DEPARTMENT OF PHILOSOPHY, LINGUISTICS
AND THEORY OF SCIENCE

CATEGORIZATION OF CONVERSATIONAL GAMES IN FREE DIALOGUE REFERRING TO SPATIAL SCENES

Axel Storckenfeldt

Thesis:	Bachelor 15 credits
Program:	Linguistics
Level:	First Cycle
Semester:	Spring/2018
Supervisor:	Simon Dobnik
Examiner:	Ylva Byrman

Abstract

This thesis examines which communicative strategies speakers use to complete a given task regarding spatial scenes, and how to systematically categorize them as conversational games.

This study expands a free dialogue *Cups corpus* in Swedish (Dobnik et al., 2016) from 794 to 985 turns and extend the existing and new data with the annotation of conversational games using a new Game Type Coding Scheme. The study defines seven distinct types of conversational games of which four show universal features believed to be applicable across different spatial tasks.

The study shows some common features between the game types, which would allow computational classification. The reason why game types are important is that they define a scope within which particular linguistic features will be manifested (Dobnik et al.,2015).

These findings would be applicable on all situated dialogue systems such as in cars, humanoids or handheld devices in order to come one step closer to more "human-like" answers to prompted questions that involve spatial reasoning, such as, "Siri, where did I park my car?"

Thesis:	Bachelor 15 credits
Program:	Linguistics
Level:	First Cycle
Semester:	Spring/2018
Supervisor:	Simon Dobnik
Examiner:	Ylva Byrman
Keywords:	Computational Linguistics, Frame of Reference, Spatial Description, Conversational Games, Discourse Structure
Word count:	11 894

Acknowledgments

Foremost, I would like to thank Simon Dobnik, who as my advisor gave me all the tools I needed to write my thesis. He patiently gave me motivation, enthusiasm and happily shared his knowledge in the field. Not only did he help with the thesis, but also opened my eyes for computational linguistics. Therefore I want to express my sincere gratitude to Dobnik who went beyond what was expected of him as a supervisor.

I also want to direct my gratitude to Joakim Stålåker who lent me his office and his staff to run my experiment, without the data gathered the results would not be as conclusive.

Last but not least my I want to mention Helena Antoni who spent an afternoon helping me test the inter-rater variance by annotating a substantial part of the data using my coding scheme, even though it was the single most boring thing she had ever done, as she expressed it.

TABLE OF CONTENTS

CHAPTER 1 – INTRODUCTION 1
1.1 Introduction 1
1.1 Purpose 2
1.2 Research Question 2
1.3 Hypotheses 2
1.4 Applications of this Research 3
CHAPTER 2 – REVIEW 4
2.1 Previous Research 4
2.1.1 HCRC Map Task Corpus 4
2.1.2 Changing Perspective 4
2.2 Theory 5
2.2.1 Conversational Moves 5
2.2.2 Conversational Games 7
2.2.3 Reliability of the Game Coding Scheme 10
CHAPTER 3 – GENERAL METHODOLOGY 11
3.1 Phase 1 – Method for Data Collection – Extension of the Cups Corpus 11
3.1.1 Purpose of the Data Collection 11
3.1.2 Task 11
3.1.3 Procedure 12
3.1.4 Participants 13
Outlier 14
3.2 Phase 2 – Method for Annotation and Analysis 14
3.2.1 Purpose of Annotation and Analysis 14
3.2.2 Procedure 15
Modified Definition of Games 16
Method for Defining Game Types 16
3.2.3 Reliability 17
Game-ID Reliability 17
Game Type Reliability 17
CHAPTER 4 – RESULT & ANALYSIS 18
4.1 Result 18
4.1.1 Game Type Coding Scheme 18
Games Related to Interactional Structure (META-games) 18

ESTABLISHING PERSPECTIVE (ESPE) 18
TASK MANAGEMENT (TAMA) 19
Clarification (CLAR) 20
MISCELLANEOUS (MISC) 22
Games Related to Finding Objects (TASK-games) 22
GLOBAL (GLOB) 22
DESCRIPTIVE (DESC) 23
SPECIFICATION (SPEC) 24
Closing Remarks on Coding Game Types 25
Summary 26
4.1.2 Reliability 27
Game ID Inter-test Reliability 27
Game Type Inter-rater Reliability 27
4.2 Analysis 28
Descriptive Analysis 28
Annotation Tags 29
CHAPTER 5 – CONCLUSION 31
5.1 Hypotheses Respondence 31
5.2 Discussion 32
5.3 Conclusion 34
5.4 Future Directions 35
CHAPTER 6 – Bibliography 36

LIST OF TABLES

Table 1: Definitions of moves in coding scheme used by the Map Task and (Houghton, 1986). 6
Table 2 Example of a game from dataset GU-SE-P5 turn 153-155. 8
Table 3 Example of a game with a nested game from dataset GU-SE-P4 turn 45-49 9
Table 4 Example of a threaded game from dataset GU-SE-P7 turn 6-12. 9
Table 5 Index of the datasets in the Swedish Cups corpus after additions. 14
Table 6 Example of a nested EsPe-game from dataset GU-SE-P7 turn 133-134. 19
Table 7 Example of a top-level EsPe-game from dataset GU-SE-P5 turn 17-18. 19
Table 8 Example of a Desc-game from dataset GU-SE-P5 turn 20-21 which could be confused with a EsPe-game. 19
Table 9 Example of two TaMa-games, the first from dataset GU-SE-P4 turn 22-24, the second from GU-SE-P2 turn 67-68. 20
Table 10 Example of a nested Clar-game from dataset GU-SE-P4 turn 46-50. 21
Table 11 Example of a top-level Clar-game from dataset GU-SE-P7 turn 51-53. 21
Table 12 Example of a Spec-game from dataset GU-SE-P2 turn 40-42 which could be mistaken for a Clar-game. 21
Table 13 Example of a Misc-game from Dataset GU-SE-P6 turn 98-102. 22
Table 14 Example of a Glob-game from dataset GU-SE-P5 turn 157-160. 23
Table 15 Example of a Desc-game from dataset GU-SE-P5 turn 36-43. 24
Table 16 Example of two Spec-games, the first from dataset GU-SE-P4 turn 60-61, the second from GU-SE-P7 turn 258-261. 25
Table 17 Example of renegotiation from dataset GU-SE-P4 turn 10-12. 26
Table 18 Example of Inquisitiveness from dataset GU-SE-P4 turn 60-61. 26
Table 19 Summary of game type categorizations. 26
Table 20 Cross tabulation showing inter-rater agreement. 28
Table 21 Descriptive analysis of game types. 28
Table 22 Cross tabulation showing distribution between occurrences of FoR viewpoint usage per game shown per game type. 30

LIST OF FIGURES

Figure 1: A representation of the complete virtual scene from (Dobnik et al., 2015). Objects marked with a number were removed from each participants view during the experiment. ... 11
Figure 2: The virtual scene from the perspective of Participant 1. 12
Figure 3: The virtual scene from the perspective of Participant 2. 12

LIST OF ABBREVIATIONS

DiET	Dialogue Experimental Toolkit
FoR	Frame of Reference
HMI	Human-Machine Interaction
NLP	Natural language processing
AI	Artificial Intelligence
L1	First Language
DA	Dialogue Act

CHAPTER 1 – INTRODUCTION

This chapter gives a brief introduction to the concept of conversational games and the problems with the earlier method for game classification. In this chapter you will also find the purpose of this study, which leads to a research question and its hypothesis.

1.1 Introduction

This study is carried out with the purpose of learning more about human interactions and how conversational participants solve a joint task involving spatial descriptions. The study borders to cognitive psychology. The field of linguistics historically has focused on each speaker individually. It is first in modern times that linguists have put focus on studying conversations as a method for achieving a joint goal or mutual intelligibility (Garrod & Anderson, 1987). This approach to scientific questions gives us a broader understanding, and a more applicable knowledge.

In order to achieve goals, interlocutors deploy certain interactional strategies which will be reflected linguistically over a sequence of turns in the dialogue. These strategies are referred to as games. Identifying games is obviously very useful, since it breaks the conversation into smaller units where similar linguistic properties may be found and examined. Dobnik et al. (2015) suggest this may be relevant for how participants assign frame of reference.

In precedent definitions the conversational games are classified solely by the initiating move. This thesis refined a coding scheme for games which classifies games in a more dynamic way, not only taking the initiating move into account, but rather by defining a set of game types based on the datasets, to classify games. This does not only give us an understanding about language in general, but also gives us a glimpse into the cognitive process of the speakers to better decipher how language is used as a tool to achieve goals in the world around them (Wittgenstein, 1953).

