback types and the degree of OTHERSHARED repetitions by the dialogue participants. The results breach this expectation. Facilitators respond with more non-positive feedback where ABOVE CHANCE OTHERSHARED repetitions are observed, and more positive feedback where NOT ABOVE CHANCE OTHERSHARED repetitions are observed, than one would expect in either case with no interaction. If significant OTHERSHARED repetitions signal mutual understanding, facilitators are more likely to respond with non-positive than with positive feedback to signals of mutual understanding. Those results suggest that the facilitators provide more encouragement where interactions are seen as difficult, and less encouragement when interactions are perceived as successful.

The results observed for the full dialogues and for each of the three question sections (as separate parts of each full dialogue) do not apply identically within each of the components of the question sections. The type of task was different from the HCRC corpus, and no party held the information, which resulted in a different dynamic of OTHER and SELF repetitions. While in the HCRC less self-repetition above chance is found, in the Full MULTISIMO dialogues the number of rejections for self-repetition and other-repetition are equivalently high. The cut sections, which are shorter, displayed a lower rate of H_0 rejection than the Full Sections; which can be partly explained by the structure of the method, that necessitates a relatively large volume of transcribed speech to distinguish actual repetitions from randomness

of their occurrence. This finding suggests that the method may be more robust if the interaction average length is at least 120 turns and 350 tokens.[2] However, interesting phenomena can be noted even at shorter lengths, in the task-related differences between the different phases of the game. In the idea generation phase (Answer) there are fewer self-repetitions, while in the idea ranking phase (Ranking), a relatively high number of self-repetitions that is not echoed for long utterances can be observed. A possible interpretation is that this could indicate a stronger will to show its preference for ranking, in particular once the elements constituting the trio to be ranked have been found, as it is a slightly easier phase.

Nonetheless, it is the relation appearing between above chance repetitions and positive and non-positive feedback that constitutes the most interesting findings of this study. It provides evidence for the general hypothesis that repetitions detected from this method reflect a degree of interactional success echoing human perception in goal-directed task-based dialogues. The state of an interaction — indicating if mutual understanding is taking place or not — is echoed in the verbal and non-verbal behaviours of an interaction facilitator. Further confirmations are given in the next section.

6.3 Different dialects of English: The AEMT and the PARDO

In contrast to the patterns found in the AEMT confirms that for familiar pairs, repetitions seem to be linked to a less successful interactional results. The numbers of rejections of H_0 show clear differences between the American corpora and the Scottish HCRC Subsets. The Information Giver in the AEMT repeated much less than the Information Followers on average than in the HCRC Subset 2 and both roles repeated themselves at a higher rate in the AEMT compared to the low rates of ABOVE CHANCE repetitions found in the HCRC Subsets and the HCRC in general. However, when compared to the task success scores, similar patterns were found for OTHERSHARED repetitions in both corpora (namely significant differences between ABOVE CHANCE and NOT ABOVE CHANCE where ABOVE CHANCE is related to less success). At first Task Attempt, significant differences appeared in both corpora in terms of OTHERSHARED repetitions, both relating higher success to NOT

[2]This estimated length correspond to the average length of the *Question* sections as can be seen in Table 3.5 and corresponds to an average of three minutes, as can be seen in Table 3.4.

ABOVE CHANCE repetitions, which provides evidence for repetitions not being an indication of interactional success for familiar pairs of speakers, even at first attempt at a task, in both dialects. It is interesting to see that this difference was only visible at first attempt but not in the subsequent attempts, where it was not found to have an impact on familiar pairs. The patterns found for self-repetitions were not significant for the AEMT while the HCRC Subset 2 showed that the Information Follower repeating themselves obtained higher task success.

The differences found could be interpreted in at least two ways: individual variations, or the differences between the two corpora (English dialect and seven years apart in the recording). The design of the method, the level granularity explored, is such that dialects of English were not expected to show variations in the results. It is found that the two American Map Tasks show more similarities with each other than with the HCRC Subsets that emulate the same familiarity conditions. The high rate of SELFSHARED rejections found in the two American corpora, i.e. female participants repeating themselves, might be a cultural difference from the HCRC.[3] That for familiar pairs, repetitions seem to be linked to less successful interactional results, led to the examination of the other side of the alignment phenomenon that is discussed in the next section: divergence in dialogues.

6.4 Under Chance: Divergence?

From the results obtained, two possible conclusions are possible: either almost no divergence[4] in consecutive turns is present in the face-to-face map task corpora analysed, or the method did not allow for its detection. The rejection rate of the UNDER CHANCE H_0 null hypothesis for Information Giver self-repetitions at lexical level simply indicate the higher possible divergence at this level of granularity. However, no significance being found between *deviation scores* for Information Giver in familiar or unfamiliar pairs do not allow for the confirmation that Under Chance repetitions might have a negative impact on Familiar pairs in the HCRC corpus. The method did not show evidence for the presence of divergence

[3]The important changes in women lives, their position in society and therefore interactional style over the period between the HCRC corpus and the AEMT, might also be the source of the found differences.

[4]Divergence, as previously defined, refers to the situation when interlocutors show explicit signs that they do not, intentionally or not, linguistically align with each other.

in the corpus, which is why this experiment is inconclusive.

6.5 Conclusion

This chapter discussed the results of the experiments presented in chapter 5. The discussions were oriented on the descriptions of the patterns found and their possible interpretations, notably the different amounts of above chance repetitions depending on each corpora studied and the impact of the familiarity in all type of task-based interactions. The next chapter summarizes those patterns and gives insights on possible future research.

Chapter 7

General Conclusion

This thesis has described a method of interaction analysis designed to provide a proxy measure of mutual understanding in goal-directed task-based dialogues based on repetitions. The specific patterns of significant repetitions in relation with task-success in various settings found by the method provide evidence that the establishment of an automatic quantifying measure of communicative success without the need to manually annotate data for understanding is possible. Non-linguistic features (task roles, familiarity, gender, eye-contact) have an important impact on the distribution of repetitions, and a careful definition of the situation is desired to assess successful communication or possible misunderstanding by observing patterns of repetitions. Confirmed patterns that could be useful in the development of dialogue management systems could be summarized as:

1. When a speaker is in the role of an information giver, self-repetition plays a crucial role in task success, while in the position of information receiver, repeating the other is more important. (See subsection 5.2.2 and subsection 6.1.1)

2. Repeating the other above chance is the sign of a higher chance of task success for unfamiliar partners, in particular at the first attempt of a task (i.e. the beginning of their process of common ground building), while it is not the case for familiar partners. (See subsection 5.2.3 and subsection 6.1.2)

3. An overwhelming presence of self-repetition in computer mediated interactions is to be noted and a lack of other repetitions for long-sequences is the sign of difficulties in

the interaction as it matches with high levels of frustrations. (See subsection 2.4.2 and subsection 6.2.1)

4. A third party facilitator provides more encouragement where interactions can be seen as difficult (partners not repeating each other above chance), and less encouragement when interactions can be perceived as successful (partners repeating each other above chance). (See section 5.4 and subsection 6.2.2)

5. Different dialects of the same language might exhibit unexpected variations in terms of repetitions patterns, but it is likely that the relation between the patterns of repetitions and task success will behave similarly. American English repeat themselves more than Scottish English, while doing the same task, and for both dialects above chance repetitions related to less successful task results when dialogue partners are familiar to each other. (See section 5.5 and section 6.3)

The extent to which repetitions may function as cues of an alignment process and provide a proxy measure of mutual understanding in task-based interactions, is validated from the five different corpus studies presented and the above patterns they exhibited. It is established that familiarity has the most striking impact among the factors, even if the cultural impact is not to be neglected, as the study undertaken to explore American English map tasks suggests. The phenomenon of alignment is taken as the sign of common ground building, which seems as essential at first encounters but of which the importance decrease once a certain level of common ground in reached (if the task does not change). The idea that alignment decrease over time has found support in empirical studies (Fusaroli et al., 2012, 2014). This indicates in particular the powerful potential of the detection of alignment in first encounter human-human interactions. A usage in call-centres, by the real-time analysis of conversations between officers and clients or emergency call directed at firefighters or police, are directions towards which future works should be directed. Conversations between pilots and air-traffic controllers, while already very codified, could also benefit from interactional measures. With the breakout of the 2019 coronavirus pandemic, phone call and video calls between general practitioners and patient multiplied, which constitute another potential usage of communicative success measure.

Conversational agents articulated by dialogue systems used for single specific tasks, such as reservations or after-sale management for example, would also benefit from such measure. If the detection of repetitions would be implemented in a dialogue system, the inclusion of metadata that correspond to the sociological factors of gender, eye-contact, task role, and most importantly familiarity, would be crucial in the evaluation of the impact of repetitions happening above chance on the success (or not) of the communication. The detection of alignment in a first interaction between unfamiliar partners could be a good indication that the communication is going well. The detection of a lack of structural alignment in computer-mediated multilingual interactions could indicate a problematic conversation. For familiar partners, further research needs to be conducted into the other factors influencing communication. Nonetheless, this thesis provides evidence that the quantification of repetitions in task-based dialogues can outline definite linguistic patterns that would be useful indicators of successful or unsuccessful communication.

