Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Bachelorarbeit Kognitionswissenschaft

Ambiguities in Dialogue: The Strategic Use of Ambiguous Utterances in Dialogues and the Rational Speech Act Model

Ella Isabel Eisemann

28. November 2019

Gutachter

Prof. Dr. Martin V. Butz
Kognitive Modellierung
Wilhelm-Schickard-Institut für Informatik
Universität Tübingen

Betreuerin

Dr. Asya Achimova
Kognitive Modellierung
Wilhelm-Schickard-Institut für Informatik
Universität Tübingen

Eisemann, Ella Isabel:
Ambiguities in Dialogue:
The Strategic Use of Ambiguous Utterances in Dialogues and the Rational Speech Act Model
Bachelorarbeit Kognitionswissenschaft
Eberhard Karls Universität Tübingen
Bearbeitungszeitraum: 01.08.2019 – 01.12.2019

Abstract

Natural languages are highly ambiguous. From a theoretical standpoint, the presence of ambiguity might pose challenges for communication, yet humans resolve ambiguity effortlessly and often without even noticing it. Recent research showed evidence for purposes of ambiguity in natural languages such as efficiency (Piantadosi *et al.*, 2012) and solving variability dilemmas (Erev *et al.*, 1991). Yet mostly very abstract and contextless scenarios were examined. This thesis sheds light on modeling human behavior in dialogues, in order to tackle this problem.

This study investigated human pragmatic reasoning about referential ambiguity in dialogues and an RSA model (Rational Speech Act Model) predicting this reasoning. An online experiment was conducted to engage in this topic in continuation of Scontras *et al.* (2019). In the course of this experiment the following was examined: (1) strategic human choices of referentially ambiguous utterances for information retrieval in dialogues, and (2) the hereby following information retrieval from the disambiguating behavior of the interlocutor. Parallel to that an RSA model modeling pragmatic reasoning for (2) was developed and tested for the experimental data. Results indicate that humans can reason rationally and pragmatically about referential ambiguity and its resolving in dialogue. Furthermore, the results imply that the RSA model fits the data well.

These findings provide broader insights into human reasoning and language use. They suggest a further investigation of human behavior with ambiguity and a further development of the RSA model.

Kurzfassung

Natürliche Sprachen sind sehr ambig (mehrdeutig). Von einem theoretischen Standpunkt aus, müsste Ambiguität (Mehrdeutigkeit) eine Herausforderung für Kommunikation sein. Jedoch wird Ambiguität von Menschen mühelos und oft sogar unbemerkt aufgelöst. Neusten Forschungen ist es gelungen zu zeigen, dass Ambiguität in Sprache mehrere Zwecke erfüllen kann, wie zum Beispiel kommunikative Effizienz (Piantadosi *et al.*, 2012) und das Lösen sozialer Dilemmas (Erev *et al.*, 1991). Da in diesen Forschungen meist abstrakte und kontextfreie Szenarios untersucht wurden, beschäftigt sich diese Arbeit mit der Modellierung von menschlichem Verhalten mit Ambiguität in Dialogen.

Die Studie untersuchte pragmatisches, logisches Nachdenken von Menschen über referentielle Ambiguität in Dialogen. Ein RSA Modell (Rational Speech Act Model) wurde entwickelt, um dieses Denken zu modellieren. Es wurde ein Online Experiment ähnlich dem von Scontras *et al.* (2019) durchgeführt, um das Thema zu beleuchten. In diesem Experiment wurden zwei Aspekte von Ambiguität in Sprache untersucht: (1) die strategische Wahl von referentiell ambigen Aussagen, um Informationen in Dialogen herauszufinden und (2) die daraus folgende Ableitung von Informationen vom disambiguierenden Verhalten des:der Gesprächspartner:in. Parallel dazu wurde ein RSA Modell für (2) entwickelt und mit den experimentellen Daten getestet. Die Ergebnisse zeigen, dass Menschen rational und pragmatisch über referentielle Ambiguität und deren Auflösung nachdenken können, sowie dass das RSA Modell dieses Verhalten und die Daten gut modelliert.