1.1 Purpose

The purpose of this study is to contribute to the field of computational linguistics by classifying the different strategies speakers use to describe spatial scenes in free dialogue. The unstructured nature of free dialogue aggravates the complexity and makes NLP harder. By structuring the data into segments where certain communicative strategies are taken, these strategies will be reflected in language, and hence segmentation and identification of these common strategies will bring greater regularity for NLP, since models could be built for each game separately.

These models could improve all situated dialogue systems, and enable them to provide more human friendly responses to prompted questions referring to spatial tasks.

1.2 Research Question

What different conversational games does interlocutors use to describe spatial scenes in free dialogue, and how can these be categorized?

1.3 Hypotheses

Besides the study of the research question, this thesis will to some extent explore the validity and tenability of six different hypotheses. These are (I) dialogue with the purpose of completing a predefined task is assumed to be more streamlined than general free dialogue; (II) this would encourage the interlocutors to implement certain strategies in order to efficiently complete the given task; (III) the interlocutors are assumed to adapt to their strategies based on one another to form an array of coherent strategies which they reuse throughout the dialogue; (IV) these strategies are expected to show resemblance across dyads who face the same task and can therefore allow a global ontology for the task; (V) similarity of contexts are assumed to encourage interlocutors to form similar strategies which fall into the same game category and could therefore be seen as universal across all spatial tasks; (VI) game types are assumed to show resemblance in linguistic features, such as FoR viewpoints, which would allow computational processing.

1.4 Applications of this Research

Knowing more about what different strategies speakers use to describe spatial scenes, we can better understand spatial description in general. With the help of an annotated corpus, such as the one produced alongside this thesis, we can further use NLP to build a classifier by training artificial intelligence to recognize these games and then do further processing on the segments of the dialogue that have been identified. The agent in a situated dialogue system would then be able to detect a game type and thereafter predict what type of prompts would be expected in return. This would give the human user a more natural and familiar answer to prompts containing spatial expressions. Dobnik et al. (2015) suggest that spatial reference is related to these conversational strategies through which interlocutors agree on the meaning of spatial descriptions, in particular to the way FoR is assigned.

Spatial reasoning is a complex task to comprehend for situated agents since the description of objects are dependent on what has previously been said in the conversation. To better understand how FoR is assigned spatial descriptions are useful for situated agents and navigation systems. These systems, such as humanoids, self-driving cars, GPS-devices and handheld devices with voice assistants such as Apples “Siri” or Googles “Google Assistant” need to be able to process this data in order to generate natural and useful descriptions and be adaptable to the communicative strategies taken by the humans.

By achieving this, dialogue systems would comprehend a broader scale of questions such as “what’s the name of the restaurant next to my office?” and would also prompt more realistic and understandable answers to questions like “Where in the parking lot is my car?”, “How do I get to the restaurant I have a reservation for?” or “Where is the closest café”.

The understanding of conversational games is a prerequisite in order to understand how a broader range of utterances describing spatial positions work together to form strategies that span over several utterances. Coding these utterances into groups depending on their goal or intent is of utter importance for NLP since it reduces variation into more manageable units. It makes it possible to computationally draw conclusions and see patterns, even in more low volume datasets since the utterances within a certain game type show resemblance and therefore also share different linguistic features.

CHAPTER 2 – REVIEW

This chapter explains the previous experiment and the corpus expanded and used for the purpose of this thesis. It also describes conversational games and conversational moves which lay the foundation for the method of discourse analysis used in this study.

2.1 Previous Research

2.1.1 HCRC Map Task Corpus

The core problem while working with natural language is that even though the vast majority of language take form as unscripted dialogue with a defined communicative goal, most of the research is based on scripted material lacking sufficient instances of the phenomena being studied. The HCRC Map Task Corpus was produced to solve that problem while boosting the occurrences of certain linguistic phenomena at the researchers interest (Anderson et al., 1991).

The Map Task involves two interlocutors sitting opposite each other with a map each. The *information giver* has a route printed on his map while the *information follower* only has a set of landmarks. The landmarks differ between the two maps, which provoke spatial communication given the task to replicate the route on the information follower's map.

The corpus containing 128 dialogues was released publicly in 1992 and has been used for numerous vertical studies, many on the subject of dialogue acts, and conversational moves/games. The researchers behind the HCRC Map Task Corpus have also published a coding manual describing a system based on the utterance function in order to code conversational games and a more high-level transaction structure (Carletta et al., 1997).

2.1.2 Changing Perspective

This thesis derives from the work by Dobnik et al. (2015). The paper describes a pilot study in which the participants were assigned a spatial task. The task is described in detail in Section 3.1.2, since this study duplicates the same experiment to gather more data.

In short, each of the two participants has an image depicting a table, on which a number of cups are placed in front of them. The images illustrate the same setting, but from two different viewpoints. *Participant 1* are able to see a few cups hidden from *participant 2* and vice versa. The goal is to, via a chat-tool, together identify all the cups that are hidden from each player. Hence, this setup is similar in design to the Map Task, with the difference being that the design

does not specify the roles of *information giver* and *information receiver*. Instead, the roles change dynamically during the conversation.

The pilot study stated a hypothesis that there would be no general preference when it comes to FoR in dialogue and that it would most likely depend on the communicative acts given the chosen conversational game. The data supported the hypothesis which leads to a need for future work in order to examine the hypothesis further with an extended dataset. Finding patterns within the games would allow the design of an ontology of these adapted game types which later on could be modeled computationally. The categorization of the games is the first step toward finding patterns in speaker's strategies, such as FoR usage.

Using FoR in NLP comes with a few challenges as stated in Dobnik et al. (2015). Firstly, there are several ways to assign the viewpoint which means that in order to computationally model these kind of spatial descriptions the situated agent needs not only to keep track of the object being referred to, but also relative to what viewpoint which means using information about the location and orientation of other objects or landmarks. Secondly, the assignment of FoR is not always clearly stated, but sometimes interpreted based on a number of factors, such as assumptions or previous conversation, and must be recovered from the perceptual context.

2.2 Theory

2.2.1 Conversational Moves

To understand conversational games we must first grasp the concept of conversational moves. Even though this thesis will not use conversational moves as a part of the game type definitions, it is crucial to understand the move categorizations since the previously used game types depend on the moves for their classification.

Looking at utterances as moves is related to the concept of speech acts. A speech act is the underlying purpose of an utterance (Searle, 1980). When asking a question or making a statement the speaker performs an act with a specific purpose. This purpose does not always have to be clearly stated in the communication, but can be retrieved from the conversation's pragmatics. This theory leads to the theory of conversational games and dialogue moves where it is the purpose/goal of the utterance, or set of utterances, that is the basis for the annotation and analysis.

All conversational games consist of a sub dialogue, a group of utterances which as a unit fulfills a purpose. This group of utterances is usually categorized by their intent. Conversational moves are thereby the smallest building blocks in this method of discourse analysis (Carletta et al., 1997).

The keystone of using conversational games is that the speaker is aware that an initiation move, for example, a question, will result in a foreseeable chain of events such as information exchange or fulfillment of a desired act (Kowtko, Isard, & Doherty, 1993).

The definitions in *table 1* are the ones used by Houghton (1986) and Carletta et al. (1997), but several later experiments have used varieties on the scheme, such as Grice and Savino (1995) who introduced the OBJECT move as a contrast to ACKNOWLEDGE for any communication that suggests a break in the game, meaning echo-questions or other irrelevant chatter. The move showed to be common in communication between interlocutors with a strong personal relation. It is also proven more common in some languages compared to other (Grice & Savino, 1995).

Table 1: Definitions of moves in coding scheme used by the Map Task and (Houghton, 1986).

Move	Category	Description
INSTRUCT	Initiation	In its purest form, the move commands the interlocutor to perform a desired action. But the move can also be used for more abstract commands such as alignment of FoR "Let's take it from Katies point of view".
EXPLAIN	Initiation	A move that states information to help the current games task or state mutually known facts such as "seen from my perspective".
CHECK	Initiation	The moves asks the interlocutor to confirm information the sender is fairly certain about, such as "You mean from my left?" The reason this example is not categorized as a question is because it doesn't request any new information.
ALIGN	Initiation	The move seeks to align with the partner, such as confirm attention, readiness or agreement. It is most often (in task orientated dialogue) used to confirm transfer of information with utterances such as "OK?"
QUERY-YN	Initiation	Any question that does not count as an ALIGN or CHECK move and could be answered by yes or no fall into this category.
QUERY-W	Initiation	Typically "wh-questions", but any question that cannot be categorized by any category counts as this move.
REPLY-Y	Response	All positive responses to yes/no questions falls into this category, even if they answer other move types than QUERY-YN.
REPLY-N	Response	Any answer such as above stated, but with a negative reply.