Future Work

The findings have confirmed existing patterns and discovered new ones, but also pointed out that research efforts attempting to unveil communication patterns are still needed. Which features are to be used in priority in the observation of repetition patterns linked to interactional success is answered to a certain degree, for the reason that if definite patterns have been found, one cannot exclude that they might constitute an over-fitting to the data used. The need for large volumes of factor-controlled yet naturally occurring speech corpora remains a limitation to quantitative analysis.

Repetition of sequences above one word and in particular of Part-Of-Speech sequences, seem to play a decisive role in the five studies within this thesis. This tends to confirm the theory of structural alignment being related to successful communication. However the process is not consistently observed in all conditions in relation with communicative success. This thesis therefore refutes alignment as being a pervasive process that leads to mutual understanding in all conditions, but rather a useful mechanism when a number of caveats are taken into account. The results suggest, across three of the studies, that the alignment phenomenon is only related to successful communication for unfamiliar partners, at the beginning of their process of common ground building. The study of other sociological factors, such as education levels, age, if an individual is a native speaker or not, determining the moment where people can be considered as familiar[1] could continue to bring light to the mechanisms of repetitions that are underlying mutual understanding. Furthermore, the generalization outside task-based interaction also remains an open question. The possible application of the method on other types of task-based interactions, such as the above mentioned, should be explored in future works. Finally, I wish to mention that interactional measures must be implemented without overlooking the ethical dimensions of data collection. Some of the factors that appear to influence communication are also considered private data under the General Data Protection Regulation (GDPR) issued by the European Union and therefore their collection should be reserved to the situations in which the success of communication is critical.

[1] Is there a definite moment or a gradual evolution that can be traced?

Appendix A

Preliminary Experiment: The Table Talk

The Table Talk corpus (Campbell, 2009) consists of three, ninety minutes round-table conversational interactions in English. The multimodal recordings happened over three days at the Advanced Telecommunication Research (ATR) Labs in Japan, in an informal setting (i.e. the participants were not wearing any devices).[1] The unscripted conversations took place among four participants, with the addition of a fifth speaker on day 2, three women (Australian, Finnish, and Japanese) and two men (Belgian and British). While they come from different cultural backgrounds, they share a common experience of living in Japan, which constitutes one of the main subjects of the conversation discussed over the sessions. The conversations were transcribed and annotated with a number of communicative elements such as laughter, and non-verbal gestures: facial expressions, head, hand, and body movements based on the MUMIN coding scheme (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007). One participant knew each speaker before the session, while the other speakers were unfamiliar with each other.

The Table Talk corpus has already been partly analysed with a previous version of the method and the results published (Vogel & Behan, 2012). The purpose of this second analysis was to observe the extent to which the addition of the levels (linguistic types of repetitions) is a useful indicator, in comparisons to the sole usage of the Token Level, to distinguish the degree of involvement in the conversation. The actual count of repetitions per n-grams levels (N1: n-grams $n=1$; N2+: n-grams $n>1$), day and speaker, can be seen in Table A.1, with the results indicating that the actual count is higher for OTHERSHARED (OS), than

[1] I am grateful to professor Nick Campbell for the granting me access to this data.

SELFSHARED (SS).

Table A.1: Count of Repetition in the Actual Table Talk dialogues

Day 1	*L1= Token*	*L2=Lemma*	*L3=Lemma+POS*	*L4=POS*	*L5=Token+POS*
OS N1	1206	1189	1012	3509	1022
OS N2+	136	116	92	331	92
SS N1	951	934	825	2172	823
SS N2+	229	220	214	425	216
Day 2					
OS N1	3530	3440	6833	14565	5818
OS N2+	408	394	553	2196	1228
SS N1	1936	1961	2768	6687	2668
SS N2+	343	304	371	507	354
Day 3					
OS N1	2072	2079	1703	7285	1731
OS N2+	270	243	198	792	198
SS N1	1758	1711	1527	4223	1543
SS N2+	503	455	419	912	420

Testing the Null Hypothesis defined in section 4.4.1, showed (see Table A.2) consistent results with previous finding using Token only: overall SELFSHARED repetitions were significantly more present in the actual conversations, as the null hypothesis was rejected at all linguistic levels and for each speaker. For other levels than token only (L2 to L5), the null hypothesis was also rejected for OTHERSHARED repetitions at the threshold of 0.05 four times (when only once at Token only level, L1), and approached significance nine times (with p-values ranging from 0.06 to 0.09). Furthermore, looking at level 3 (L3: Lemma+POS) and level 5 (L5: Token+POS) for OTHERSHARED, there is an increase of rejections or "close to" rejections of the null hypothesis. It could have been expected that the count of repetition for the Lemma level would be higher than the Token level in the actual dialogues, however, as can be seen in Table A.1 it is not always the case. It has to be noted that this corpus of naturally flowing conversations contained many disfluencies and a certain amount of Japanese words transcribed in *romaji* that could have influenced the count as these words were not taken into account in the lemmatisation step or recognised by the POS tagger.

As previously mentioned, the dataset only contains three dialogues so the result that "speakers repeating each other more on the second and third day at other linguistic level than Token Only", cannot be said to be robust, but it is possible to notice that there is a gradual increase in the number of H_0 rejections over the three days of the experiment. It would not be appropriate to state the usefulness of the linguistic levels, but nonetheless, there is an

Table A.2: Rejection of H_0 for the Table Talk dialogues per Speakers (Sp.), for OTHERSHARED (OS) and SELFSHARED (SS)

	OS Day 1	*SS Day 1*	*OS Day 2*	*SS Day 2*	*OS Day 3*	*SS Day 3*
L1 Sp. 1 (n)	0.45353	< 0.001 ***	0.0218 *	< 0.001 ***	0.208	< 0.001 ***
L2 Sp. 1 (n)	0.5764	< 0.001 ***	0.162	< 0.001 ***	0.89502	< 0.001 ***
L3 Sp. 1 (n)	0.491	< 0.001 ***	0.6694	< 1e-04 ***	0.0678 .	< 0.001 ***
L4 Sp. 1 (n)	1	< 0.001 ***	1	< 0.001 ***	1	< 0.001 ***
L5 Sp. 1 (n)	0.4667	< 1e-04 ***	0.33087	< 1e-05 ***	0.03043 *	< 1e-04 ***
L1 Sp. 2 (d)	0.33607	< 0.001 ***	0.7715	< 0.001 ***	0.231	< 0.001 ***
L2 Sp. 2 (d)	0.5755	< 0.001 ***	0.403	< 0.001 ***	0.08560 .	< 0.001 ***
L3 Sp. 2 (d)	0.4109	< 0.001 ***	0.0600 .	< 1e-04 ***	0.0691 .	< 0.001 ***
L4 Sp. 2 (d)	1	< 0.001 ***	1	< 0.001 ***	1	< 0.001 ***
L5 Sp. 2 (d)	0.36	< 1e-04 ***	0.2126	< 1e-05 ***	0.09104 .	< 1e-04 ***
L1 Sp. 3 (k)	0.11381	< 0.001 ***	0.8746	< 0.001 ***	0.6938	< 0.001 ***
L2 Sp. 3 (k)	0.3003	< 0.001 ***	0.974	< 0.001 ***	0.72959	< 0.001 ***
L3 Sp. 3 (k)	0.0663 .	< 0.001 ***	0.7476	< 1e-04 ***	0.3688	< 0.001 ***
L4 Sp. 3 (k)	1	< 0.001 ***	1	< 0.001 ***	1	< 0.001 ***
L5 Sp. 3 (k)	0.0795 .	< 1e-04 ***	0.22978	< 1e-05 ***	0.53492	< 1e-04 ***
L1 Sp. 4 (y)	0.67346	< 0.001 ***	0.1001	< 0.001 ***	0.4737	< 0.001 ***
L2 Sp. 4 (y)	0.7877	< 0.001 ***	0.701	< 0.001 ***	0.1767	< 0.001 ***
L3 Sp. 4 (y)	0.5159	< 0.001 ***	0.0175 *	< 1e-04 ***	0.0643 .	< 0.001 ***
L4 Sp. 4 (y)	1	< 0.001 ***	0.99435	< 0.001 ***	1	< 0.001 ***
L5 Sp. 4 (y)	0.6386	< 1e-04 ***	0.00642 **	< 1e-05 ***	0.08227 .	< 1e-04 ***
L1 Sp. 5 (g)	NA	NA	0.9507	< 0.001 ***	NA	NA
L2 Sp. 5 (g)	NA	NA	1	< 0.001 ***	NA	NA
L3 Sp. 5 (g)	NA	NA	0.9551	< 1e-04 ***	NA	NA
L4 Sp. 5 (g)	NA	NA	1	< 0.001 ***	NA	NA
L5 Sp. 5 (g)	NA	NA	0.84878	< 1e-05 ***	NA	NA

indication at this early stage that linguistic levels provide different, and possibly additional, information.

As an initial finding from this small dataset, it is possible to conclude as a preliminary finding that speakers behaviour displays different repetition patterns at different linguistic levels in casual conversations.

Appendix B

Step-by-Step Method

This appendix describes a detailed step-by-step procedure of the method used, including the script and a description of their usage and function. The method uses perl,[1] R[2] and Shell[3] scripts alternately. Two main data frames per corpus are created during the process that are used for the various statistical tests of which the results are presented in this document. This succession of steps, which has been modified and improved with each successive experiment are still only semi-automated. The finalized version of the method will compress all twelve steps into one.