Mit jenen Ergebnissen ermöglicht diese Arbeit weitere Erkenntnisse über menschliches Denken und Sprachgebrauch und regt weitere Forschung über menschliches Verhalten mit Ambiguität und eine Weiterentwicklung des RSA Modells an.

Acknowledgments

On the way to the completion of this thesis many people offered advice and helped me getting to the point of presenting this work. I would like to thank Asya Achimova for her great supervision and assistance. She helped me wherever I had questions and supported me for the whole time.

I would like to thank Prof. Martin Butz for giving me the opportunity to develop this thesis on this very interesting field of research, and for his encouraging feedback along the journey.

Thanks to Michael Schiller and Sara Fiedler for supporting me along the way and in the long hours in the library.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The key feature of human language is in some way all about understanding what the other person is trying to convey and not about what they are literally saying. To get to this understanding is one goal of computational linguistics and psycholinguistics. Many obstacles have to be overcome before this goal can be reached. One of them is the understanding of use, purpose and resolution of ambiguity in natural languages[1].

Traditionally, the study of language is separated into different areas, such as semantics, syntactics and pragmatics. Pragmatics in this distinction analyses language in the context of a situation in which utterances are made. This incorporates the relation between speaker and listener and the speaker's knowledge, predictions and beliefs. The phenomenon of ambiguity, which refers to "a word or sentence which expresses more than one meaning", is falling under the category of pragmatics (Crystal, 2015). Ambiguity can make it hard for a listener to comprehend what the speaker truly wanted to convey. Therefore, it is an important topic in linguistic research. Ambiguity occurs in various form and mostly arises in context. Puns and many jokes for example only are funny because the ambiguity in the pun or joke can be resolved into two meanings. For example in "7 days without pizza makes one weak." the word "weak" sounds the same as "week" and therefore creates the pun (see also Gao and Ren, 2013).

Computational cognitive modeling is an approach in cognitive science and other disciplines to predict human behavior like the understanding and production of puns. Also other cases of intentional use of ambiguity can be predicted with cognitive models. This modeling aims to understand behavioral data and, more generally, the mind and brain by building computational models of the cognitive processes that produce the behavior.

This thesis aims to contribute to the literature by providing a cognitive model suitable for a simulation for the exploitation of ambiguity resolving in dialogues[2]. In doing so, this approach offers a tentative answer to the question of how much of

[1] Natural languages are languages that occur and are used as native tongues by a group of speakers on earth.

[2] *Dialogues* are defined here as utterance-response-chains which happen in a certain context. Abstractly a dialogue is an exchange of ideas and opinions between two or more people.

human reasoning is modelable with Bayesian probabilistic modeling. The second contribution is to provide insights into the mechanisms of human reasoning as well as into ambiguity and its purpose in natural languages.

This topic is located at the intersection of psycho-linguistics, computational linguistics and cognitive science with a cognitive modeling approach. With the conduction of an experiment and the development of a model, the expectations were met and the datafitting was successful. The findings offer implications for research on the handling, use and modeling of ambiguity, as well as the power of Bayesian probabilistic models in general. They additionally point out which approaches appear promising for the future.

In the following, previous research on ambiguity and RSA models (Rational Speech Act Models) will be discussed, while questions, that still remain unanswered, will be elaborated. Then, the research questions that were taken up in this thesis will be set forth, alongside a depiction of how the model and experiment have been constructed and conducted. Subsequently the results of the model fitted to the data will be presented and discussed.

Chapter 2

Foundations

2.1 Ambiguity

Ambiguity and vagueness are terms which are often used conterminous. However, despite the fact that the distinction is sometimes difficult, there is an important difference: ambiguity means that one expression has multiple distinct and unrelated meanings. Vagueness means that one expression has one imprecise meaning which cannot be grasped completely. Vagueness and ambiguity often occur together (see Wasow, 2015). Following the idea of Gärdenfors (2014) who proposes that meanings are regions in one or multiple conceptual spaces, ambiguous expressions refer to multiple regions in multiple spaces and vague expressions refer to a not clearly confined region in one conceptual space (see Wasow, 2015).