REPLY-W	Response	Any complex replies to questions that does not fall into the REPLY-N or REPLY-Y moves.
CLARIFY	Response	A move that makes information additions that is not strictly asked for, such as an elaborated reply to a yes/no question. "The blue cup, with a white ear."
ACKNOWLEDGE	Response	Any utterance that just verifies that the preceding move is understood. Such as "mm". This moves is most frequently used in face-to-face conversations.
READY	Ready	These move work as a marker that the speaker have closed a game and is prepared for a new. Such as "Right" or "OK".

2.2.2 Conversational Games

Conversational games is a concept within the field of discourse analysis. It is a system where a series of utterances are grouped into a game and, depending on the purpose of the game, assigned a game type. A conversational game initiates by a so called *initiation-move* and encompasses all utterances until the desired outcome for the game has been fulfilled or the game has been abandoned (Carletta et al., 1997).

Conversational Games derive from the philosopher Ludwig Wittgenstein (1953) who regularly referred to his concept of *language game* as a way of describing that language is not separate from reality but is "consisting of language and the actions into which it is woven". The center of Wittgenstein's idea, is that language games serve to establish a link between language conceptualized as a calculus to the actual reality interpreted, described by it. The term derives from the concept that conversation is being a part of an activity, such as a game.

The concept of conversational games was presented in an essay written by Lewis (1969). It presents an approach to discourse analysis where discourse is seen as a way for the interlocutor to manipulate the conversational partner by transferring certain signals with the goal to influence the proceeding communication.

The organization of dialogue into games is a way to approach the fundamental problem, posed by Labov: "the fundamental problem of discourse analysis is to show how one utterance follows another in a rational, rule-governed manner – in other words, how we understand coherent discourse" (1972).

The first studies that included conversational games were designed to develop a theory of how conversation is used to achieve non-linguistic goals. The first studies were carried out by Power

(1979) and also Houghton (1986). Both used their findings to model an AI in the form of a pair of robots with the added abilities to use conversational games to solve tasks, where they used communication to engage each other in, predictable acts initiated by linguistic utterances. The first research with a more sociolinguistic approach was presented by Sinclair and Coulthard (1975) who studied the situations in classrooms to find what functions utterances had, how they followed each other, and how turns were distributed. They laid the foundation for game types which have been used in several studies later on.

These conversational building blocks reflect the intent with the dialogue being analyzed since they are coded depending on the purpose of the utterances, or more precise, the intent. A game can be initiated by a question, instruction or any utterance which holds an intent to provoke an action from the conversational partner. In *table 2* we see a typical game which starts with the intent of finding out if the opposing interlocutor has marked "the three cups". Depending on the context, the intent could also be to encourage *participant 2* (P2) to mark the cups if *participant 1* (P1) suspects that this has not been done. The game ends with an ACKNOWLEDGE-move which signals that the purpose is fulfilled and that the game is finished.

Table 2 Example of a game from dataset GU-SE-P5 turn 153-155.

Participant		**Game no.**
P1	har du ritat till tre muggar? *Have you marked the three cups?*	Game 1
P2	Jepp. Två röda muggar och en takeaway. *Jupp, Two red cups and a takeaway.*	Game 1
P1	okej bra *OK, good*	Game 1

Games can also be nested within other games if they pause the intended goal of the parent-game with a subservient purpose such as a sub-dialogue to repair or clarify missing information, or ask clarification-questions. The structure of nested games are always clear and mutually understood among the interlocutors and typically contribute to and lead back to the top-level intent of the game (Carletta et al., 1997).

The example in *table 3* shows how a game is initiated with the intent of finding white cups with lids. The opposing participant replies with a description of a targeted mug. P2 wants to check

if he understood the description and therefore pauses *game 1* to ask a clarification question. When the intent of the nested game is fulfilled *game 1* is resumed.

Table 3 Example of a game with a nested game from dataset GU-SE-P4 turn 45-49

Participant		**Game no.**
P2	… tar vi vita med lock? *… let's take white ones with lid?*	Game 1
P1	mellan den blå och gula, framför Katie, ser jag en mugg med lock och utan handtag *I see a cup with lid and without an ear between the blue and the yellow, in front of Katie.*	Game 1
P2	Står den lite längre bort från Katie (lite mer mot mitten) än den gula och den blå? *Is it a bit further away from Katie (a bit closer to the middle) then the yellow and the blue?*	Game 2 (nested)
P1	lite mot mitten inte exakt mellan den blåa och gula *A bit towards the middle, not exactly between the blue and yellow*	Game 2 (nested)
P2	OK, den muggen kan jag se. *OK, I can't see that mug.*	Game 1

A game can also be threaded, which occurs when the chronological order of a game is broken and several games are active simultaneously. In *table 4* we see two games, *game 1* with the purpose of establishing a common perspective and possibly find landmarks which can later be used to align FoR, and *game 2* with the purpose of comparing the total number of cups by colors.

Table 4 Example of a threaded game from dataset GU-SE-P7 turn 6-12.

Participant		**Game no.**
P2	framför mig står ett bord *There's a table in front of me.*	Game 1
P1	Katie står till höger om mig *Katie is standing on my right.*	Game 1
P2	jag ser tre blåa muggar *I can see three blue cups*	Game 2
P2	tre röda *Three red.*	Game 2
P1	mhm, ett bord för mig också *mhm, a table for me too.*	Game 1
P2	två gula *Two yellow.*	Game 2
P1	grön gräsmatta *Green lawn.*	Game 1

In the Map Task Coding Scheme, the opening of a game is coded with the name of the initiating move as the game type, i.e. the category of the game. The game then concludes all utterances up until the utterance which fulfills the goal or the utterance abandoning the game. For successful games this is usually signaled with the ACKNOWLEDGE-move. There are, however, a few exceptions. The initiation move can be proceeded by a READY-move. Also, not all initiation moves start a new game if they serve the purpose of continuing an already opened top-level game. After the annotation of the game's opening and closing, the level on which they occur are marked as either top-level or embedded at unspecified depth (Carletta et al., 1997).

2.2.3 Reliability of the Game Coding Scheme

Since the Map Task Coding Scheme is widely used in a variety of research the question about the annotation reliability has already been raised. The reliability is essential in order to gain credibility for the method itself but also to prove that the coding scheme is applicable for others apart from the scheme developers (Krippendorff, 1980).

Krippendorff (1980) practices three different reliability tests to assess the strength of a method. The test measures stability, reproducibility, and accuracy. Stability, also named intertest variance, shows how stable coder's judgement is over time. Reproducibility, also named inter-coder variance, measures to what extent two coders align. Accuracy describes to what degree the defined standard is followed by coders other than the scheme developer.

The accuracy test on the Map Task corpus where four coders annotated the same data shows that the agreement on where a conversational game begins is fairly high (70%, N=203) although there are some uncertainties regarding where the games end (Carletta et al., 1997). The result is explained partly by consequent differences between coders and partly by confusion about nested games. In those cases where all coders participating in the reliability test agreed on where the game began, agreement regarding where the game ended reached 65% pairwise agreement.

The coding schemes stability was tested by letting the same highly trained person code the same data two months apart. The agreement for the games starting point was 90% (N=49) and for the endings 89% (N=49). This implies that an experienced coder can develop a reliable sense of the structure with experience, even though it is hard to write down explicit instructions that cover all possible scenarios in free dialogue (Carletta et al., 1997).

CHAPTER 3 – GENERAL METHODOLOGY

This chapter is divided into two phases. Section 3.1 explains the experiment for the data collecting used to expand the corpus. Section 3.2 describes the annotation where the data is divided into games which later is analyzed to find patterns in order to be able to categorize them into game types. This chapter also narrates a slightly modified definition of conversational games which will be used throughout the study.

3.1 Phase 1 – Method for Data Collection – Extension of the Cups Corpus

3.1.1 Purpose of the Data Collection

The purpose of the data collection is to extend the *Cups corpus* described in Dobnik et al. (2015) in connection with their study of frame of reference assignment in free situated dialogue. The data gathering will extend the Swedish corpus discussed in Dobnik et al. (2016).

The extended corpus will be available for other related work in the field of semantics and pragmatics of dialogue and their computational modelling.