[1]perl 5, version 18, subversion 2 (v5.18.2) built for x86-64-linux-gnu-thread-multi Copyright (C) 1987-2013, Larry Wall. `http://www.perl.org/` last accessed 22.06.2018

[2]R version 3.3.3 RC (2017-02-27 r72279) – "Another Canoe" Copyright (C) 2017 The R Foundation for Statistical Computing Platform: x86-64-pc-linux-gnu (64-bit) `https://www.r-project.org/` last accessed 22.06.2018

[3]GNU bash, version 4.3.11(1)-release (x86-64-pc-linux-gnu) Copyright (C) 2013 Free Software Foundation, Inc. License GPLv3+: GNU GPL version 3 or later `http://gnu.org/licenses/gpl.html`

Table B.1: Step-by-Step Method description

Step	Description
Step 1: Corpus preparation	Create a folder /DIALOGUE containing the transcripts. Each line must have the dialogue participant identifier (ID) at the start of the turn. For each dialogue or dialogue section, create two version, one with a tabulation between the participant identifier and the dialogue turn, and one version with only a space between the two just mentioned. Remove punctuation. Refer to file: USAGE-CORPUS-eg
Step 2: Creation of subdirectories	Create a folder that will contain the experiment itself, each dialogue or dialogue section containing five subdirectories: mkdir -p S01/AnalysisLemma S01/AnalysisLemma+POS S01/AnalysisPOS S01/AnalysisToken S01/AnalysisToken+POS
Step 3: Corpus Labelling	Scripts using the TreeTagger: make sure have the correct language. file-path for each dialogue, directing to the corpus files without tabulation. Place yourself in the root directory and execute: perl ./preprocesstreetag-lemma-CORPUS.pl perl ./preprocesstreetag-lemma-POS-CORPUS.pl perl ./preprocesstreetag-POS-CORPUS.pl perl ./preprocesstreetag-Token-POS-CORPUS.pl This step will need to check the correctness of language used as well as filepath at each new usage. Remove first line of all created files with a pipeline(USAGE-CORPUS-lemma-eg)
Step 4: Normalization of pronouns	Execute iyp-treat.pl script (files with tabulation for L1 token level, and output files from labelled corpus for other levels) perl ./iyp-treat.pl -i S01Full.txt Move files to correct subdirectory: mv S01Full.txt.iy.p0.c1 S01/AnalysisToken
Step 5: Assign time stamps	The time stamp assigned to each turn is random in its length but follow a chronological order: dialog-treat-time.pl perl ./dialog-treat-time.pl -i S01/AnalysisToken/S01Full.txt.iy.p0.c1
Step 6: Turn 10 times randomization	Execute randomization script to correct output file: perl ./transcriptrandomizer.5.pl -i >S01/AnalysisLemma/S01FullLemma.iy.p0.c1.fmtd This operation does not support the use of pipeline, therefore, use shell script: sh ./randomizeToken.sh containing the execution for each file.
Step 7: Concatenation	Concatenation of actual and randomization files: cat S01/S01Full/AnalysisLemma/*.parsed >S01Full/mergedS01FullLemma.data
Step 8: Post Treatment	Addition of a column containing the levels identifiers: perl ./CORPUSPostProcessingLevel.pl (Addition of the level) Then create first main dataframe by concatenation: cat Dialogues/*Level.data >mergedCORPUS.data This dataframe contain each speaker turn with his count of OTHERSHARED and SELFSHARED repetitions, number of token, reality (Actual dialogue (0) or randomization), Ngram length, and level.
Step 9: R, format.	Removing unused columns and factorization of speakers ID (for following steps): >createCORPUSDial.R
Step 10: Statistical Model	Tukey Test for OTHERSHARED and SELFSHARED repetitions: >Pvalue.FullCorpus.R The output files gives for each dialogue, n-gram length, level and speaker, the result P-Values, Odds Ratio and Confidence Intervals of the tests.
Step 11: Extraction of *p*-values	Use the shell script: extractionPValues.sh that will execute all the perl files extracting the p-values from the previous step output files. Then merge them: cat *OS.txt >mergedPvalueOS.txt cat *SS.txt >mergedPvalueSS.txt
Step 12: Creation of final dataframe	Create file CORPUSDialData in excel file by merging the previous step files with their corresponding, dialogue ID, n-gram length, level and speaker ID. CORPUSDialData.csv

Appendix C

Reconstruction of Deviation Scores

The following tables (C.1, C.2, C.3, C.4) gives the precise citation that have been used for the replication of the HCRC Map Task deviation scores. The Table C.4 give the Scores found following the different counting methods of which the description is given in section 4.5.1.

Table C.1: Description found in direct reference to the HCRC Map Task

Source	Textual description of Deviation score found for Reconstruction	Notes
http://groups.inf.ed.ac.uk/maptask/maptask-description.html — Last consulted: Fri Sep 12, 2019	Measuring task performance The main measure of task performance that has been used for the Map Task is in terms of how far the route that the follower has drawn deviates from the route shown on the giver's map. To reconstruct it, using the original A3 size maps, trace the giver's route on acetate marked with a one centimetre square grid, and impose it over the follower's map. The deviation score is the number of squares between the two routes. [...] The method was first described in print by A. H. Anderson, A. Clark, and J. Mullin (1991) Introducing information in dialogues: How young speakers refer and how young listeners respond. Journal of Child Language, 18, 663-687.	This description is the main basis used to reconstruct the deviation scores counting methods described in section 4.5.3.
Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... & Sotillo, C. (1991). The HCRC map task corpus. Language and speech, 34(4), 351-366. — Extracted from pages: 353 & 364	In the Map Task, however, the overall success achieved by any pair of speakers is measurable in terms of the deviation between the original route found on the map of the instruction Giver and that reproduced by the Instruction Follower. To measure such route deviations, a 1 cm grid is used on which the route is represented by filled grid squares. A deviation score in grid cells gives an objective non-linguistic estimate of communicative success. [...] This approach has already been used with earlier versions of the Map Task to determine components of communicative success in young speakers (Anderson, Clark, and Mullin, 1991). [...] more detailed documentation on the Glasgow HCRC Database, [...] and on the design of the Map Task Corpus itself (McAllister, Sotillo, Bard, and Anderson, 1990).	The description is ambiguous as it is not known if it is the deviation scores or the map task technique as a whole that is concerned.

Table C.2: Description found in direct reference to the HCRC Map Task

Source	Textual description of Deviation score found for Reconstruction	Notes
Anderson, A., Clark, A., & Mullin, J. (1991). Introducing information in dialogues: Forms of introduction chosen by young speakers and the responses elicited from young listeners. Journal of Child Language, 18(3), 663-687. Other citation format found: Introducing information in dialogues: How young speakers refer and how young listeners respond. A. Anderson, A. Clark, J. Mullin - Journal of Child Language, 18, 663-687, 1991 — Extracted from pages: 667 *to* 669	In the dialogue task which we employ, the two subjects both have copies of a simple schematic map. One subject has a route shown on her version of the map, and her task is to describe this so that her listener can follow the route. Both subjects are warned that the maps have been drawn by different explorers and that there may be some differences between them. The children are encouraged to talk freely to one another and to ask questions if they do not understand what their partner is saying. In this task then, one of the main requirements for the speaker is to discover whether her listener shares her knowledge of any feature she wishes to discuss. [...] Method – Subjects and procedureIn this study, the following subjects were tested: 33 pairs with a mean age of 7;9 (range 7;5-8;7), 26 pairs with a mean age of 10; 2 (range 9; 6-10;8) and 26 pairs with a mean age of 13;2 (range 12;8-13;7). [...] Approximately half the subjects in each group were male, half female. Subjects tackled one map, then swapped instruction-giver and instruction-follower roles and tackled a second different map task. [...] There were four different sets of maps, each with the following characteristics: the maps showed ten different features directly relevant to the route, and another eight peripheral features which could be used as landmarks. [...] Two features were shown only on the instruction follower's map. All features were labelled. This procedure produced a database of 170 compatible dialogues for analysis.Materials In this research, we employed a dialogue task, the map task, developed by Brown, Anderson, Yule & Shillcock (1984). As described above, both subjects had copies of a simple schematic map. One member of the pair was randomly assigned to the role of instruction giver (I.G.) and only that copy of the map had the route shown on it. The task was to instruct the partner (instruction follower-I.F.) how to draw the route on his or her copy. The children were instructed as follow:To the instruction giver You and your partner both got a map of the same place. Your map has got a route on it. It's the ONLY SAFE ROUTE through all the dangers. Your partner hasn't got a route on her/his map. Your job is to describe the route to your partner so that (s)he can draw it on her/his map. You must describe it Exactly because it's the ONLY SAFE ROUTE. The maps have been drawn by different explorers, so they might not be quite the same; there might be some differences.To the instruction follower You and your partner have both got a map of the same place. Your partner's map has got a route on it. It's the ONLY SAFE ROUTE through all the dangers. (S)he's going to tell you what the route is. Your job is to draw the route on your map. Listen carefully to what your partner says, and ask questions if there's anything you're not sure about. You must draw it EXACTLY because it's the ONLY SAFE ROUTE. The maps have been drawn by different explorers, so they might not be quite the same; there might be some differences. Do you understand what you're supposed to do?	No mention of the scores at all, even if it is the paper that is cited by many authors as the reference for deviation scores. This work often quoted as the reference for a task score is therefore false, or at least misleading.

