MSc Artificial Intelligence
Master Thesis

How should we call it?
Introducing the PhotoBook Conversation Task and Dataset
for Training Natural Referring Expression Generation in Artificial Dialogue Agents

by
Janosch Haber
Student-ID: 10400192

July 2, 2018

36 European Credits
January 2018 - July 2018

Supervisor:
Dr. Raquel Fernández
Co-supervisor:
Dr. Elia Bruni

Assessor:
Dr. Wilker Aziz

University of Amsterdam
Abstract

The past few years have seen an immense interest in developing and training computational agents for visually-grounded dialogue, the task of using natural language to communicate about visual input. While the resulting dialogue agents often already achieve reasonable performance on their respective task, none of the models can produce consistent and efficient outputs during a multi-turn conversation. We propose that this is primarily due to the fact that they cannot properly utilise the dialogue history. Human interlocutors on the other hand are believed to collaboratively establish a shared repository of mutual information during a conversation. This common ground then is used to optimise understanding and communication efficiency. We therefore propose that implementing a similar representation of dialogue context for computational dialogue agents is a pivotal next step in improving the quality of their dialogue output.

We believe that one of the main reasons why current research seems to eschew modelling common ground is that it cannot be assessed directly. In order to address this problem and gain crucial insights into its workings, we propose to first investigate the generation of referring expressions: Being an indirect representation of a referent object, they too are not absolute but conventions established with a specific conversation partner. By tracking the development of referring expressions during a conversation we therefore obtain a proxy for the underlying processes of the emerging common ground.

In order to develop a computational dialogue agent that can utilise the conversation’s common ground, we propose to implement a data-driven, modular agent architecture in an end-to-end training framework. With this setup, the dialogue agent is expected to learn the correct usage of referring expressions from recorded dialogue data directly. Opting for this approach requires a large amount of dedicated dialogue training data that has never been collected before. To initiate this new track in dialogue modelling research, we therefore introduce a novel conversation task called the PhotoBook task that can be used to collect rich, human-human dialogue data for extended, goal-oriented conversations. We use the PhotoBook task to record more than 2,500 dialogues stemming from over 1,500 unique participants on crowd-sourcing platform Amazon Mechanical Turk (AMT). The resulting data contains a total of over 160k utterances, 130k actions and spans a vocabulary of close to 12k unique tokens. An extensive analysis of the data validates that the recorded conversations closely resemble the dialogue characteristics observed in natural human-human conversations. We therefore argue that this data provides a pivotal new repository to be used in further research which has the potential to significantly improve the dialogue output consistency, efficiency and naturalness of artificial dialogue agents.
**Keywords**

Computational Dialogue Agents  
Visually-Grounded Dialogue  
Partner-Specificity  
Referring Expression Generation  
Referring as Collaborative Process

**Acknowledgements**

First and foremost I would like to thank my thesis supervisor Raquel Fernández for her dedication to this project, providing invaluable insights and feedback on the work presented here. I greatly appreciate the freedom and responsibility I was given in the development of the Photo-Book task and dataset, which I sincerely believe hold a great potential for future research in artificial dialogue agents.

I also want to thank my co-supervisor Elia Bruni for contributing his vast experience with computational model architectures, leading the discussions on how to approach the development of an artificial dialogue agent suitable for learning the usage of referring expressions.

I would like to thank Lieke Gelderloos for her frequent feedback, labour-intensive data annotations and knowledgeable contributions to the brainstorming sessions. Likewise I want to thank Aashish Venkatesh and Tim Baumgartner for reserving time for this project while working on their own theses.

I’m grateful to Facebook’s Jack Urbanek who provided real-time updates to the ParlAI framework each time my prototypes ran into problems never encountered before.

I want to thank Joop Pascha for spending hours and hours listening to my ideas and plans as well as creating the typesetting template for this thesis.

And, lastly, thanks to all my friends and fellow students who took on the drudgery of reviewing and spell-checking this thesis.

The data collection described in this thesis has been funded by a Facebook ParlAI Research Award granted to Raquel Fernández¹

In this thesis I use the right margin for illustrative figures, examples, references and citations of research central to the matter treated in the main content. Vital work here is listed with its full bibliographic information in order to facilitate the recognition of known work, other resources are referred to by their main authors. A full overview of references can be found at the end of this document.

¹ [https://research.fb.com/announcing-the-winners-of-the-facebook-parlai-research-awards/](https://research.fb.com/announcing-the-winners-of-the-facebook-parlai-research-awards/)
Contents

1 Introduction ........................................ 6
  1.1 Overview ....................................... 8
2 Linguistic Background .............................. 9
  2.1 Common Ground ................................ 9
  2.2 Referring as Collaborative Process ............... 10
  2.2.1 Conceptual Pacts .......................... 14
  2.3 Referring Expression Generation (REG) .......... 15
3 Related Work ....................................... 16
  3.1 Visual Dialogue ................................. 17
    3.1.1 ReferIt .................................. 18
    3.1.2 VQA ..................................... 18
    3.1.3 VisDial .................................. 19
    3.1.4 Cooperative Visual Dialogue ................. 21
    3.1.5 GuessWhat?! ................................ 22
    3.1.6 CoDraw ................................... 23
    3.1.7 FlipDial ................................... 24
    3.1.8 MutualFriends .............................. 25
  3.2 Dialogue Personalisation .......................... 27
    3.2.1 History-based Personalisation ................. 27
    3.2.2 Knowledge Base Personalisation ............... 28
  3.3 Context-aware Image Descriptions ................. 28
  3.4 Available Dialogue Datasets ....................... 30
    3.4.1 Free Chat Dialogue ........................ 30
    3.4.2 Task-Oriented Dialogue ...................... 31
4 The PhotoBook Dialogue Task ....................... 33
  4.1 Design Requirements ............................ 34
    4.1.1 Dialogue Modality ........................ 34
    4.1.2 Number of Interlocutors ....................... 35
    4.1.3 Task Symmetry ................................ 35
    4.1.4 Grounding of Dialogue ....................... 36
    4.1.5 Referent Types ................................ 37
    4.1.6 Re-Occurrence Rates of Referent Objects .... 37
  4.2 Task Setup ...................................... 37
    4.2.1 The PhotoBook Image Sets ..................... 38
    4.2.2 Specification of Games ....................... 39
  4.3 Implementation ................................... 39
    4.3.1 Task Instructions .......................... 40
    4.3.2 Warming-Up Round ........................ 41
    4.3.3 Game Round Mechanics ....................... 42
    4.3.4 Feedback Questionnaire ...................... 42
5 The PhotoBook Dataset ............................. 44
  5.1 Game Statistics ................................ 44
    5.1.1 Completion Times .......................... 45
    5.1.2 Game Scores ............................... 48
    5.1.3 Number of Messages ........................ 50
    5.1.4 Number of Tokens .......................... 51
    5.1.5 Efficiency vs. Task Performance .............. 52
    5.1.6 Messages and Tokens in Subsequent Games .... 53
    5.1.7 Order of Selections ........................ 54
    5.1.8 Evaluation Scores ........................... 56
  5.2 Linguistic Analysis .............................. 56
    5.2.1 Vocabulary ................................ 56
    5.2.2 POS-Tags .................................. 58
    5.2.3 Determiners ................................ 61
    5.2.4 Participant Contribution ....................... 63
    5.2.5 Dialogue Acts ............................... 65
    5.3 Dialogue Samples ............................... 66
6 Learning Referring Expression Usage ............... 72
  6.1 Automatically Extracting Referring Expressions .... 72
  6.2 Proposed Model Architecture ..................... 75
  6.3 Evaluation ..................................... 76
    6.3.1 Task Efficiency ............................ 77
    6.3.2 Qualitative Analysis ......................... 77
7 Conclusion ......................................... 79

Figures ............................................... 87

Tables ............................................... 90

Appendices .......................................... 92
  A The PhotoBook Task Design ....................... 92
    A.1 Number of Interlocutors ....................... 92
    A.2 Referent Types ............................... 92
    A.3 Initial Pilot Runs ............................ 94
    A.4 Determining Image Set Properties ............... 95
    A.5 Main Object Categories in the PhotoBook Image Sets .... 96
    A.6 Turn-taking ................................ 97
    A.7 The ParlAI dialogue framework .................. 98
    A.8 HIT Preview ................................ 98
    A.9 Qualifications ................................ 100
A.10 Matching Participants . . . . . . 100
A.11 Worker Payment . . . . . . . . . 101
A.12 Recording Conversations . . . . 101
B Dataset Analysis . . . . . . . . . . . . 103
B.1 Universal POS-Tags . . . . . . . . 105
C Evaluation . . . . . . . . . . . . . . 105
Introduction

The past few years have seen an immense interest in developing and training computational agents for visually-grounded dialogue, the task of using natural language to communicate about visual input. Growing from image processing research and empowered by the recent breakthroughs in the Deep Learning community, the field quickly moved from its initial set of object recognition\(^1\) and image captioning\(^2\) tasks towards incorporating more and more language-based elements. Current challenges include using natural language to distinguish and refer to certain elements of a visual input,\(^3\) posing and answering questions about them\(^4\) and constructing simple visual scenes from verbal descriptions.\(^5\)

It is not by coincidence that all of these approaches to developing artificial dialogue agents involve some form of multi-modal conversation task. Most applications of artificial dialogue agents currently are - and are expected to be - in goal-oriented settings. Often functioning as a human-machine interface,\(^6\) in those settings dialogue agents developed using free chat training data have two major shortcomings: Firstly, being trained on generic dialogue data they oftentimes fail the specific requirements of a particular application. And secondly, evaluating free chat output is inherently difficult.

Traditionally, a large part of statistical natural language processing research was focused on translation problems. As here samples from both domains have a one-on-one correspondence, computational models can be trained in a supervised fashion by minimising the difference between a predicted translation and a gold-standard sample. By contrast, training data-driven dialogue agents cannot draw on this feature: If they are designed to produce a next utterance based on the current dialogue state, there is a practically infinite number of completely different utterances that could be correct. Simply comparing produced samples with those recorded in the training data therefore no longer indicates model performance in a meaningful way - a different training regime is required here.\(^7\)

Using a conversation task to approach language mitigates the output evaluation issue: The quality of an utterance now can be assessed independently of the recorded samples by evaluating the model’s downstream task performance instead. If for example an utterance containing an object description causes the other participant to correctly identify the referent object, it can be considered correct as it achieved its intended goal. Visually-grounded conversation tasks therefore require dialogue agents to produce utterances that

1. exhibit a structure and style which are correct and as natural as possible (syntax),

\(^1\) Ren et al. (2017)  
\(^2\) Fortunato et al. (2017)  
\(^6\) Serban et al. (2015)  
2. hold an information content which correctly represents the visual input (*semantics*), and
3. efficiently forward the process of reaching an intended conversational goal. (*pragmatics*)

Through this setup, visually-grounded dialogue tasks are one of the first language tasks to incorporate pragmatics in their (indirect) evaluation criteria.

Analysing dialogue output of the current line of state-of-the-art dialogue agents developed for visually-grounded conversation tasks however reveals that they drastically degrade in performance when applied to multi-turn conversations. In this thesis we argue that this deterioration is primarily due to the fact that they appear to be incapable of adequately using the dialogue history by missing a representation of the conversation's *common ground*. Common ground is the pivotal concept of seminal linguistic theories that aim to explain the observation that throughout a conversation the interlocutors’ descriptions of referent objects are shortening - while the correct referents can still be identified with an unchanged accuracy. They propose that at the beginning of a conversation speakers are likely to use longer, more intricate descriptions of referent objects in order to ensure that they are correctly understood by their interlocutor. Once this initial *referring expression* is accepted and grounded in the conversation’s common ground, a mutual understanding is confirmed. This means that whenever one of the speakers now needs to re-refer to this specific referent, he or she can optimise the referring expression without a loss in its indicative power. Description details de facto migrate from the speakers’ utterances to the mutual repository.

In this thesis we propose to utilise the computational counterpart of common ground to enrich visually-grounded dialogue agents and ameliorate their multi-turn dialogue output. As a first step in this process, we focus on learning a correct and efficient usage of referring expressions. Developing a data-driven model for this task however requires large amounts of specific human-human dialogue samples which have not been collected before. In order to overcome this problem, we developed a novel dialogue task called *The PhotoBook Task*. In this task, participants are paired for a five-round game during which they repeatedly refer to a controlled set of referent images. Until publication we collected more than 2,500 dialogues stemming from over 1,500 unique participants on crowd-sourcing platform Amazon Mechanical Turk (AMT). The resulting data contains a total of over 160k utterances, 130k actions and spans a vocabulary of close to 12k unique tokens.

The resulting dialogues contain a large amount of referring expressions established during a single conversation - while the visual context and therefore also a large part of their partner-specific common ground is controlled by the task setting. A linguistic analysis of the data revealed that the recorded conversations closely resemble the characteristics re-

---


9 See Serban et al. (2015) for an elaborate study of available dialogue datasets.
ported by the previously mentioned dialogue experiments of Krauss and Weinheimer (1964)\textsuperscript{10} and Clark and Wilkes-Gibbs (1986).\textsuperscript{11} We therefore argue that this data provides a promising new repository for the intended research on common ground in artificial dialogue agents. We conclude this thesis with a detailed outline of the next steps towards a data-driven approach to end-to-end training of a computational dialogue agent for the PhotoBook task.

1.1 Overview

The remainder of this thesis is structured as follows: In Chapter 2 we review the most prominent results of linguistic research concerning the processes involved in human-human conversations, especially with respect to establishing referring expressions, and use them to motivate our objective of extending current computational dialogue agents with the ability to use referring expressions. Chapter 3 then covers an in-depth overview of the current approaches to modelling visually-grounded dialogue agents and dialogue agent personalisation, concluded with an analysis of their shortcomings with regard to their consistency and efficiency in multi-turn dialogues.

In Chapter 4 we present the PhotoBook Conversation Task, a novel dialogue task that can be used to collect rich, human-human dialogue data for extended, goal-oriented conversations, as well as function as a training framework for the development of an artificial dialogue agent for this specific task. Chapter 5 gives an overview of the specifics of the dataset collected with the PhotoBook task, showing that the dialogues elicited by the task indeed closely resemble those characteristics observed by the seminal works on human-human conversation and therefore should provide a solid base for developing a computational model of common ground.

Chapter 6 proceeds by showing that referring expressions recorded in the PhotoBook dataset can be linked to their respective referent image with high precision and recall using a simple heuristics baseline. Discussing evaluation metrics that are currently used to assess dialogue tasks. Introducing a completely new conversation objective, we also introduce a range of novel evaluation metrics that are proposed to function as proxies in order to evaluate model performance in tasks related to the usage of referring expressions. In Chapter 7 we then conclude this thesis with a clear-set path for future work on common ground in artificial dialogue agents.


2 Linguistic Background

Since the middle of the 1960’s, linguistics research has developed a number of hypotheses concerning our use of language in (task-oriented) dialogue. One of the main findings that resulted from this work can be summarised by the observation that while language is a vast, universal tool to communicate almost anything, it also tends to become highly efficient when applied in conversations where speakers share common information or a common context.

We aim to utilise this insight in the course of this project in order to propose improvements to the current design of computational dialogue agents. To do so, the following sections aim to give a more detailed overview of the central linguistic theories and experiments concerning language usage in goal-oriented conversation settings and how they might apply to artificial dialogue agents as well.

2.1 Common Ground

Grice famously expressed the observation that natural language may be used as an efficient tool to optimise interactions in his 1975 Cooperative Principle. In this principle he states that communication among rational agents can be seen as a cooperative interaction where all agents involved follow certain maxims in order to maximise the utility of the interaction.¹

For the specific case of verbal communication in natural language, Grice presented a set of four of these interaction maxims:

1. **Quantity.** Make your contribution as informative as is required (for the current purposes of the exchange) and do not make your contribution more informative than is required.

2. **Quality.** Do not say what you believe to be false and do not say that for which you lack adequate evidence.

3. **Relation.** Be relevant.

4. **Manner.** Avoid obscurity of expression and avoid ambiguity. Be brief (avoid unnecessary prolixity) and be orderly.

We will see in Chapter 3 that current state-of-the-art task-oriented dialogue agents are designed mostly with regard to the second and third maxims: Utterances produced by the agents should be relevant to what their task is - and they should be correct. In visually-grounded dialogue, this translates to producing utterances that are relevant in the sense that they cover what is depicted in the input scene and correct in the sense that they state accurate information about it (e.g. the right number of

people pictured or the right colour of the car in the foreground). We propose to enhance this current line of models towards a more natural language usage by directing the focus towards the quantity maxim: Are utterances produced by a dialogue agent efficient, i.e. does their level of detail match the requirements of the task?

One of the main reasons why analysing dialogue output based on this criterion is not a common standard yet is that whether or not the conveyed level of information is appropriate cannot be assessed in an absolute manner - but is determined by the amount of context information available to the interlocutor. Evaluating visually-grounded dialogue outputs in terms of the quantity maxim therefore means that the entire dialogue context and history need to be considered on top of the visual conversation context.

One central concept in discussing context-dependent relevance is the notion of common ground. It was formally introduced by Stalnaker (1978),\textsuperscript{3} based on previous research covering concepts such as common knowledge, mutual knowledge,\textsuperscript{4} and joint knowledge,\textsuperscript{5} and can be defined as the sum of the mutual, common, or joint knowledge, beliefs and suppositions of two interlocutors.\textsuperscript{6} The common ground between two speakers therefore always depends on the individual speakers, it is partner-specific. While some of the components of the common ground between two speakers are truly unique to the specific pairing of speakers, some components also are shared by a wider group of speakers, the so-called cultural common ground.\textsuperscript{7} It covers for example a shared language, knowledge about cultural practices or the jargon of a shared occupation, hobby or interest. Without any form of common ground, communication becomes effectively impossible. Because we are interested in the characteristics of conversations however, from here on out we will assume a certain minimum level of cultural common ground that allows speakers to communicate with one another in an everyday fashion. What (abstract) information is contained in the partner-specific common ground on the other hand eludes an exhaustive investigation. In order to still assess its function, research on common ground oftentimes focuses on one feature of conversations that heavily draws on the partner-specific part of the common ground between interlocutors: Referring Expressions.

2.2 Referring as Collaborative Process

Referring expressions often are defined as any expression used in an utterance to refer to something or someone (or a clearly delimited collection of things or people), i.e. used with a particular referent in mind.\textsuperscript{9} This definition builds on the triangle of reference or semiotic triangle displayed in Figure 2.1, one of the most widely accepted ways to conceptualise referring expressions. Published by Ogden and Richards\textsuperscript{10} in 1923, it shows an indirect link between a referring expression indicated referent object. The actual reference passes through an internal reference or thought state. When generating a referring expression, this referent state needs to adequately represent the referent object in order to subsequently produce...
a correct symbolisation in the referring expression. When processing a referring expression as listener this order is inverted; a listener first needs to translate the referring expression into a reference that indicates the actual referent object. Being this indirect representation of a referent object, referring expressions also are not absolute, but conventions of a specific group of interlocutors.

One of the most common forms of referring expressions that we use in everyday conversations are *proper nouns* - or names. Oftentimes there is no direct link between a name and the object it refers to, but by convention we are able to effortlessly link the two. Referring by name is highly efficient as it often only requires a single word or a small number of words to indicate otherwise complex referents. As an example, compare *Apple* to *an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services*, as it is described on Wikipedia

Referring expressions can be under- and over-specified. In the first case, an interlocutor cannot correctly resolve the reference because he or she misses necessary information to disambiguate referents; in the second case the speaker conveys more information than necessary because the common ground between the speakers would already allow the listener to correctly identify the referent object based on a part of it. Seen as a collaborative process, the parties involved in the conversation are involved in re-formulating proposed referring expressions to optimise them with regard to the current dialogue context. Krauss and Weinheimer (1964) were the first to empirically research the generation of referring expressions in a visually-grounded conversation task. In their small-scale lab experiment, two participants were given an identical set of six cards, each containing six objects arranged in a 3x2 grid. Three of the six figures (the so-called *redundant objects*) appear in the same position on all cards. The remaining three (the *discriminating objects*) appear in permuted positions. Without seeing their partner’s cards, the participants’ task then was to match pairs of cards based on the positions of the discriminating objects by talking to one another. Arguing that some objects also already have common or *popular* referring expressions attached to them through a shared cultural background, Krauss and Weinheimer

---

**Figure 2.1: Triangle of reference or semiotic triangle as proposed by Ogden and Richards (1923).**


chose to use self-designed, ambiguous line drawings that evade such common references (see Figure 2.2), forcing the participants to develop their own referring expressions. The experiment was completed by five different pairs of participants.

Analysing the resulting dialogues, Krauss and Weinheimer found that the most commonly employed strategy was to initially refer to figures through combinations and variations of common objects (e.g., an upside-down martini glass in a wire stand). They then noted that in those later turns the participants apparently coded the figure's description into a shortened version of the initial reference phrase (e.g., martini). As an effect, the length of referring expressions quickly reduces with the number of recurrences until the speakers converge on a certain reference. For most speaker dyads, this convergence occurred after three to six encounters of a given figure (see Figure 2.3). This means that under a stable task-success rate, speakers collaboratively optimised the utterance length to task success ratio - as would be expected based on the cooperative principle.

During the next years, Krauss and his team continued investigating this effect, observing for example that speakers do not shorten their referring expressions when dictating to a tape recorder, but it took until Clark and Wilkes-Gibbs revisited it in their 1986 seminal paper indicatively titled Referring as a collaborative process for the field to gain momentum. In their paper, Clark and Wilkes-Gibbs presented an elaborate model of reference expression generation based on a new set of dialogue experiments: In their task one of the participants (the director) has to instruct a second participant (the matcher) to recreate a grid of 12 different Tangram figures (see Figure 2.4). Being separated by an opaque screen, the participants again could only communicate verbally. This task was repeated with the same director-matcher pairs for six rounds and completed by eight different pairs of participants.

Clark and Wilkes-Gibbs observed a similar effect to that reported by Krauss and Weinheimer, with referring expressions adapting and notably shortening over consecutive game rounds. For a broad picture of this shortening process, they present the following string of utterances used by a director to refer to the Tangram figure depicted in Figure 2.5 in subsequent trials of the task:

1. All right, the next one looks like a person who’s ice skating, except they’re sticking two arms out in front.
2. Um, the next one’s the person ice skating that has two arms?
3. The fourth one is the person ice skating, with two arms.
4. The next one’s the ice skater.
5. The fourth one’s the ice skater.
6. The ice skater.

As it is in this example, Clark and Wilkes-Gibbs note that directors in all cases described a figure in the first trial, while referring to figures in all subsequent trials. Additionally, referring expressions moved from

\[\text{Figure 2.2: Example figure cards shown to participants in the experiments of Krauss and Weinheimer (1964).}\]

\[\text{Figure 2.3: Length of referring expressions used by the five subject pairs as a function of the number of repetitions in the experiments of Krauss and Weinheimer (1964).}\]

\[\text{Figure 2.4: The twelve Tangram figures used in the experiments of Clark and Wilkes-Gibbs (1986).}\]

\[\text{Figure 2.5: Tangram figure from the conversation task by Clark and Wilkes-Gibbs (1986).}\]
non-standard noun phrases in early trials (Um, the next one’s the person ice skating that has two arms?) to standard ones in the later trials (The ice skater). These two characteristics lead to a significant reduction of utterance length over trials (Figure 2.6). Specifically, directors used an average of 41 words per figure in the first trial while only 8 in the sixth trial. The number of speaking turns per trial develops in a similar way, ranging from an average of four turns per figure in the first trial to a single utterance in the last three trials.