Here we summarize the task and the procedure as outlined in Dobnik et al. (2016, 2015)

3.1.2 Task

The task is based on the Map Task scenario with the difference in design being that the interlocutors dynamically changes their roles as conversation leader versus follower during the task.

Figure 1: A representation of the complete virtual scene from (Dobnik et al., 2015). Objects marked with a number were removed from each participants view during the experiment.

Each of the two participants in the experiment is shown an image of a virtual scene of a table on which a number of cups are placed which vary in shape and color. The image of the virtual scene was designed in SketchUp (previously Google SketchUp) and show, beside the table, avatars for the two participants facing each other with the table between them. The setup

also involves a third person named “Katie” who is explained to the participants as the silent observer. The image shown to each participant is a snapshot from their avatar’s point of view, see *Figure 2 and 3*.

The snapshots show the same scene, except that some cups are hidden from each participant’s view but which the other can see. Their joint task is to collaborate by to find and mark the complete set of hidden cups on a provided physical print of the scene.

3.1.3 Procedure

The participants are welcomed and get a chance to greet each other to establish a first impression if they are not familiar with each other beforehand. Each participant is given a written briefing description, stating the rules and the goal of the task, which they read and then have opportunity to ask questions about. This procedure is designed to avert the risk of the researcher subconsciously influencing the participants by the so called experimenter expectancy effect (Colman, 2009) and to assure that all participating dyads are given the same information over time. It does also allow for several data collectors to work simultaneously.

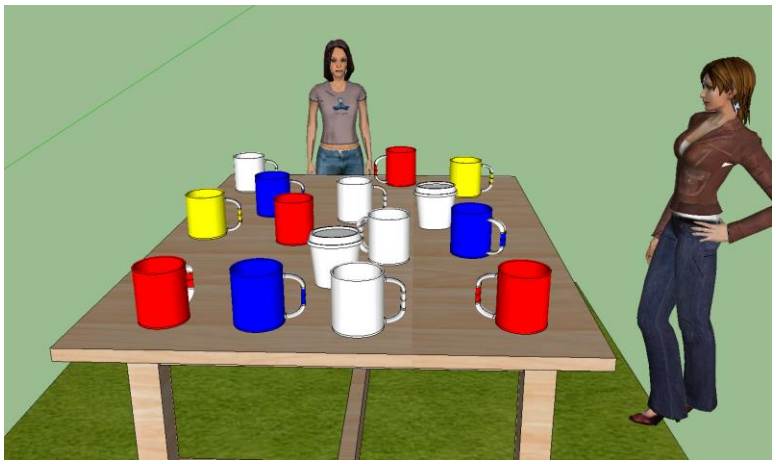

Figure 2: The virtual scene from the perspective of Participant 1.

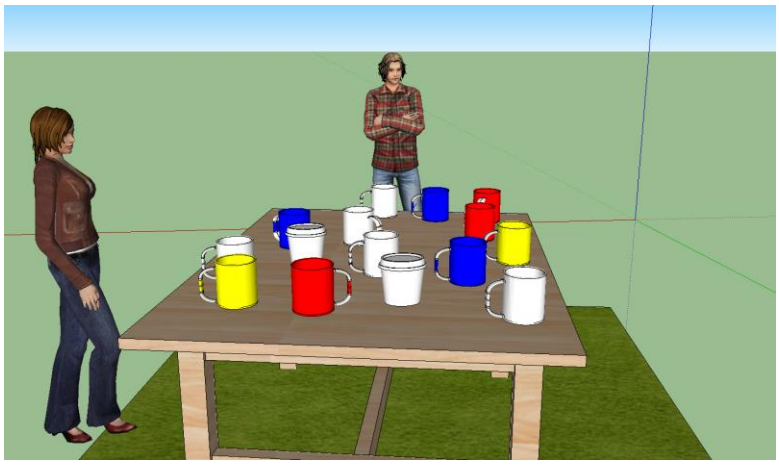

Figure 3: The virtual scene from the perspective of Participant 2.

The participants are then seated in front of a computer and their views are separated by a wall. The computer screen display a snapshot of the spatial scene from their avatars point of view and the DiET chat tool (Healey & Mills, 2009) consisting of a window that allows text chat

interaction between the participants. They are also provided a print of their snapshot and a pen to mark down the objects they have found.

The data collection progresses until both interlocutors believe that their task is completed successfully or the time limit of one and a half hour occurs. Afterwards they are thanked for the participation and are debriefed about the purpose of the study and asked to sign an agreement that they approve that their data will be used for research purposes.

3.1.4 Participants

The participants were selected on a voluntary basis. The experiment was announced through social media, word of mouth and email within the University of Gothenburg. The participants were asked to sign up in pairs as it was believed that familiarity will contribute to more natural descriptions. If they did not, a partner was assigned to them. There were two restrictions on participation: (i) both participants were to have Swedish as a L1, not excluding participants with several L1's; (ii) they should not be familiar with the study beforehand.

Five people signed up for the data gathering experiment, of which two dyads met the criteria for participation (1 female, 3 male; mean age: 24.75 years; age range: 22-27 years). The first dyad produced 118 turns, of which 77 were usable for the purpose of this study. The experiment took 82 minutes. The second dyad produced 114 turns during 39 minutes. In total the new additions contribute to the existing corpus with 232 turns. In total the corpus holds 985 turns from 12 contributors divided into 6 dyads.

Table 5 summarizes the datasets in the Swedish *Cups corpus*. In contrast to the previous papers, the datasets are named with a new naming convention where, for example, GU-SE-P1 stands for *Gothenburg University, Swedish, Pair 1*. The new dialogues were named GU-SE-P1 and GU-SE-P2 respectively since the names P1 and P2 were previously not used.

Table 5 Index of the datasets in the Swedish Cups corpus after additions.

Dataset	Native Language	Duration (min)	Length (turns)	
GU-SE-P1	SE	≈80	77	(New)
GU-SE-P2	SE	≈40	114	(New)
GU-SE-P4	SE	≈30	75	
GU-SE-P5	SE	≈60	163	
GU-SE-P6	SE	≈60	248	
GU-SE-P7	SE	≈60	308	
6 dyads			**985**	**2 New**

Outlier

Due to one dyad's misinterpretation of the instructions, turns 7 to 47 were excluded from the dataset GU-SE-P1. The dyad discovered their mistake themselves during the experiment and corrected it, and therefore the remaining part of the dataset can be used.

3.2 Phase 2 – Method for Annotation and Analysis

3.2.1 Purpose of Annotation and Analysis

The old datasets are annotated with a set of tags describing each utterance's characteristics in terms of spatial descriptions such as dialogue acts and viewpoints for perspective. The annotation process is described in Dobnik et al. (2015) and also Dobnik et al. (2016) which describe the Swedish corpus. Due to lack of time and lack of volunteering experienced annotators, the new datasets are not annotated with dialogue acts. Since the practice of the Game Type Coding Scheme presented in Section 4.1.1 does not rely on the conversational moves of each utterance the dialogue acts will not be used. The annotation of FoR viewpoints will only be used for the analysis in Section 4.2.

Frame of reference can be seen as the coordinate system a speaker uses to describe location. A spatial expression does not need a viewpoint, it can be understood by previous utterances or social conventions. The utterance *"the cup to the right"* does not use a FoR viewpoint, where as *"the cup to your right"* uses the opposing partner as a viewpoint.

The tags relating to the viewpoints of the FoR are annotated by the viewpoint the spatial description uses, and can be; one of the participants, the silent observer Katie, an object, or extrinsic, which means that the viewpoint is external, for example a coordinate system, such as cardinal directions.

In this work, we extend the annotation for the complete corpus by assigning each utterance an ID unique to each game in order to specify the utterance's affiliation to its conversational game. The games are also categorized with a game type depending of the nature of the game. The games are also labeled for their types, as discussed below. Segmentation of dialogue into games will allow us a more structured view on the dialogue, for example by analyzing each game type by itself or within NLP by build a classifier to recognize characteristics of each game type such as the preferred FoR in a given situation based on the type of conversational game in use.

3.2.2 Procedure

The dataset from each dyad was annotated individually. It was first read in its entirety, and subsequently read turn by turn where each turn was assigned an ID depending on each utterances game affiliation.

The definition of a game is adopted from the original Game Coding Scheme published in the HCRC Dialogue Structure Coding Manual (Carletta et al., 1997). The HCRC Dialogue Structure Coding Manual serves as a good reference, since it is the most commonly used and a later development on the earlier work from Houghton (1986), Kowtko (1993) and Powers (1979). The adoption of the scheme allows more accurate comparison between this study and other studies using the same scheme.