Table C.3: First reference to a scoring system in relation to a map task method

Source	Textual description of Deviation score found for Reconstruction	Notes
Teaching Talk: Strategies for Production and Assessment. Gillian Brown, Anne Anderson, Richard Shillcock & George Yule (Cambridge: Cambridge University Press, 1984). — Extracted from page: 70	Chapter 4 section 5 Co-operative tasks [...] One such task, which is very popular with pupils, is where the information required to complete a task is distributed between two pupils and they have to cooperate to complete the task. Thus speakers A may have in front of him a map with a safe route marked across it. Speaker B may have a map of the same island but one said to be 'made by an earlier explorer' which contains some of the features on A's map, but not all, and contains, in addition, three features which are not marked on A's map. A is asked to describe to B the safe route across the island so that B can draw it in on his map. The reason a task like this is difficult to grade or assess is because the behaviour of one member of the pair depends so much on the behaviour of the other member.	Earliest description found of the map task technique
Teaching Talk: Strategies for Production and Assessment. Gillian Brown, Anne Anderson, Richard Shillcock & George Yule (Cambridge: Cambridge University Press, 1984). — Extracted from page: 111	5.13 Co-operative tasks: The map task, described in Chapter 4, represents a means of making pupils aware of some of the very sophisticated skills involved in communicating with a partner who is able, as outside school, to ask questions and comment on what the other is saying to him. Pupils performing in this task may only be assessed as pairs, so it seems to be more appropriate to use the assessment of the task in order to stimulate pupils to consider the skills involved, [...]. One simple means of assessing co-operation in the transfer of information in the map task, is to inspect the route which one of the partners draws on the unmarked map, form his partner's instructions. The object of the exercise was to replicate the route on the marked map. It will then be possible, when the pupils compare their two maps, for gross differences to be remarked upon. Remember that it is not draughtsmanship which is being assessed, but the ability to recognise that the other person needs to be told whether to go right or left, or up or down at crucial parts of the maps. The route which was described may be scored simply by awarding a mark for each feature of the map which the route goes to correctly, given that the route passes that features on the correct side and in the correct direction. It will be appreciated that much depends on the precise design of the pairs of maps used. Maps with conflicting features increase the need for sensitive co-operation. The discovery by one partner that the other partner does not share a crucial feature, or that the shared feature is displaced, may be scored in a simple yes/no way; sometimes it is clear from the route produced that a misunderstanding was never resolved, sometimes it is necessary to listen to the taped performance in order to discover the confusion. It is possible to assess aspects of the taped performance. Listening once to the tape will reveal points in the dialogue at which the pupil who is describing the route is interrupted by his partner, often to request him to slow down, to repeat something, to go back to a certain point on the route, to give extra information, to confirm something which his partner only suspects, and so on. By stopping the tape briefly at these points, it should be possible for the pupils themselves to keep note of how adequately the information is conveyed.	Earliest mention of a score system using the maps. It is not a deviation score that is proposed, however, but more a feature correctness score.

Table C.4: Mentions and description from other authors, in other map tasks

Source	Textual description of Deviation score found for Reconstruction
The DCIEM Map Task Corpus: Spontaneous dialogue under sleep deprivation and drug treatment. Bard, E. G., et al. *Speech Communication.* (1996), 20(1-2), 71-84.	Maps drawn by Followers were first analyzed for accuracy by means of a route deviation score. Each map was overlaid with a 1 cm grid on which the squares covering the Giver's printed route were blacked out. The deviation score was the number of grid squares which would (a) cover the parts of the Follower's route still visible and (b) fill the space between those visible sections and the model route.
Alignment and task success in spoken dialogue. D. Reitter, J.D. Moore. *Journal of Memory and Language* 76, (2014), 29–46.	The Map Task consists of re-tracing a defined route according to the interactive description provided by the other interlocutor. So, task performance is measured in terms of how far the route that the follower has drawn deviates from the route shown on the giver's map. To compute this for each dialogue, the developers of the Map Task corpus overlaid the giver's map on the follower's map and computed the area covered in between the paths (PATHDEV).
Predicting success in dialogue. Reitter, David, and Johanna D. Moore. *Association for Computational Linguistics* (ALC), (2007). p.808.	The Map Task provides us with a precise measure of success, namely the deviation of the predefined and followed route. Success can be quantified by computing the inverse deviation between subjects' paths. Both subjects in each trial were asked to draw "their" respective route on the map that they were given. The deviation between the respective paths drawn by interlocutors was then determined as the area covered in between the paths (PATHDEV).
Is disfluency just difficult? Bard, Ellen G., Robin J. Lickley, and Matthew P. Aylett. *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech.* (2001).	Deviation score is the mismatch in cm2 between the model route on IG's map and the route ultimately drawn on the IF's.

Table C.5: Deviation Score per Method of Counting

DialID	M1	M2	M3	M4	HCRCDevS	DialID	M1	M2	M3	M4	HCRCDevS
q1nc1	84	99	119	104	78	q1ec1	148	170	197	175	135
q1nc2	221	237	269	253	204	q1ec2	287	321	363	329	227
q1nc3	55	63	69	61	40	q1ec3	92	117	133	108	74
q1nc4	56	66	81	71	53	q1ec4	129	147	168	150	152
q1nc5	51	62	72	61	35	q1ec5	121	151	172	142	105
q1nc6	35	46	56	45	34	q1ec6	52	59	67	60	53
q1nc7	28	32	39	35	18	q1ec7	121	145	164	140	104
q1nc8	95	105	115	105	69	q1ec8	97	130	143	110	85
q2nc1	51	63	69	57	51	q2ec1	128	141	162	149	120
q2nc2	100	116	131	115	90	q2ec2	104	125	138	117	97
q2nc3	56	63	79	72	44	q2ec3	115	129	161	147	105
q2nc4	101	107	124	118	104	q2ec4	119	138	154	135	142
q2nc5	51	61	81	71	41	q2ec5	61	67	83	77	66
q2nc6	94	123	144	115	99	q2ec6	78	92	111	97	89
q2nc7	30	36	42	36	20	q2ec7	40	44	48	44	28
q2nc8	27	37	44	34	19	q2ec8	72	85	93	80	58
q3nc1	232	263	281	250	105	q3ec1	159	193	220	186	140
q3nc2	96	115	130	111	114	q3ec2	197	239	261	219	191
q3nc3	55	68	84	71	56	q3ec3	24	33	40	31	28
q3nc4	42	49	60	53	45	q3ec4	133	145	166	154	154
q3nc5	129	145	176	160	139	q3ec5	29	35	38	32	20
q3nc6	86	100	122	108	99	q3ec6	75	79	102	98	89
q3nc7	66	85	92	73	55	q3ec7	77	93	102	86	57
q3nc8	51	58	64	57	42	q3ec8	50	61	69	58	37
q4nc1	83	106	121	98	74	q4ec1	200	232	249	217	204
q4nc2	112	138	159	133	83	q4ec2	127	144	157	140	117
q4nc3	73	77	97	93	75	q4ec3	35	39	49	45	11
q4nc4	31	33	45	43	45	q4ec4	72	88	111	95	73
q4nc5	30	37	46	39	21	q4ec5	26	29	38	35	26
q4nc6	39	41	53	51	37	q4ec6	37	42	57	52	49
q4nc7	40	49	55	46	28	q4ec7	52	67	78	63	43
q4nc8	39	47	52	44	20	q4ec8	48	58	67	57	30
q5nc1	75	107	121	89	64	q5ec1	139	168	191	162	146
q5nc2	262	293	324	293	201	q5ec2	197	246	267	218	187
q5nc3	59	66	85	78	29	q5ec3	78	94	105	89	66
q5nc4	92	102	122	112	90	q5ec4	6	7	8	7	4
q5nc5	36	44	50	42	17	q5ec5	45	52	56	49	28
q5nc6	78	93	110	95	65	q5ec6	40	52	65	53	38
q5nc7	132	162	186	156	108	q5ec7	157	209	221	169	154
q5nc8	44	54	65	55	28	q5ec8	132	147	165	150	102
q6nc1	83	97	111	97	83	q6ec1	55	69	78	64	38
q6nc2	75	87	94	82	56	q6ec2	112	119	130	123	81
q6nc3	170	209	240	201	161	q6ec3	85	105	131	111	97
q6nc4	121	156	181	146	108	q6ec4	47	51	64	60	35
q6nc5	194	242	271	223	178	q6ec5	63	72	89	80	60
q6nc6	47	49	74	72	52	q6ec6	78	96	108	90	32
q6nc7	68	83	94	79	52	q6ec7	51	58	61	54	54
q6nc8	167	193	209	183	145	q6ec8	83	86	91	88	67
q7nc1	179	228	234	185	157	q7ec1	101	122	133	112	86
q7nc2	63	74	90	79	56	q7ec2	12	13	17	16	7
q7nc3	99	111	130	118	104	q7ec3	33	52	60	41	27
q7nc4	20	20	25	25	35	q7ec4	51	54	60	57	57
q7nc5	41	47	63	57	34	q7ec5	19	27	38	30	24
q7nc6	19	21	26	24	26	q7ec6	9	9	9	9	7
q7nc7	56	72	78	62	48	q7ec7	45	51	60	54	27
q7nc8	43	52	59	50	53	q7ec8	46	53	65	58	38
q8nc1	43	50	68	61	25	q8ec1	51	60	76	67	37
q8nc2	152	166	190	176	112	q8ec2	132	164	194	162	135
q8nc3	52	66	78	64	54	q8ec3	12	13	19	18	11
q8nc4	48	58	76	66	76	q8ec4	43	52	69	60	54
q8nc5	45	63	76	58	40	q8ec5	18	20	25	23	25
q8nc6	53	68	88	73	62	q8ec6	37	44	61	54	47
q8nc7	64	74	85	75	47	q8ec7	51	60	73	64	43
q8nc8	58	65	67	60	49	q8ec8	32	37	42	37	19