Abstracting from these observations, the authors propose a model comprised of three processes involved in reaching mutual agreement about a specific referring expression: Its initialisation, refashioning and acceptation. In the initialisation step, an initial object description is proposed. With the task setup of Clark and Wilkes-Gibbs, this is done by the director who describes the figures in his or her grid following the order of their arrangement. Being unsure of how well the matcher will understand this initial description of an abstract Tangram figure, initial descriptions often are elaborate but purposely provisional in nature, inviting the matcher to participate in its formulation process. This initiates the second process in which participants refashion an initial description to a more suitable, efficient one that is understood and unambiguously grounded by both participants. The refashioning process can contain multiple steps and is concluded through a mutual acceptance of the resulting canonical reference by both speakers, either by explicitly agreeing to its usage or by simply using an expressions without further alterations.

Clark and Wilkes-Gibbs stress that it is central to view this process as a collaborative one as participants minimise the collaborative effort: If speakers were adhering to classical theories of minimal effort, their focus should be on optimising their contribution to the conversation instead of the efficiency of the conversation as a whole. In the recorded dialogues, directors however oftentimes use elaborate initial descriptions. Through this choice they aim to reduce the amount of refashioning required until the referent object can be grounded by the matcher and a canonical reference can be formed. According to Clark and Wilkes-Gibbs, this minimises the collaborative effort as the simplification or narrowing of a proposed expression is far less costly than establishing a grounded referring expression in the first place.

Regarding the setup of referring expressions, the authors differentiate between temporary and permanent properties: Some of the attributes used in a referring expression are constant to the object described, while others are relative to the current display of the object (like for example its position in the grid). In contrast to temporary properties, permanent characteristics are a salient, distinctive and highly recognisable part of the common ground. Optimising collaborative effort therefore should favour the usage of permanent properties to refer to figures instead of the temporal properties that are easy to access for the director. Clark and Wilkes-Gibbs argue that this claim is supported by the dialogue data collected, where 90% of all referring expressions used permanent attributes only, 7% used a combination of permanent and temporal

---

Figure 2.6: Average words spent per figure during the six trials of an experiment run in Clark and Wilkes-Gibbs (1986).


17 Schegloff et al. (1977); Levelt (1983), see Clark and Brennan (1991) for a comprehensive overview of costs proposed to result from grounding a referring expression during a conversation.
features and just 2% were solely based on temporal aspects. These in turn often resulted from figures of special interest in the previous trial (e.g. The first one from last round, the one we had wrong last round).

2.2.1 Conceptual Pacts

In 1996 Brennan and Clark extended the collaborative reference generation model by the aspect of partner-specificity. In this extension, the canonical references developed during a specific conversation are proposed to be the result of a conceptual pact between the speakers. This pact is based in their common ground and therefore does not readily transfer to conversations with other individuals. In other words: Referring expressions developed in one conversation cannot simply be used and understood in subsequent conversations with other speakers. They support this claim through dialogue data obtained from a variation on the Tangram experiment where participants were given grids containing pictures of 12 everyday objects. Six of the 12 displayed items were common to both participant’s sheets, and six of them different. The participants were then asked to identify the common objects. For the partner-specificity experiment, two sets of these grids were made: Set A contained same base-category distractor images for two target objects selected from the set of common images (e.g. a sneaker and a high-heel shoe for the depicted target shoe (a loafer), set B did not contain any same base-category images. During a full experiment run, a director completed four trials with a first matcher, using set A, and then either continued with the same matcher or was paired with a new matcher for four trials with set B. The experiment was completed by 10 same-partner and 10 switch-partner groups.

The idea behind this setup is that during the first four trials participants will develop a canonical reference to the target that is more specific than the object’s base-category (i.e. the referring expression shoe will not be distinctive enough given the context of the two other shoes). Assuming that this canonical reference is a conceptual pact between the two speakers, Brennan and Clark expect that directors are more likely to adhere to that referring expression when continuing with the same matcher during the B part than when being paired with a new matcher. This indeed was the case, with directors continuing to use referring expressions established during part A in the first trial of part B in 48% of the cases in the same-partner setup and only 18% of the cases in the switch-partner setting. In the subsequent trials, directors with the same partner also were more likely to continue to rely on established conceptual pacts while directors in the switch-partner setting often refashioned referring expressions to incorporate the base-category reference (15% to 23% for same-partner, 20% to 55% for switch-partner). This phenomenon contradicts the basic quantity maxim in the sense that in the B set base category descriptions can be used, which would be more efficient and therefore the preferred choice. Directors choosing to adhere to the established conceptual pact and trading in some utterance length efficiency therefore indicate that conceptual pacts have a stronger influence on
perceived conversation efficiency than changing an established canonical reference to a simpler one. This was later coined \textit{lexical entrainment}, closely related - but not equal - to the phenomenon of \textit{lexical convergence} that can be observed in human-human conversations as well.\textsuperscript{19}

\subsection{2.3 Referring Expression Generation (REG)}

Referring Expression Generation (REG) is the computational counterpart to the linguistic research on referring expressions. As a sub-field of Natural Language Generation (NLG), REG primarily is concerned with the generation of definite descriptions for referent identification, focusing on content selection and linguistic realisation.\textsuperscript{20} According to a comprehensive study of Krahmer and van Deemter (2012), REG can be traced back to the very beginnings of Natural Language Processing (NLP) itself, with Winograd’s SHRDLU program using a simplistic REG algorithm.\textsuperscript{21} Not much is known anymore about the specifics of many other approaches to REG before the 1990’s, but with Reiter and Dale’s \textit{Greedy Heuristic}\textsuperscript{22} and \textit{Incremental Algorithm}\textsuperscript{23} a new chapter in REG research started. Algorithms developed during the next ten years primarily built on an incremental preference over objects attributes: When describing a referent object and the context makes it necessary to elaborate the base description, additional information should be given according to this order of preference. The problem of finding an optimal description under this restriction can be re-formulated as a search problem in a restricted space and is \textit{NP-hard}.\textsuperscript{24} Many of the later REG research therefore was concerned with extending its coverage, using relational descriptions, context dependencies and the salience of attributes. All however stayed within the search framework, venturing into graph search, constraint satisfaction and novel approaches to knowledge representation.\textsuperscript{25} Concluding their survey, Krahmer and van Deemter designate the most central challenges that the field currently faces, containing among other things the questions how to generate suitable referring expressions for interactive settings and how to incorporate information encoded in visual input, lamenting that for neither aspect sufficient amounts of dedicated human data are available. With this project we hope to give a new input to these ongoing issues.


\textsuperscript{25} Krahmer and van Deemter (2012)
From the study of the linguistic background covering human-human dialogue presented in the previous chapter, we concluded that a conversation’s common ground is a central aspect in developing what we perceive as a natural dialogue. Following this model, during a conversation human interlocutors are believed to collaboratively establish a shared repository of mutual information which they can access in order to refine their utterances. Grice proposes that we do so because communication can be seen as an optimisation problem where information needs to be conveyed correctly but efficiently. While common ground is difficult to access directly, experiments on goal-oriented dialogues supported this theory by showing that for instance referring expressions used to indicate previously encountered referent objects shorten with each re-occurrence - without changing its identification accuracy.

Incorporating common ground into artificial dialogue agents therefore is a crucial step in increasing the consistency and naturalness of their output when they are to be applied for leading longer conversations with a specific dialogue partner. In order to assess the current state of the field, in this chapter we will cover research on artificial dialogue agents, mainly focusing on an analysis of the most prominent approaches to developing dialogue agents for visually-grounded conversation tasks (Section 3.1). In this specific set of tasks the conversation goal is linked to a visual context, i.e. identifying certain referent objects from a set of distractors, and thus can be evaluated more easily than in other multi-modal dialogue settings. Investigating their respective dialogue output, we conclude that while some models already reach reasonable performance levels on their respective task, these tasks themselves often are too simple to require and elicit dialogue output that resembles a natural, multi-turn conversation between human interlocutors.

Section 3.2 then covers research that identified and addresses a similar problem: The recently emerged field of dialogue personalisation. These approaches aim to mitigate inconsistency in a dialogue agent’s output by conditioning generated utterances on an additional source of information, like for example an agent’s persona representation. This additional repository of information already resembles what we might aim for when modelling common ground for computational agents, but one thing that all of the current approaches to dialogue personalisation have in common is that they consider (previously collected) information about a single agent only. When considering the usage of partner-specific referring expressions, we on the contrary propose that the mutual information established during a conversation is a central element.
A third aspect covered very briefly in Section 3.3 then is context-aware image captioning, closely related to the process of referring expression generation: When describing a referent image in the context of other, similar images, which features should be mentioned in order to improve the likelihood of a correct identification? This however only applies to the first reference to a referent image as later references should be conditioned on the established common ground.

Concluding that the implementation of common ground into computational dialogue has not been attempted, we propose that this is an interesting issue to investigate in order to improve their output performance. Developing a data-driven dialogue agent for multi-turn conversations inevitably requires respective dialogue data. We therefore conclude this chapter with a summary of available datasets that currently are or could be used for training computational dialogue agents (Section 3.4), noting however that none of them is suitable for the intended application. To mitigate this problem, we will introduce a new dialogue task and dataset in Chapters 4 and 5.

3.1 Visual Dialogue

Visual Dialogue is a multi-discipline problem setting in the field of Artificial Intelligence (AI) research. In visual dialogue agents communicate about visual input, either in a free chat setting or - more often so - to solve a multi-modal conversation task. Research on visual dialogue grew from the Computer Vision (CV) community, starting with object recognition tasks that quickly transferred into image captioning and image understanding tasks with the emergence of Deep Neural Networks (DNNs) around 2013.¹ Solving visual dialogue tasks with an artificial agent requires elements from the diverse fields of Computer Vision, Natural Language Processing and Knowledge Representation and Reasoning. Research on visual dialogue therefore often is seen as a central contribution towards strong artificial intelligence.² A thorough analysis of computational models developed for this set of early visually-grounded language tasks however revealed that oftentimes a simple, coarse scene-level understanding of images paired with even simpler n-gram statistics suffices to generate reasonable image captions.³ These one-sided tasks like sole image captioning therefore fall short of the requirements for a more AI-complete task.⁴

Addressing this shortcoming, during the past few years a number of more involved visually-grounded language tasks were introduced, all of which incorporate a second agent to further the field towards actual visually-grounded dialogue. In the following subsections we present the currently most prominent of these visually-grounded dialogue tasks, each requiring participants to use more and more complex multi-modal knowledge in order to be completed well. And while some artificial dialogue agents developed for these tasks manage to perform reasonably well already, most tasks are still too simple to require natural, multi-turn dialogue, which as a result is not learned to be produced by the models.

¹ Fang et al. (2015); Chen and Zitnick (2015); Donahue et al. (2017); Mao et al. (2014); Karpathy and Fei-Fei (2017); Vinyals et al. (2015)
³ Antol et al. (2015)
⁴ Shapiro (1992); Yampolskiy (2013)
3.1.1 ReferIt

Kazemzadeh et al. (2014)\textsuperscript{5} presented the first large-scale dataset that allowed for studying the generation of referring expressions in visually-grounded dialogue settings. In order to collect this kind of data, they developed a simple conversation task called ReferIt in which a photograph is shown to two participants. One participant is assigned the role of the instructor. In the instructor’s display, one of the objects in the photograph is highlighted, and he or she is asked to describe that object with a single sentence. Based on that description, the other participant (the guesser) then has to select the object that he or she believes to be the indicated referent.

Through implementing the ReferIt task on a website and making it publicly accessible, the authors collected more than 130k successful games, addressing close to 100k distinct objects in almost 20k different photographs of the ImageCLEF IAPR image retrieval dataset\textsuperscript{6} with the SAIAPR TC-12 object annotations.\textsuperscript{7}

Being a single-turn task, in ReferIt interactions the interlocutors cannot collaborate in the development of referring expressions as there is no feedback opportunity or long-term pairing. This means that task is not suitable for investigating actual dialogue.

3.1.2 VQA

Antol et al. (2015)\textsuperscript{8} introduced visual question answering (VQA) as a conversation task to collect visually-grounded but open-ended dialogue data in the form of question-answer pairs. Arguing that answering visually grounded questions requires more than the coarse-level scene understanding which appears to be sufficient for traditional image captioning, they suggest that solving this particular task indeed relies on a number of intelligence-related capabilities like fine-grained object detection, activity recognition and knowledge-base and commonsense reasoning.

To create a suitable dataset for training computational agents for the VQA task, Antol et al. first tasked human participants to ask a specific question about a displayed image or clip-art scene, collecting three questions each for a total of more than 200k MS COCO\textsuperscript{9} images and 50k clip-art scenes\textsuperscript{10} through crowd-sourcing the experiments online. In a second task, they then collected 10 answers from different participants for each of those 760k questions, resulting in a total set of around 10M question-answer pairs. While the resulting dataset contains a large variety of question types (see Figure 3.1), answers are mostly short (1 word 89.32\%, 2 words 6.91\%, 3 words 2.74\%). Almost 80\% of all answers are either yes or no answers.

In order to evaluate task performance, Antol et al. developed two different testing modalities: Open-ended and multiple-choice. In the open-ended task setting, an answer to a question is considered correct if at least 3 of the human participants gave the exact same answer. Although this is a very strict and limiting requirement, the authors argue that this is necessary because more semantic metrics such as Word2Vec\textsuperscript{11}

\textsuperscript{5} Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. (2014). Referit game: Referring to objects in photographs of natural scenes. In EMNLP

\textsuperscript{6} Grubinger et al. (2006)

\textsuperscript{7} Escalante et al. (2010)


\textsuperscript{11} Mikolov et al. (2013)
and GloVe\textsuperscript{12} might group together concepts that for this task need to be differentiated (e.g. ‘left’ and ‘right’), and basic word-overlap features like BLEU\textsuperscript{13} or ROGUE\textsuperscript{14} are typically only reliable for longer sentences. As reported before, the vast majority of answers in the VQA dataset however only contains a single word. In the multiple-choice setting, the gold-label answer is presented with a set of false candidate answers. The task of the model then is to rank these candidate answers given an input image and question, which should favour the correct answer.

As this basic VQA task setup also does require nor contain interaction between participants, it too falls short of our more strict requirements for multi-turn visually-grounded dialogue.

3.1.3 VisDial

Arguing in a similar vein, Das et al. (2017a)\textsuperscript{15} generalise the previously presented visual QA approach to open-ended, goal-driven visual question answering by introducing a two-stage image guessing task: In their VisDial setup, one of the participants (the questioner) only gets a description of the image that the other player (the answerer) sees on his or her screen. The questioner then needs to ask questions about that image until he or she is confident that he or she could identify the image shown to the answerer from a panel of similar distractor images. Because here both participants can send free-form messages, the VisDial task elicits actual dialogue outputs.

By crowd-sourcing their task on AMT, the authors collected dialogues consisting of 10 QA pairs each for more than 120k images from the MS COCO train- and validation set. Compared to other image question-answering datasets like VQA, Visual 7\textsuperscript{W}\textsuperscript{16} or Baidu mQA\textsuperscript{17}, Das et al. note that the VisDial dataset does not contain a visual priming bias that often is present in the traditional VQA task setups: Research by Zhang et al. (2016)\textsuperscript{18} showed that formulating questions about an image while it is displayed leads to a strong bias towards questions about objects visible in the scene. As an example: Simply answering ‘yes’ to all questions in

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Distribution of first n-grams for (left to right) VisDial questions, VQA questions and VisDial answers. N-grams extend from the centre, the width of a portion indicates its relative frequency in the data.}
\end{figure}

\textsuperscript{12} Pennington et al. (2014)
\textsuperscript{13} Papineni et al. (2002)
\textsuperscript{14} Lin (2004)
\textsuperscript{16} Zhu et al. (2016)
\textsuperscript{17} Gao et al. (2015)
the VQA dataset that start with ‘Do you see a...’ results in an accuracy of 87%. Because questioners do not see the image while asking questions about it in the VisDial task, this setup reduces the bias and leads to more open-ended question types. In the resulting VisDial dataset, the most frequent start to a question therefore is ‘is’, compared to ‘what’ in VQA, and answers include uncertainty as questioners sometimes ask specific questions about details not (entirely) visible in the image (see Figure 3.1).

Based on the collected data, Das et al. then also established a number of performance baselines using different neural answerer models for their specific task setting. All models follow an encoder-decoder setup where the encoder converts the 1) input image \(I\), 2) dialogue history \(H\) and 3) current question \(q_t\) into a latent vector space representation. The decoder then maps that representation back into the natural language output space (see Figure 3.2). In order to evaluate performance, a ranking task is proposed where the gold-label answer to a question again are enriched with a set of additional distractor answers. Measuring the mean reciprocal rank (MRR) on reordering the candidate answers, they obtain best performance with a memory network encoder (MN),\(^{19}\) outperforming simple neural network approaches.

Being the first visually-grounded language task to elicit actual human-human multi-turn dialogue data, the VisDial dataset warrants further investigation as to whether it can be used to model the interlocutor’s common ground through it as well. As questioners however try to cover as much detail of the image as possible in order to correctly identify it in the second part of the task, object re-references are quite rare in the resulting dialogues. This characteristic makes it inherently difficult to assess the common ground through them.

---

\(^{19}\) Bordes et al. (2016)
3.1.4 Cooperative Visual Dialogue

Remarking that their previous approach treats dialogue as a static supervised learning problem rather than an actual interactive agent learning problem,\(^{20}\) Das et al. (2017b)\(^ {21}\) extended their approach by introducing a reinforcement learning (RL) setting to interactively train agents for the VQA task of the VisDial dataset. To do so, they identify two key challenges to this task. Firstly, the questioner (labelled Q-BOT) must be able to ground the initial description to estimate which images from a given pool of candidates match this description and then ask follow-up questions in order to build a ‘mental model’ of what the image shown to the answerer (A-BOT) looks like. Note that because here the pool of candidates is potentially infinite, the task cannot simply be solved by eliminating all candidates until the actual match is found. And secondly, the A-BOT needs to have some kind of mental model of Q-BOT’s understanding of the image in order to supply answers that are precise enough to allow discrimination between similar candidate images.

To train the models, the authors apply a collaborative RL approach where Q-BOT and A-BOT complete a fixed number of rounds in which Q-BOT states a question, A-BOT answers it and Q-BOT generates a feature vector of the resulting mental model image. The bots then receive a reward that is inversely proportional to the Euclidean distance between the estimated feature vector and the target image’s actual feature vector (see Figure 3.3). This way an unlimited number of ‘game plays’ can be simulated during training. Natural language for questions and answers is generated and parsed by LSTM networks pre-trained on VisDial dialogues.

Das et al. show that their RL approach to collaboration in visual question answering is principally feasible by applying the model to a toy problem domain with simple geometric shapes and a code like language. Not being confined to using actual natural language, the bots here are able to maximise the reward after about 400 training iterations. In the real-world task, the full RL model significantly outperforms a supervised pre-training baseline and models with frozen agents in generating feature representations that are closer to the target image. When testing


![Figure 3.3: Policy networks for Q-BOT and A-BOT as presented by Das et al. (2017b). At each round of the dialog, Q-BOT generates a question based on its state encoding and sends it to A-BOT. A-BOT then updates its state encoding based on the received question, generates an answer and returns it to Q-BOT. Q-BOT then predicts an image representation and receives a reward based on the distance to the actual image shown to A-BOT.](attachment:image.png)
the generated language output, they however had to conclude that - despite appearing more informative - the RL model only lead to minor improvements in ranking candidate replies to questions when evaluated on the VisDial test-set. They propose that this due to the fact that in end-to-end RL training, successful dialogues do not necessarily need to be human-like. The generated outputs therefore also do not necessarily improve the performance on this particular ranking task.

3.1.5 GuessWhat?!

In GuessWhat?!, De Vries et al. (2017b) invert the ReferIt setup: Like in ReferIt, one of the participants is assigned the role of the oracle, who sees an image with a highlighted object, while the other participant is the questioner who gets to see the same image, but has to guess the highlighted object. To do so, the guesser can now send free-form questions to the oracle, who is restricted to a three-way yes/no/NA answer selection. A dialogue is considered successful when the questioner selects the object highlighted in the oracle’s image. Crowd-sourced on Amazon Mechanical Turk (AMT), they collect more than 150k dialogues composed of more than 80k question-answer pairs based on images from the MS COCO dataset. On average, there are 5.2 questions per dialogue and 2.3 dialogues per image (see Figure 3.4).

When investigating the participant’s strategy, the authors found that the number of questions posed before selecting an object grows with a function on the number of displayed objects that is somewhere between linear and logarithmic. A linear scaling would imply that questioners just list all displayed objects, while a logarithmic increase implies an optimal strategy that always exactly halves the number of potential objects. Human strategy therefore lies somewhere in between those two extremes (see Figure 3.5). Investigating this phenomenon a little further, De Vries et al. used a Dynamic Topic Model23 to divide the object references in the questioner’s utterances into abstract and concrete descriptions. The results in Figure 3.6 show that participants tend to start dialogues with more abstract object categories while turning to concrete object descriptions for later questions.

Based on the collected dataset, they then train three different modules to establish a baseline for the object identification task of GuessWhat?! Based on their roles, these modules are dubbed oracle (providing answers given the knowledge of the highlighted object), question generator and object guesser.

The oracle model concatenates 1) the visual features obtained from the image as a whole, 2) the visual features of the highlighted object’s crop, 3) an encoding of the spatial information of the crop, 4) the MS COCO object annotation of the highlighted object and 5) an LSTM encoding of the question into a vector and feeds it to a single hidden-layer MLP with a softmax over the three answer options.

The guesser combines an LSTM encoding of the previous dialogue and the FC8 features of the image and calculates a softmax over all

![Figure 3.4: Number of questions per dialogue in the GuessWhat?! dataset by De Vries et al. (2017b).](image)

![Figure 3.5: Human solution strategy efficiency in the GuessWhat?! task as compared to linear or logarithmic strategies (image from De Vries et al. (2017a)).](image)

![Figure 3.6: When trying to identify the selected object, questioners move from abstract object classes to concrete descriptions (image from De Vries et al. (2017a)).](image)

23 Blei and Lafferty (2006)
annotated objects in the image, which are encoded using their object category and spatial annotation provided by the MS COCO metadata. An ablation study however revealed that just using the dialogue state’s encoding outperforms the full model by about one percent (test error 38.7%; human 9.2%, random 82.9%).

The question generator finally poses the most challenging part of the setup as it requires both high-level visual understanding to ask meaningful questions as well as a long-term context in order to produce a sequence of relevant questions. In order to approximate this behaviour, they implement an extension of a hierarchical recurrent encoder model (HRED)\textsuperscript{24} that also incorporates an image’s VGG features.