However, a few minor adaptations have been made. The changes do not affect the definition of a game per se, but simplifies the annotation and makes it easier to handle computationally. The original definition marks the beginning and the end of each game and thereafter notes if the game is occurring at top-level or as nested. Here, each utterance is marked with the corresponding game-ID. The game-ID is unique to each game. This method addresses the discovered problem with threaded games mentioned in Section 2.2.2. All the utterances with the same ID are defined as the game, regardless of the chronological order in which they appear in the chat-log. This method also takes away the need to mark down at what level the game

occurs since it is a nested game by definition if it is preceded and followed by a game with the same game-ID.

Modified Definition of Games

Below, we give the complete, modified definition of a game that will be used throughout the study.

A *game* is a set of utterances directed to a particular goal. A game begins with a locutionary act with a particular goal and ends when that goal is fulfilled, when it does not need additional information to be fulfilled, or if the game is abandoned, or the subject of discussion changes in such a way that it does not help the goal intended for the game.

A game is considered *nested* if a set of utterances fulfill the criteria for a game, but do not occur at top-level but within another game. This frequently happens when an interlocutor asks a question for clarification purposes and there after resumes the game.

When two games are active simultaneously independently of each other they count as *threaded*. This occurs when the chronological order in the dataset gets disrupted. The disruption most likely occurs as a result of participants speaking beside each other and keeping several games active simultaneously.

Method for Defining Game Types

Below we describe the method, used to assign games with a game type. The definitions of the new game types can be found under Section 4.2.1.

In order to categorize the strategies used, based on the conversational dynamics, the games were assigned a game type based on the underlying purpose and not by the move initiating the game such as in the original coding scheme.

We assigned all utterances with a game-ID based on the definition of a game. Thereafter, we treated each game as a unit and sorted them with other games where the speaker had the same goal. In those cases where several sets of characteristics were common within the group of games, we divided the group further. We then assigned each game with a game type based on the common nature of each group of games.

These common features within the group makes up the foundation of the definition. We then coded the whole corpus, strictly based on the new definitions where after all ambiguous cases

were marked with a comment which we later used to revise the definition until a definition that worked for the whole corpus was found.

When the game types were set, they were analyzed to find patterns significant for each game type. This was done as a test to see if the game types seemed substantial and if there were features that possibly could be used to annotate datasets computationally.

3.2.3 Reliability

The assigning of game-ID and the game type categorization calls for different reliability checks. Since this study uses the coding scheme from HCRC Map Task, general reliability research regarding the delimitation of games has already been carried out and presented in Section 2.2.3.

Game-ID Reliability

Since the game-ID annotations in this study were made by the same person, what is interesting is to examine the intertest reliability of this coder. Earlier research has showed that the most experienced coders can achieve up to 90% intertest reliability (Carletta et al., 1997).

This study's coder annotated the datasets GU-SE-P4 to GU-SE-P7 containing 794 turns. The same data was annotated again, one month later, by the same coder after which the annotations were compared and the percentage of agreement was calculated.

Game Type Reliability

Since the new categorizations are developed within this study the reproducibility needs to be examined in order to measure the credibility of the definitions. An inter-rater comparison will in this case not only measure how clear the differences between the game types are, but also how clear and comprehensible the definitions are.

Another coder was therefore given the game type definitions and asked to annotate dataset GU-SE-P5 and GU-SE-P4, which were chosen at random. The datasets contained 238 turns divided into 67 games from previous annotation, as mentioned in section 3.2.2. The weighted kappa was calculated using IBM SPSS (Statistical Package for the Social Sciences). Weighted kappa measures pairwise agreement on categorizations corrected for agreement by chance due to the bias of annotators to particular categories.

CHAPTER 4 – RESULT & ANALYSIS

This chapter presents the definitions of the discovered game classifications as a Game Type Coding Scheme. The coding scheme is written as a stand-alone manual which could be used by other coders. The chapter also presents the reliability for the coding scheme. The analysis interprets the annotations made with the coding scheme to find patterns within each of the game types, which could be used to build an automatic classifier.

4.1 Result

4.1.1 Game Type Coding Scheme

The Game Type Coding Scheme is the product of the thesis where the categorizations of the conversational games are presented and defined. The idea is that the coding scheme could be used on other similar corpora with or without modification. The definitions are stated as comprehensive as possible to encourage usage by other researchers.

The classifications derive from the data and are therefore driven by the linguistic phenomena being examined. In this corpus, the examined phenomena are the interaction between participants in terms of how they generate referring expressions in perceptual dialogue, in particular how they negotiate and assign FoR. Hence, we divide the game types into two overarching categories; META-games, relating to interactional dynamics, and TASK-games, relating to the descriptions of scenes.

Games Related to Interactional Structure (META-games)

The game types are divided into two categories, TASK-games and META-games. All game types that do not directly contribute to the main task of referring to and finding the objects falls in the overarching category META-games. The category contain game types providing information and structure to the dialogue.

Without these games, interaction would not be possible. Participants need a way to negotiate how they will proceed with the dialogue and also initiate clarification and repair if their interaction becomes out of sync.

ESTABLISHING PERSPECTIVE (ESPE)

An ESPE-game is used to establish a common ground or align FoR. This game appears either as top-level, often before other games, or as a nested game to agree on the perspective used in preceding utterances as discussed below.

A nested EsPe-game is not a Clar-game since the initiator of the EsPe-game does not need clarification for something referred to or directly related to in the game of which the EsPe-game is nested in. A Clar-game needs to be directly related to the parent-game.

Not all games containing referential alignment need to be EsPe. All games can contain this type of phrase. The game requires the goal of the game to be to establish perspective or an agreement on the suggested perspective.

Below there are three examples: One nested EsPe-game shown in *table 6*, one occurring at top-level shown in *table 7* and a Desc-game which contains an explicit description of perspective and therefore could be confused with an EsPe-game, shown in *table 8*.

Table 6 Example of a nested EsPe-game from dataset GU-SE-P7 turn 133-134.

Participant		Game Type
P2	kan du börja med från katie? *Can you start with Katies perspective?*	EsPe
P2	det är lättare för mig att hålla reda på *It's easier for me to keep track of.*	EsPe

Table 7 Example of a top-level EsPe-game from dataset GU-SE-P5 turn 17-18.

Participant		Game Type
P2	Ska vi börja från din ända av bordet? *Shall we start from your side of the table?*	EsPe
P1	Ja *Yes.*	EsPe

Table 8 Example of a Desc-game from dataset GU-SE-P5 turn 20-21 which could be confused with a EsPe-game.

Participant		Game Type
P1	från mitt vänster till höger: röd, blå, vit och röd *From my left to right: red, blue, white and red.*	Desc
P2	Jag ser bara tre muggar längs med din kant. Röd, blå och vit. *I can only see three cups along your side. Red, blue and white.*	Desc

Task Management (TaMa)

Strategic dialogue on how to approach the task is classified as a Task Management-game. The game does not directly contribute to the main goal of describing and finding objects. However it plays a crucial role to the conversational dynamics, by establishing a common tactic

in order to solve the task, dialogue about the game in general, or the overall progress of the dyads. *Table 9* shows two examples of TAMA-games with different characteristics.

Table 9 Example of two TAMA-games, the first from dataset GU-SE-P4 turn 22-24, the second from GU-SE-P2 turn 67-68.

Participant		**Game Type**
P1	ok vilken färg tar vi nu? *Ok, what color do we take next?*	TAMA
P2	Blå kanske? *Blue, maybe?*	TAMA
P1	Ok *Okay*	TAMA

P2	Ok men jag kan beskriva exakt hur mina vita muggar står. Låter bra? *OK, but I can describe exactly where my white cups are. Sounds good?*	TAMA
P1	Gör så. *Do so.*	TAMA

Clarification (CLAR)

CLARIFY-games hold the purpose of repairing some type of miscommunication. They can also be used to request a clarification of an interpretation where the describer is fairly certain about the outcome. CLAR-games most often occurs as nested games and in these cases directly relates to a miscommunication or an ambiguity in the parent-game. If a game asks for new information that cannot be seen as missing from the parent-game, such as a follow-up question, it is by definition not a CLAR-game, but a part of the top-level game since it doesn't break or paus the top-level game.

Below we give three examples, one nested CLAR-game shown in *table 10*, one on top-level in *table 11*, and also a SPEC-game which could be confused with a CLAR-game in *table 12*.

Table 10 Example of a nested CLAR-game from dataset GU-SE-P4 turn 46-50.