Table C.6: HCRC MapTask Conditions

DialID	EyeCon.	Fam	Gender	Partner	Time	Task	DialID	EyeCon.	Fam	Gender	Partner	Time	Task
q1nc1	n	U	M	F	18:33	1	q1ec1	e	U	F	M	04:25	1
q1nc2	n	U	F	F	15:26	1	q1ec2	e	U	F	F	05:34	1
q1nc3	n	F	F	F	09:50	2	q1ec3	e	F	M	F	08:15	2
q1nc4	n	F	F	M	06:47	2	q1ec4	e	F	F	F	03:37	2
q1nc5	n	U	F	F	05:49	3	q1ec5	e	U	F	F	06:05	3
q1nc6	n	U	F	M	03:40	3	q1ec6	e	U	M	F	02:37	3
q1nc7	n	F	F	F	11:11	4	q1ec7	e	F	F	M	07:35	4
q1nc8	n	F	M	F	06:42	4	q1ec8	e	F	F	F	04:03	4
q2nc1	n	U	M	M	03:37	1	q2ec1	e	U	M	F	06:21	1
q2nc2	n	U	M	M	08:48	1	q2ec2	e	U	F	M	04:18	1
q2nc3	n	F	M	M	10:25	2	q2ec3	e	F	F	F	04:40	2
q2nc4	n	F	M	M	05:39	2	q2ec4	e	F	M	M	09:58	2
q2nc5	n	U	M	M	05:12	3	q2ec5	e	U	M	F	05:12	3
q2nc6	n	U	M	M	03:59	3	q2ec6	e	U	F	M	05:35	3
q2nc7	n	F	M	M	06:21	4	q2ec7	e	F	F	F	03:59	4
q2nc8	n	F	M	M	08:10	4	q2ec8	e	F	M	M	05:31	4
q3nc1	n	U	M	F	04:18	1	q3ec1	e	U	M	M	08:50	1
q3nc2	n	U	F	F	08:58	1	q3ec2	e	U	M	M	06:11	1
q3nc3	n	F	F	F	05:31	2	q3ec3	e	F	M	M	07:28	2
q3nc4	n	F	F	M	05:15	2	q3ec4	e	F	M	M	06:14	2
q3nc5	n	U	F	F	04:11	3	q3ec5	e	U	M	M	04:15	3
q3nc6	n	U	F	M	03:06	3	q3ec6	e	U	M	M	03:55	3
q3nc7	n	F	F	F	03:34	4	q3ec7	e	F	M	M	03:57	4
q3nc8	n	F	M	F	04:48	4	q3ec8	e	F	M	M	04:47	4
q4nc1	n	U	M	F	06:42	1	q4ec1	e	U	M	F	03:32	1
q4nc2	n	U	F	M	09:19	1	q4ec2	e	U	F	M	10:02	1
q4nc3	n	F	F	F	14:41	2	q4ec3	e	F	F	F	08:14	2
q4nc4	n	F	M	M	10:07	2	q4ec4	e	F	M	M	07:43	2
q4nc5	n	U	M	F	10:32	3	q4ec5	e	U	M	F	07:18	3
q4nc6	n	U	F	M	09:13	3	q4ec6	e	U	F	M	05:18	3
q4nc7	n	F	F	F	09:48	4	q4ec7	e	F	F	F	08:00	4
q4nc8	n	F	M	M	07:15	4	q4ec8	e	F	M	M	07:17	4
q5nc1	n	F	F	F	06:52	1	q5ec1	e	F	M	M	08:12	1
q5nc2	n	F	F	F	08:17	1	q5ec2	e	F	F	M	05:33	1
q5nc3	n	U	F	F	04:05	2	q5ec3	e	U	M	M	04:31	2
q5nc4	n	U	F	F	08:05	2	q5ec4	e	U	M	F	04:51	2
q5nc5	n	F	F	F	04:06	3	q5ec5	e	F	M	M	04:23	3
q5nc6	n	F	F	F	06:03	3	q5ec6	e	F	M	F	04:15	3
q5nc7	n	U	F	F	04:15	4	q5ec7	e	U	F	M	03:51	4
q5nc8	n	U	F	F	04:55	4	q5ec8	e	U	M	M	04:36	4
q6nc1	n	F	F	F	05:32	1	q6ec1	e	F	F	F	08:51	1
q6nc2	n	F	M	M	13:04	1	q6ec2	e	F	F	M	13:37	1
q6nc3	n	U	M	F	05:40	2	q6ec3	e	U	M	F	07:33	2
q6nc4	n	U	F	M	07:47	2	q6ec4	e	U	F	F	04:41	2
q6nc5	n	F	F	F	04:54	3	q6ec5	e	F	F	F	06:07	3
q6nc6	n	F	M	M	06:22	3	q6ec6	e	F	M	F	04:07	3
q6nc7	n	U	M	F	05:18	4	q6ec7	e	U	F	F	04:09	4
q6nc8	n	U	F	M	05:24	4	q6ec8	e	U	F	M	06:25	4
q7nc1	n	F	F	F	05:59	1	q7ec1	e	F	M	M	06:45	1
q7nc2	n	F	F	F	10:31	1	q7ec2	e	F	M	M	16:07	1
q7nc3	n	U	F	F	06:36	2	q7ec3	e	U	M	M	07:26	2
q7nc4	n	U	F	F	07:42	2	q7ec4	e	U	M	M	07:52	2
q7nc5	n	F	F	F	07:04	3	q7ec5	e	F	M	M	05:27	3
q7nc6	n	F	F	F	02:43	3	q7ec6	e	F	M	M	19:00	3
q7nc7	n	U	F	F	04:49	4	q7ec7	e	U	M	M	05:21	4
q7nc8	n	U	F	F	03:19	4	q7ec8	e	U	M	M	13:45	4
q8nc1	n	F	F	F	06:19	1	q8ec1	e	F	M	M	09:50	1
q8nc2	n	F	M	F	03:17	1	q8ec2	e	F	M	M	04:16	1
q8nc3	n	U	F	F	04:57	2	q8ec3	e	U	M	M	05:13	2
q8nc4	n	U	F	M	04:49	2	q8ec4	e	U	M	M	10:27	2
q8nc5	n	F	F	F	10:01	3	q8ec5	e	F	M	M	05:39	3
q8nc6	n	F	F	M	03:42	3	q8ec6	e	F	M	M	10:22	3
q8nc7	n	U	M	F	04:50	4	q8ec7	e	U	M	M	04:50	4
q8nc8	n	U	F	F	06:11	4	q8ec8	e	U	M	M	04:36	4

List of Figures

List of Tables

Glossary

SHARED

Refers to a token repeated either by another person or the same person.

OTHERSHARED

Refers to the repetition of a token uttered by another person 54–58, 78–83, 86, 88–90, 93–95, 100–102, 107–110, 113–115, 118, 122–126, 133, 134, 137

SELFSHARED

Refers to the repetition of a token uttered by the same person 54–58, 78, 80–83, 86, 88–90, 92–94, 100, 101, 107–109, 112–115, 118, 122–124, 127, 134, 137

deviation score

Refers to a measure of successful task management in a map task, that is the sum of squares found between the routes of an Information Giver and an Information Follower, with the maps placed on top of each other and divided into a grid of one centimetre squares 35, 40, 50, 51, 64, 67, 78–80, 83, 86–89, 107, 127

Dialogue Details

Eye-Contact

eye-contact

Refers to subjects that have eye contact with each other during the corpora recordings used as data in this thesis 28, 35, 39, 64, 83, 94, 95, 123, 124

E Abbreviation of “eye-contact" that is used in figures and tables of this thesis .... *see also* eye-contact

Eye Abbreviation of “eye-contact" that is used in figures and tables of this the-

sis .... 39, 81, *see also* eye-contact

w/ EC Abbreviation of “eye-contact" that is used in figures and tables of this thesis .... *see also* eye-contact

no eye-contact

Refers to subjects that do not have eye contact with each other during the corpora recordings used as data in this thesis 29, 35, 39, 50, 51, 64, 83, 86, 105, 123

nE Abbreviation of “no eye-contact" that is used in figures and tables of this thesis .... *see also* no eye-contact

NoEye Abbreviation of “no eye-contact" that is used in figures and tables of this thesis .... 39, 81, *see also* no eye-contact

w/o EC Abbreviation of “no eye-contact" that is used in figures and tables of this thesis .... *see also* eye-contact