Together, these three modules can simulate an entire game, in which with the best parameter settings the model was able to determine the highlighted object with an accuracy of 38.6% (human-human accuracy 84.4%). Like Das et al. they then also frame the problem in a reinforcement learning setting, specifically training the question generator to learn how to strategically ask questions and how to assess when enough information is available to guess the highlighted object.\textsuperscript{25} This improved selection accuracy to 53%.

While the three-part GuessWhat?! model architecture is an interesting approach to further investigate, with one of the participants being limited to yes/no replies only, the GuessWhat?! task itself is by definition unsuitable for assessing the common ground between speakers.

3.1.6 CoDraw

Kim et al. (2017)\textsuperscript{26} extend the previously presented visual dialogue task setups by adding an explicit action through which they aim to better assess the language grounding. In their setup called CoDraw, two participants are paired on crowdsourcing platform AMT. One of the participants (the teller) is a set of Drawer’s clip arts, and CD is a set of Drawer’s clip arts. W and H denote the width and height of drawing canvas, respectively.

The other participant (the drawer) sees a blank canvas with a drag-and-drop selection of candidate objects in different sizes and orientations (see Figure 3.7). The task of the participants then is to recreate the scene shown to the teller on the drawer’s canvas by chatting with one another. The resulting dataset contains almost 10k dialogues with a total of 138k utterances. Games took an average of 6 minutes, with a median of 7 rounds and a mean resulting scene similarity score of 0.4 out of 5.

In order to train a model for both the teller and the drawer, Kim et al. propose two different setups: In the simple sequential single attention model, the teller sequentially selects an object in the scene and creates an utterance using attention on this object in the clip-art representation vector. The drawer module for this model is trained on all train-set turns where the drawer only modified or added a single object in the scene (about 37% of all turns). Because this setup does not utilise the dialogue history, the authors then also introduce a dynamic, multi-attention model that gives memory to both the teller and the drawer (see Figure 3.8).

\textsuperscript{24} Serban et al. (2015)


\textsuperscript{27} Zitnick and Parikh (2013); Zitnick et al. (2013)

Figure 3.7: Participant’s AMT interface for the drawer in the CoDraw task setup by Kim et al. (2017).
Using the conversation’s context encoding as additional input to the attention module, the focus on different objects in the scene occurs naturally, shifting with the proceeding state of the dialogue.

Comparing different input sentence representations for the proposed models, their results show that the simple sequential model works best using a bag-of-word representations, outperforming the more complex dynamic multi-attention model. Both models however significantly under-perform the human reference, with the simple model producing reasonable drawing outputs but no actual co-operative dialogues and the more complex model producing richer dialogues but unrecognisable output drawings.

The dialogue data collected through the CoDraw task also is principally suitable for investigating the common ground establishing among the participants. Through the drawer’s interactions with the referent objects, the visual language grounding is comparatively clear; The authors however note that utterances in the recorded dialogues cannot be mapped directly to the drawer’s actions as tellers for example tend to mention multiple features in a single message to increase efficiency and drawers often wait with drawing until they have a general idea of the teller’s scene. With respect to the about five or six object re-references that according to the observations of Clark and Wilkes-Gibbs are needed until a referring expression is accepted as canonical reference, CoDraw dialogues are very unlikely to reach those numbers, with only some of the objects being re-referred to when placing or reordering other elements with respect to their position.

3.1.7 FlipDial

FlipDial by Massiceti et al. (2018) finally is the most recent addition to the visual dialogue family, introducing a generative model to simultaneously produce utterances for both roles in a visual question answering task. With this setup, for the first time a single agent can generate questions as well as answers and therefore model an entire VQA dialogue by itself.

Building on the VisDial dataset, Massiceti et al. use convolutional neural networks (CNNs) to encode entire dialogues directly, breaking with the traditional approach of explicitly modelling a dialogue’s sequentiality through hierarchical recurrent models. In order to encode sentences as CNN input, the authors treat a stack of fixed-sized word
embeddings of the sentence’s tokens as a regular, single-channel image. On this ‘image’, convolutions can be applied in the common manner, generating a sentence embedding. Extending this approach to entire dialogues, sequences of sentences can be viewed as 3D stacks of those sentence ‘images’, creating the counterpart of a multi-channel image. Applying convolution to such a dialogue representation can therefore directly create a complete dialogue’s embedding and implicitly captures the sequential nature of the input. In analogy to the common meaning of channels in visual processing, Massiceti et al. refer to this process as the colouring of visual dialogue.

For the generative part of the framework, conditional variational auto-encoders (CVAEs)\(^{29}\) are used to generate utterances conditioned on an image input and a given dialogue context, and to recognise an image referred to in a given utterance. In their experiments, this is either used to simply answer visually-grounded questions like in the traditional VQA tasks, or to generate entire dialogue blocks directly. For the latter, the authors also investigated the effect of adding an auto-regressive module\(^{30}\) to the generator, explicitly enforcing sequentiality in the generation of the dialogue blocks. Combining this CVAE framework with the convolutional approach to input encoding, Massiceti et al. present a three-stage prior-encoder-decoder architecture (Figure 3.9).

While they can evaluate traditional VQA performance based on the ranking task proposed by Das et al. (2017a), evaluating full dialogue blocks requires a new evaluation method because here the questions also are generated by the model itself instead of given by a test-set. As this means that comparison with gold-standard becomes impossible, Massiceti et al. propose to evaluate dialogue quality by directly comparing the ELBO’s cross entropy and Kullback-Leibler divergence terms, and introduce two novel similarity heuristics, question relevance and latent dialogue dispersion.

Results show that in the traditional VQA task, the CVAE setup underperforms the baseline presented by Das et al. (MRR of 0.45 compared to 0.52 in VisDial). In the full dialogue evaluation, the auto-regressive decoder leads to the lowest cross-entropy error and KL divergence while the non-regressive model stays closer to the recorded dialogue structure as measured by both the question relevance and latent dialogue dispersion metrics.

3.1.8 MutualFriends

Not actually visual dialogue but otherwise very related to the kind of dialogue that we are interested in is the data collected through the MutualFriends task by He et al. (2017).\(^{31}\) In this task, two participants are given different lists of fictional friends with a number of attributes (School, Major, Company, etc.). In both lists there is only a single common entry (the mutual friend), which the participants are asked to collaboratively determine by chatting with one another. Through crowdsourcing a large-scale, online dialogue data collection run with this task (see Figure 3.10 for a screenshot of the task’s interface), He

\(^{29}\) Kingma and Welling (2013); Sohn et al. (2015)

\(^{30}\) Gulrajani et al. (2016); Oord et al. (2016)

et al. collect a dataset of 11k human-human dialogues, which, in contrast to many prior settings, are task-oriented but still contain open-ended dialogue acts.

The main goal of the computational model trained on this dataset then is to capture conversational implicature that is part of human solution strategies to the dialogue task. For example: By asking ‘do you have any friends from Columbia?’ in the context of this particular task, a speaker also conveys the information that he or she has at least one friend from Columbia. This information would be lost when interpreting the utterance as being a request only. In order to model both the structured and open-ended context that is present in the data, they propose a novel Dynamic Knowledge Network (DynoNet) architecture that models the agent-specific dialogue state as a private knowledge graph. Each node of this graph holds a context embedding for a specific item, attribute or entity, which is recursively updated every time that this node or a related one is mentioned during a conversation. New utterances are then generated by an attention-based mechanism\(^{32}\) over the node embeddings (see Figure 3.11).

The resulting model reaches a task success rate (.25) that is a lower but comparable to a rule-based baseline (.29, human performance: .38). Generated utterances however are a little longer and more diverse and the order of mentioned attributes as well as the number of entities and attributes mentioned are closer to those captured in the human-human dialogues.

The dialogue data collected through the MutualFriends conversation task has two major drawbacks when considered for learning how model the conversation’s common ground: References to listed individuals are strongly primed by the textual input in the shown dataset. Referring expressions thus are highly optimised and will not be refined if they were to be re-used - losing the central tool available for tracking common ground. What is more, references to individual are only partial, mentioning properties that group individuals instead. The data thus does not actually contain object references or re-references in the form that we require for our analysis.

---

\(^{32}\) Bahdanau et al. (2014)
3.2 Dialogue Personalisation

The second area of research that is closely related to our interest in keeping track of the common ground between speakers is that of dialogue personalisation. Instead of using a conversation’s history to facilitate the formation of referring expressions, dialogue personalisation focuses on using (outside) information about an interlocutor in order to make dialogues more engaging. It therefore also builds on some form of model about a specific conversation partner that functions as an additional input to a language processing unit, priming or conditioning dialogue output. This field of research gained popularity after a survey by Serban et al. (2015), which found that the personalisation of dialogue systems is ‘an important task, which so far has not received much attention.’ At the moment, there are two main approaches to personalisation of dialogue: Either information about speakers is collected from the dialogue history, or relevant information is provided by an external knowledge base.

3.2.1 History-based Personalisation

Li et al. (2016) represent the first approach, extending the sequence-to-sequence neural conversational model of Vinyals and Le (2015) with an agent’s persona model in order to improve utterance coherence. They motivate this step by showing that while language models trained on large dialogue datasets can produce syntactically and semantically correct utterances, each output is in fact an independent consensus response from the training data. Responses to similar questions will therefore often lead to non-committal or inconsistent replies as they do not follow a consistent agent representation. In order to establish an agent’s persona model, they train two different SEQ2SEQ embeddings for interactions in human-human dialogues: A Speaker model which represents the personality of an agent with respect to information conveyed about it, and a Speaker-Addressee model that adapts the agent’s writing style to the personality of the interlocutor. In every step of an utterance generation process, these persona embeddings are fed as an additional input to the decoder state and condition the next word generations. Through this addition, Li et al. (2016) report a relative decrease in utterance perplexity by about 10% compared to a state-of-the-art regular LSTM baseline, and a relative increase in BLEU score of more than 20% when tested on a dataset of Twitter conversations.

Zhang et al. (2018) extended this approach by collecting a new dataset to explicitly investigate persona-based chat. To do so, they first task participants to create personas by writing at least five descriptive profile sentences about an imaginary character. They then pair participants randomly, assigning each of them one of the more than 1k collected personas and ask them to chat while playing that character. This way they collect more than 10k natural, persona-based dialogues with a total of more than 160k utterances. In order to train a conversational agent on the collected data, Zhang et al. propose a novel Generative Profile Memory Network architecture that encodes each of the profile entries as individual memory representations in a memory network. This model improves


37 Sutskever et al. (2014)

utterance perplexity by about 10% over a SEQ2SEQ baseline when encoding the speaker’s persona. When only encoding the interlocutor’s persona or both personas, this effect is less prevalent.

3.2.2 Knowledge Base Personalisation

Joshi et al. (2017)\(^{39}\) represent the second approach to dialogue personalisation, proposing to use external knowledge base information in order to improve dialogue quality. The authors argue that dialogue agents which are employed as human-machine interfaces should utilise information about the human interlocutors in order to optimise the interaction and seamlessly integrate into everyday routines. To demonstrate this, they focus on extending the task-oriented dialogue agents for restaurant reservations presented in the extended bAbI tasks.\(^{40}\) To improve the interaction quality in these tasks, Joshi et al. propose to establish a database of the users’ dietary preferences. This additional data then should always be included in the search for suitable restaurant reservation along with the default fields like type of cuisine and location to personalise the interaction. Using the Memory Network architecture introduced to the bAbI tasks by Weston (2016),\(^{41}\) they proposed a novel split memory architecture where two separate memory modules keep track of the dialogue state and KB information, respectively. With this architecture, they observed a performance gain of 10% in one out of three KB tasks, but also a decrease of more than 14% in one of the simpler tasks that do not require the external user data. The authors largely blame this decrement on the difficulty of training the split memory network and leave it for future work to further investigate this problem setting.

3.3 Context-aware Image Descriptions

Besides the research concerned with the overall development and training of data-driven dialogue agents for visually-grounded language tasks presented in Section 3.1, a first few papers have been published that - like us - focus on specific aspects of the visual dialogue problem domain. One of these aspects covers context-aware image descriptions, the approach of conditioning utterances about images or objects in them based on the feature sets of both the referent and the set of distractors.

Realising the importance of good referring expressions in the series of object identification tasks, Mao et al. (2016)\(^{42}\) for example propose a novel model for generating context-sensitive object descriptions. Compared to a baseline description generator that simply maximises \(p(S|R,I)\), the probability of a referring expression \(S\) given a referent object \(R\) in an image \(I\), they propose to train the generator through Maximum Mutual Information (MMI) training instead. With this approach, the generator maximises

\[
\text{MI}(S, R) = \log \frac{p(S|R)}{p(R)p(S)} = \log \frac{p(S|R)}{\sum_{R'} p(S|R')} \quad (3.1)
\]

to account for the probability that other objects in the scene \(R'\) could be identified with the proposed expression as well.


\(^{40}\) Bordes et al. (2016)


In order to train this model, Mao et al. present a new dataset of context-aware descriptions gathered through a slight modification of the ReferIt task: Like in ReferIt they collect single-sentence descriptions for objects highlighted in a photograph - but instead of using the ImageCLEF IAPR dataset they base their data collection only on those images from the MS COCO repository that contain between two and four instances of the same object type as the highlighted object. Through this method of selecting only images that contain referent objects together with a number of similar distractors, they require participants to use more specific referring expressions to disambiguate the referent objects than in the ReferIt task. As a result, sentence length increased from an average of 3.61 to 8.43 words. The resulting dataset labelled Google Refexp contains a total of over 100k referring expressions to 54k objects in 26k images.

To evaluate generator performance, Mao et al. used a Turing Test-like setting on Amazon Mechanical Turk: They displayed an image with a highlighted object together with two candidate descriptions - one of which was actually generated by the model - to human assessors, asking them to select the better one. As a result, assessors rated generated descriptions better than or equal to human descriptions in 20.4% of the cases, compared to 15.9% for the baseline model. Arguing that human evaluation however is not scalable, they also assessed downstream performance by evaluating the object recognition performance of their comprehension module, increasing precision from 0.77 to 0.81.

In a similar vein, Vedantam et al. (2017) introduce the task of discriminative image captioning, a task where image captions should not just be a literal coverage of the image’s content but follow a more pragmatic approach and highlight those elements that disambiguate a referent image from similar distractors. Like Mao et al., they introduce a context-aware model called the introspective speaker (IS) for this task, maximising

\[
\lambda \log p(S | C_t, I) + (1 - \lambda) \log \frac{p(S | C_t, I)}{p(S | C_d, I)}
\]

where \(C_t\) is the concept of the target image and \(C_d\) a distractor concept.

To evaluate the introspective speaker, Vedantam et al. use a two annotation forced choice (2AFC) setup, displaying a generated description together with its target and distractor image to human assessors who are asked to select the image that they believe is referred to by the shown expression. Using a set of simple distractor images, the expressions generated by the introspective speaker lead to a selection accuracy of 89.0% (baseline 74.6%), using a set of more semantically similar images selection accuracies were 74.1% for the IS model and 52.5% for the baseline, respectively.

Cohn-Gordon et al. (2018) recently extended this approach by using a character-level generative model instead of the word-level implementation of Vedantam et al., reducing the softmax over candidate emissions from over 20,000 to a under 30. Claiming it is more efficient, their version as of yet however underperforms the higher-level approaches.

---


3.4 Available Dialogue Datasets

As described in Chapter 2, our main goal is to extend the current generation of artificial dialogue agents for visually-grounded conversation tasks by explicitly adding the ability to use referring expressions when addressing recurring visual inputs. To this end, the model will need to keep track of the common ground with its interlocutor and use the current dialogue state to correctly determine how to (re-)refer to a displayed object.

When considering a supervised regime to train such a dialogue agent, we require a dataset of human-human dialogue which contains a sufficiently large number of object re-references that can unambiguously be linked to their respective referent objects. After having touched upon some of the current approaches to modelling (visual) dialogue, we will therefore now briefly analyse the available (visually-grounded) dialogue data-sets and determine whether they qualify for this end - or whether we need to collect a dedicated dataset ourselves, like many of the presented projects did. In this analysis, the datasets are grouped by the type of dialogue and context modality.

3.4.1 Free Chat Dialogue

One of the two main categories of dialogue types is free chat. In free chat, conversations are either recorded during regular everyday interactions or after speakers are asked to discuss a given topic. In their study of available dataset corpora, Serban et al. (2015) further divide free chat dialogues into spontaneous, written and scripted dialogues.

A seminal example of spontaneous free chat is the Switchboard corpus. To collect it, participants were asked to discuss a given topic while calling one another on the telephone. Other large-scale spontaneous dialogue corpora like the BNC corpus for which conversations from business meetings or radio shows we collected and subsequently transcribed and annotated.

In the written category, Serban et al. list text-chat dialogue corpora that were obtained from platforms like Twitter, Reddit or Ubuntu FAQ sites, as well as for example the CATAN and CARDS corpora that contain the chat output of participants playing online versions of the games Settlers of Catan and CardsWorld.

The scripted section lastly contains dialogue datasets extracted from scripted dialogues, like movie and TV scripts and subtitles.

Depending on the collection method, free chat dialogues can contain a wide range of topics or just cover a very specific domain. How a conversation is held also widely differs, depending on whether speakers are talking face-to-face, talk through the phone, use a text-chat application or actually reply to a thread of a message board over the course of multiple days. What combines them however is that dialogue arises naturally based on each specific dialogue modality and conversation type. This often unbounded flow of conversation is what makes these dialogues interesting for linguistic research, emphasising their indicative

---


Afantenos et al. (2012)

Djalali et al. (2012)
expressiveness of human-human conversations. On the other hand, this unbounded character of free chat dialogues is also seen as one of the principal reasons why free chat dialogue data is only seldom used to develop artificial dialogue agents: In most cases, dialogue agents are developed to fulfil a certain task as human-machine interface - which free chat dialogue training data can only insufficiently prepare them for. With respect to our focus on using referring expressions, in free chat dialogues it is difficult to ground the occurring referring expressions in a meaningful way as no set of referent objects is clearly defined. What is more, it is not very likely that free chat dialogues contain a large number of re-references, further decreasing its suitability for the task at hand. A better candidate for obtaining this kind of data is defining a conversation task to collect specifically tailored task-oriented dialogue data.

3.4.2 Task-Oriented Dialogue

All of the dialogue datasets collected and used for the data-driven dialogue agents presented in Section 3.1 are based on task-oriented dialogues. In task-oriented dialogues, participants are specifically tasked to solve a problem or reach some kind of goal that requires communication of some sort. One of the first approaches to collect task-oriented dialogue data on a larger scale is the seminal HCRC Map Task Corpus by Thompson et al. (1993) who recorded and transcribed conversations of two participants navigating with different maps. In 2016 Zarrieß et al. presented PentoRef, a transcribed and annotated collection of a range of visually-grounded tasks using geometric shapes, totalling more than 20k utterances.

Collecting data for a specific conversation task oftentimes is not as easy or as cheap as just recording or collecting free chat conversations. Participants have to come to the lab, be instructed, recorded and compensated for their participation. The range of available large-scale task-oriented dialogue datasets therefore is much smaller than that of free chat datasets - and the datasets themselves are smaller as well, with one of the largest containing little over 700 dialogues. This recently changed however when crowd-sourcing dialogue data online became an option to researchers. With Lasecki et al. (2013) being one of the first to explore this possibility (collecting a mere 16 chat dialogues), a number of large-scale dialogue datasets have been collected since by pairing participants online to complete conversation tasks.

To the best of our knowledge, the previously presented ReferIt, VQA, VisDial, GuessWhat?! and CoDraw datasets represent the complete set of visually-grounded dialogue datasets collected by crowd-sourcing a conversation task. Investigating these datasets with respect to their applicability to train a dialogue agent to use referring expressions, we however had to draw the conclusion that none of these conversation tasks requires the participants to re-refer to the elements of the visual input in the required manner. To summarise, through its task design, ReferIt ‘conversations’ for example contain only a single reference to
one highlighted object which should be descriptive enough for the other player to correctly disambiguate that referent object. VQA on the other hand contains independent question-answer pairs from speakers who never actually were paired in the first place. GuessWhat?! dialogues are longer interactions of a fixed speaker dyad, which principally allows for a common ground to be established and partner-specific referring expressions to form. The oracle here however is limited to its three-way answer selection while the guesser asks about different aspects of the visual input scene in order to find the highlighted object. The guesser therefore is not likely to often re-refer to the same objects in the photograph - and the limited reply selection of the oracle prevents collaboration altogether.

In the VisDial and CoDraw task finally, participants have a natural language conversation about the elements depicted in an image or clipart scene, rendering them to be viable candidates for the task of extracting referring expressions. Besides the fact that participants in both cases are only paired for the duration of discussing a single image - with the result that the number of object re-references will be low here, too, what makes these datasets not particularly suitable for training collaborative referring expression generation is that in both tasks only one of the participants sees the reference image while the other has to either form a mental image of it to identify it from a selection of distractors later, or to replicate it by placing the candidate objects on his or her empty canvas. Strictly speaking, the participants in these tasks therefore never actually refer to the same referent object.

To conclude, a number of approaches emerged to model artificial dialogue agents for visually-grounded conversation tasks While some of them already show reasonable performances on the specific tasks, the tasks themselves often are too simple to require or elicit natural multi-turn dialogues. Available large-scale dialogue datasets on the other hand either are too small or not suitable to train a more involved data-driven dialogue agent drawing on the interlocutor’s common ground. In order to progress in this direction, a novel dialogue task thus will have to be developed through which training a natural usage of referring expressions becomes possible.
4 The PhotoBook Dialogue Task

In the previous chapters we established that a conversation’s common ground plays a central role in the process of interlocutors adapting and refining their utterances to optimise mutual understanding and communication efficiency (Chapter 2). Current data-driven approaches to computational dialogue agents for visually-grounded conversation tasks however do not utilise a representation of the common ground yet, oftentimes resulting in inconsistent and unnatural dialogue outputs during multi-turn conversations (Chapter 3). We therefore proposed to investigate the implementation of a computational counterpart of common ground for artificial dialogue agents as a next step in improving their dialogue output quality. Because the collaborative formulation of referring expressions draws heavily on the established common ground, we proposed to start with an approach to learning the usage of those partner-specific referring expressions - which is not yet possible due to a lack of respective training data (Section 3.4). As a result, one of the main contributions of this project is the collection of a dedicated dataset.

In this chapter we present the PhotoBook Task, a visually-grounded conversation task that elicits free-form but goal-directed dialogue which naturally contains a large number of referring expressions to a controlled set of referent objects. The PhotoBook Task is implemented in Facebook’s ParlAI dialogue framework, enabling us to use it for experiments in a traditional lab setting as well as to crowd-source human-human dialogue data collection on Amazon Mechanical Turk (AMT). At the time of writing we collected 2,500 multi-round, written chat dialogues from more than 1,500 unique participants, containing over 160k utterances and roughly 60k referring expressions to a set of 360 reference images. We will present this PhotoBook dataset in Chapter 5.