Participant		Game Type
P1	Mellan den blå och gula, framför Katie, ser jag en mugg med lock och utan handtag. *I see a cup with a lid, and without an ear, between the blue and yellow, in front of Katie.*	DESC
P2	Står den lite längre bort från Katie (lite mer mot mitten) än den gula och den blå? *Is it located a bit further away from Katie (a bit closer to the middle) then the yellow and blue?*	CLAR
P1	lite mot mitten inte exakt mellan den blåa och gula *A bit towards the middle, not exactly between the blue and yellow*	CLAR
P2	OK *Okay*	CLAR
P2	Den muggen kan jag inte se *I can't see that cup.*	DESC

Table 11 Example of a top-level CLAR-game from dataset GU-SE-P7 turn 51-53.

Participant		Game Type
P1	sa du att den första var av take away-typen? *Did you say the first one was of the take-away type?*	CLAR
P2	Nä *Nope*	CLAR
P2	den andra *The other one*	CLAR

Table 12 Example of a SPEC-game from dataset GU-SE-P2 turn 40-42 which could be mistaken for a CLAR-game.

Participant		Game Type
P1	Ser du två röda bredvid varandra? *Do you see two red ones next to each other?*	SPEC (not CLAR)
P2	Precis, en är längst ner i hörnet och en är precis nedanför den (från mitt håll sett) *Exactly, one is down in the corner and one right below it (seen from my perspective)*	SPEC (not CLAR)
P1	Så är det inte för mig. Jag har en röd mugg i varje hörn från där jag står. *That's not how I see it. I have a red cup in each corner from where I stand.*	SPEC (not CLAR)

MISCELLANEOUS (MISC)

A MISC-game is a game with a goal that does not relate to describing or finding objects, such as social chatter, greetings or other conversational glue which does not focus directly on the main goal, but facilitates the task on a social level to establish familiarity.

Table 13 shows a MISC-game where two participants show satisfaction over their progress in the experiment.

Table 13 Example of a MISC-game from Dataset GU-SE-P6 turn 98-102.

Participant		Game Type
P1	fan va fint asså *Damn, that's nice!*	MISC
P2	vi är grymma *We rock*	MISC
P1	fet med bra *much so*	MISC
P2	tror du det går att göra karriär såhär... *You think one can make a career doing this?*	MISC
P1	Japp *Yup*	MISC

Games Related to Finding Objects (TASK-games)

TASK-games are the overarching category for all games directly related to solving the main task, with the assistance of META-games. The main task in this corpus, is finding and describing objects.

GLOBAL (GLOB)

GLOBAL-games are not involving any describing or alignment of views but rather reasoning on a global level. The game is related to the overall task of finding and describing objects in general and not a specific portion of the view. In difference of the TAMA-game, the goal for a GLOB-game is directly related to solving the main task. TAMA is solely used to align tactics.

In *table 14* we give an example of a GLOB-game where the participants try to summarize the overall progress of their task.

Table 14 Example of a GLOB-game from dataset GU-SE-P5 turn 157-160.

Participant		Game Type
P2	Ok. En ny recap. Jag har fem röda, tre blå, fem vita, två gula och tre takeaways. *Ok, a new recap. I have five red, three blue, five white, two yellow and three take-aways.*	GLOB
P2	Med de jag har ritat in. *With the ones I marked down.*	GLOB
P1	okej, jag ska räkna mina *Okay, I will count mine.*	GLOB
P1	ja det verkar stämma *Yes, that seems correct.*	GLOB

DESCRIPTIVE (DESC)

A DESCRIPTIVE-game can be described as a systematical search for a mismatch. One interlocutor acts as a describer of the scene as they perceive it while the other conversational partner acts as a follower who is looking to spot any inconsistencies in the description that might suggest a mismatch. The goal for the game is to find an inconsistency in the described view. The goal is fulfilled when the follower raises an issue about a mismatch. The leader/follower role may change within the game since the follower often explains the same region as they see it, if no mismatches are found. Since the goal stays the same, this occurs within the same game.

This normally leads to a SPECIFICATION-game where the search, and therefore the DESC-game, is completed in favor of a new game where the description of the missing object is negotiated.

Table 15 shows an example of a DESC-games which shows how the roles can change during the game. P1 is first describing his view, which after P2 describes the same view from his perspective.

Table 15 Example of a DESC-game from dataset GU-SE-P5 turn 36-43.

Participant		Game Type
P1	okej, nästa rad mot mitten *Okay, next row towards the middle.*	DESC
P1	från mitt håll står det en take-away bakom den vita muggen *From my perspective I can se a take-away behind the white cup.*	DESC
P1	snett vänster om *Diagonally to the left.*	DESC
P2	Ok. Här det en vanlig vit mugg strax till höger om den vita närmast dig. *OK. Here there is a regular white cup a bit to the right of the white closest to you.*	DESC
P2	Till höger och innåt bordet då. *To the right and towards the table, that is.*	DESC
P1	okej, den ser jag *Okay, I can see that one.*	DESC

SPECIFICATION (SPEC)

A SPECIFICATION-game is a game with the purpose of aligning the interlocutors' views to be focused on the same object or geographical area. The describer has a specific object or location in mind and tries to share that information with the follower.

The goal of a SPEC-game is not to describe the scene head-first in order to maybe find a mismatch, but to establish a common visual focus. The game's goal is therefore fulfilled when the follower understands which object or placement the describer tries to refer to.

In *table 16* we give two examples of SPEC-games with slightly different characteristics.

Table 16 Example of two SPEC-games, the first from dataset GU-SE-P4 turn 60-61, the second from GU-SE-P7 turn 258-261.

Participant		Game Type
P2	Kan du förklara var den vita med lock på din sida står igen? *Can you repeat where the white one with a lid on your side were?*	SPEC
P1	den står inåt mitten mellan den vita och blåa, kanske lite närmare den vita *It's towards the middle between the white and the blue, maybe a bit closer to the white.*	SPEC

P1	på den första rad ska det finnas en mugg jag inte ser mellan den röda och vita? *On the first row, there should be a cup which I can't see. Between yo red and white?*	SPEC
P1	*din **your*	SPEC
P2	mm, en take away *Mhm, a wake-away*	SPEC
P1	ok, npterat *OK, noted.*	SPEC

Closing Remarks on Coding Game Types

Since the definition of a game is that it is started with a specific goal in mind the games need to be categorized depending on that goal, since the whole sequence of utterances defined as a game belong to the same game type. However, the responding interlocutor can react to the games' initiating move and renegotiate the goal of the games. The establishing of the goal is usually made within the first two utterances of the game. If the first utterance of a new game provokes a renegotiation in the second utterance, the first utterance is not treated as an independent game but rather as a part of the following game with a renegotiated goal.

A question can also be used to initialize a game. Getting the question answered is not necessarily the goal, as the question can also have a performative function. It is then the goal for the desired action that sets the game type.

Below we give two examples. *Table 17* shows a goal renegotiation where the first utterance by itself could be annotated as a part of a GLOB-game but is not, since it is followed by the renegotiation shown on the second line in *Table 17*.

The second example, *table 18*, shows a question with the purpose of obtaining a description about the placement of the white cup. The goal is not receiving a statement answering the question if the opposing interlocutor is physically able to provide such a description or not but to initiate a performative function. The game is therefore a SPEC-game.

Table 17 Example of renegotiation from dataset GU-SE-P4 turn 10-12.

Participant		Game Type
P1	jag ser totalt 3 blåa muggar *I can see a total of 3 blue cups.*	TAMA (not GLOB)
P2	Ska vi försöka markera de röda muggarna som vi inte kan se först på pappret? *Shall we start by marking down the red cups we can't see first?*	TAMA
P1	Ok. *Okay.*	TAMA

Table 18 Example of Inquisitiveness from dataset GU-SE-P4 turn 60-61.

Participant		Game Type
P2	Kan du förklara var den vita med lock på din sida står igen? *Can you explain where the white with lid on your side were again?*	SPEC (not TAMA)
P1	den står inåt mitten mellan den vita och blåa, kanske lite närmare den vita *It's in the middle bewteen the white and the blue, maybe a bit closer to the white.*	SPEC

Summary

For an easier overview and as a quick reference guide during annotation the above defined game types are summarized in *table 19* with their name, code and a brief description.

Table 19 Summary of game type categorizations.