Familiarity

familiar

Refers to subjects familiar with each other in the data and results from these subjects 35, 39, 83, 86, 88, 89, 92, 105, 117–119, 121, 123, 124, 126, 127, 129, 131

F Abbreviation of “familiar" that is used in figures and tables of this thesis .... *see also* familiar

Fam Abbreviation of “familiar" that is used in figures and tables of this thesis .... 39, 81, *see also* familiar

unfamiliar

Refers to subjects unfamiliar with each other in the data and results from these subjects 5, 29, 30, 35, 39, 51, 83, 87–89, 92, 105, 118, 119, 123, 124, 127, 129, 132, 133

U Abbreviation of “unfamiliar" that is used in figures and tables of this thesis .... *see also* unfamiliar

UnFam Abbreviation of “unfamiliar" that is used in figures and tables of this thesis .... 39, 81, *see also* unfamiliar

Gender

female Refers to female subjects of the data and results from female subjects 28, 35, 39, 43, 45, 49–51, 81, 83, 86, 98, 105, 123, 127

♀ Abbreviation of “female" that is used in figures and tables of this thesis .... *see also* female

male Refers to male subjects of the data and results from male subjects 28, 39, 43, 45, 49, 51, 81, 83, 86, 98, 123

♂ Abbreviation of “male" that is used in figures and tables of this thesis .... *see also* male

Role

Information Follower

Refers to subjects with the role of following instructions in the collection and/or data and results from Information Follower subjects — also found to be referred to as Instruction Follower in literature. 37, 50, 51, 64, 66–68, 72, 78, 79, 81, 83, 107, 118, 122, 126, 127

IF Abbreviation of “Information Follower" that is used in figures and tables of this thesis .... 37, 39, 40, 44, 50, 51, 58, 66, 68–70, 78, 79, 81, 83, 86, 93–95, 106, 108–110, 112–115, 120, 122, 123, 142, 145, *see also* Information Follower

Information Giver

Refers to subjects with the role of giving instructions in the collection and/or data and results from Information Giver subjects — also found to be referred to as Instruction Giver in literature. 28, 29, 37, 50, 51, 64, 67, 68, 78, 81, 83, 107, 113, 114, 118, 122, 126, 127

IG Abbreviation of “Information Giver" that is used in figures and tables of this thesis .... 37, 39, 40, 44, 50, 51, 58, 66, 68, 78–81, 83, 86, 93–95, 106, 108–110, 112–115, 118, 120, 122, 123, 142, *see also* Information Giver

Bibliography

Albert, S. (2017). Research methods: Conversation analysis. In M. F. Shober, D. N. Rapp, & A. M. Britt (Eds.), *The Routledge Handbook of Discourse Processes* (pp. 99–108). Routledge. doi: 10.4324/9781315687384

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, *41*(3), 273–287. doi: 10.1007/s10579-007-9061-5

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307. doi: 10.1038/d41586-019-00857-9

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*(4), 351–366. doi: 10.1177/002383099103400404

Anderson, A. H., Clark, A., & Mullin, J. (1991). Introducing information in dialogues: forms of introduction chosen by young speakers and the responses elicited from young listeners. *Journal of Child Language*, *18*(3), 663–687.

Ansaldo, U., Don, J., & Pfau, R. (Eds.). (2010). *Parts of Speech: Empirical and theoretical advances*. John Benjamins. doi: 10.1075/bct.25

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., ... Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, *27*(2), 108-119. doi: 10.1002/per.1919

Bard, E. G., Sotillo, C., Anderson, A. H., Thompson, H. S., & Taylor, M. M. (1996). The DCIEM Map Task Corpus: Spontaneous dialogue under sleep deprivation and drug treatment. *Speech Communication*, *20*(1-2), 71–84. doi: 10.1016/S0167-6393(96)00045-3

Behan, L. (2010). *The structure of conversation real or random?* (Unpublished Bachelor's Thesis). Trinity College Dublin, The University of Dublin.

Beňuš, Š., Levitan, R., & Hirschberg, J. (2012). Entrainment in spontaneous speech: the case of filled pauses in supreme court hearings. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 793–797). IEEE. doi: 10.1109/CogInfoCom.2012.6421959

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal Behavior*, *12*(2), 120–138. doi: 10.1007/BF00986930

Bernieri, F. J., Reznick, J. S., & Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology*, *54*(2), 243–253. doi: 10.1037/0022-3514.54.2.243

Bernieri, F. J., & Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony. In Feldman, R. S. and Rimé, B. (Ed.), *Fundamentals of Nonverbal Behavior: Studies in Emotion and Social Interaction.* (pp. 401–432). Cambridge University Press.

Bock, K. (1989). Closed-class immanence in sentence production. *Cognition*, *31*(2), 163–186. doi: 10.1016/0010-0277(89)90022-X

Bolinger, D. L. (1961). Syntactic Blends and Other Matters. *Language*, *37*(3), 366–381. doi: 10.2307/411078

Branigan, H., Lickley, R., & McKelvieDavid. (1999). Non-Linguistic Influences on Rates of Disfluency in Spontaneous Speech. In *Proceedings of ICPhS XIV (14th International Congress of Phonetic Sciences)* (pp. 387–390). San Francisco, California, USA.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25. doi: 10.1016/S0010-0277(99)00081-5

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355–2368. doi: 10.1016/j.pragma.2009.12.012

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57. doi: 10.1016/j.cognition.2011.05.011

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Nass, C. (2003). Syntactic alignment between computers and people: The role of belief about mental states. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 186–191). Boston, Massachusetts, USA: Cognitive Science Society.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two Minds, One Dialog: Coordinating Speaking and Understanding. In *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 53, pp. 301–344). Academic Press. doi: 10.1016/S0079-7421(10)53008-1

Brennan, S. E., & Hanna, J. E. (2009). Partner-Specific Adaptation in Dialog. *Topics in Cognitive Science*, *1*(2), 274–291. doi: 10.1111/j.1756-8765.2009.01019.x

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

Briggs, W. M. (2012). It is Time to Stop Teaching Frequentism to Non-statisticians. *arXiv:1201.2590*. Retrieved from `https://arxiv.org/abs/1201.2590`

Brooker, C., & Harris, O. (2013). *Be right back.* Endemol Shine UK.

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

Cacciamani, S., Cesareni, D., Martini, F., Ferrini, T., & Fujita, N. (2012). Influence of participation, facilitator styles, and metacognitive reflection on knowledge building in online university courses. *Computers & Education*, *58*(3), 874-884. doi: 10.1016/j.compedu.2011.10.019

Campbell, N. (2009). An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data. In *Proceedings of INTERSPEECH'09: the 10th Annual Conference of the International Speech Communication Association* (pp. 2159–2162). Brighton, United Kingdom: ISCA.

Cappella, J. N. (1991). Mutual Adaptation and Relativity of Measurement. In B. M. Montgomery & S. Duck (Eds.), *Studying Interpersonal Interaction* (Vol. 1, pp. 103–117). Guilford Press.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., & Anderson, A. H. (1997). The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, *23*(1), 13–32.

Charny, J. E. (1966). Psychosomatic manifestations of rapport in psychotherapy. *Psychosomatic Medicine*, *28*(4), 305–315.

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

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC, US: American Psychological Association. doi: 10.1037/10096-006

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of memory and language*, *50*(1), 62–81. doi: 10.1016/j.jml.2003.08.004

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. doi: 10.1016/0010-0277(86)90010-7

Clayes, E. L., & Anderson, A. H. (2007). Real faces and robot faces: The effects of representation on computer-mediated communication. *International Journal of Human-Computer Studies*, *65*(6), 480–496. doi: 10.1016/j.ijhcs.2006.10.005

Colman, M. (2012). *Quantifying mutual-understanding in dialogue* (Unpublished doctoral dissertation). Queen Mary University of London.

Colman, M., Eshghi, A., & Healey, P. (2008). Quantifying ellipsis in dialogue: an index of mutual understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 96–99). Association for Computational Linguistics.

Colman, M., & Healey, P. (2011). The Distribution of Repair in Dialogue. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 1563–1568). Austin, Texas, USA: Cognitive Science Society.

Condon, W. S., & Ogston, W. D. (1966). Sound film analysis of normal and pathological behavior patterns. *Journal of Nervous and Mental Disease*, *143*(4), 338–347.

Condon, W. S., & Sander, L. W. (1974). Synchrony Demonstrated between Movements of the Neonate and Adult Speech. *Child Development*, *45*(2), 456–462.

Cowan, B. R., Doyle, P., Edwards, J., Garaialde, D., Hayes-Brady, A., Branigan, H. P., ... Clark, L. (2019). What's in an accent? The impact of accented synthetic speech on lexical choice in human-machine dialogue. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (pp. 1–8). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3342775.3342786

Crawley, M. J. (2005). *Statistics: An introduction using R*. John Wiley & Sons.

Curl, T. S. (2005). Practices in other-initiated repair resolution: The phonetic differentiation of 'repetitions'. *Discourse Processes*, *39*(1), 1–43. doi: 10.1207/s15326950dp3901_1

Cushing, S. (1994). *Fatal words: Communication clashes and aircraft crashes*. University of Chicago Press.