Designing a conversation task to focus on the generation of referring expressions in a visually-grounded dialogue is far from trivial. As indicated in Chapter 2, a large number of factors are believed to influence the way we communicate. In order to control at least those factors that will potentially have the largest impact on the form and content of the dialogues resulting from the PhotoBook conversation task, we have to carefully draft the different aspects of the data collection experiment to correctly account for them - while keeping in mind that the application of the task is in a large-scale online setting. To show the sometimes opposing requirements of these two characteristics, Section 4.1 covers design decisions based on a comprehensive study of previous literature and related work on dialogue datasets, while Section 4.2 details the final
task design based in practical insights from a series of pilot runs. Section 4.3 of this chapter then details the most important aspects of the task implementation as well as an explanatory example experiment run.

4.1 Design Requirements

The foundation of the PhotoBook task design lies in the previously presented seminal experiment setups of Krauss and Weinheimer and Clark and Wilkes-Gibbs. Both tasks make it necessary that participants refer and re-refer to visually-grounded objects throughout the task, encouraging a natural development of partner-specific referring expressions. Other aspects however differ, like for example whether speakers have a certain, fixed role in the conversation, or how abstract the visual objects are. In addition, the intended large-scale crowd-sourcing setup of our dialogue collection task prevents us from implementing some of their design details altogether. The following subsections will therefore take an in-depth look at those conversation aspects that can be controlled by the task setup and through them define the specific setup of the PhotoBook conversation task.

4.1.1 Dialogue Modality

As a first basic aspect of the dialogue task design, we chose written chat messages to be the sole channel of communication for task participants. With this choice we select a different option than the seminal experiments on referring expressions which are based on transcriptions of verbal dialogue, but argue that written chat currently is the only reasonable choice for an automatic collection of a large-scale dialogue dataset.

While verbal communication arguably is the most natural form of human-human conversations, capturing it means that either participants need to come to the lab to complete the task - while being instructed and recorded by the researchers - or participants need to be able to record themselves when choosing for an online crowd-sourcing setting. Both alternatives however have their distinctive drawbacks: The lab setting simply is inhibitory expensive when scaled to a number of dialogues that is sufficiently large for training data-driven computational agents. Using voice chat in an online data collection setting on the other hand appears to be a viable option considering the quality of audio recording-and playback devices of modern laptop computers and smartphones. Lane et al. (2010) and Manuvinakurike and DeVault (2015) however found that during conversation tasks previously conducted on Mechanical Turk, in many cases the quality of recordings obtained from participants’ devices still was far from optimal. Low quality recordings in turn complicate both, the difficulty of the task for the participants as well as an automatic transcription of the dialogues. Manually transcribing the recorded conversations on the other hand in large part defeats the purpose of crowd-sourcing the data collection process in the first place.

Using written chat as communication medium directly returns the participant’s interactions in text form and therefore removes the need for


4 Gallardo et al. (2017); Chiu et al. (2017)
transcriptions altogether. With regard to the naturalness of written chat as dialogue modality, Meredith (2014)\(^5\) shows that the characteristics of communications through written chat lie in a spectrum between those of spoken conversations and traditional written text, with interlocutors utilising the affordances of the medium wherever advantageous to the interaction. Especially with respect to (self-)repairs that were found to be an integral part of spoken conversations,\(^6\) Meredith and Stokoe (2014)\(^7\) show that quasi-synchronous written chats exhibit characteristics similar to those of traditional spoken conversation, leading them to the conclusion that ‘online interaction should be treated as an adaptation of an oral speech-exchange system’. In addition, written chat emerged as the preferred way of communication in many online settings, which, based on the average AMT user demographics (see for example Difallah et al. (2018)),\(^8\) the task’s participants should be well accustomed to.

Considering all of the above, we believe that keeping spoken dialogue as communication medium will introduce more noise to the data collection process than we are willing to accept as a trade-off for a decrease in naturalness which appears to be limited. Like the other previously presented large-scale data collection experiments, we therefore decided to use written chat in the data collection experiment setup.

### 4.1.2 Number of Interlocutors

Regarding the number of dialogue participants, we decided on a simple speaker dyad. Involving more interlocutors in the data collection experiment would result in a more complex task design and dialogue output while not introducing any novel aspects of referring expression generation.\(^9\) Besides that, additional participants in a crowd-sourced online experiment setting are a factor of uncertainty that is best to be avoided with regard to the higher risk of one of the participants being disconnected from or abandoning a conversation. All of the previously mentioned large-scale dialogue data collections therefore use speaker dyads as well.

### 4.1.3 Task Symmetry

One aspect in which the seminal experiments of Krauss and Weinheimer and Clark and Wilkes-Gibbs differ is the symmetry of the conversation task: The Card Matching task is a symmetrical task while the Tangram Ordering task of Clark and Wilkes-Gibbs is asymmetrical. The difference between them is that in symmetric tasks no specific roles are assigned to the interlocutors - whereas in asymmetrical tasks there are. With exception of the MutualFriend task, all previously collected large-scale human-human dialogue data-sets are based on asymmetric data collection tasks. Arguing however that not restricting the dialogues through the assignment of specific speaker roles leads to more natural conversations, we decided to use a symmetrical task setting for the PhotoBook task instead.

One consequence of asymmetric task settings that can be observed in

---


\(^9\) See Appendix A.1 for more details.
the dialogue data resulting from the aforementioned large-scale data collection experiments is that the dialogue output is oftentimes heavily dominated by a certain speaker role and its pragmatics: In identification tasks for example, conversations consist mostly of object descriptions by the instructor. The other participant here only notifies the instructor when he or she believes to have identified the correct referent object - and incidentally asks a clarification question. He et al. (2017) therefore argue that these task settings do not lead to a dialogue output that sufficiently represents the rich dialogue states and nuanced, pragmatic dialogue acts that define natural human-human conversations. And as a consequence, they also do not contain the collaborative aspect of developing referring expressions that we aim to investigate either.

The original card matching task of Krauss and Weinheimer does not impose specific roles on the participants: It is a symmetric conversation task where both speakers are asked to match up their cards with those of their interlocutor by comparing the discriminating figures. The same holds for the MutualFriend task, which attempts to bridge the gap between goal-oriented and open-ended dialogue by designing a collaborative task-oriented dialogue setting where both interlocutors simply are asked to find the common entry in the given person register. Because these tasks do not impose a role on the different speakers, dialogue develops naturally based on the requirements of the task. In most cases, this means that both interlocutors contribute to the conversation in a roughly equal fashion.\(^{11}\)

4.1.4 Grounding of Dialogue

One thing that all of the presented goal-oriented dialogue data collection tasks have in common is that they require speakers to ground large parts of their dialogue in order to complete the task. By for example asking participants to determine marked or common objects, they will have to indicate and refer to certain objects in order to communicate about them with their interlocutor.

As mentioned in Chapter 2, for this research we will focus on visually-grounded dialogue, requiring a task setup that puts referent objects in the visual domain. Compare this for example with the MutualFriends task, where conversations are grounded in textual knowledge-base entries: When trying to identify the common entry, speakers here use the pre-defined characteristics indicated in the column headers to describe (groups of) displayed individuals. The textual input therefore strongly influences the resulting referring expressions and particularly affects those elements of the the generation process that are concerned with proposing and establishing an initial referring expression to introduce a novel object. Using visual referent objects relieves this effect because here referring expressions are not primed by any text input: The speaker introducing a referring expression to a novel referent object needs to propose an initial expression solely based on his or her interpretation of the common ground with his or her interlocutor. From this proposal they then collaboratively develop a canonical expression - which develops


\(^{11}\) He et al. (2017). Note however that this does not necessarily need to be the case as speakers can also choose to take a more active or passive role in the conversation as a part of the collaborative solution strategy: In the MutualFriends task, one player could for example describe all entries displayed in his or her list of friends with the other player simply noting the common one. Albeit not the most efficient, this would be a valid strategy.
to be as easily and uniquely identifiable as one derived from a textual primer. Both the traditional experiments by Krauss and Weinheimer and Clark and Wilkes-Gibbs therefore use visual grounding for their task setups - and for us it was an unequivocal choice as well.

4.1.5 Referent Types

As for the specific type of visual referent objects, there still is a wide range of options to consider. We mentioned in Chapter 2 that Krauss and Weinheimer noted that some objects already have fixed, *popular* referring expressions that speakers are likely to use when encountering them. Since we are interested in how partner-specific referring expressions are developed during a conversation, we need to use some type of referent object that evades such a popular expression instead. As a result, we decided to use sets of photographs depicting everyday scenes.\(^\text{12}\)

While these photographs mostly do not have a popular referring expression attached to them, a major disadvantage compared to other object types like geometrical shapes is that for images it is harder to control what exactly they display. Identifying a referent object in a photograph therefore might be trivial or difficult based on the similarity of distractors present in the picture.\(^\text{13}\) Displaying instead a set of images and asking participants to select a referent image from the set mitigates this effect because here the presence and similarity of distractor images can be largely controlled.\(^\text{14}\) Aiming for natural dialogues with developing referring expressions, this option therefore was the best choice available.

4.1.6 Re-Occurrence Rates of Referent Objects

In order to obtain suitable data for training computational models to learn a natural usage of referring expressions, we want the collected dialogues to contain as many and as complete as possible samples of referring expression generation processes. To this end, it is of utmost importance to set up the task in a way that it requires the participants to re-refer to a previously seen image sufficiently often.

As we have seen in Chapter 2, the experiments by Krauss and Weinheimer and Clark and Wilkes-Gibbs show that referring expressions for their abstract referent objects tend to converge after about five to six references to a given object. Completing a full set of six rounds of the Tangram experiment however took about 25 minutes, which would be quite long for an collaborative task crowd-sourced online. Considering this, we decided to aim for about five referring expressions per object so as not to lose valuable data but keep task completion times short as well.

4.2 Task Setup

Now that we covered and decided on the basic principles of the data collection experiment, we need to design the actual task in such a manner that it represents those principles. To summarise, the data collection experiment should be an *online, symmetric, two-party, goal-oriented, multi-*
As a result, in the PhotoBook task two participants are paired to complete a conversation task. Both of them see a randomly ordered grid of six similar images of everyday scenes. Some of the displayed images are the same for both participants (the common images) and some of them are displayed to one of the participants only (the different images). Three of the images in each display are highlighted through a yellow bar under the picture. The participants then are tasked to mark these highlighted images as either common or different by chatting with their partner. When they are confident that their selections are correct, they can continue to the next round. A full game consists of five rounds; each round of a game contains images from a specific set of similar images only - some of them recurring multiple times throughout a game.

Through a careful setup of the image sets and the selection of highlighted images, each referent image is highlighted exactly five times during a game. As a consequence, dialogues resulting from the PhotoBook conversation task also contain up to five references to each of the referent images and thus provide a valuable resource for investigating collaborative referring expression generation with respect to the conversation’s common ground.

### 4.2.1 The PhotoBook Image Sets

The images used in the PhotoBook task are taken from the MS COCO 2014 Trainset. Images in MS COCO were collected from the Flickr image repository, which contains labelled photos predominantly uploaded by amateur photographers. The pictures therefore oftentimes are snapshots of everyday situations, placing objects in their natural and frequently rich context (hence the name Common Objects in Context) instead of showing an iconic view of objects. In the MS COCO Trainset, these images are manually annotated with the outlines of the depicted objects as well as their object categories. We can use this information to select similar pictures that complicate the process of finding a disambiguating image description and as a result lead to more interesting dialogues.

With the final setup of the PhotoBook task, a game requires exactly 12 unique but similar images. In order to obtain these image sets, we grouped together all images that passed a custom similarity criterion:

1. The two objects that cover the largest area of an image need to be the same for all images in a set, and
2. the area covered by each of those main objects needs to be at least 30,000 pixels.

Additionally, we require that

1. all images need to be in landscape format
2. all images need to be a colour photographs, and

---

15 Lin et al. (2014)

16 https://www.flickr.com/

17 Compare for example the otherwise highly popular ImageNet dataset by Deng et al. (2009)
With this grouping criterion, we obtained 36 sets of at least 20 images, each of which is annotated with the two main object categories. After inspection, we decided to keep a subset of 30 of those sets because photographs in the other six sets were so similar that producing a disambiguating description would likely be too difficult of a task for participants of our data collection experiment. For the remaining 30 sets of images we manually checked the quality of the pictures, removed any blurry or otherwise odd pictures and then randomly selected 12 of the remaining ones for the final PhotoBook Image sets. The complete dataset thus contains 360 unique images.

The most prominent object category in our image sets is person, which is one of the two main categories in 19 out of the 30 sets. In total there are 26 unique object categories in the image sets (see Figure 4.1 for an overview of how often an object category appears as main category in our dataset and Appendix A.5 for a full list of categories). Investigating the similarity of the photographs per set, we embedded images through their hidden state encoding in the pen-ultimate layer of a pre-trained Deep Residual ResNet-152 Network and calculated pairwise cosine similarity for all in-set and per-set combinations. As can be seen in Figure 4.2, the average in-class similarities range from 0.68 to 0.83 with a mean of 0.75 and each are higher than their respective per-set similarity with other sets (range 0.51-0.80, mean = 0.6).

### 4.2.2 Specification of Games

As a next step, we developed a simple function to select which images of a set should be shown to which participant in which round of a task. In this function, all 12 images in a set of similar photographs are randomly indexed and assigned to a participant’s display in a specific round based on the schema displayed in Table 4.1. With this schema, each photograph is displayed exactly five times while the order of images and the order of rounds can be randomised to prevent participants from detecting patterns in the display. Each of these sets then is duplicated and assigned a different selection of highlighted images to obtain the 60 base image sets of the PhotoBook task. While most highlighted images recur five times during a game, they can be highlighted for both participants in the same round. As a result, any given image is highlighted in an average of 3.42 rounds of a game (see Table 4.2 for the highlighting schema).

### 4.3 Implementation

The PhotoBook conversation task itself is implemented in Facebook’s ParlAI framework. ParlAI is designed to be a unifying framework for the development of dialog models, building towards a single general model for computational dialogue agents. According to their website, currently already more than 20 dialogue tasks and dialogue agents are implemented in ParlAI. One of the central features of the ParlAI framework is that it is designed to be a unifying framework for the development of dialog models, building towards a single general model for computational dialogue agents.

Table 4.1: Assignment of image IDs to the different participants and rounds of a game schema. The order of rounds and the arrangement of images on the participant’s display can be randomised without effect to the game setup.

See Appendix A.4 for a detailed background of the task setup.

Miller et al. (2017)
framework is its modular architecture that allows interfacing different components. Participants of a dialogue task - in ParlAI called agents - for example can be chosen to be a local user on a terminal window, a previously developed computational dialogue model or a user of crowdsourcing platform Amazon Mechanical Turk. We use this last option for the collection of the PhotoBook dataset.

In order to control the quality of collected dialogues, we require AMT users called workers to be native English speakers located in the US, Canada, Great Britain, Australia or New Zealand. Following Kazemzadeh et al. (2014)\(^{22}\) we also require workers to have completed at least 100 other tasks on AMT with a minimum acceptance rate of 90\%.\(^{23}\) Workers are paired through an algorithm based on whether or not they have completed the PhotoBook task before and which of the different games they have played. In order to prevent biased data, workers can complete a maximum of five games.\(^{24}\)

4.3.1 Task Instructions

To make the task more accessible to human participants, we followed

\(^{22}\) Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. (2014). Referit game: Referring to objects in photographs of natural scenes. In EMNLP

\(^{23}\) See Appendix A.9 for details.

\(^{24}\) See Appendix A.10 for a detailed overview of the pairing algorithm.
Table 4.2: Schema of referent image highlighting in the PhotoBook task. The left part of the table indicates whether a given image is highlighted for one of the two participants (A and B) in a given game round in either game 1 or 2. T indicates the total count of highlights (which is 5 always), H counts the highlights per game and R the number of rounds that an image is highlighted in.

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>T</th>
<th>1</th>
<th>2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>T</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
<td>5</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

the approach popularised by von Ahn et al. (2006)\textsuperscript{25} and cast the task formulation as a game. The participants are instructed that both of them will get to see a digital photo book that has five pages with each six images on them. The photo books they get will be different, but some of the photographs on each page will be the same. Their task is to determine which of the highlighted photographs on each page are also displayed on their partner’s page by chatting with them.\textsuperscript{26} In order to facilitate the subsequent processing of the resulting dialogue data, in the instructions we also ask participants to specifically

- Try to find the common and different photos as quickly as possible
- use correct and grammatical English and do not use abbreviations or chat language
- only describe a single photo per message, and
- click on the common or different label as soon as they identify it

4.3.2 Warming-Up Round

During an initial series of pilot runs we observed that for new participants the first round of their first game took significantly longer than any other ones. Although we do expect that participants get more efficient over time, we argue that this effect is largely related to the fact that participants need to get familiar with the task’s mechanics when it is the first time they are exposed to it. In order to control for this effect, we therefore added a warming-up round with three images per participant (two of them highlighted) for each pair of new participants (see Image 4.3). This strongly reduced the completion time of new participants’ first game rounds.


\textsuperscript{26} See Appendix A.8 for details about the instructions.
4.3.3 Game Round Mechanics

The PhotoBook task AMT user interface is designed in such a way that the six images per round are displayed in a $2 \times 3$ grid, with a coloured bar under each image: If the image is highlighted, this bar is yellow and contains a radio button option for the common and different labels (see Figure 4.4(a)). If they are not highlighted for a player, the bar is greyed out and empty. The submit button is deactivated as long as not all highlighted images have been labelled. As soon as both players submitted their selection, a feedback page is shown where the bars under the highlighted images either colour green to indicate a correct selection or red to indicate a wrong one (see Figure 4.4(b)). The radio buttons are disabled in the feedback screens so players cannot revise their selection - they can however communicate about their mistakes or pass any other feedback to their partner.

The title of a page indicates the current page number so participants can always check their progress; the text input field is limited to a maximum of hundred characters to prevent listings of multiple images or overly elaborate descriptions - which if necessary can be conveyed in a number of subsequent messages.

4.3.4 Feedback Questionnaire

In order to facilitate a qualitative analysis of dialogue agents developed for the PhotoBook task, we also collect a gold-standard benchmark of participant’s self-reported satisfaction scores. These scores later can be
compared with those obtained by pairing human participants with an artificial dialogue agent in order to assess it in a Turing Test-like setting. Following He et al. (2017), we ask participants to rate three statements on a five-point Likert scale, ranging from strongly agree to strongly disagree:

1. Overall collaboration with my partner worked well.
2. I understood the descriptions of my partner well.
3. I understood the descriptions of my partner well.

Besides that, participants are given the opportunity to provide optional free-text feedback of any kind.

Through the careful setup of the PhotoBook task, with design choices being grounded in both a comprehensive study of underlying linguistic theories and insights from extensive pilot testing, we believe to have formulated a visually-grounded dialogue task that allows for the collection of natural human-human dialogues relevant for studying the development of the speakers’ common ground. When completing the task, participants will use a large amount of re-references to previously encountered images. This facilitates tracking the development of canonical referring expressions during a conversation and proxy the wider concept of common ground.

In the next chapter we will validate these claims through an extensive analysis of the dialogue data obtained from crowd-sourcing the PhotoBook task online. Chapter 6 will then proceed by proposing the setup of a data-driven computational agent to learn the correct usage of referring expressions based on the PhotoBook data.

---

(a) Example screenshot of the PhotoBook AMT game interface for one of the two participants.

(b) Example screenshot of the PhotoBook AMT feedback display.

---

Figure 4.4: Example screenshots for a participant’s display during a game round and feedback screen.


Through crowd-sourcing the PhotoBook conversation task with the ParlAI framework on Amazon Mechanical Turk, we collected human-human dialogues for a total of 2,506 completed games, stemming from 1,514 unique workers. The resulting data contains a total of 164,615 utterances, 130,322 actions and spans a vocabulary of 11,805 unique tokens.

In this chapter we will analyse the PhotoBook dataset statistics with respect to overall characteristics related to the participant’s task performance (Section 5.1), as well as its more linguistically motivated properties like the participants’ language usage (Section 5.2). Through these analyses we show that the collected data indeed contains samples of collaborative reference generation processes that were never before collected on this scale and therefore validate the PhotoBook conversation task setup detailed in Chapter 4. Besides a validation of the data collection process itself, we also hope to show dataset features that will inspire future research into developing data-driven dialogue agents capable of using referring expressions, for which we will give a number of footholds in Chapter 6.

5.1 Game Statistics

Starting with the dataset characteristics that mostly relate to task performance, the following sections will each cover an in-depth analysis of different aspects of the collected data, assembling a thorough overview of the PhotoBook dataset.

Having collected 2,506 complete games based on our set of 60 unique game sets as described in Section 4.2.1, each game set was played between 15 and 72 times, with an average of 41 games per set (see Figure 5.1 for a detailed distribution).

![Figure 5.1: Coverage of the 60 unique game sets in the PhotoBook conversation task.](image-url)
The PhotoBook task was completed by pairs of a total of 1514 unique workers, of which 472 only completed a single game, 448 completed between two and four games and 594 the maximum of five games (see Figure 5.2 for a full overview). Registering participants by their unique Mechanical Turk Worker ID allows us to track a specific participant not only during a single game but also over the course of his or her up to five games. With this metadata available to us, we can use the collected dialogues to investigate both of the main trends previously observed by the seminal works on partner-specific referring expression generation: 1) the optimisation of communication efficiency during the five rounds of a single game - which mainly depends on the collaboration of the paired participants, and 2) the optimisation of a given participant’s communication efficiency over the course of consecutive games. In what follows, most of the metrics will thus be investigated with respect to these two paradigms.

5.1.1 Completion Times

Completing a full five-round game took an average of 14.2 minutes, with times ranging from 4 minutes to some scattered outliers of up to 53 minutes (see Figure 5.3). For these calculations, the first message sent by any of the participants is considered the beginning of the conversation, while the moment that the last player clicks the submit button in the fifth round is considered its end. After that, participants were shown the feedback of round five and asked to fill in the questionnaire. Overall participation times therefore are slightly longer than the game times reported here.
ing a task, but with non-significant decreases after the second game it mostly hints at the possibility that first games take a special role in this optimisation process.

Considering that the data contains a large number of participants who only completed a single game - which might indicate that they did not like the task, had a bad partner or encountered difficulties during this first game, we also investigated filtering the data to only consist of those participants that continued after completing their first game, i.e. who completed a minimum of two games. In that case, first games take about 20 seconds less (difference is non-significant) while the subsequent game times naturally remain the same. This means that the reduction in game completion time between a participant’s first and second game also drops by about 20 seconds, but remains a significant decline (p-value 9.19e-15).

As a third variant, we then also filtered out all games of participants who did not complete a full series of five games and evaluated the remaining
five sets of 292 games each. This filtering further reduces the game times for participants’ second to fourth games (see Figure 5.5 and Table 3), with the effect that now there is a slight increase in game duration (although non-significant either) between the fourth and fifth game. The fact that even in the specific data of only those participants who completed a full set of five games we can observe a significant decrease in game time between their first and second games further supports the observation of individual participant’s optimising their task solving efficiency through multiple trials. The slight increase in game duration between the last two games on the other hand is a good indication that the limit of five games per participant is a reasonable constraint for our data collection experiment, as participants appear to have converged on a certain game strategy by then.

Investigating the completion times of individual game rounds, we can observe the same effect as reported by Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986), with durations being significantly reduced each round (see Figure 5.6 and Table 1 for more details).