There are different types of ambiguity discussed in linguistics. There is *lexical/semantic ambiguity* where a word has more than one possible meaning (e.g. "bank"). Then, there is *structural/grammatical/syntactic ambiguity* where the construction of the expression implies more than one possible interpretation (e.g. "Visiting relatives can be boring"). Then, there is *scope ambiguity* that arises from multiple possible relations of scope-bearing elements. In the example "Every man loves a woman" these would be quantifier phrases ("every man") and indefinites ("a woman"). And then, there is *referential ambiguity* where a linguistic expression might potentially refer to more than one referent (object, person, etc.). This type of ambiguity can occur within expressions when pronouns are involved (e.g. "The girl told her mother of the theft. *She* was very upset"), but also between an expression of a full noun phrase and its context (e.g. A situation were two cows are standing on a meadow and somebody says "Look at the cow!") (see Hirst, 1987). In the literature still more forms of ambiguity can be found (see Macagno and Bigi, 2018).

This thesis focuses on referential ambiguity and its resolution within context.

2.2 Ambiguity in Communication

Reading about ambiguity in most cases quite directly leads to Grice and his conversational maxims, underlying the *Cooperative Principle*, which was proposed in a lecture at Harvard University in 1967 (Grice, 1975, p. 46). Following the idea to find parameters to define good conversation behavior, he wrote: “Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.” In the cooperative principle Grice (1975) introduces four super-maxims. Following these should lead to the conversational behavior which he supposes to be necessary for good human linguistic interaction: The maxim of quantity, the maxim of quality, the maxim of relation and the maxim of manner. Under the maxim of manner Grice comprehends: “Be perspicuous. Avoid obscurity of expression. Avoid ambiguity. Be brief (avoid unnecessary prolixity). Be orderly” Grice (1975, p. 46). The notion of avoiding ambiguity as schemed in the Cooperative Principle is quite intuitive at first sight. Why should somebody make an ambiguous utterance if they could just say what they mean? Therefore, humans should avoid ambiguity as much as possible when using language with the intention to communicate cooperatively. Interestingly this does not correspond to human behavior (Wasow, 2015). In fact humans use deliberately ambiguous language and even humans, which are highly concentrated on communication, actively choose ambiguity. This might seem drastic but actually most words in dictionaries have multiple (related and not-related) meanings and are therefore ambiguous per se (Wasow, 2015). There are also other naturally ambiguous traits in many human languages for example some grammatical structures which allow different interpretations (Hirst, 1987). One possible interpretation of this phenomenon would be that this vast presence of ambiguity in language is basically just a bug and maybe the humankind is just not smart enough. Or one could agree with Chomsky and ask oneself:

> The use of language for communication might turn out to be a kind of epiphenomenon. I mean, the system developed however it did, we really don’t know. And then we can ask: how do people use it? [...] If you want to make sure that we never misunderstand one another, for that purpose language is not well designed, because you have such properties as ambiguity.
> (Chomsky *et al.*, 2002, p.107).

There is also a third solution for this ostensible paradox. Ambiguity could have, and actually has, purpose for cooperative communication. But this cooperativeness might not be visible at first sight. Already in 1949 Zipf developed *The Principle Of Least Effort*:

> For convenience, we shall use the term least effort to describe the [...] least average rate of probable work. We shall argue that an individual’s

> entire behavior is subject to the minimizing of effort. Or, differently stated, every individual's entire behavior is governed by the Principle of (collective) Least Effort. (p. 6)

Applied to ambiguity in language speakers would then try to minimize the overall effort of speaking with the goal of transmitting the message. Speakers therefore need to consider how much context and other factors can explain and what additional information has to be given. Oriented on Zipf, Piantadosi *et al.* (2012) argue that indeed (only) ambiguity permits efficiency and thus must be inherent to every communicative efficient system, if the context contains information about meaning. They state:

> [W]e argue that ambiguity can be understood by the trade-off between two communicative pressures which are inherent to any communicative system: *clarity* and *ease*. A *clear* communication system is one in which the intended meaning can be recovered from the signal with high probability. An *easy* communication system is one which signals are efficiently produced, communicated, and processed. (p. 281)

Erev *et al.* (1991) conducted social experiments looking at linguistic communities. In their research social dilemmas were often solved, while thereby creating the best possible outcome for the community, by uttering ambiguous and vague statements. Imprecise statements were understood differently and this benefited society under certain circumstances by leading to heterogeneous choices. This behavior often helped the community but not in every case the individual.

Conflict avoidance and politeness can also be considered as an important part and maybe even as a precondition for human cooperation (Brown, 2015). Good behavior means to take part in the socio-cultural conversation-"dance": to be polite, to be friendly, to only talk about appropriate topics, to not be too curious and to ask only questions which are appropriate in the situation. If one still wants to find out about something, then ambiguous questions or statements can solve this dilemma. If for example person X meets someone at a party and the discussion is about politics, probably X wants to know first what position the interlocutor Y has. X could ask directly what this person thinks about, for example, Angela Merkel but maybe this is not the best move, as X still wants to be polite. Instead X could say something like "Merkel has been chancellor for quite a long time, right?". This would leave open what X thinks about Merkel while not being perceived as a direct question. It would however probably provoke a positioning of Y (saying either something like "Yeah, your right she did a great job!" or "Ugh, yes you are right, there needs to be change!").

These examples and many more can present a purposeful use of ambiguity (see Wasow, 2015). Scontras *et al.* (2019) demonstrated in two experiments that humans are able to pragmatically reason about why and how others resolve ambiguity and extract information from that reasoning. Further, the study revealed that speakers are able to prepare the situation for the use of the previously explained strategy. This leaded to speakers strategically selecting those utterances, which were most likely to add information to their knowledge of the interlocutor's preferences. These most information-provoking utterances were the ambiguous ones.

Generally speaking different theories about the presence of ambiguity in natural languages exist and especially more recent research focused mainly its the purposes. Ambiguity has in fact a variety of different purposes for communication. The one to find out about the interlocutors beliefs, preferences and states of mind by producing ambiguous expressions and observing the disambiguation process of the interlocutor will be the main focus in this thesis.

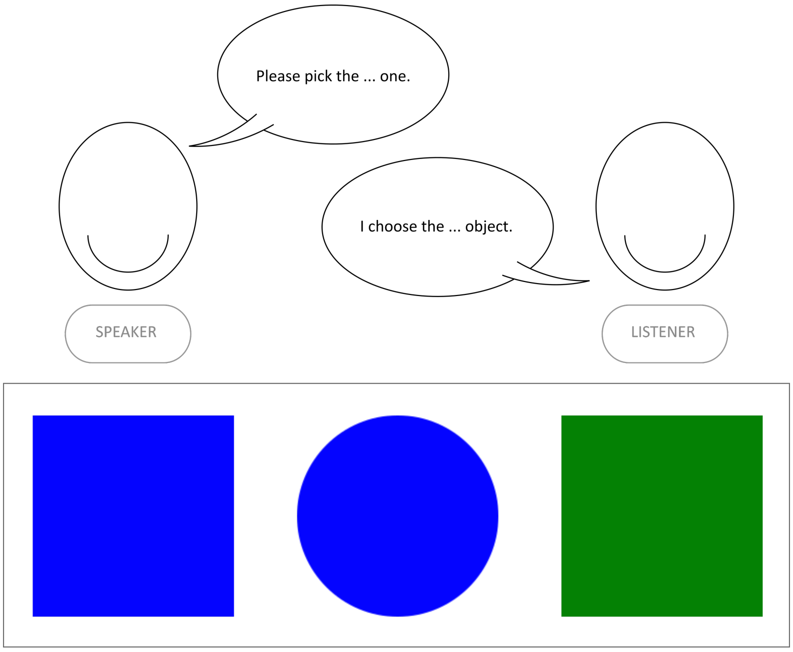


Figure 2.1: Basis mechanism of most reference games used in the empirical research around ambiguity and the RSA-model.