Davies, B. L. (1997). *An empirical examination of cooperation, effort and risk in task-oriented dialogue* (Unpublished doctoral dissertation). University of Edinburgh.

Davies, B. L. (2006). Testing dialogue principles in task-oriented dialogues: An exploration of cooperation, collaboration, effort and risk. *Leeds Working Papers in Linguistics and Phonetics*, *11*, 30–64.

Davis, M. (2001). *Engines of logic: Mathematicians and the origin of the computer*. USA: W. W. Norton & Co., Inc.

Dubuisson Duplessis, G., Clavel, C., & Landragin, F. (2017). Automatic Measures to Characterise Verbal Alignment in Human-Agent Interaction. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2017* (pp. 71–81). Saarbrücken, Germany: Association for Computational Linguistics.

Duran, N., Dale, R., & Galati, A. (2016). Toward Integrative Dynamic Models for Adaptive Perspective Taking. *Topics in Cognitive Science*, *8*(4), 761–779. doi: 10.1111/tops.12219

Feldman, R. (2007). Parent–infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions. *Journal of Child psychology and Psychiatry*, *48*(3-4), 329–354. doi: 10.1111/j.1469-7610.2006.01701.x

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, *23*(8), 931–939. doi: 10.1177/0956797612436816

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157. doi: 10.1016/j.newideapsych.2013.03.005

Fusaroli, R., & Tylén, K. (2016). Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance. *Cognitive Science*, *40*(1), 145–171. doi: 10.1111/cogs.12251

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, *25*(3), 203–219. doi: 10.1016/0022-1031(89)90019-X

Gamallo, P., & Garcia, M. (2013). *FreeLing e TreeTagger: um estudo comparativo no âmbito do Português* (Tech. Rep.). Universidade de Santiago de Compostela: Centro Singular de Investigação em Tecnologias da Informação (CITIUS).

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218. doi: 10.1016/0010-0277(87)90018-7

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. doi: 10.1198/000313006X152649

Giles, H., Coupland, J., & Coupland, N. (Eds.). (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press. doi: 10.1017/CBO9780511663673

Glas, N., & Pelachaud, C. (2015). Definitions of Engagement in Human-Agent Interaction. In *Proceedings of The First International Workshop on ENgagement in HumAN Computer IntEraction (ENHANCE) @ACII 2015* (pp. 944–949). IEEE. doi: 10.1109/ACII.2015.7344688

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, *5*(10), 3–8. doi: 10.3102/0013189X005010003

Goffman, E. (1996). *Behavior in public places*. Simon and Schuster.

Gumperz, J. J. (1982). *Discourse strategies* (Vol. 1). Cambridge University Press.

Han, Y., & Gmytrasiewicz, P. (2019). IPOMDP-Net: A Deep Neural Network for Partially Observable Multi-Agent Planning Using Interactive POMDPs. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 6062–6069). Association for the Advancement of Artificial Intelligence. doi: 10.1609/aaai.v33i01.33016062

Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, *97*(3), 363–386. doi: 10.1037/0033-2909.97.3.363

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological science*, *15*(6), 409–414. doi: 10.1111/j.0956-7976.2004.00693.x

Haspelmath, M., & Michaelis, S. M. (2017). Analytic and synthetic: Typological change in varieties of European languages. In I. Buchstaller & B. Siebenhaar (Eds.), *Language Variation-European Perspectives VI: Selected papers from the Eighth International Con-*

ference on Language Variation in Europe (ICLaVE 8), Leipzig 2015 (Vol. 19, pp. 3–22). Amsterdam: Benjamins. doi: 10.1075/silv.19.01has

Hayakawa, A., Luz, S., & Campbell, N. (2016). Talking to a System and Talking to a Human: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In *Proceedings of INTERSPEECH'16: the 17th Annual Conference of the International Speech Communication Association* (pp. 1422–1426). San Francisco, CA, USA: ISCA. doi: 10.21437/Interspeech.2016-1623

Hayakawa, A., Luz, S., Cerrato, L., & Campbell, N. (2016). The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Paris, France: European Language Resources Association (ELRA).

Hayakawa, A., Vogel, C., Luz, S., & Campbell, N. (2017). Perception Changes With and Without the Video Channel: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. In *Proceedings of 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 401–406). Debrecen, Hungary: IEEE. doi: 10.1109/CogInfoCom.2017.8268279

Healey, P. G., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running repairs: Coordinating meaning in dialogue. *Topics in Cognitive Science*, *10*(2), 367–388.

Healey, P. G., Purver, M., & Howes, C. (2014). Divergence in dialogue. *PloS ONE*, *9*(6), e98598. doi: 10.1371/journal.pone.0098598

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi: 10.1017/S0140525X0999152X

Heritage, J. (2009). Conversation analysis as social theory. In *The new blackwell companion to social theory* (p. 300-320). John Wiley Sons, Ltd. doi: 10.1002/9781444304992.ch15

Howes, C., Healey, P. G., & Purver, M. (2010). Tracking lexical and syntactic alignment in conversation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2004–2009). Austin, Texas, USA: Cognitive Science Society.

Hutchby, I., & Wooffitt, R. (2008). *Conversation analysis*. Polity Press. Cambridge.

Hymes, D. H. (1982). Toward linguistic competence. *Philadelphia: University of Pennsylvania, Graduate School of Education*, *23*(4), 217–239.

Jokinen, K. (2009). Gaze and gesture activity in communication. In C. Stephanidis (Ed.), *Proceedings of the 5th International Conference of Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments (UAHCI'09), HCI International* (pp. 537–546). Springer Berlin Heidelberg.

Jonze, S. (2013). *Her.* Warner Bros. Pictures.

Kadar, J. S. (1993). Infants' vocalization in different types of interactions. In M. Kalliopuska (Ed.), *Proceedings from the 3rd Fenno-Hungarian Conference on Developmental Psychology* (pp. 2–13). Lahti, Finland: University of Helsinki.

Koutsombogera, M., & Vogel, C. (2018). Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 2945–2951). Paris, France: European Language Resources Association (ELRA).

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*(448), 1372–1381. doi: 10.1080/01621459.1999.10473888

LaFrance, M., & Broadbent, M. (1976). Group rapport: Posture sharing as a nonverbal indicator. *Group & Organization Studies*, *1*(3), 328–333. doi: 10.1177/105960117600100307

LaVoie, L. M. (2002). Subphonemic and suballophonic consonant variation: The role of the phoneme inventory. *ZAS Papers in Linguistics*, *28*, 39–54.

Lester, J., Branting, K., & Mott, B. (2004). Conversational agents. In M. P. Singh (Ed.), *The practical handbook of internet computing* (pp. 220–240). Chapman & Hall/CRC Press.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8*, 339–359. doi: 10.1007/BF00258436

Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics*, *41*(5), 791–824. doi: 10.1515/ling.2003.026

Loewen, S. (2012). The role of feedback. *The Routledge handbook of second language acquisition*, 24–40.

Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 265–284). Amsterdam: Benjamins.

Meyer, D., Zeileis, A., & Hornik, K. (2006). The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. *Journal of Statistical Software, Articles*, *17*(3), 1–48. doi: 10.18637/jss.v017.i03

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis, 5th Edition*. John Wiley & Sons.

Newlands, A., Anderson, A. H., & Mullin, J. (2003). Adapting communicative strategies to computer-mediated communication: an analysis of task performance and dialogue structure. *Applied Cognitive Psychology*, *17*(3), 325–348. doi: 10.1002/acp.868

O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., & Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction. *Interacting with Computers*, *8*(2), 177–192. doi: 10.1016/0953-5438(96)01027-2

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393. doi: 10.1121/1.2178720

Pennycook, A. (2010). *Language as a local practice*. Bristol, UK: Multilingual Matters.

Phinney, W. C. (2015). *Science Training History of the Apollo Astronauts* (Tech. Rep.). Houston, TX, United States: NASA Johnson Space Center. Retrieved from `Availableonlineathttps://www.hq.nasa.gov/alsj/PhinneySP-2015-626.pdf`

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459. doi: 10.1037/0033-2909.134.3.427

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. doi: 10.1017/S0140525X04000056

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ramseyer, F., & Tschacher, W. (2006). Synchrony: A core concept for a constructivist approach to psychotherapy. *Constructivism in the human sciences*, *11*(1-2), 150–171.

Ramseyer, F., & Tschacher, W. (2010). Nonverbal Synchrony or Random Coincidence? How to Tell the Difference. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, & A. Nijholt (Eds.), *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers* (pp. 182–196). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10 .1007/978-3-642-12397-9_15

Reidsma, D., Nijholt, A., Tschacher, W., & Ramseyer, F. (2010). Measuring Multimodal Synchrony for Human-Computer Interaction. In S. Alexei, D. Fellner, D. Thalmann, & O. Sourina (Eds.), *2010 International Conference on Cyberworlds* (pp. 67–71). Singapore: IEEE. doi: 10.1109/CW.2010.21

Reitter, D. (2008). *Context effects in language production: Models of syntactic priming in dialogue corpora*. (Doctoral dissertation) University of Edinburgh.

Reitter, D., & Moore, J. D. (2007). Predicting Success in Dialogue. In *ACL 2007 — Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 808–815). Prague, Czech Republic: Association for Computational Linguistics.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46. doi: 10.1016/j.jml.2014.05.008

Reitter, D., Moore, J. D., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 685–690). Austin, Texas, USA: Cognitive Science Society.