These significant decreases in round duration strongly support the collaborative model of referring expression generation. Completing a game round means that the participants have grounded the referring expressions for the highlighted images in order to establish whether they are shown to both or just one of them (common or different). If they manage to do so correctly in less and less time in each subsequent round of the game, it appears that they are developing a more efficient way to do so. In order to validate that this intra-game optimisation is a general trend in the dialogues collected through the PhotoBook task, we filtered out games of participants who only concluded a single game and participants who did not complete all five games here as well. In the former case, round durations still decrease for each subsequent game round, but the difference between the completion times of round four and five
is no longer significant (p-value 0.1). If only considering games of participants who completed all five games on the other hand, completion times increase by about half a second between round four (p-value 0.9).

After manual inspection of the data it became clear that this increase is very likely due to thank-you messages, compliments on good collaboration and greetings that participants sent before what they might have thought was ending the HIT with the last submission. In order to account for this, we also considered the last selection of a round to mark its end, because after it no game-deciding actions for that particular round are taken anymore. Note however that this filters out any discussion and potential repairing of referring expressions that takes place during the feedback phase. As a result, we again see a decrease in completion times between round four and five, even though non-significant (p-value 0.9, see also Figure 5.7).

That round completion times are converging after four game rounds of the PhotoBook task for those participants who were dedicated to the task and completed all five games validates our game setup with its specific reference schema. Arguing however that we can observe a very similar overall characteristics in the recorded dialogues, during the further analysis in this chapter we will use the full dataset if not indicated otherwise.

5.1.2 Game Scores

With three highlighted images per player per round, during a full game of five rounds 30 labelling decisions have to be made. On average, participants correctly label 28.62 out of these 30. 39% of all games reached full scores, and only 1.2% of all completed games did not reach our pre-set minimum score of 24 points (see Figure 5.8). What is remarkable is that a higher percentage of games was completed with 28 points than with 29 points. This might be related to the fact that miss-identification of common highlighted images causes a loss of two
points, but requires further analysis to exactly pinpoint a cause as Figure 5.9 shows that the distribution of scores per game round decreases with an increasing number of errors - and therefore does not support this two-error hypothesis.

Figure 5.10 depicts the average scores obtained by pairs of participants for the five subsequent game rounds of the PhotoBook task. Scores increase with each round, but only significantly so between the first and second round (p-value 0.004). This, too, points at the collaborative referring expression generation process: When a new set of players is paired, they will need to establish and shape their common ground, adjusting to the new pairing - which can cause miss-identifications of images if their initial descriptions and references are not properly grounded yet. Figure 5.11 shows that more than three quarter of all errors are due to common images being labelled as different while only a quarter of the errors stems from different images being labelled as common. Miss-identifying a common image as different indicates a non-successful grounding of the used referring expression, labelling a different image as common on the other hand suggests that the interlocutor had a similar image on his or her page, which he or she falsely identified as referent object for the reference used by the speaker.

Investigating the game scores obtained in subsequent games (Figure 5.12(a)), we can observe a slight drop in game score between the first and second game, followed by a continuous increase in score until they surpass the first game’s score in the fourth one. All of these relative changes in game score however are below significance level when compared to the respective previous game score (see Table 5). Filtering all games of participants that did not complete the full set of five games here results in a convex graph, with the third game scoring lowest (Figure 5.12(b), Table 6). Here again however all relative changes with respect to the previous game’s scores are non-significant.
5.1.3 Number of Messages

One of the main observations of the seminal experiments of Krauss and Weinheimer and Clark and Wilkes-Gibbs was that the average number of utterances needed to identify a given referent object decreased during the first few iterations of the task and then slowly converged after five re-occurrences of a given referent object. With the PhotoBook dataset, we can investigate this phenomenon using data that is two orders of magnitude larger than the data used for the original experiments.

Because in the PhotoBook conversation task the dialogue modality changed from direct, spoken interactions to a written chat interface, we will consider messages sent by participants as their utterances. This means that we neither split sentences contained within a message into multiple utterances, nor do we concatenate any subsequent messages to a single utterance. We argue that this best represents the actual characteristics of an utterance which formally is defined as *The string of sounds or written symbols produced by a speaker between two pauses. An utterance can consist of a single word or several sentences.*¹ Figure 5.13 shows the average and median number of messages sent per game round for the five succeeding rounds of a game. Considering that participants have to label and therefore refer to six images per round, the average number of messages per referent object respectively are 2.46, 2.21, 2.09, 2.03 and 2.14. Note however that in some cases referent images are shown to both participants (i.e. they are common in that round) and therefore rounds include a maximum of six object references, instead of exactly six. The reported number of messages per referent ratios therefore are artificially low and cannot directly be compared to those of for example Clark and Wilkes-Gibbs (1986), where each of the twelve referent figures (except the last one) needed to be referred to in each round.

While we as expected observe a significant decrease in message counts from the first to the fourth round (see Table 7), in the dialogues resulting from the PhotoBook conversation task the last round of interaction shows


---

Figure 5.13: Average and median number of messages sent per game round in the PhotoBook conversation task dataset.
a significant increase in messages sent. As noted before, especially this
fifth game round however exhibits a large amount of non-game related
or meta-level messages. In order to control for this characteristic, we here
also filter all messages sent after each round’s last image selection. As a
result, the average number of messages sent per round now significantly
decreases with each round (decrease from round four to five with p-value
0.04, see Table 8 and Figure 5.14). If not noted otherwise, for all further
analysis we will retain this filtering of messages.

![Figure 5.14: Average and median number of messages sent per game round after filtering out all messages sent after the last selection in each round.](image)

(a) Average number of messages sent (total for both participants) per game round. (b) Median number of messages sent (both participants) per game round.

5.1.4 Number of Tokens

Next to the observation that participants use less utterances per trial,
both original papers on the effect of recurring references to the same
referent objects also observed a shortening in utterances - or more specif-
ically in the referring expressions conveyed in the utterances. As it
requires no further linguistic annotation of the data, we will here start
with an investigation of the development of the raw token counts per
message. Word tokens were extracted with Python’s Natural Language
Toolkit (NLTK) `word_tokenize`. Figure 5.15 shows the total number
of tokens used by pairs of participants during the five game rounds of
the PhotoBook task, each decreasing significantly with respect to the
previous round (see Table 9).

In the previous section we however established that the number of
messages also significantly reduces each round, so in order to evaluate
the average number of tokens per utterance, we need to normalise for
this effect. Figure 5.16 now shows an almost linear decline in average
token/message ratios through the subsequent game rounds (Pearson
Correlation Coefficient of -0.36, with p-value ≪ 0.05). With a decrease of
an average of 0.70, 0.74, 0.64 and 0.51 tokens per message respectively,
all of these average ratios are significantly different than the preceding
ones (see Table 10 for details). Five game rounds of the PhotoBook
conversation task therefore appear to not be enough for participants to
converge on a constant token/message ratio.
Having observed that the number of messages and tokens decreases during a game while per-round scores slightly increase, we now check the resulting points per token (Figure 5.17) and points per message (Figure 5.18) ratios. While the former shows an almost linear increase throughout the five rounds of a game (Pearson Correlation Coefficient of 0.36 with p-value $\ll 0.05$), the increase in points per message ratio slowly decreases during a game. All of the relative increases in both statistics are statistically significant (see Tables 15 and 16).

Because spending less utterances and words on the referring process does not lead to a reduction or stagnation in task performance, but rather to an increase in the relative score ratio instead, from these results we can conclude that the grounding of developing canonical expressions indeed becomes stronger through the course of a conversation, leading to a more secure identification of referent objects given even less elaborate expressions.
5.1.6 Messages and Tokens in Subsequent Games

Like comparing game scores over the course of the up to five games completed by a participant, we can also investigate the differences in the total number of messages sent and tokens uttered per game. For this analysis, we will only consider games of those 594 participants that completed a full set of five games, while noting that in this case the differences with the unfiltered set of games and that filtered by only removing games of players who completed only one game are not significant. Figure 5.19 shows a decrease in total message count per game for all five consecutive games, while none of the relative differences however is statistically significant (see Table 17).

For the total token counts per messages we observe a similar but stronger effect, with a statistically significant average decrease of more than 70 tokens between first and second games (p-value 3.67E-13, see Figure 5.20 and Table 18). This indicates that even with the warming-up
round before the first game that is meant to reduce the first-time effect of working on a novel task, participants apparently still also use the first game round in order to get accommodated to the task and converge on a certain strategy which then is optimised during subsequent games.

5.1.7 Order of Selections

Next, we will investigate whether participants have preferences for the order in which to refer to the displayed images. We will do so considering two factors: 1) Whether an image was selected and therefore referred to in a previous round and 2) How similar an image is with respect to the other images displayed to a participant.

Figure 5.21 shows the distribution of games where the $n^{th}$ selection per game round was either an image that was selected in a previous round or a selection of an image that was selected for the first time during that round. While an in-depth comparison of these statistics per round is a non-trivial endeavour because the percentage of previously selected pictures increases throughout a game (given the limited set of candidate images in a game), the graph clearly shows that within a
single game round participants tend to first select (and therefore refer to) images that were selected in a previous round already, and only later turn to those images that need to be selected for the first time. This preference holds throughout all later rounds of the game.

In order to assess the similarity of a selected image with respect to the other images on the display, we again extracted image feature vectors from the pen-ultimate layer of ResNet-152 and calculated the pairwise cosine similarities between a selected image and all others on the player’s display with them. We then ranked each player’s selection by their mean similarity with the other images and averaged those ranks for the three selections per player per round. Figure 5.22 shows the results of this process, which indicate that for the first image description in each round the participants appear to favour that image which is least similar to the distractor set. After that there however appears to be a slight preference for the most similar one.

Figure 5.21: Order in which images are marked during the later rounds of the game: For each nth selection per round, in what part of the recorded dialogues was the selected image selected in a previous round (dark) and in what part is that image selected for the first time in the game (light).

Figure 5.22: Average similarity rankings for the three selections per player per game round. Dark blue indicates that the selected image was the least similar of the set of highlighted pictures, red indicates the most similar highlighted image.
5.1.8 Evaluation Scores

Lastly, we analyse the outcome of the questionnaire shown to the participants of the PhotoBook Dialogue collection task before submitting the HIT. The statement Overall collaboration with my partner worked well was scored with an average of 4.50 out of 5, I understood the descriptions of my partner well scored 4.49 and My partner seemed to understand me well was rated with an average of 4.55 (see Figure 5.23 for the detailed distributions). While all ratings are quite similar, participants seem to rate their partner’s understanding of proposed referring expressions higher than their own understanding of their partner’s expression (the ratings for statement My partner seemed to understand me well are significantly different from the previous two, with p-values 0.0001 and 3.09E-05, respectively).

![Participant’s ratings of the statement](a). Overall collaboration with my partner worked well.

![Participant’s ratings of the statement](b). I understood the descriptions of my partner well.

![Participant’s ratings of the statement](c). My partner seemed to understand me well.

Figure 5.23: Participant’s ratings on the three statements of the evaluation page.

5.2 Linguistic Analysis

Besides the overall dataset statistics, we now want to investigate its more linguistics related characteristics. To do so, in the following sections we will look in more detail at the participant’s vocabularies, their usage of word types over time, speaker contribution and a first study of dialogue acts in the collected data. We conclude this chapter with an inspection of some example dialogues and re-occurring phenomena.

5.2.1 Vocabulary

After manually correcting the 200 most frequent typos and filtering all words collected in the Banned Word List, the PhotoBook dataset contains a total of over 1M tokens with close to 12k unique tokens. When ordering word tokens based on their occurrence frequencies, we observe a strong decline in token counts, resembling Zipf’s Law (see Figure 5.24 (a) and (b)). The 25 most frequent tokens cover 50% of the entire data, the 2000 most frequent tokens 98% (see Figure 5.24 (c) and (d)).

In order to proxy the notion of referring expression in a more representative manner than the previously reported token statistics, we filtered out all word tokens registered in NLTK’s stopword list to obtain the content

\(^3\) http://www.bannedwordlist.com/

words only. Figure 5.25 shows that even after this filtering, the word
token statistics display a similar distribution to that of the overall token
counts, significantly decreasing between each consecutive game round
(see also Table 11). Further investigating a potential change in referring
expressions throughout a game, we normalised the obtained content
word counts by the total token count to show the relative ratio of content
words. This results in an almost linear increase of content token over
total token ratio throughout a game (Pearson Correlation Coefficient of
0.21 with p-value ≪ 0.05, see Figure 5.26 and Table 12). With referring
expressions getting shorter, content words thus appear to be favoured to
remain.

A similar effect can be observed in the rate of content words over total
token counts that pairs of participants used over the course of their five
game completions: Figure 5.27 shows a significant increase between the
first and second game (p-value 3.73E-13), which then starts to converge
at an average of just under 0.56 content words per token. Participants
therefore appear to not only collaboratively refine referring expressions
during a game by leaving out information that became irrelevant due
to the common ground, they also adapt their game strategy by using an
Figure 5.26: Average and median content word/total token ratio per game round after filtering out all messages sent after the last selection in each round.

Figure 5.27: Average and median content word over total token count ratios for all participants who completed a full set of five games.

Lastly, we also investigated the average amount of novel tokens introduced to the conversation during each of the five game rounds. Figure 5.28 shows the results of this analysis, with the decrease of novel tokens following roughly a negative logarithmic function (see Table 14 for details). This result also points at the previously reported trend that references tend to converge on a canonical expression during a conversation: Once an initial proposal is accepted, it is refined by removing details instead of adding novel information.

5.2.2 POS-Tags

To further analyse what kind of content words are retained during the refinement of initial referring expressions, we used the NLTK POS-Tagger to annotate utterances with one of 12 part-of-speech tags from the Universal tag set.\(^4\) Figure 5.29 (a) displays the distribution of the

\(^4\) http://universaldependencies.org/u/pos/.

For an overview of the POS-Tags, see Appendix B.1.
Figure 5.28: Average and median number of new word tokens introduced to the conversation in each of the five game rounds. Data filtered by messages sent after the last selection in each round.

(a) Average number of novel word tokens introduced during the five game rounds (both players).

(b) Median number of novel word tokens introduced during the five game rounds (both players).

The top 10 POS-Tags per game round of the PhotoBook conversation task. As expected, absolute counts decrease during a game as messages are shortening. When analysing the amount of their respective decrease between the first and last game round (Figure 5.29 (b)), we observe that the usage of pronouns is reduced most, with almost 70% less pronouns used in round five than in round one. The usage of verbs and determiners both are reduced by about 60%, followed by numerals and adpositions. The decrease in the usage of nouns is the smallest, with a reduction of a little over 40%.

Figure 5.29: POS-Tag distributions during the subsequent rounds of the PhotoBook Dialogue Task. Subfigure (a) and (b) show POS-Tag frequency counts and their decrease between the first and the last game. Subfigure (c) and (d) display the relative changes.

In order to account for the decreasing number of tokens per game round, we then also looked at the relative changes in word type usage by dividing token counts by the respective sum of tokens per round. The resulting Figure 5.29 (c) is less indicative than the absolute one, but reveals that the relative frequency of nouns increases while verbs and adjectives are used less during later game rounds. In order to make this analysis more specific, we here investigate the amount of change between the first and last game rounds as well (Figure 5.29 (d)). Displaying the decrease of usage, this graph indicates that the relative decrease in
pronouns is most pronounced, with a relative decrease of more than 30\%.
Determiners (20\%), verbs (17\%), adpositions (8\%) and numerals (2\%) are also used less frequently. All other word classes show a relative increase in usage, starting with adjectives and punctuation marks (10\%), followed by adverbs (12\%), conjunctions (13\%) and finally nouns (20\%). We can thus conclude that when interlocutors refine their referring expressions during a game, nouns are likely to remain in the expressions while particularly pronouns, verbs and determiners are dropped. These word classes however also may play a special role here as an inspection of the data revealed that during the beginning of a game, participants often tend to use expressions starting with ‘I have the …’ or ‘Do you have …’ which are omitted later. We will further investigate this in the following Section 5.2.5 about dialogue acts.

Besides analysing the change of used word classes during a game, we also investigated any potential trends in word class usage over the course of a participant’s five games. For this analysis we again first filtered out all games of participants who did not complete the full set of five games. As a result, Figure 5.30 (a) shows the absolute occurrence counts of the top-10 most frequent POS-Tags during a participant’s five games and Figure 5.30 (b) the decrease in frequency counts between the first and fifth game. As with the in-game decreases, the frequency of nouns stays most similar with a decrease of 13\%. The follow-up however changes for the inter-game statistics: Adpositions and conjunctions are used only about 17\% less in the last game, followed by adjectives and punctuation marks. Most decreases are registered for adverbs and pronouns (31\%) and numerals (28\%).

Relative changes in word class usage over the course of a participant’s five games are difficult to infer from Figure 5.30 (c), but the individual decreases between first and last game are more enlightening (Figure 5.30 (d)): Like in the in-game statistics, usage of nouns increases relative to the total word class occurrences, in this case by close to 10\%. The most drastic differences with respect to the in-game statistics on the other hand are adpositions and adverbs. The relative usage of adpositions decreases during a game but increases by an average 5\% over the course of a participant’s five games. An even stronger opposite effect can be
observed for adverbs, whose usage increases by more than 10% during a game but decreases by 13% over the course of a participant’s five games - being the strongest decline recorded.

As a conclusion, participants appear to adjust their individual task solving strategy by using a relatively larger proportion of nouns and conjunctions during later games than they did in the first, while using far less adverbs, pronouns, numerals, determiners or verbs. This gives a first indication of how referring expressions are likely to develop as participants accommodate to the task.

5.2.3 Determiners

Having grouped word tokens by their respective word classes, we can also perform an in-depth analysis of the distribution of specific surface words per word class. We will leave most of these detailed analyses to future work and here only briefly touch upon one of the word classes, the determiners. Krauss and Weinheimer as well as Clark and Wilkes-Gibbs observed that in their tasks participants always described an object during the first trial while referring to it in the following ones. One simple way to visualise this trend in a rough, high-level manner is looking at determiners: Descriptions of objects that a participant cannot be sure of to also be visible to his or her partner are bound to use an indefinite determiner. When later referring to a grounded referent object however, the determiner should indicate that exactly that referent object is addressed and therefore be definite. Figure 5.31 (a) shows the frequency counts of the top ten most frequent determiners in the five game rounds of the PhotoBook task. Indefinite determiner ‘a’ clearly is the most frequent in the first round, covering over 50% of all determiner counts in that round. The two most frequent definite determiners, ‘the’ and ‘that’, together cover 25% of the distribution. ‘That’ mostly is used in replies stating ‘I also have that.’ or ‘I don’t have that.’ to indicate common or different pictures. ‘No’ eludes a precise analysis as it is used in utterances like ‘no firetrucks here.’, where it functions as an indefinite determiner, but also in utterances like ‘no umbrella lady this time.’ as a reply to ‘the umbrella lady again’, which is a referring expression and ‘no’ thus replaces a definite determiner.

When analysing the changes in occurrence counts between the first
and last game round (Figure 5.31 (b)), we observe strong decreases of over 80% for most determiners except ‘that’ and ‘no’, whose usage still decreases, and an increase in absolute counts for definite determiners ‘the’ (13%) and ‘this’ (30 %). The trade-off between definite and indefinite determiners becomes even more apparent in the comparison of relative distributions (Figure 5.31 (c)): The relative usage of ‘a’ drops from over 50% in round one to under 20% in round five, ‘the’, ‘that’ and ‘this’ on the other hand climb from 25% to over 50%. The relative increase in usage for this ‘the’ between the first and last round is close to 200%, ‘this’ even increases with close 250% (see Figure 5.31 (d) for details).

From this rough analysis we can therefore expect to see a similar change from descriptions to references in the data collected.

As a next step we extracted all token tri-grams succeeding a determiner in the recorded utterances. The distribution of subsequent words is large, exhibiting a frequency decrease per rank roughly following Zipf’s law as well (see Figure 5.32).

Figure 5.33 displays the ten most frequent succeeding words together with their most frequent follow-ups. For this diagram we can infer that participants overall particularly tend to mention persons (‘man’, ‘guy’, ‘woman’, ‘girl’), which is not surprising given the previously mentioned prevalence of the person object category as one of the main objects in the image sets (see Section 4.2.1). Colours also appear to be a frequent choice in addressing a referent picture. In order to further investigate high-level differences in descriptions and references, we split the n-grams into two groups, one succeeding indefinite determiners (Figure 5.34 (a)) and those following a definite determiner (Figure 5.34 (b)). In the former case the overall distribution is very similar, we again see the a prevalence of person and colour mentions. In the latter case we however can observe some differences to the other two distributions with tri-grams starting with ‘one’ and ‘same’ entering the top 10. Both of them indicate references to previously seen images (i.e. ‘the one with the big teddy bear’ or ‘the same red bus as last round’).

For a more structural analysis of referring expressions captured in the PhotoBook dialogue data, we will need to annotate the data first. Chapter
6 will cover a first attempt to do so automatically, expert annotations however are the preferred choice and therefore should be considered for future work.

5.2.4 Participant Contribution

Through not assigning specific roles to the participants and not determining a specific game strategy beforehand, participants themselves have to collaboratively develop and agree on how to organise solving the task. We will look at specific strategies for turn-taking and reference grounding observed in the data in Section 5.3, and here primarily investigate their contribution statistics. Figure 5.35 (a) shows the average difference in the number of messages sent per speaker during the five rounds of the PhotoBook task. This difference in messages sent appears to decrease - but as we have seen that the number of messages sent in total decreases with every round, variances for these statistics are so high that decreases are non-significant (Pearson Correlation Coefficient of -0.03 with p-value 0.037). To further investigate speaker contribution, we follow an approach related to that of Xu and Reitter (2016)\(^5\) and determine a conversation leader and follower. With these roles assigned, we can investigate whether contribution patterns change over times. To do so, we assign that speaker who has the higher message count in the first round of a game to be the leader and record the further message counts based on this role assignment. Figure 5.35 (b) shows the resulting graph, indicating that on average speakers who sent more messages in the first round appear to remain leaders in all subsequent rounds as well, converging towards the end of the game. Here however so little messages are sent that through the game setup (presence of a referent image needs

to be either affirmed or denied) convergence is almost inevitable. As an alternative approach to analyse speaker contribution we therefore also determined whether roles swapped during a subsequent round by determining the conversation leader for each round independently and counting the number of swaps per game round. Leader and follower roles actually swap in at least one round of 61% of the recorded conversations. As can be seen in Figure 5.35, roles swap in 19% of the recorded conversations in round 2, 15% in round 3, 22% in round 4 and 23% in round five.

![Distribution of Top 10 n-grams succeeding an Indefinite Determiner](image1)

![Distribution of Top 10 n-grams succeeding a Definite Determiner](image2)

(a) n-grams succeeding an indefinite determiner. The inner circle covers 36% of the data.

(b) n-grams succeeding a definite determiner. The inner circle covers 30% of the data.

Figure 5.34: n-grams succeeding indefinite and definite determiners.

(a) Absolute difference in messages sent per speaker per game round.

(b) Messages sent by the conversation leader (blue) and follower (red) per game round.

(c) Number of leader and follower role swaps per game round.