Name	Code	Definition
TASK MANAGEMENT	TAMA	Strategic negotiation about how to approach the task.
ESTABLISHING PERSPECTIVE	ESPE	Negotiating a common viewpoint or perspective.
MISCELLANEOUS	MISC	Social chatter such as greetings.
CLARIFY	CLAR	All types of repair and clarification of miscommunication.
GLOBAL	GLOB	Reasoning or other strategies describing the scene at a global level.
DESCRIPTIVE	DESC	Systematically searching to find a mismatch, comparing each other's view.
SPECIFICATION	SPEC	Specifying the location or characteristics of a specific item.

4.1.2 Reliability

Game ID Inter-test Reliability

The inter-test reliability, measured as agreement in game-ID by the same coder one month apart, was 78 % N=794. In 85 % of the games that differed, the latter annotation was favorable upon review.

When compared to the results of the HCRC Map Task inter-test reliability where the expert coder got 90 % agreement, this thesis' results must be seen as a substantial agreement since the coder lacks experience when it comes to annotating the duration of games.

Since the latter annotation was favorable to a significant degree of the cases, the results clearly show that the sense of the games evolves with practice. One would expect higher agreement to correlate with the coder's experience.

Game Type Inter-rater Reliability

The agreement on game type for the two coders reached κ=0.74 (N=67) using Cohen's weighted kappa. Kappa values range between -1 – 1 where values between 0.6 – 0.8 count as a substantial agreement and values between 0.8 – 1 translates to almost perfect agreement.

Since the agreement is substantial, the cross tabulation in *table 20* can be used to analyze what part of the definitions that might need improvement and what game types that are most sensitive to confusion.

The most common mismatch involves games coded as SPEC by the first coder being coded as DESC by the second. In total, 11 out of 14 mismatches involves either the SPEC- or the DESC-game, which is an overrepresentation even though the mentioned game types constitutes 46 % of the datasets. The weighted kappa for only the games coded as SPEC or DESC by the scheme developer measured κ=0.71 (N=33), nearly as high as the overall data. Since the kappa is almost as substantial for this selection, no conclusion can be reached regarding the definitions inadequacy despite the frequency of mismatches.

The cross tabulation of mismatches is shown in *table 20* where "Coder One" represents the scheme developer and "Coder Two" represents the second coder in the inter-rater test.

Table 20 Cross tabulation showing inter-rater agreement.

Coder One * Coder Two Crosstabulation

Count

		Coder Two							
		CLAR	DESC	ESPE	GLOB	MISC	SPEC	TAMA	Total
Coder One	CLAR	7	0	0	1	0	3	0	11
	DESC	0	14	0	0	0	0	0	14
	ESPE	2	1	2	0	0	0	0	5
	GLOB	0	2	0	6	0	0	0	8
	MISC	0	0	0	0	3	0	0	3
	SPEC	1	4	0	0	0	14	0	19
	TAMA	0	0	0	0	0	0	7	7
Total		10	21	2	7	3	17	7	67

4.2 Analysis

Descriptive Analysis

The descriptive analysis in *table 21* shows statistical data over the length of the different game types, measured in utterances and frequency, in the corpus as a whole. The duration of games related to finding object (TASK-games) are longer than games related to interactional structure (META-games). The average length of META-games is 3.5 compared to 5.2 for TASK-games. The difference is distinct since no META-game game type is longer than any TASK-game game type.

Table 21 Descriptive analysis of game types.

	DESC	SPEC	CLAR	GLOB	ESPE	MISC	TAMA	Total
Mean Length	5.5	5.1	3.4	4.7	2.9	3.9	3.9	**4.2**
Median Length	5	5	3	4	2	4	4	**4**
Number of turns	305	236	139	127	44	67	67	**985**
Number of games	55	46	41	27	15	17	17	**218**
Percentage of games	25 %	21 %	19 %	12 %	7 %	8 %	8 %	**100 %**

Annotation Tags

The datasets previously collected by Dobnik et al. (2016, 2015) are annotated with FoR viewpoints described in Section 3.2.1.

By merging the annotation tags containing FoR viewpoints per game, we get a variable that shows us the distribution between occurrences of FoR viewpoint usage per game shown per game type. In *table 22* we can see a cross tabulation which shows the *observed count* and *Expected* count for the occurrences of FoR-viewpoints divided by game type. The *expected count* shows the expected distribution if the variables were independent.

For example, there are eight CLAR-games with zero turns tagged with a FoR-viewpoint, and nine games containing one FoR-viewpoint.

Running Pearson's Chi Square test we get $\chi^2_{54}=104$ ($p<0.0005$; $N=176$), which tells us that there is very significant evidence against independence between the tested variables. Thus we can reject the null hypothesis.

The game types TAMA and MISC (with two exceptions) lacks utterances containing reference frames. This is expected since these META-games do not contain games describing objects and their location but rather discusses the approach and tactics of the task.

Note that the result for the META-game type ESPE include some games containing FoR viewpoints. This is expected since the goal for an ESPE-game is to establish a perspective that in many cases involves the need for specifying a frame of reference. The game type CLARIFY shows similar features as the TASK-games since the utterances in a CLAR-game often are nested within TASK-games and therefore needs to refer to reference frames to repair the utterances in the parent-game. We expect that CLARIFY would contain more FoR definitions then other META-games.

Table 22 Cross tabulation showing distribution between occurrences of FoR viewpoint usage per game shown per game type.

GameType * FoR Crosstabulation

			FoR										
			0	1	2	3	4	5	6	7	8	9	Total
GameType	CLAR	Count	8	9	5	8	2	1	0	0	0	0	33
		Expected	11,8	6,0	3,9	5,3	2,3	1,5	0,8	0,6	0,2	0,8	33,0
	DESC	Count	4	6	6	10	5	4	3	3	1	3	45
		Expected	16,1	8,2	5,4	7,2	3,1	2,0	1,0	0,8	0,3	1,0	45,0
	ESPE	Count	7	3	1	2	0	0	0	0	0	0	13
		Expected	4,7	2,4	1,6	2,1	0,9	0,6	0,3	0,2	0,1	0,3	13,0
	GLOB	Count	13	1	3	1	2	0	0	0	0	0	20
		Expected	7,2	3,6	2,4	3,2	1,4	0,9	0,5	0,3	0,1	0,5	20,0
	MISC	Count	13	1	0	0	0	0	0	0	0	0	14
		Expected	5,0	2,5	1,7	2,2	1,0	0,6	0,3	0,2	0,1	0,3	14,0
	SPEC	Count	5	12	6	7	2	3	1	0	0	1	37
		Expected	13,2	6,7	4,4	5,9	2,5	1,7	0,8	0,6	0,2	0,8	37,0
	TAMA	Count	13	0	0	0	1	0	0	0	0	0	14
		Expected	5,0	2,5	1,7	2,2	1,0	0,6	0,3	0,2	0,1	0,3	14,0
Total		Count	63	32	21	28	12	8	4	3	1	4	176
		Expected	63,0	32,0	21,0	28,0	12,0	8,0	4,0	3,0	1,0	4,0	176,0

CHAPTER 5 – CONCLUSION

This final chapter will briefly discusses the hypotheses presented in Section 1.3. Thereafter a general discussion will follow which leads up to the conclusion and suggestions for future work.

5.1 Hypotheses Respondence

This study hypothesized that conversational games in free dialogue referring to spatial scenes can be categorized based on the game characteristics related to the task. The research question is examined by defining the new Game Type Coding Scheme. The coding scheme presents a set of categories that are found in all dialogues in the corpus. This set of categories can be annotated with high agreement between different coders using the coding scheme. This supports the hypothesis (II) that interlocutors implement certain strategies in order to efficiently complete the task, and hypothesis (IV) that these strategies are expected to show resemblance across dyads since the same game types are applicable for the whole corpus.

Since META-games occur across all dyads and are by definition not specific to the given task, but provides interactional structure to the dialogue, they are expected to be universal across all spatial tasks. This supports hypothesis (V) that similarity in context forces interlocutors to form a certain set of universal strategies.

Datasets from participants unfamiliar with each other are observed to show a cleaner structure with less threading and fewer MISC-games. Hypothesis (I), that dialogue with the purpose of completing a predefined task is more streamlined, is therefore supported for participants unfamiliar with each other. Participants familiar with each other show greater resemblance to what would be expected of truly free dialogue.

The practice of semantic coordination where dyads use invented lexical entries without the need for explanation has been observed. The observation that the opposing interlocutor reuses the same game type after a strategy has been introduced, such as in *table 15*, supports hypothesis (III) that dyads form an array of coherent strategies throughout the dialogue.

The high Chi Square value for the data presented in *table 22* states that there is a correlation between the use of FoR viewpoints and game type. Hypothesis (VI), stating that game types are assumed to show resemblance in linguistic features, is only partially supported by the data since our analysis does not present specific features for each game type.