Reverdy, J., Hayakawa, A., & Vogel, C. (2018). Alignment in a multimodal interlingual computer-mediated map task corpus. In H. Koiso & P. Paggio (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 55–59). Paris, France: European Language Resources Association (ELRA), Workshop on Language and body in real life & Multimodal Corpora 2018.

Reverdy, J., Koutsombogera, M., & Vogel, C. (2020). Linguistic repetition in three-party conversations. In A. Esposito, M. Faundez-Zanuy, F. C. Morabito, & E. Pasero (Eds.), *Neural approaches to dynamics of signal exchanges* (pp. 359–370). Singapore: Springer Singapore. doi: 10.1007/978-981-13-8950-4_32

Reverdy, J., & Vogel, C. (2017a, September). Linguistic Repetitions, Task-based Experience and A Proxy Measure of Mutual Understanding. In *Proceedings of CogInfoCom 2017: the 8th IEEE International Conference on Cognitive InfoCommunications* (pp. 395–400). Debrecen, Hungary: IEEE. doi: 10.1109/CogInfoCom.2017.8268278

Reverdy, J., & Vogel, C. (2017b). Measuring Synchrony in Task-Based Dialogues. In *Proceedings of INTERSPEECH'17: the 18th Annual Conference of the International Speech Communication Association* (pp. 1701–1705). Stockholm, Sweden: ISCA. doi: 10.21437/Interspeech.2017-1604

Rieger, C. L. (2003). Repetitions as self-repair strategies in English and German conversations. *Journal of Pragmatics*, *35*(1), 47–69. doi: 10.1016/S0378-2166(01)00060-1

Riley, M. A., Richardson, M., Shockley, K., & Ramenzoni, V. C. (2011). Interpersonal Synergies. *Frontiers in Psychology*, *2*, 38. doi: 10.3389/fpsyg.2011.00038

Roque, A., & Traum, D. (2008). Degrees of Grounding Based on Evidence of Understanding. In D. Schlangen & B. A. Hockey (Eds.), *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 54–63). Columbus, Ohio, USA: Association for Computational Linguistics.

Rosenthal, R. (1963). On the social psychology of the psychological experiment: 1, 2 the experimenter's hypothesis as unintended determinant of experimental results. *American Scientist*, *51*(2), 268–283.

Rothwell, C. D. (2018). *Recurrence quantification models of human conversational grounding processes: Informing natural language human-computer interaction* (Unpublished doctoral dissertation). Wright State University.

Saito, A., Iio, T., Kimoto, M., Hagita, N., Shiomi, M., & Shimohara, K. (2018). Lexical entrainment in interaction with two robots. In *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)* (pp. 32–37). Nadi, Fiji: IEEE. doi: 10.1109/APWConCSE.2018.00014

Schegloff, E. A. (1992). Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *American Journal of Sociology*, *97*(5), 1295–1345. doi: 10.1086/229903

Schegloff, E. A. (1993). Reflections on Quantification in the Study of Conversation. *Research on Language and Social Interaction*, *26*(1), 99–128. doi: 10.1207/s15327973rlsi2601_5

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

Schmid, H.-J. (2015). Does gender-related variation still have an effect, even when topic and (almost) everything else is controlled? *Change of Pardigms–New Paradoxes: Recontextualizing Language and Linguistics*, *31*, 327–346. doi: 10.1515/9783110435597

Schneider, A., & Luz, S. (2011). Speaker Alignment in Synthesised, Machine Translated Communication. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2011* (pp. 254–260). San Francisco, California, USA: ISCA.

Serban, I. V., Lowe, R., Charlin, L., & Pineau, J. (2016). Generative Deep Neural Networks for Dialogue: A Short Review. *CoRR*, *abs/1611.06216*. Retrieved from `https://arxiv.org/abs/1611.06216`

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., & Pineau, J. (2015). A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *CoRR*, *abs/1512.05742*. Retrieved from `https://arxiv.org/abs/1512.05742`

Shriberg, E. (1996). Disfluencies in Switchboard. In *Proceedings of the International Conference on Spoken Language Processing* (Vol. 96, pp. 11–14). Philadelphia. PA.

Sotillo, C. F. (1997). *Phonological reduction and intelligibility in task-oriented dialogue* (Unpublished doctoral dissertation). University of Edinburgh.

Stalnaker, R. C. (1978). Assertion. In P. Cole (Ed.), *Syntax and semantics 9: Pragmatics* (pp. 315–332). New York: Academic Press.

Stivers, A. (2004). "No no no" and Other Types of Multiple Sayings in Social Interaction. *Human Communication Research*, *30*(2), 260–293. doi: 10.1111/j.1468-2958.2004.tb00733.x

Sun, X., & Nijholt, A. (2011). Multimodal Embodied Mimicry in Interaction. In A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, & A. Nijholt (Eds.), *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues: COST 2102 International Conference, Budapest, Hungary, September 7-10, 2010* (pp. 147–153). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-25775-9_14

Sundberg Cerrato, L. (2007). *Investigating communicative feedback phenomena across languages and modalities* (Unpublished doctoral dissertation). KTH Royal Institute of Technology, Stockholm.

Swerts, M., Koiso, H., Shimojima, A., & Katagiri, Y. (1998). On Different Functions of Repetitive Utterances. In *Proceedings of ICSLP 1998: the Fifth International Conference on Spoken Language Processing* (pp. 1287–1290). Sydney, Australia: ISCA.

Tannen, D. (2007). *Talking voices: Repetition, dialogue, and imagery in conversational discourse, 2nd Edition*. Cambridge University Press.

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

Toma, C. L. (2014). Towards Conceptual Convergence: An Examination of Interpersonal Adaptation. *Communication Quarterly*, *62*(2), 155–178. doi: 10.1080/01463373.2014.890116

Traum, D. (2017). Computational approaches to dialogue. In E. Weigand (Ed.), *The Routledge Handbook of Language and Dialogue* (pp. 143–161). Routledge.

Traum, D. R. (1994). *A Computational Theory of Grounding in Natural Language Conversation.* (Unpublished doctoral dissertation). Department of Computer Science, University of Rochester, New York.

Tree, J. E. F., & Meijer, P. J. (1999). Building Syntactic Structure in Speaking. *Journal of Psycholinguistic Research*, *28*(1), 71–90. doi: 10.1023/A:1023239604158

Truong, K. P., & Heylen, D. (2012). Measuring prosodic alignment in cooperative task-based conversations. In *Proceedings of INTERSPEECH'12: the 13th Annual Conference of the International Speech Communication Association* (pp. 843–846). Portland, Oregon, USA: ISCA.

Turnbull, W. (2003). *Language in action: Psychological models of conversation*. Psychology Press. doi: 10.4324/9780203360859

Ursi, B., Oloff, F., Mondada, L., & Traverso, V. (2018). Diversité des répétitions et des reformulations dans les interactions orales: Défis analytiques et conception d'un outil de détection automatique. *Langages*(4), 87–104. doi: 10.3917/lang.212.0087

van Dolen, W., de Ruyter, K., & Carman, J. (2006). The role of self- and group-efficacy in moderated group chat. *Journal of Economic Psychology*, *27*(3), 324–343. doi: 10.1016/j.joep.2005.05.007

Varges, S. (2006). Overgeneration and ranking for spoken dialogue systems. In *Proceedings of the Fourth International Natural Language Generation Conference* (pp. 20–22). USA: Association for Computational Linguistics.

Vicaria, I. M., & Dickens, L. (2016). Meta-Analyses of the Intra- and Interpersonal Outcomes of Interpersonal Coordination. *Journal of Nonverbal Behavior*, *40*(4), 335–361. doi: 10.1007/s10919-016-0238-8

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

Vogel, C., & Behan, L. (2012). Measuring synchrony in dialog transcripts. In Esposito A., Esposito A.M., Vinciarelli A., Hoffmann R., Müller V.C. (Ed.), *Cognitive Behavioural Systems: Lecture Notes in Computer Science* (Vol. 7403, pp. 73–88). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-34584-5_6

Ward, N., & Devault, D. (2015). Ten Challenges in Highly-Interactive Dialog System. In *2015 AAAI Spring Symposium Series* (pp. 104–107).

Whittaker, S. (2003). Theories and methods in mediated communication. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 243–286). Lawrence Erlbaum Associates Publishers.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 1556–1559). Genoa, Italy: European Language Resources Association (ELRA).

Wolf, F. M. (1986). *Meta-Analysis: Quantitative Methods for Research Synthesis* (Vol. 59). Sage.

Xia, Z., Levitan, R., & Hirschberg, J. (2014). Prosodic Entrainment in Mandarin Chinese and English: A Cross-Linguistic Comparison. In N. Campbell, D. Hirst, & D. Gibbon (Eds.), *Proceedings of the 7th International Conference on Speech Prosody 2014* (pp. 65–69). Dublin, Ireland: ISCA. doi: 10.21437/SpeechProsody.2014-1

Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, *101*(5), 1160–1179. doi: 10.1109/JPROC.2012.2225812

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? *CoRR*, *abs/1801.07243*. Retrieved from `https://arxiv.org/abs/1801.07243`