Figure 5.35: Speaker contribution statistics.
5.2.5 Dialogue Acts

As a last aspect we want to highlight the diversity of the dialogue acts contained in the PhotoBook dataset. In order to proxy an in-depth analysis of the task-specific dialogue acts, we will investigate the different types of messages sent by their initial n-grams.

50% of all messages start with one of the words *I* (28238), *no* (21379), *yes* (17635), *do* (9867) or *ok* (5578). The set of initial n-grams however also has a long tail, indicating the diversity of utterances.

Figure 5.36 shows the distribution of the top-10 first n-grams. From this distribution, we can already conclude participants mainly use *descriptions* of referent images - either indicating that an addressed image is visible in their display or asking whether their interlocutor has the addressed image in his or her screen - and thus implicitly stating that they in fact also have it. This behaviour is not surprising given their objective in the PhotoBook task. Equally unsurprising is the large part of *confirmations* or *rejections*, starting with either *yes* or *no*, or more elaborate replies like *I don't have...*. Not restricted in their utterances, replies however frequently also start with colloquial variants such as *yep, yup, yeah, same or nope*. Participants also seem to initiate many of their utterances with *signal words* like *ok* to indicate that they agree with some previous statement or process and use a significant amount of utterances for *game mechanics*, such as indicating that they have marked all their images and are ready to proceed to next round. One aspect that is more difficult to cover in this manner is the wide range of *clarification questions* posed to identify referent objects with previously ambiguous referring expressions.

To get a more detailed picture of the different dialogue acts, we also investigated the most frequent n-grams following a specific message

![Distribution of Top-10 initial n-grams](image-url)
opening. Figure 5.37 shows the top-10 distributions for messages starting with either I or You. More than half of all messages starting with I continue with have, splitting into references re-referring to a previously seen image (the) or introducing new images (a). I have that one and similar utterances are used to affirm a common image. About a quarter of the I message end in negative answers (I don’t have...) followed by a much smaller portion of game mechanic messages, stating I am done. Other top-10 sequences indicate insecure replies (I think...) or alternative versions of image descriptions (I see..., I also have...).

The distribution of You messages is more diverse. The biggest fraction appears to be an artefact of the greetings at the end of the game, as in A: Thank you! Have a good one - B: You too! Next are questions containing a referring expression, omitting the Do: [Do you have a/the/any/that... , followed by strategy messages like You can go (first) or You good?

Do (Figure 5.38 (a)) shows a similar behaviour, with you... covering almost 90% of the data. Much part of the remainder here is filled by negative replies, omitting the I: (I] don’t have that/the/it...).

Variants of OK appear to indicate a new section of the conversation. While many come alone, most of them are followed by either an image reference (ok, I have... or utterances related to the game mechanics (ok, I’m done, ok, that’s it).

5.3 Dialogue Samples

To conclude this chapter, we want to present some samples from the dialogues recorded during the completion of the PhotoBook task in order to further understanding of the data beyond its statistics.

As a first aspect, we will focus on reference chains. As introduced by Krauss and Weinheimer and Clark and Wilkes-Gibbs (see Section 2.2),
when participants collaboratively establish a canonical referring expression to a given object, the process of its development can be displayed as a reference chain. Because in the PhotoBook task participants often re-refer to a small set of referent images, the resulting dialogues for a large part consist of reference chains of different types spread throughout the five rounds of a game. In the samples, lines (-) indicate a different game round.

Collaborative referring expression generation. One of the central assumptions of this project is that referring expressions are formed collaboratively, with each of the participants responsible to refine referring expressions to increase their indicative power and efficiency. This is a phenomenon that distinctively features in the collected data:

A: Man with dog on lap looking at his computer
B: I don’t have that, but could it be a TV in yours? Mine has a man sitting with his dog watching TV
A: yes to sorry
- B: Dog in the man’s lap as he watches TV
A: nope
- A: man with dog on lap looking at tv
B: I have that

A proposes an initial referring expression including computer. B considers her display and is able to partially ground the image - but with a TV instead of a computer. She therefore describes that image and proposes an alteration to the initial proposal if indeed that this the image referred
to by A. A confirms that and accepts TV as a replacement for computer.

In some cases the interlocutor asks for clarification to ground an initial referring expression while details from the extended description do not recur in subsequent references (i.e. they are directly encoded in the common ground):

B: I have a child in sunglasses riding a motorcycle with a woman behind him
A: I think I have that one, there are more motorcycles in the background?
B: Yes there are more in the background, he is wearing a white shit and I can make out DC on it
  A: and the kid with sunglasses and woman riding behind him
  B: I don’t have that one
  -
  B: kid with woman behind him
  A: yep got that one again too

And lastly, incidentally participants explicitly establish a canonical expression as soon as an initial proposal is grounded by their interlocutor (especially veteran participants who completed a number of tasks before):

A: Guy looking in mirror with bike helmet
B: yep
A: I’m gonna call it mirror from now on
B: gotcha
  -
B: guy in mirror again
A: No mirror
  -
A: Mirror
B: nope
  -
A: Mirror
B: nope

Leading to less natural dialogues, we again stress the importance of our pre-set limit of a maximum of five games per participant.

We can also see cases of self-repair where speakers adorn their proposed expression without their interlocutor explicitly asking for it in order to improve its expressiveness:

A: Do you a umbrella between a man and woman?
A: opening
B: no
  -
B: i have the man and the women again
B: with the umbrella opening. do you?
A: I don’t have it

Omitting image details. One aspect already observed before is that
certain word types like verbs and adverbs tend to be omitted more often than for example nouns:

A: Do you have a boy with a teal colored shirt with yellow holding a bear with a red shirt?
B: Yes
- B: boy with teal shirt and bear with red shirt?
A: Yes!
- A: Teal shirt boy?
B: No

One-sided reference chains. In some cases, only one of the participants repeatedly sees a specific referent image while it is not in the display of his or her interlocutor. Still the referring expression for that referent image oftentimes is directly accepted by the interlocutor and further refined by the initiating speaker through the course of the conversation:

A: How about a lady smiling and waving with a silver umbrella?
B: no
- A: How about the lady waving and smiling?
B: no
- A: The lady smiling and waving?
B: no
- A: Lady smiling and waving?
B: yep

In some cases, the grounding then however appears to not be fully successful:

A: what about a man with a shaved head riding a green motorcycle?
A: this is my last one
B: I don’t have that one either
- A: I see the guy on the green motorcycle again, with shaved head
B: I do not have the shaved head guy
- A: green motorcycle guy again
B: What is the green motorcycle picture
A: the guy has shaved head and sunglasses
B: Okay, I don’t have that one
- B: shaved head green motorcycle guy
A: I do have the shaved head guy again

B here seems to require the shaved head in order to correctly ground the expression, so when A solely gives the green motorcycle guy, she asks for
clarification - and later continues using the *shaved head* as an additional indicator.

As a second aspect, we will look into more detail on the task solving strategies recorded in the data. This mostly concerns turn-taking and meta-level discussions.

**Turn-Taking.** As participants can send messages freely, turn-taking becomes an aspect of the participant’s task solving strategy. In the recorded dialogues, we mostly see one of two alternatives: Either participants agree on a strict order:

A: *hello, i have a boy in a hat in front of an elephant*
B: *two man, one man touching elephant*
A: *if the man is wearing a blue shirt, than yes*
B: *wait, man wearing hat with lot of peoples means yes*
A: *wait no. a single young boy in a red hat in front of a single elephant.*
B: *nno*
B: *first finish yours than mine okay. ask now*
A: *an elephant tail locked with another elephants snout, half a man in the shot wearing a green shirt*
B: *yes*
A: *an elephant painting a picture, a blue man in a tan hat watching*
B: *yes*
A: *done, ask*
B: *green t shirt man touching elephant, with blue color jeans*
A: *no*
B: *done*

which they continue during the next rounds of the game, or they extend replies with references as well:

B: *I have a rider looking into a side mirror*
A: *the side mirror one, is he wearing a helmet and a bright green vest?*
B: *yes*
A: *I have that one*
- A: *ok, final one for me is the side mirror guy again*
B: *I don’t have side mirror guy*
- B: *Okay I have side mirror*
A: *no side mirror this time*
- A: *side mirror guy*
B: *I do not have side mirror guy*

**Differences between first and last rounds.** We observed before that last rounds contain significantly less messages and word tokens than first rounds. To get a better understanding of the cause of this effect, consider these samples of the conversation of two participants in the first and last rounds:

See Appendix A.6 for background details on turn-taking.
round of a single game:

**B:** I have a picture of a banana tangled up in cables sitting in front of a laptop

**A:** I think I see that one, are there like 4 bananas?

**B:** yes i think 4 bananas

**B:** i also have one with a vase of red flowers in front of a laptop

**A:** Ok, I have that one

**A:** I don't see a red vase

**A:** I have one that's a small stuffed cat sitting on a coffee cup in front of a laptop

**B:** my vase is glass but with red flowers

**B:** I have the cat and a coffee mug

**B:** it is a cute cat

**A:** lol :)

**B:** so you don't have a vase of flowers?

**A:** ok last one is two plates of food with some red sauce in between the plates

**A:** nope. no vases here

**B:** i do not have two plates of food

**A:** ok

**A:** so I think just the cat and the bananas are the ones in common

**B:** i think so too

**B:** i have the green laptop

**A:** no green laptop this time

**A:** I have the margarita one again though

**B:** i have the margarita one too

**B:** i also have the beer and muffin

**A:** got that one too

**A:** what about the 2 plates and sauce in the middle?

**B:** i do not have two plates and sauce

**A:** ok

**B:** and you do not have the green laptop correct?

**A:** no green laptop

**B:** sounds good

**A:** that's all 3?

**B:** yes yes
Collecting the PhotoBook dataset is the foundation and a pivotal first step in developing data-driven artificial dialogue agents for partner-specificity in (visually-grounded) dialogue. In this chapter we present a number of suggestions for further research, starting with an initial investigation in automatically extracting reference chains from the collected data in Section 6.1, followed by an overview of proposed evaluation metrics for dialogue agents developed for the PhotoBook task (Section 6.3) and a first proposal for the setup of an end-to-end model architecture based on the collected dialogue data (Section 6.2).

6.1 Automatically Extracting Referring Expressions

One of the main objectives for the design of the PhotoBook task was to set up a conversation task that elicits visually-grounded dialogue that is rich of extensive collaborative reference processes. An important aspect herein is that the referring expressions used by the participants should be unambiguously linked to their respective referent objects in the recorded dialogue data as well. In the PhotoBook task, this linking is facilitated by the setup: As participants are asked to label the highlighted images as either common or different, they first describe or refer to the target images and then select their label. By explicitly asking participants to mark an image as soon as they identified it, we additionally aim to increase the closeness of a used referring expression and the selection of the referent image. By this means we hope to prevent situations like those observed by Kim et al. in the CoDraw task where participants often first discuss the overall layout of the scene - already referring to multiple referent objects - before the drawer selects a first element to be placed. In those cases, correctly linking referring expressions and referent objects often cannot yet be done automatically.

In order to evaluate whether the PhotoBook task produced dialogue data that simplifies the extraction and linking of referring expressions, we developed a validation set consisting of 50 random dialogues. In this validation set, we manually processed the recorded dialogues and annotated utterances containing referring expressions with the ID of their respective referent image. In order to do so consistently, we established the following set of annotation guidelines:

1. All utterances that contain any words describing objects in a referent image should be marked as containing a referring expression to that image. Object descriptions here should be seen as independent of the specific dialogue act that the utterances perform. Specifically, this may include:
• **Image descriptions.** Examples: *green tank lady* and *Do you have 2 men standing in front of a truck, 1 holding a stop sign?*

• **Clarifications.** Example: The second utterance in the interchange
  
  A: *Just one more... larger man in black and blue coat with elephant and man on left?*
  
  A: *The larger man is on the right... and the man in white near the middle.*

• **Confirmations and Acknowledgements** that are extended with a reference to the target image. Examples: *yes to hand shake* and the second utterance of

  A: *Do you have a picture of a man in a blue shirt with the elephant to the right?*
  
  B: *Is the shirt striped and the man wearing a hat? If so, yes!*

• **Rejections** - if they are descriptive of the content of the interlocutor’s referent image. Example: The second utterance in

  A: *I also got soccer boys again*
  
  B: *no soccer boys*

2. Referring expressions that should explicitly **NOT** be annotated for this task include:

• **Expressions containing only single pronouns** - since they are not descriptive of the image’s content. Example: *I don’t have that or I’ve had those for like 4 rounds. not this time though*

• **Meta-Descriptions of referent images.** Example: *haha, yeah that’s a confusing picture*

• **Statements that do not cover the contents of a specific referent image.** Examples: *mmm..no I don’t have another one with a kid or Hmmmm, I have two girls with phones*

3. If multiple referent images are mentioned in the same utterance, the utterance should be marked as containing a referring expression for the first image only.

To assess inter-annotator agreement, ten of the validation set transcripts were independently annotated by two reviewers and compared based on the binary decision whether or not an utterance was marked as containing a referring expression. The selected dialogues contained a total of 1015 utterances, of which 505 were unanimously marked as referring expressions, 494 as negative samples and 16 decisions were disagreed on. This results in a naive agreement of 98.4% with a Cohen’s Kappa Coefficient of 96.9%, indicating an almost perfect agreement.¹

With this annotated validation set we then tested the closeness of referring expressions and their respective referent image. To do so, we selected the last n utterances per speaker before a recorded image selection and checked whether these candidate utterances contain all referring expressions to that referent image which were uttered during that specific game round. The resulting graph in Figure 6.1 shows that with a history cutoff of two utterances per speaker the coverage already is over 90%, reaching 98% for a cutoff of three utterances per speaker. We

¹ Cohen (1960), Landis and Koch (1977)
can thus conclude that in the collected dataset referring expressions indeed are quite close to the selection of their respective referent image, facilitating the grounding of references during processing.

We then continued by developing a simple heuristics to automatically annotate referring expressions in all other dialogue transcripts. Using a similar approach as He et al. (2017), for each game transcript we first collect sets of candidate utterances that are likely to contain referring expressions to the highlighted images. We do by adding the last three utterances per speaker before a recorded image selection to that image’s candidate set. Utilising the observation that referring expressions shorten over time while keeping identifiers grounded in the interlocutors’ conceptual pacts, we then extract referring expressions based on a similarity function: Iterating over a referent image’s candidate utterances, we select those utterances that are proposed to contain a reference to it by comparing the utterance’s similarity to the set of previously used references. Because this process needs to be bootstrapped for the first candidate set, the five image captions given by the MS COCO metadata are used to start the matching process.

To calculate a candidate utterance’s similarity score, we sum its pairwise similarity scores with each of its previous references during a given game. To put more emphasis on the most recently used referring expressions that are proposed to be most similar to the next one, we reverse the list of previous expressions and use an exponentially decaying weight when summing the similarity scores. The weight factor of this decay function is one of the parameters of the heuristics. The second parameter is a minimum score threshold. When the sum of similarity scores is calculated, it is normalised by the size of the set of previous reverences and compared to a minimum threshold: If it passes the threshold, the utterance is added as a referring expression - and considered as most recent reference for the next candidate set.2

The similarity between two utterances lastly is assessed either based on exact word overlap (WO) of the bag-of-words representations of the two samples or the cosine similarity of their average word embedding based on Word2Vec.3

Figure 6.2 displays heatmaps of the effects of the weight decay and minimum threshold parameters in the matching heuristics’ performance for both the word overlap and cosine similarity metrics. Comparing the precision and recall heatmaps, the reciprocity of the two parameters becomes notable: If the weight decay factor is low, most focus is put on the recent utterances while older expressions weight only little. If the minimum score threshold then is set high, only utterances very similar to the last referring expression can pass - increasing precision but drastically decreasing recall. Weight decay factors on the other hand mean that older references carry a large amount of weight as well - which means that low score thresholds lead to high recall but significantly lower precision. Figures (a) and (d) show that for both metrics there is a range of combinations that lead to a similar maximum F1 score on the

\[ F_1 \]

See Appendix 1 for an algorithmic description of the matching process.

Mikolov et al. (2013)
annotated validation set, with the best setting of parameters for the word overlap metric outperforming the best setting for cosine similarity by about 3%.

With this baseline approach already reaching high precision and recall, we argue that a more refined method to automatically link referring expressions and their respective referent images is decidedly feasible. Using referring expressions however is only one of the aspects in producing the full dialogues recorded in the PhotoBook task. Approaching the full task therefore requires a more involved architecture. As mentioned previously, we propose to use a modular architecture to develop a computational dialogue agent for the PhotoBook task.

6.2 Proposed Model Architecture

Drawing on the review of the current line of state-of-the-art dialogue agents in Section 3.1, we propose to approach the task of learning a natural usage of referring expressions by training a dedicated dialogue agent in an end-to-end fashion using the PhotoBook conversation task. Specifically, we propose develop a modular agent architecture to predict the common or different labels for highlighted images based on the visual context of the participant’s display (consisting of six images and the highlight indications) and the previous dialogue history.

Following the approach of De Vries et al. (2017a), the agent archite-
ture should contain a *Guesser* module that is responsible for identifying an indicated referent image given an utterance of its interlocutor and the full dialogue history. This guesser module then should either output a guess on which of the images in its display is the correct referent and its confidence - as the image referred to can be different and thus only be shown to the other player - or guess probabilities for all visible images and forward this information to a *Decider* module that is specifically trained to make the guessing decision. In the former case, decisions can either be forced at the end of each turn, or collectively at the end of a game round.

For the *Describer* module we propose an approach similar to the context-aware image captioning architecture reviewed in Section 3.3, extending the model with the dialogue history in addition to the visual context as in the current version.

Lastly, the agent architecture also requires a *Reply* module that can confirm or deny image descriptions and ask clarification questions. This is closely related to the question generator of De Vries et al. (2017a) but also should be able to take on replies.

Besides the exact implementation of the different modules, we identified two additional aspects that demand further attention: How to combine the visual information available in the display into a sensible representation for the different modules, and, most importantly, whether passing the dialogue history to the modules already is a sufficient representation of the conversation’s common ground. As we argue that an inadequate use of dialogue history is one of the main causes for the decline in artificial dialogue agent output quality, future work on this issue should thoroughly investigate potential improvements in its encoding or the introduction of a dedicated, comparable to those presented in Section 3.2.

### 6.3 Evaluation

We briefly mentioned in Chapter 1 that evaluating dialogue data output is a challenging endeavour as at any point in a conversation the number of possible correct replies is practically infinite. One of the central functionalities of the conversation tasks reviewed in Chapter 3 and the PhotoBook Task presented in Chapter 4 therefore is to provide an additional means of assessment: If utterances produced by an agent lead to downstream task success, these utterances can be viewed as ‘correct’ with respect to their pragmatic role. The end-to-end training of the model architecture proposed in the previous section also heavily builds on the task performance assessment to learn how to incorporate the current dialogue state and conversation context to generate a next utterance. Using the recorded human-human dialogue data to model both sides of a conversation however means that not only utterance generation needs to be learned, but the correct identification of target images from given descriptions and references needs to be learned as well. The quality of this recognition module takes an equal role in determining model performance: If it requires referring expressions to be close to those recorded in the human training data, training can
become inhibitably slow or even impossible as the generator needs to learn to produce a close resemblance of the recorded ‘gold-label’ utterance - which might not be possible given the range of options. If on the other hand identification module is too strong and establishes a code to collaboratively solve the task with the descriptor (consider for example the toy experiment by Das et al. (2017b)\(^5\)), the generated outputs diverge too much natural language and the agent can no longer be used as human-machine interface. In order to evaluate this gap between machine-machine and human-machine cooperation, Chattopadhyay et al. (2017) developed GuessWhich?!, where similar to the VisDial setup an AI agent describes images to a human partner who has to select them from a grid of distractor images. In their experiments, they showed that models that performed better in machine-machine settings performed worse when paired with a human partner, stressing the discrepancy in communication.

6.3.1 Task Efficiency

The review of linguistic research on human conversations in Chapter 2 shows that participants usually not only are able complete the task, but also tend to become more efficient in doing so when repeating the task with the same partner. Arguing that this improvement in their interaction is largely due to the developing common ground, assessing how well a computational dialogue agent manages to utilise its encoding of the current dialogue history and conversation context therefore should also incorporate conversation task efficiency. In Chapter 5 we presented a first rough indication of task efficiency by calculating the number of messages sent or tokens used per correct label selection. While facilitating automatic evaluation, these metrics however cover only low-level dialogue properties. A more involved but potentially more indicative method could be to register the number of clarification questions asked by an interlocutor or the amount of adjustments made to a proposed referring expression before it has obtained its canonical form.

Focusing on this development of referring expressions as well, Aina et al. (2017)\(^6\) present three metrics covering the length decrease in referring expressions, their lexical alignment and form alignment to evaluate the dynamics of the refinement process. Relating these metrics to task success, lexical entrainment was found to be the most reliable predictor of task success - with the peculiarity that intermediate levels of lexical alignment best predict good task performance.

6.3.2 Qualitative Analysis

Besides task success and task efficiency, assessing the produced dialogue outputs in a more qualitative fashion can provide more intuitive insights into the models’ working. Currently the only reliable method to do so is in experiment settings where independent human assessors rate produced utterances in comparison to human samples, like for example employed by the previously presented works of Das et al. (2017a); He


et al. (2017). Liu et al. (2016) recently published an empirical study of the agreement between human ratings and those produced by unsupervised evaluation metrics like BLUE, ROUGE, METEOR and different embedding-based metrics, showing that all of them correlate very weakly with the human judgements.

To conclude, automatic evaluation of artificial dialogue agent’s outputs still is an open issue, with dedicated metrics covering only parts of its diverse aspects and more generic methods showing only low agreement with human ratings. Task performance therefore currently remains one of the few automatic options for assessment.

In this chapter we showed that the dialogue data collected through the PhotoBook task facilitates an automatic extraction of referring expressions together with their referent images. Being able to do so with high precision and recall using a simple heuristics baseline, we strongly believe that training a dialogue agent to learn a natural usage of referring expressions should be possible when training based on the PhotoBook data in an end-to-end training regime. In Section 6.2 we presented a first proposal to approach such an agent architecture.

8 Papineni et al. (2002)
9 Lin (2004)
10 Lavie and Agarwal (2007)
Conclusion

In this thesis we present a novel visually-grounded conversation task called the PhotoBook Task, together with a large repository of human-human dialogue data collected by crowd-sourcing the conversation task online. The PhotoBook Task requires participants to identify common and different images by chatting with one another. In contrast to previous setups, we extend the standard image identification setting to five consecutive rounds with the same conversation partner and let referent images re-occur in different rounds. As a result, participants naturally develop canonical expressions for the referent images - a phenomenon which has not been captured in large-scale dialogue data before.

Dialogues in the PhotoBook dataset cover everyday visual input scenes and are rich in collaborative reference generation processes. We argue that this is a pivotal feature of the PhotoBook task design as referring expressions heavily draw on the common ground between two interlocutors. Common ground in turn is hypothesised to be a central element in producing natural and efficient utterances during a conversation. A review of current state-of-the-art computational dialogue agents on the other hand reveals a drastic decrease in dialogue output quality when applied to multi-turn dialogue settings. We therefore argue that this decline in performance is primarily due to the lack of an adequate representation of the conversation’s common ground and propose that extending the current line of artificial dialogue agents with the computational counterpart of a common ground representation seems to be a promising next step in their development.