5.2 Discussion

A great deal of the corpus is not orderly due to the nature of free dialogue. For example, when participants initiate games with a new goal during other active games, which most often leads to threading, confusion and CLAR-games. They also tend to fail to make their goals clear to the opposing interlocutor, which leads to a growing confusion. The nature of uncontrolled data makes it difficult to create a reliable coding scheme for game structures. This challenge of reliability is well known (Carletta et al., 1997).

This challenge explains the lack of a standardized coding schemes where the game type is based on a set of more complex variables than the initiating move. In free dialogue the conversations are adapted to solve a specific task and therefore, the strategies depend on the task the participants are engaged in. There are specific games related to the task carried out in the data collection experiment. The properties of these games are related to how we specify the task. Had this been done differently, it would most certainly affect the games and therefore the game types. A game therefore reflects a strategy of how to solve a specific task, and the same strategy might not be found in conversations solving different tasks. However, there seems to be some ubiquitous META-games which the participants deploy to direct the conversation. Since their definition is not based on the specific task examined, they are believed to be universal for all spatial tasks even though they might show different characteristics depending on the task, such as CLAR-games which would be expected to show slightly different features depending on the type of conversation they repair. The findings in the analysis comparing META-games to TASK-games partially support the hypothesis (V) which states that similarity in context is assumed to force interlocutors to adapt similar, universal strategies.

The game type categorizations could also be done differently, for example based on the grounds of the linguistic phenomena being investigated. In this study the categorization is based solely on the corpus. Another strategy would be to base the categorization of games on the FoR assigned, if that is the goal of investigation. A set of game types like that would describe the frequency of the studied phenomena and would not be a descriptive representation of the data by itself.

The demarcation between what utterances belong to what games is not trouble-free either. The definition partly relies on the coder, who must develop a sense for the definition. This is shown

in the results of the inter-test variance test, where the second annotation was favorable in 85 % of the cases where the annotation differed. The game definition might also be biased by the task used to collect the data. Coders might for instance interpret some utterances as starts of new games instead of a part of the ongoing game, depending on the nature of the task. This adds complexity to the definition of a universal set of game types, since the games analyzed might be of different type depending on the task.

The frequency of the games types are observed to differ depending on the dyads. For instance, the occurrence of threaded games seems to depend on the familiarity of participants. Dyads unfamiliar with each other tend to have a formal, less chatty conversation and focuses to a higher degree on the task at hand. Therefore, they keep the chat "cleaner" to some extent, which supports the hypothesis (I) that a conversation with the purpose of completing a predefined task is more streamlined than general free dialogue. This supports Grice and Savino's (1995) observation that dyads with strong personal relations tend to use more "irrelevant" utterances, often annotated as OBJECT-moves, more frequently. This phenomena is not statistically examined in this thesis, but the observation might suggest that familiarity between participants affects more aspects in task related dialogue than previously thought.

Participants seem to practice semantic coordination, when facing a task. Several dyads used invented words throughout the dialogue to streamline their conversation without needing to explain the meaning of the invented lexical entry. Examples such as "Lockmugg" (Eng: *lidmug*), "K" (abbreviation for "Katie") and "HH" (Abbreviation for "Högra Hörnet", Eng: *Right Corner*) were encountered repeatedly by both the inventor and the opposing participant. This supports hypothesis (III) that speakers adapt strategies depending on their conversational partner. This conforms with earlier work regarding "the principle of least collaborative effort" which states that speakers work together to find mutually acceptance for definite references (Clark & Wilkes-Gibbs, 1986). Also, Clark (1996) who describes how speakers reuse referring expressions to ambiguous objects to create a "conceptual pact".

The correspondence between the game types and the annotated tags shown in *table 22* show that there may be some correlations between game types, and perceptual and discourse features even though this has not been shown conclusively here nor has been tested statistically. Since this thesis uses the tags annotated for the purpose of another study, the tags are not chosen specifically for this study. To prove this relationship further, a broader range of linguistic

features need to be tested with correlation analysis. We did not base the game types directly on the questions of the previous study, but created the definitions independently. This avoids a circular investigation since the hypotheses are not directly reflected in the game types.

Since the definitions in the new Game Type Coding Scheme do not rely on conversational moves there is no need to first annotate each utterance by its conversational move before annotating game types. This saves times, especially for small corpora where there are not enough data to train a dialogue act tagger to automatically tag dialogue acts that correlates to the conversational moves.

5.3 Conclusion

We have extended the Swedish *Cups corpus* and formed a new Game Type Coding Scheme that defines seven distinct game types. The results from the inter-rater variance test shows that conversational games from free dialogue referring to spatial tasks can be categorized into game types in such a way that a good level of agreement between coders using this scheme can be achieved.

The game types, which directly reflect strategies to solve the task, are due to their definition, believed to be specific for each given task and, therefore, need to be adapted or expanded for each use of the coding scheme on corpora collected from different tasks. The META-games seem to be universal, and therefor believed to directly apply on all spatial conversations.

The research provides a method for free dialogue to be divided into more manageable units which allows computational modelling of spatial descriptions with a higher accuracy, since the game types provide subsets of data with reduced variation.

5.4 Future Directions

To further test the accuracy of the new Game Type Coding Scheme, we would need to measure pairwise agreement on game types of a number of coders annotating the data. By not using the scheme developer's result in the comparisons, we would get a more objective evaluation of the precision of the definitions. The scheme developer may not only use the definitions stated in the coding scheme, but also, subconsciously using other, more complex definitions which are not represented in the written definition and therefore not accessible for other coders. Unfortunately, lack of volunteers and time prevented using several external annotators during this research.

In order to build a classifier to computationally categorize game types, we need to annotate the utterances in the corpus with additional features to find more substantial correlations. We also need more example instances, which means that the corpus need to be expanded even further.

To examine the usability of the coding scheme in general, it needs to be tested on different corpora from participants solving different spatial tasks. For one, the HCRC Map Task could be used to examine the universal features of the META-games by coding all TASK-games as "TASK" while using the definitions in Section 4.1.1 for the META-games. If the definitions match the data in this corpora, it would strengthen the assumption that the META-games are universal. There is also a need to specify a method for defining TASK-games which are specific for each task that the dialogue is based around.

The coding scheme also needs to be tested in other languages to show to what degree the strategies deployed by interlocutors are based on language or culture specific cognition.

The effect of familiarity among participants needs to be examined further since the results suggest that it might affect more aspects than previously believed. Also, familiarity with the experiment leader might affect the results.

CHAPTER 6 – Bibliography

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. https://doi.org/10.1177/002383099103400404

Carletta, J., Isard, S., Anderson, A. H., Doherty-Sneddon, G., Isard, A., & Kowtko, J. C. (1997). The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, *23*(1), 20.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Colman, A. M. (2009). *Dictionary of Psychology* (3rd ed.). Oxford University Press.

Dobnik, S., Howes, C., Demaret, K., & Kelleher, J. D. (2016). Towards a computational model of frame of reference alignment in Swedish dialogue. In J. Björklund & S. Stymne (Eds.), *Proceedings of the Sixth Swedish language technology conference (SLTC)* (pp. 1–3). Ume\a a: Ume\a a University.

Dobnik, S., Howes, C., & Kelleher, J. D. (2015). Changing perspective: Local alignment of reference frames in dialogue. In C. Howes & S. Larsson (Eds.), *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 24–32). Gothenburg, Sweden.

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

Grice, M., & Savino, M. (1995). Intonation and communicative function in a regional variety of Italian, 14.

Healey, P. G. T., & Mills, G. J. (2009). A Dialogue Experimentation Toolkit. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*(31).

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

Houghton, G. (1986). *The production of a language in dialogue : A computational model.* (Ph.D.). University of Sussex.

Kowtko, J. C., Isard, S. D., & Doherty, G. M. (1993). Conversational Games Within Dialogue. In *University of Edinburgh*.

Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. Beverly Hills, Calif.: Sage.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: Philadelphia.

Lewis, D. K. (1969). *Convention: a philosophical study*. Cambridge, Mass.: Harvard UP.

Power Richard. (1979). The organisation of purposeful dialogues. *Linguistics*, *17*(1–2), 107. https://doi.org/10.1515/ling.1979.17.1-2.107

Searle, J. R. (1980). *Speech act theory and pragmatics*. Dordrecht: Reidel.

Sinclair, J. M. (1975). *Towards an analysis of discourse: the English used by teachers and pupils*. London: Oxford UP.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.