By implementing the PhotoBook conversation task in Facebook’s ParlAI framework and crowd-sourcing data collection on Amazon Mechanical Turk, we collected a total of over 2,500 dialogues. An extensive analysis of the resulting data shows that the recorded conversations closely resemble the characteristics observed in natural human-human conversations: Participants tend to collaboratively optimise their interaction during the five rounds of a single task, as well as independently refine their strategy over the course of completing multiple instances of the task. We propose to use this unique, dedicated data to train a data-driven model to complete the PhotoBook task in and end-to-end setting, enabling it to learn a natural usage of referring expressions which cannot be achieved by current model architectures.

Considering all of this, we argue that the work presented in this thesis is an elaborate and solid foundation for a new approach to investigating common ground for artificial dialogue agents that we believe has the
potential to significantly improve their dialogue output consistency, efficiency and naturalness in multi-turn conversations.
References


Figures

2.1 Triangle of reference or semiotic triangle as proposed by Ogden and Richards (1923).

2.2 Example figure cards shown to participants in the experiments of Krauss and Weinheimer (1964).

2.3 Length of referring expressions used by the five subject pairs as a function of the number of repetitions in the experiments of Krauss and Weinheimer (1964).

2.4 The twelve Tangram figures used in the experiments of Clark and Wilkes-Gibbs (1986).

2.5 Tangram figure from the conversation task by Clark and Wilkes-Gibbs (1986).

2.6 Average words spent per figure during the six trials of an experiment run in Clark and Wilkes-Gibbs (1986).

3.1 Distribution of first n-grams for (left to right); VisDial questions, VQA questions and VisDial answers. N-grams extend from the centre, the width of a portion indicates its relative frequency in the data.

3.2 Model pipeline of the best-performing system of Das et al. (2017a). Image and question features are concatenated and compressed to 512 nodes through a fully-connected layer; previous QA-pairs are encoded as facts in a memory bank and combined by a weighted sum using attention over the elements. Answers are generated by a decoder fed with the combination of current and previous input features.

3.3 Policy networks for Q-BOT and A-BOT as presented by Das et al. (2017b). At each round of the dialog, Q-BOT generates a question based on its state encoding and sends it to A-BOT. A-BOT then updates its state encoding based on the received question, generates an answer and returns it to Q-BOT. Q-BOT then predicts an image representation and receives a reward based on the distance to the actual image shown to A-BOT.

3.4 Number of questions per dialogue in the GuessWhat?! dataset by De Vries et al. (2017a).

3.5 Human solution strategy efficiency in the GuessWhat?! task as compared to linear or logarithmic strategies (image from De Vries et al. (2017a)).

3.6 When trying to identify the selected object, questioners move from abstract object classes to concrete descriptions (image from De Vries et al. (2017a)).

3.7 Participant’s AMT interface for the drawer in the CoDraw task setup by Kim et al. (2017).

3.8 Model pipeline for the dynamic multi-attention architecture of Kim et al. (2017). The utterance generated by the teller is a product of an attention mechanism over displayed objects conditioned on the encoding of the current dialogue state. In a similar manner, the drawer interprets a received utterance through attention on available objects conditioned on its encoding of the dialogue state.

3.9 Prior-Encoder-Decoder architecture as presented by Massici et al. (2018). The top pipeline shows the combined conditional encoder and prior network, the middle one the default conditional decoder and the bottom one the auto-regressive conditional decoder architecture.

3.10 AMT Task interface of the MutualFriend task by He et al. (2017).

3.11 Model pipeline for the Dynamic Knowledge Network architecture of He et al. (2017). Knowledge graph nodes are embedded through relevant utterance embeddings and utterances are generated by an attention-based mechanism over those node embeddings.

4.1 Frequency of occurrence of the different main object categories in the PhotoBook Image sets.
4.2 Heatmap of the pairwise cosine similarity of image embeddings for the different PhotoBook Image sets. ........................................... 40
4.3 Screenshot of the PhotoBook task's display for one of the participants during the warming-up round. ........................................... 42
4.4 Example screenshots for a participant's display during a game round and feedback screen. ........................................... 43
5.1 Coverage of the 60 unique game sets in the PhotoBook conversation task. ................................. 44
5.2 Percentages of participants completing a given number of games. ........................................... 45
5.3 Histogram of Game Completion Times in the PhotoBook conversation task, with mean 14.2 minutes (red) and median 13.08 minutes (orange). ........................................... 45
5.4 Game Completion Times for the nth game per participant. Statistics obtained from the unfiltered dataset. ........................................... 46
5.5 Game Completion Times for the nth game per participant. Data filtered to games of participants who completed a full series of five games only. ........................................... 46
5.6 Completion Times per Game Round in the PhotoBook Task. ........................................... 47
5.7 Completion Times per Game Round in the PhotoBook Task after filtering out all messages sent after the last selection. ........................................... 48
5.8 Percentage of games that scored a total of \( n \) points, where \( n \) is indicated in the outer ring. ........................................... 48
5.9 Distribution of scores per game round. ........................................... 48
5.10 Average score per game round (scores are out of six). ........................................... 49
5.11 Distribution of incorrectly labelled images: different means that common images were marked different, and vice versa. ........................................... 49
5.12 Total Game Scores for consecutive completions of the PhotoBook conversation task. ........................................... 49
5.13 Average and median number of messages sent per game round in the PhotoBook conversation task dataset. ........................................... 50
5.14 Average and median number of messages sent per game round after filtering out all messages sent after the last selection in each round. ........................................... 51
5.15 Average and median number of tokens used per game round after filtering all messages sent after the last selection in each round. ........................................... 52
5.16 Average and median token/message ratio per game round after filtering out all messages sent after the last selection in each round. ........................................... 52
5.17 Average and median points per token ratios for subsequent game rounds after filtering out all messages sent after the last selection in each round. ........................................... 53
5.18 Average and median points per message ratios for subsequent game rounds after filtering out all messages sent after the last selection in each round. ........................................... 53
5.19 Average and median message counts (total of both players) for all participants who completed a full set of five games. ........................................... 54
5.20 Average and median token counts (total of both players) for all participants who completed a full set of five games. ........................................... 54
5.21 Order in which images are marked during the later rounds of the game: For each \( n \)th selection per round, in what part of the recorded dialogues was the selected image selected in a previous round (dark) and in what part is that image selected for the first time in the game (light). ........................................... 55
5.22 Average similarity rankings for the three selections per player per game round. Dark blue indicates that the selected image was the least similar of the set of highlighted pictures, red indicates the most similar highlighted image. ........................................... 55
5.23 Participant’s ratings on the three statements of the evaluation page. ........................................... 56
5.24 Token frequencies and data coverage per occurrence rank. ........................................... 57
5.25 Average and median number of content words per game round after filtering out all messages sent after the last selection in each round. ........................................... 57
5.26 Average and median content word/total token ratio per game round after filtering out all messages sent after the last selection in each round. ........................................... 58
5.27 Average and median content word over total token count ratios for all participants who completed a full set of five games.

5.28 Average and median number of new word tokens introduced to the conversation in each of the five game rounds. Data filtered by messages sent after the last selection in each round.

5.29 POS-Tag distributions during the subsequent rounds of the PhotoBook Dialogue Task. Subfigure (a) and (b) show POS-Tag frequency counts and their decrease between the first and the last game. Subfigure (c) and (d) display the relative changes.

5.30 POS-Tag distributions in a participant’s five consecutive completions of the PhotoBook Dialogue Task. Subfigure (a) and (b) show POS-Tag frequency counts and their decrease between the first and the last round. Subfigure (c) and (d) display the relative changes.

5.31 Top-10 determiner distributions during the subsequent rounds of the PhotoBook Dialogue Task. Subfigure (a) and (b) show determiner frequency counts and their decrease between the first and the last round. Subfigure (c) and (d) display the relative changes.

5.32 Occurrence frequencies of word tokens following a determiner sorted by rank.

5.33 n-grams succeeding a determiner. The inner circle covers 30% of the data.

5.34 n-grams succeeding indefinite and definite determiners.

5.35 Speaker contribution statistics.

5.36 Top-10 initial n-grams. The inner circle covers 59% of the data.

5.37 Initial n-grams succeeding I and You.

5.38 Initial n-grams succeeding OK and Do.

6.1 Coverage of referring expressions with dialogue history cutoff at n turns before an image selection.

6.2 Heuristics performance grid for the Word Overlap (a-c) and Cosine Similarity (d-f) metric. Lighter is better.
### Tables

1. **Game Round Completion Times** in seconds, their respective differences and the t-test p-values showing significant decreases with respect to the previous round.

2. **Game Completion Times** in seconds, their respective differences and the t-test p-values for up to five subsequent games per participant.

3. **Game Completion Times** in seconds, their respective differences and the t-test p-values for five consecutive games from participants who completed the full limit of five games.

4. **Average Round Scores** in points out of 6, their respective differences and the t-test p-values.

5. **Average Game Scores** in points out of 30, their respective differences and the t-test p-values for up to five subsequent games per participant.

6. **Average Game Scores** in points out of 30, their respective differences and the t-test p-values for five consecutive games from participants who completed the full limit of five games.

7. **Total number of messages sent per game round**, their respective differences and the t-test p-values. Statistics obtained from the unfiltered dataset.

8. **Total number of messages sent per game round**, their respective differences and the t-test p-values. All messages sent after the last selection in each round were filtered out.

9. **Total number of tokens used per game round**, their respective differences and the t-test p-values. All messages sent after the last selection in each round were filtered out.

10. **Token/Message ratios** of subsequent game rounds in the PhotoBook conversation task. All messages sent after the last selection in each round were filtered out.

11. **Total number of content words used per game round**, their respective differences and the t-test p-values. All messages sent after the last selection in each round were filtered out.

12. **Content Token/Total Token ratios** of subsequent game rounds in the PhotoBook conversation task. All messages sent after the last selection in each round were filtered out.

13. **Average unique token count per game round** after removing all messages sent after the last selection in round 5.

14. **Number of new word tokens per subsequent game round**, their respective differences and the t-test p-values. All messages sent after the last selection in each round were filtered out.

15. **Points per Token ratios** for subsequent game rounds. All messages sent after the last selection in each round were filtered out.

16. **Points per Message ratios** for subsequent game rounds. All messages sent after the last selection in each round were filtered out.
17 Total number of messages sent per consecutive game, their respective differences and the t-test p-values for five consecutive games from participants who completed the full limit of five games. . . . . . . . . . 105

18 Total number of tokens uttered per consecutive game, their respective differences and the t-test p-values for five consecutive games from participants who completed the full limit of five games. . . . . . . . . . 105
Appendices

A The PhotoBook Task Design

A.1 Number of Interlocutors

Clark and Carlson (1982)\(^1\) had investigated multi-party dialogues before, noting that in settings with more than one interlocutor, speakers seem to adhere to a responsibility of keeping all participants informed, carefully curating the accumulation of common ground with each of them according to whether they are direct addressees or side-participants. Based on a review of a range of fundamental dialogue research, Branigan (2006)\(^2\) also concluded that two-party and multi-party dialogues qualitatively are the same, but oftentimes are compared incorrectly: Other dialogue dimensions like the type and the goal of the conversation significantly impact a dialogue’s form as well - but frequently are not sufficiently controlled for when analysing the effects of increasing the interlocutor count, causing skewed results.

A.2 Referent Types

The following list gives an overview of the different visual object types available to us and the primary arguments for or against using them in our data collection task setup:

1. **Geometrical shapes** of different sizes and colours probably are the most simple form of visual object. In their simplicity however it is trivial to fully and uniquely describe them with a few words only. This means that there is less need for adapting referring expressions during the course of a conversation because initial descriptions are likely to be quite optimal already. This obviates the collaborative generation process we want to capture in our data; Simple shapes thus completely disqualify for the task at hand.

2. **Polyminoes**\(^3\) are the infinite set of shapes that can be constructed by connecting any number of square elements at at least one side. If the figure contains exactly five square elements, it is called a Pentomino. Pentominoes are the most common -mino version used in previous research on referring expression generation (see for example Schlangen and Fernández (2007)\(^4\) or Siebert and Schlangen (2008)\(^5\)). With this restriction of exactly five elements, the basic set of Pentominoes contains just 12 shapes (see figure 1), but can be extended by mirroring the shapes (resulting in a set of 18 distinct Pentominoes) and rotating the figures (resulting in a set of 63 so-called fixed Pentominoes). These larger sets of shapes often are not trivial to distinguish and describe anymore, leading to a full development of the referring generation process in conversations addressing them. We however argue that

---


these abstract figures are likely to elicit referring expressions that are closer to an encoding of the figures (similar for example to their official naming convention) instead of the initially more descriptive referring expressions we encounter in natural conversations about real-world objects. For a large-scale data collection experiment, we thus think that this is not the best option either.

3. **Line drawings** can range from outlines or sketches of actual objects over more abstract figures as used by Krauss and Weinheimer - where parts of the drawing might form recognisable elements - to completely random scribblings. Depending on the level of abstraction, line drawings are easy or difficult to describe and recognise - and therefore demand different levels of effort spent on properly grounding a novel object. As Krauss and Weinheimer argue, if the level of abstraction is too low, the depicted object might elicit popular referring expressions. On the other hand, we believe that figures that are too abstract will make the referring task quite complicated while also leading to dialogues that are not representative of interesting real-world conversations anymore either. For our data collection, line drawings therefore seem inappropriate as visual input domain.

4. **Tangram Figures** were used for the seminal dialogue collection experiments by Clark and Wilkes-Gibbs. They, too, can be arranged to depict shapes with the outlines of actual objects or just abstract forms, while generally speaking being more abstract than line drawings. Consequently, here the same argument holds as for the line drawings: We believe that a conversation task about abstract geometrical figures does not produce dialogue data that is relevant enough for a wider study of computational referring expression generation.

5. **Cliparts** are used in the VQA and CoDraw conversation tasks. The main advantage of using clipart figures is that they can be arranged to form a highly natural but controllable visual input that is less complex than actual photographs. Precisely this was done in CoDraw, providing the describer with a manually created clipart scene to be explained to the drawer, leading to relatively natural visual scene descriptions.

We thoroughly considered using cliparts as well, but arrived at the conclusion that designing an interactive interface like CoDraw for a symmetrical task setting that still elicits dialogue output which can be controlled to focus on re-references to given objects is too difficult of a task to apply in the large-scale online setting we aim for. Changing the task to a simple identification task on the other hand would render clipart scenes too trivial to disambiguate due to the high level of detail compared to the low number of visible objects they contain. We therefore argue that cliparts neither are a perfect fit for our data collection experiment.

6. **Photographs** lastly are the most complex sample of this range of (two-dimensional) visual objects - but also are closest to the actual real-world visual inputs that an embodied agent might get. Their complexity for the moment renders interactive tasks like CoDraw...
infeasible, but allows for a range of other tasks such as identifying scenes or objects in them. One difficulty that photographs introduce - and which does not exist for all other afore-mentioned object types - is that in photographs it is much harder to control what exactly is depicted in them, i.e. what are the characteristics of the central objects and what kind of (and how many) background objects are there. We propose to mitigate this issue by assembling a number of sets of photographs with certain mutual or at least similar properties. By presenting those images alongside one another, we regain some control over the complexity of the visual input to the conversation task.

Especially considering the arguments against the other visual input types, we decided that using those sets of photographs therefore provides us with the best options to design a visually-grounded task that elicits natural dialogue with a focus on how referring expressions are formulated.

A.3 Initial Pilot Runs

For a first series of pilot runs we used the following task setup: Two participants were placed at a computer in separate rooms. Each of them was given a set of five presentation slides that each displayed a number of photographs with a common theme (like for example persons with umbrellas). One of the images per subsequent slide was common, the remainder of images was different for both participants. The order of images on the slides was randomised for both participants and the common image changed per slide. The participants then were asked to detect the common image in each corresponding pair of slides using a written chat interface. Modelled closely to the MutualFriends setup, this quickly revealed two key insights:

**Position Bias.** Determining a single common image out of a set of displayed photographs leads to strong position-based effects in the order of addressing them: While He et al. noticed that in the MutualFriends task participants often start by mentioning those attributes that are most informative for eliminating large parts of the knowledge base at once, (larger) sets of photographs appear to provide a less structured view of the attributes they contain and therefore oftentimes evade such a strategic grouping. Instead, participants tend to describe images sequentially, either starting at the the top-left image, a photo they recognise from earlier rounds or one that has salient, distinctive features.9 This way, if the common image is addressed and identified early during a game round, the remaining other images are not referred to at all in that round. And when keeping in mind that we want to collect data of the complete referring expression generation process of as many referent objects as possible, this is clearly a counterproductive effect.

**Number of Distractors.** When referring to a specific photograph from the displayed set, all other images practically function as distractor images that influence the amount of information necessary for the initial proposal of a referring expression. Displaying too few distractors makes

---

it trivial to spot disambiguating features that can be used to propose simple initial referring expressions - which then do not need to be refined much more during the remainder of the conversation. Displaying too many similar images on the other hand makes the task of establishing a canonical referring expression for all of them quite involved again - and decreases the likelihood of re-reference considering the position bias effect just mentioned. It also leads to practical difficulties as displaying more than nine pictures on a regular laptop screen together with a chat interface renders the photographs so small that many details cannot be easily recognised anymore: A qualitative analysis of the dialogue data obtained from these first pilot runs lead us to the conclusion that displaying no more than five to six images per round provides the best trade-off between disambiguation complexity and image size.

A.4 Determining Image Set Properties

Instead of identifying a single common image out of a set of images, next we investigated a task setup where participants were asked to mark all displayed images as either common or different. This setup requires the participants to refer to each unique image at least once per round - given that there are varying numbers of common and different images in each round, as it otherwise simplifies to the original task as soon as the (or all of the) common image(s) are detected. While this leads to dialogues with precisely controllable numbers of references to a given set of objects, this also makes the task much more time-intensive.

Because we now need different numbers of common images per round, in each round some of the images will be shown to both participants, and some will be shown to only one of them. This means that we need strictly more images in an experiment’s input set than those that are displayed to each of the participants in a single round. If for example we display six images per round, of which two are common, this means that the set of unique pictures for that round already needs to be ten (two times four distractors and two common images). We then want to collect at least four re-occurrences for each of those ten images, so we continue by designing a second round that for example has four common images and four distractors (two per participant). As a result, only a maximum of eight of the ten images from the first round re-occur in round 2. Continuing this process for all subsequent rounds, getting exactly five occurrences for each of the images would require much more than five game rounds. Game completion times on the other hand then would be much longer than we think is sensible for an online task.

In order to balance these two issues, we decided on the following setup: Each full task should consist of five rounds, where in each round six images out of a set of twelve are shown to each of the participants. By exactly controlling which images are shown to which of the participants in each round, we can precisely control the number of common and different images - and their total occurrence count. With this setting, we managed to design the task in such a way that each of the twelve images in a set is displayed exactly five times during a task and experiments with
this setting took pilot participants an acceptable 15 minutes on average. The caveats of this setup, however, is that some of the five occurrences of an image are in the same round of the task. In these cases the referent image is common and displayed to both participants at the same time - which often means that only one of the participants refers to it. As a consequence, we can no longer guarantee an exact five referring expressions per referent object per game.

Manually inspecting the referring expressions generated during those runs however revealed that this lower number of references oftentimes did not prohibit a canonical referring expression from being established - although it might not be fully optimised yet. Considering this a reasonable setup, we decided to investigate this setup in a first series of actual online experiments, only to quickly discover that the online participants spent significantly more time on the task than the UvA students we asked to do our initial pilot experiments. In the actual online setting, full tasks took between 20 and 45 minutes, with an average of 26 minutes (Sample size = 15). As all of these are far beyond what we deem acceptable for a large-scale online data collection experiment, we developed an alternative way of limiting the number of referent objects while not changing the task’s dynamics: By highlighting three images on each participant’s display and asking them to mark only those images as either common or different, we practically halve the effort of a game round - while not making the task any easier or harder. And with a careful selection of those subsets of images to mark in each round, we keep an almost equal number of references to a given referent image while halving the number of images that will be referred to during a game. This means that each set of twelve images now needs to be used in two games with different highlights in order to get referring expression for all of them.

A.5 Main Object Categories in the PhotoBook Image Sets

<table>
<thead>
<tr>
<th>bus + truck</th>
<th>person + cake</th>
</tr>
</thead>
<tbody>
<tr>
<td>car + motorcycle</td>
<td>person + car</td>
</tr>
<tr>
<td>car + truck</td>
<td>person + couch</td>
</tr>
<tr>
<td>chair + couch</td>
<td>person + dog</td>
</tr>
<tr>
<td>couch + laptop</td>
<td>person + elephant</td>
</tr>
<tr>
<td>dining-table + bowl</td>
<td>person + motorcycle</td>
</tr>
<tr>
<td>dining-table + cake</td>
<td>person + oven</td>
</tr>
<tr>
<td>dining-table + couch</td>
<td>person + refrigerator</td>
</tr>
<tr>
<td>dining-table + cup</td>
<td>person + suitcase</td>
</tr>
<tr>
<td>dining-table + laptop</td>
<td>person + surfboard</td>
</tr>
<tr>
<td>dining-table + refrigerator</td>
<td>person + train</td>
</tr>
<tr>
<td>person + bed</td>
<td>person + truck</td>
</tr>
<tr>
<td>person + bench</td>
<td>person + tv</td>
</tr>
<tr>
<td>person + bicycle</td>
<td>person + umbrella</td>
</tr>
<tr>
<td>person + boat</td>
<td>teddy-bear + person</td>
</tr>
</tbody>
</table>
A.6 Turn-taking

One last option of the task setup that we investigated through an initial set of trials was whether to allow participants to write and receive messages simultaneously (free chat), or to limit the chat to be turn-based. With the latter, participants need to organise turn taking themselves, with the former turns are alternating after sending a message or exceeding a pre-set time limit. In a first set of pilot runs using free chat we observed a number of instances where participants sent messages almost at the same time, resulting in confusing situations where they were unsure to which of their messages a reply of their interlocutor referred to. While this could be compared to cross-talk in spoken dialogue settings, it has a different character because spoken cross-talk situations normally do not lead to confusion as to which of a speaker’s own utterances an interlocutor’s reply refers to. This is mainly due to the fact that in spoken conversations speakers realise cross-talk relatively quick and do not tend to complete their utterance - or their partner does not fully register the utterance because she was talking herself.\textsuperscript{10} In the written case however, utterances are almost always complete when sent.\textsuperscript{11}

Noting this source of confusion, the first series of online pilots was implemented with a turn-based chat interface to prevent this kind of simultaneous message sending. On the other hand, this meant that a certain participant had to start a round, which we selected by random. In the resulting dialogues, the starting player often dominated that specific round by iteratively describing his or her pictures, while the other player just replied with very short yes or no answers, unable to actively participate in the principally symmetrical task. When the starting participants had marked all their highlighted images, they sometimes even forgot that their partners also had to mark their images - resulting in conversations getting stuck with the second player unable to communicate during the starting player’s turn. Besides observing this de-facto subversion to a asymmetric dialogue, we detected that the turn-based chat seemed to cause participants to get detached from the task when waiting too long for a reply, oftentimes again increasing the completion times of the task.

To counteract these effects, we ran a second series of online pilot runs without the turn-based chat limitation - which strongly reduced completion times and increased participant’s self-reported satisfaction with the task. In addition to that, we observed in the resulting data that exactly those situations where participants were unsure about which of their messages their partner’s reply belonged to, were from that moment on resolved by either developing a clearer game strategy or extending replies with referring expressions as well.\textsuperscript{12} Instead of just replying with yes or no, many of the participants started re-referring to a referent object in their replies as well in order to indicate which of the referent objects the reply belonged to. Enabling free chat therefore oftentimes caused referent objects to be referred to more often than they were actually displayed during a game - which further moderates the previously mentioned limitation of having only a maximum of five displays per image per experiment run.


\textsuperscript{11} Examples of cross-talk in written chat dialogues:

(1) Simultaneously describing the same image

A: do you have a person with a long sleeve red shirt, next to a guy with a cowboy hat
B: man on left with hat and grey suit talking to girl with red shirt and gray cap on
A: yes that one , do u have it
B: yes

(2) Messages being overlooked due to simultaneous sending

A: I have the boy on the bench!
B: I have the woman taking a picture of her fridge now.
B: I don't have him on this page.
A: And those orange apron pizza people
B: Do you have the woman taking a picture of the fridge now?
A: I have the woman taking the pic of the fridge too

\textsuperscript{12} Examples of game strategies preventing or disambiguating cross-talk:

(1) Explicit turn taking

B: (I’ll wait for your description)
A: Same picture from before, surfboard with shark and time clock on it?
B: No that’s not here this time.

(2) Explicit referent object mentions

B: A birthday cake with candles
A: no birthday cake
A.7 The ParlAI dialogue framework

The general setup of the ParlAI framework consists of three main modules: worlds, teachers and agents.

Worlds are the environments that conversation can take place in. The most simple environment would be a two-party, written text exchange environment, which can be extended to include multi-party conversations, use different dialogue modalities or enrich the context of a conversation through additional inputs.

Teachers define how the environment is filled, for example by specifying one of the previously mentioned dialogue tasks, or by loading input text resources, images and instructions for participants. In the case of the PhotoBook data collection task, the teacher module selects an image set to display to the participants, specifies which of the images is shown to which participant in which round of the task and whether they are highlighted or not. It also gives instructions to the participants, notifies them about game events like submissions or a partner’s missing selections and checks if all highlighted images are marked so that participants can continue to the next round.

Agents finally are the actors in those environments, receiving and sending messages in the modality specified by the environment, following the regime set out by the teacher. ParlAI comes with a number of pre-specified agents like for example the LocalHuman agent which relays incoming (written-text) messages to a terminal window, or the RepeatLabel agent that just echoes any received text input as its own output. The agent module has two main purposes: First and foremost it is meant as an interface between a custom dialogue model and the ParlAI framework. If a dialogue model is implemented as an agent in ParlAI, by specifying a certain world and/or selecting a specific teacher module, the framework can be used as a testbed for any of the previously mentioned dialogue tasks - or any self-designed one. And secondly, agents can be used as interface to crowd-sourcing platform Amazon Mechanical Turk (AMT). Through a fully customisable web user interface, Mechanical Turk users (called workers) can also fill in the role of an agent in a running ParlAI dialogue world, interacting with the other agents in the environment. Those other agents in turn can then either also be Mechanical Turk workers, with a teacher tasking participants to fulfil a specific task in order to collect novel dialogue data, or they could be an artificial dialogue agent that is then assessed on its performance in a Turing-Test like setting.

A.8 HIT Preview

When the PhotoBook task environment (called a world in ParlAI) is initialised, it publishes a specified number of Human Intelligence Tasks (HITs) titled Game: Detect common images by chatting with another player on Amazon Mechanical Turk. This means that in the AMT list of HITs and search function the PhotoBook task can be found by users (which AMT calls workers). The PhotoBook task entry can be extended through clicking on a small plus sign next to the title, showing a description that
lays out the high-level setup of the photo book comparison game and informs workers that participating in the HIT will take about 12 minutes (see Figure 2 for a full print of the descriptions).

From the HIT description, participants can either view a preview page with extended instructions or directly accept the HIT. In an initial series of runs we observed that almost all participants however skipped the preview and directly entered the HIT, so we decided to move the task’s full description to a set of instruction slides at the beginning of a HIT and just copy the most important details to the preview page as well. See Figure 3 for a print of the full preview page.
A.9 Qualifications

A last element that is displayed in the extended HIT descriptions are the qualifications needed to work on the PhotoBook task. Qualifications are AMT’s tool to filter the worker pool and exclude poor workers from HITs. AMT comes with a number of system qualifications like a worker’s location, number of completed HITs and overall acceptance rate (the rate of HITs that were deemed completed correctly by their respective requesters). For the location qualification we decided to include the default native English speaking countries only, which includes AU, CA, GB, NZ and US. Based on we then decided to set the minimum acceptance rate to 90% on more than 100 completed HITs. In addition to that we also defined two new, custom qualifications: The first qualification is assigned to a worker if he or she completed a first game of the PhotoBook task. Workers with this qualification will no longer do the warming-up round if they accept a new HIT. The second qualification is assigned to a worker if he or she completed five full games. Workers with this qualification can no longer participate in the PhotoBook task. We set this limitation because we are interested in how human interlocutors collaborate to establish canonical referring expression for the displayed visual input. Manual inspection of the first batch of collected data however revealed that frequent participants stopped to do so naturally in later games but instead started to develop a sort of code for referent objects that they directly proposed to or forced on their partners. While this meant that they could quickly and correctly complete a game, the resulting dialogue data is far less interesting for training computational dialogue agents. Through setting this maximum number of games we also wanted to prevent the dataset to be biased on certain participants like for example GuessWhat?! where the top ten participants each are responsible for more than 2,000 dialogues or CoDraw where the five most contributing participants produced more than a quarter of the data.

A.10 Matching Participants

In order to collect unbiased samples of the referring expression generation process, we aim to prevent both, 1) participants completing the same game multiple times (as here they could re-use referring expressions that worked well during the last one) and 2) specific pairs of participants completing multiple games (as they might have established some kind of strategy or code already). On the other hand, we are interested in the degree of the partner-specificity of the established canonical expressions, which can be tested by interchanging one of the interlocutors while remaining in the same game setting (compare for example Brennan and Clark (1996)). To accommodate both aspects, we decided to try to have participants that re-enter the lobby after a completed game to be matched with a new partner that also has played before, but never the specific last game of the re-entering participant. They will then be assigned the same game ID as the participant’s previous one - but with randomised speaker labels, round orders and image positions. We call this setup a second game that can be used to analyse whether and how participants

---

13 Kazemzadeh et al. (2014)
14 De Vries et al. (2017a)
15 Kim et al. (2017)
re-use descriptions from previous encounters with a specific set of referent images. In order to maximise the number of second games, we encourage workers to continue playing by paying them a bonus of 0.25 USD for each 2nd, 3rd, 4th and 5th game. Other than trying to match players for second games, we make sure that no participant plays the same game more than twice and that participants that played before are never paired again.

In order to indicate to the participants that in each game they play with a different partner, we automatically select player names based on the number of games that their partner has completed. As a result, participants will see a different partner’s name in each round. The list of names that can be selected from are randomly picked, gender-neutral names.

A.11 Worker Payment

The HIT description also details the worker’s payment. We want to provide fair payment to workers, which we calculated based on an average wage of 10 USD per hour\footnote{Hara et al. (2018), DynamoWiki}. An initial set of runs resulted in an average completion time of 12 minutes, which indicated an expected expense of about 2 USD per participant per game. More experienced workers however managed to complete a full game in six to ten minutes, meaning that for them we would often surpass the 10 USD/h guideline based on this calculation. Other workers - especially new ones - took up to 25 minutes for the first game, which means that they on the other hand would be strongly under-paid with a rigid per-game payment strategy. To mitigate this effect, we developed the following payment schema: Each worker that completes a full game is payed a base-amount of 1.75 USD - which is indicated in the HIT description. If the game took longer than ten minutes, the participants are payed a bonus amount of 0.10 USD per minute, up to an additional bonus payment of 1.50 USD for 25 or more minutes. In order to not encourage workers to play slowly, we only inform them about this bonus at the end of a HIT. With this bonus and the 20% AMT fee on each transaction, we expect an average cost of about 5 USD per game.

We specified the following guideline for disconnects: If one of the participants disconnects during the first two rounds, the other participant can return the HIT without any consequences and the data is discarded. If a participant disconnects in a later round, the remaining worker is compensated in full but the data is discarded as well. Players are disconnected automatically if they are idle for more than 2:30 minutes.

A.12 Recording Conversations

While participants communicate through the AMT user interface, all their messages are relayed through ParlAI and can be parsed an processed in the framework. As our main goal is data collection, we record all user messages and actions and save them to file. To do so, the task world logs all incoming messages in a custom conversation\_log. The conversation
log holds six static fields that contain information with regard to the
game as a whole, being 1) a unique game ID constructed from a concate-
nation of the worker’s assignment IDs (which are always unique due to
the pairing function), 2) the image set ID, 3) the player IDs, 4) the player
labels, 5) the player’s assignment IDs, 6) the player’s feedback. Besides
those static fields, it also contains a list of round_log objects. Each round
log in turn has three static fields, 1) the round number 2) the displayed
images and 3) the player’s scores, and list of message_log objects. Messages finally are logged with 1) their timestamp, 2) the turn number, 3)
the speaker label, 4) the speaker’s ID and 5) the message content. In
order to record agent actions as well, all selections and button clicks are
also parsed and stored as messages with specific tags like <selection>,
<feedback> and <next_round>. After each game round, the current state
of the conversation log is written to file, directly encoding the object
structure in a JSON file.\footnote{https://www.json.org/}
### B Dataset Analysis

#### Table 1: Game Round Completion Times in seconds, their respective differences and the t-test p-values showing significant decreases with respect to the previous round.

<table>
<thead>
<tr>
<th>Round</th>
<th>Duration</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>179.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>139.59</td>
<td>-39.99</td>
<td>1.37E-52</td>
</tr>
<tr>
<td>3</td>
<td>120.94</td>
<td>-18.65</td>
<td>8.61E-16</td>
</tr>
<tr>
<td>4</td>
<td>108.03</td>
<td>-12.91</td>
<td>4.29E-09</td>
</tr>
<tr>
<td>5</td>
<td>101.20</td>
<td>-6.83</td>
<td>3.93E-03</td>
</tr>
</tbody>
</table>

#### Table 2: Game Completion Times in seconds, their respective differences and the t-test p-values for up to five subsequent games per participant.

<table>
<thead>
<tr>
<th>Game</th>
<th>Duration</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>954.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>826.63</td>
<td>-139.07</td>
<td>9.06E-22</td>
</tr>
<tr>
<td>3</td>
<td>810.28</td>
<td>-16.34</td>
<td>2.71E-01</td>
</tr>
<tr>
<td>4</td>
<td>782.80</td>
<td>-27.48</td>
<td>7.82E-02</td>
</tr>
<tr>
<td>5</td>
<td>778.69</td>
<td>-4.12</td>
<td>8.10E-01</td>
</tr>
</tbody>
</table>

#### Table 3: Game Completion Times in seconds, their respective differences and the t-test p-values for five consecutive games from participants who completed the full limit of five games.

<table>
<thead>
<tr>
<th>Game</th>
<th>Duration</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>954.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>826.63</td>
<td>-139.07</td>
<td>9.06E-22</td>
</tr>
<tr>
<td>3</td>
<td>810.28</td>
<td>-16.34</td>
<td>2.71E-01</td>
</tr>
<tr>
<td>4</td>
<td>782.80</td>
<td>-27.48</td>
<td>7.82E-02</td>
</tr>
<tr>
<td>5</td>
<td>778.69</td>
<td>-4.12</td>
<td>8.10E-01</td>
</tr>
</tbody>
</table>

#### Table 4: Average Round Scores in points out of 6, their respective differences and the t-test p-values.

<table>
<thead>
<tr>
<th>Round</th>
<th>Score</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5.71</td>
<td>0.05</td>
<td>4.33E-03</td>
</tr>
<tr>
<td>3</td>
<td>5.73</td>
<td>0.02</td>
<td>3.26E-01</td>
</tr>
<tr>
<td>4</td>
<td>5.73</td>
<td>0.01</td>
<td>7.22E-01</td>
</tr>
<tr>
<td>5</td>
<td>5.74</td>
<td>0.01</td>
<td>7.02E-01</td>
</tr>
</tbody>
</table>

#### Table 5: Average Game Scores in points out of 30, their respective differences and the t-test p-values for up to five subsequent games per participant.

<table>
<thead>
<tr>
<th>Game</th>
<th>Score</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>28.51</td>
<td>-0.05</td>
<td>5.33E-01</td>
</tr>
<tr>
<td>3</td>
<td>28.53</td>
<td>0.02</td>
<td>7.68E-01</td>
</tr>
<tr>
<td>4</td>
<td>28.68</td>
<td>0.15</td>
<td>5.20E-02</td>
</tr>
<tr>
<td>5</td>
<td>28.70</td>
<td>0.02</td>
<td>7.99E-01</td>
</tr>
</tbody>
</table>

#### Table 6: Average Game Scores in points out of 30, their respective differences and the t-test p-values for five consecutive games from participants who completed the full limit of five games.

<table>
<thead>
<tr>
<th>Game</th>
<th>Score</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>28.56</td>
<td>-0.12</td>
<td>1.95E-01</td>
</tr>
<tr>
<td>3</td>
<td>28.52</td>
<td>-0.04</td>
<td>6.80E-01</td>
</tr>
<tr>
<td>4</td>
<td>28.62</td>
<td>0.10</td>
<td>2.77E-01</td>
</tr>
<tr>
<td>5</td>
<td>28.75</td>
<td>0.13</td>
<td>1.32E-01</td>
</tr>
</tbody>
</table>

#### Table 7: Total number of messages sent per game round, their respective differences and the t-test p-values. Statistics obtained from the unfiltered dataset.

<table>
<thead>
<tr>
<th>Round</th>
<th># Messages</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>13.28</td>
<td>-1.51</td>
<td>6.61E-29</td>
</tr>
<tr>
<td>3</td>
<td>12.56</td>
<td>-0.71</td>
<td>4.18E-09</td>
</tr>
<tr>
<td>4</td>
<td>12.20</td>
<td>-0.37</td>
<td>1.45E-03</td>
</tr>
<tr>
<td>5</td>
<td>12.87</td>
<td>0.67</td>
<td>4.93E-08</td>
</tr>
</tbody>
</table>
### Table 8: Points per Message Ratios for Subsequent Game Rounds

<table>
<thead>
<tr>
<th>Round</th>
<th># Messages</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.62</td>
<td>-1.20</td>
<td>2.55E-27</td>
</tr>
<tr>
<td>2</td>
<td>10.42</td>
<td>-0.54</td>
<td>4.20E-08</td>
</tr>
<tr>
<td>3</td>
<td>9.89</td>
<td>-0.28</td>
<td>3.21E-03</td>
</tr>
<tr>
<td>4</td>
<td>9.61</td>
<td>-0.20</td>
<td>4.17E-02</td>
</tr>
<tr>
<td>5</td>
<td>9.42</td>
<td>-0.14</td>
<td></td>
</tr>
</tbody>
</table>

### Table 9: Points per Token Ratios for Subsequent Game Rounds

<table>
<thead>
<tr>
<th>Round</th>
<th># Tokens</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94.92</td>
<td>-16.35</td>
<td>4.52E-49</td>
</tr>
<tr>
<td>2</td>
<td>78.57</td>
<td>-10.95</td>
<td>1.14E-28</td>
</tr>
<tr>
<td>3</td>
<td>67.62</td>
<td>-8.10</td>
<td>6.27E-19</td>
</tr>
<tr>
<td>4</td>
<td>59.52</td>
<td>-6.00</td>
<td>3.29E-02</td>
</tr>
<tr>
<td>5</td>
<td>53.52</td>
<td>-3.51</td>
<td>1.34E-15</td>
</tr>
</tbody>
</table>

### Table 10: Token/Message Ratios of Subsequent Game Rounds in the PhotoBook Conversation Task

<table>
<thead>
<tr>
<th>Round</th>
<th>T/M Ratio</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.25</td>
<td>-0.71</td>
<td>5.65E-25</td>
</tr>
<tr>
<td>2</td>
<td>7.55</td>
<td>-0.74</td>
<td>6.17E-28</td>
</tr>
<tr>
<td>3</td>
<td>6.81</td>
<td>-0.64</td>
<td>9.78E-22</td>
</tr>
<tr>
<td>4</td>
<td>6.17</td>
<td>-0.51</td>
<td>3.87E-15</td>
</tr>
</tbody>
</table>

### Table 11: Token/Message Ratios of Subsequent Game Rounds in the PhotoBook Conversation Task

<table>
<thead>
<tr>
<th>Round</th>
<th>C/T Ratio</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.48</td>
<td>-0.02</td>
<td>1.80E-09</td>
</tr>
<tr>
<td>2</td>
<td>0.49</td>
<td>-0.02</td>
<td>9.12E-11</td>
</tr>
<tr>
<td>3</td>
<td>0.51</td>
<td>-0.02</td>
<td>2.55E-08</td>
</tr>
<tr>
<td>4</td>
<td>0.53</td>
<td>-0.01</td>
<td>3.87E-05</td>
</tr>
</tbody>
</table>

### Table 12: Average Unique Token Count per Game Round after Removing All Messages Sent after the Last Selection in Round 5

<table>
<thead>
<tr>
<th>Game</th>
<th># Unique Tokens</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57.73</td>
<td>-7.46</td>
<td>2.30E-46</td>
</tr>
<tr>
<td>2</td>
<td>50.27</td>
<td>-5.96</td>
<td>2.69E-35</td>
</tr>
<tr>
<td>3</td>
<td>44.31</td>
<td>-4.32</td>
<td>1.36E-21</td>
</tr>
<tr>
<td>4</td>
<td>39.98</td>
<td>-3.80</td>
<td>3.30E-06</td>
</tr>
<tr>
<td>5</td>
<td>32.18</td>
<td>-3.24</td>
<td></td>
</tr>
</tbody>
</table>

### Table 13: Number of New Word Tokens per Subsequent Game Round, Their Respective Differences and the t-test p-values. All Messages Sent after the Last Selection in Each Round Were Filtered Out

<table>
<thead>
<tr>
<th>Round</th>
<th># New Tokens</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51.14</td>
<td>-30.66</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>2</td>
<td>20.49</td>
<td>-10.95</td>
<td>1.83E-21</td>
</tr>
<tr>
<td>3</td>
<td>11.51</td>
<td>-8.97</td>
<td>1.16E-09</td>
</tr>
<tr>
<td>4</td>
<td>7.54</td>
<td>-2.04</td>
<td>1.20E-23</td>
</tr>
<tr>
<td>5</td>
<td>5.50</td>
<td>-1.35</td>
<td></td>
</tr>
</tbody>
</table>

### Table 14: Points per Token Ratios for Subsequent Game Rounds. All Messages Sent after the Last Selection in Each Round Were Filtered Out

<table>
<thead>
<tr>
<th>Game</th>
<th>Points/Token</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.07</td>
<td>-0.02</td>
<td>8.26E-34</td>
</tr>
<tr>
<td>2</td>
<td>0.09</td>
<td>-0.02</td>
<td>2.42E-32</td>
</tr>
<tr>
<td>3</td>
<td>0.11</td>
<td>-0.02</td>
<td>5.10E-23</td>
</tr>
<tr>
<td>4</td>
<td>0.14</td>
<td>-0.02</td>
<td>4.68E-15</td>
</tr>
</tbody>
</table>

### Table 15: Points per Message Ratios for Subsequent Game Rounds. All Messages Sent after the Last Selection in Each Round Were Filtered Out

<table>
<thead>
<tr>
<th>Game</th>
<th>Points/Message</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.54</td>
<td>-0.06</td>
<td>1.17E-29</td>
</tr>
<tr>
<td>2</td>
<td>0.60</td>
<td>-0.03</td>
<td>8.75E-09</td>
</tr>
<tr>
<td>3</td>
<td>0.65</td>
<td>0.02</td>
<td>8.31E-05</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
<td>0.02</td>
<td>3.24E-03</td>
</tr>
</tbody>
</table>

### Table 16: Points per Message Ratios for Subsequent Game Rounds. All Messages Sent after the Last Selection in Each Round Were Filtered Out
### Table 17: Total number of messages sent per consecutive game, their respective differences and the t-test p-values for five consecutive games from participants who completed the full limit of five games.

<table>
<thead>
<tr>
<th>Game</th>
<th># Messages</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>65.91</td>
<td>-1.45</td>
<td>1.34E-01</td>
</tr>
<tr>
<td>3</td>
<td>65.24</td>
<td>-0.67</td>
<td>4.31E-01</td>
</tr>
<tr>
<td>4</td>
<td>65.18</td>
<td>-0.05</td>
<td>9.49E-01</td>
</tr>
<tr>
<td>5</td>
<td>65.51</td>
<td>0.32</td>
<td>7.01E-01</td>
</tr>
</tbody>
</table>

### Table 18: Total number of tokens uttered per consecutive game, their respective differences and the t-test p-values for five consecutive games from participants who completed the full limit of five games.

<table>
<thead>
<tr>
<th>Game</th>
<th># Tokens</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>475.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>403.87</td>
<td>-72.08</td>
<td>3.67E-13</td>
</tr>
<tr>
<td>3</td>
<td>383.06</td>
<td>-20.81</td>
<td>1.85E-02</td>
</tr>
<tr>
<td>4</td>
<td>378.63</td>
<td>-4.43</td>
<td>6.09E-01</td>
</tr>
<tr>
<td>5</td>
<td>378.62</td>
<td>-0.02</td>
<td>9.98E-01</td>
</tr>
</tbody>
</table>

### B.1 Universal POS-Tags

SectionADJ: adjective  
ADP: adposition  
ADV: adverb  
AUX: auxiliary  
CCONJ: coordinating conjunction  
DET: determiner  
INTJ: interjection  
NOUN: noun  
NUM: numeral  
PART: particle  
PRON: pronoun  
PROPN: proper noun  
PUNCT: punctuation  
SCONJ: subordinating conjunction  
SYM: symbol  
VERB: verb  
X: other

### C Evaluation

```plaintext
for each game transcript:
    initialise empty candidate dictionary
    for each selection in a game_round, for each game_round in a game transcript:
        add the previous 3 utterances per speaker to the candidate dictionary
    initialise empty reference dictionary
    for each image ID, list of candidates in candidate dictionary:
        referring expressions = list of COCO captions of the image
        for each utterance in candidate group, for each candidate group in candidates:
            invert referring expressions list
            for RE in referring expressions:
                score += weight factor * similarity(utterance, RE)
                weight factor *= weight factor
            divide score by the length of the referring expressions list
            if score > threshold:
                add utterance to local referring expressions list
                add utterance to global reference dictionary
```

Listing 1: Heuristics for matching referring expressions and their respective referent image.