2.3 Cognitive Modelling and RSA

Before diving into the research on cognitive modeling and RSA models (computational) cognitive modeling needs to be defined:

> Research in computational cognitive modeling, or simply computational psychology, explores the essence of cognition (broadly defined, including motivation, emotion, perception, and so on) and various cognitive functionalities through developing detailed, process-based understanding by

> specifying corresponding computational models (in a broad sense) of representations, mechanisms, and processes. It embodies descriptions of cognition in computer algorithms and programs, based on computer science (Turing 1950).
> (Sun, 2008, p.2)

Box 1: RSA-model by Goodman and Frank (2016), Details

The formalization of the RSA-model is based on few Bayesian equations. In this version the following four equations constitute the framework.
The handled variables are u the utterance and s the state of world referred to by u.
The pragmatic listener infers the state of the world (in this case the relevance of the objects) over which utterance was chosen by the speaker.

$$P_L(s|u) \propto P_S(u|s)P(s) \tag{2.1}$$

The speaker chooses the utterance rationally to maximize the utility gain $U(u;s)$. The degree to which the speaker maximizes their utility is given by α. $U(u;s)$ describes how certain a literal listener is about s after hearing u.

$$P_S(u|s) \propto \exp(\alpha U(u;s)) \tag{2.2}$$

$$U(u;s) = \log P_{Lit}(s|u) \tag{2.3}$$

The simple listener is a non-pragmatic listener who interprets the utterance literally and updates their beliefs accordingly. ($[[u]]$ is one instance of u.)

$$P_{Lit}(s|u) \propto \delta_{[[u]](s)} P(s) \tag{2.4}$$

Cognitive modeling is a very promising and successful approach to understand and predict human behavior. When talking about predicting language use and social behavior the topic of use of ambiguity and disambiguation comes up. The RSA-modeling framework developed by Frank and Goodman (2012) investigates the prediction of pragmatic reasoning in language games. This framework in its earliest version allows accurate prediction of disambiguation, rationally solving referential ambiguity and using referential expressions. Frank and Goodman used Bayesian inference to model the behavior of listener and speaker in the reference games. Goodman and Frank (2016) then published an updated and empirically tested version of the RSA-modeling framework showing empirically that the model predictions were accurate. For more details on their model see Box 1. The basis of the framework are the Gricean maxims or the cooperative principle. In this tradition one of the mechanisms reviewed was disambiguation and the use of disambiguating utterances based on (predicted) contextual information. The purposeful use of ambiguity, clearly also a feature of human language, instead was not reviewed.

To take up this rather new research topic - the strategic use of ambiguity and ambiguous utterances and the modeling of these phenomena with the RSA-modeling framework - Scontras *et al.* (2019) successfully varied the RSA-framework to meet the new requirements. For details on their version see Box 2.

The general scenario displayed in the reference games in most of the current research on modeling human behavior with ambiguity is a variation of this situation: two entities (a speaker, and a listener) and a couple of objects. A schema of this game is sketched in Figure 2.1. The speaker then refers to one or more of the objects by uttering one of the features present in the objects. After that, the listener picks one of the objects, preferably one referred to previously by the speaker. The task for the speaker can vary, as well as the speakers reasoning. Also, the listeners reasoning differs in the variations of the reference games. More actions can also be added.

While not using reference games like described here, Nichols (nd) developed and tested a dialogic RSA model (dRSA). He accurately predicted human behavior in a dialogue-like game with this dRSA model. The model also considered the order of utterances and was able to find the optimal sequence of utterances. Unfortunately, only little is known about this research, wherefore most of the possible insights remain unknown. Still his work supports the presumption that a model based on RSA can predict human behavior also in dialogues.

Box 2: simple RSA model by Scontras *et al.* (2019), Details

After testing both versions of the RSA model - the fully pragmatic one (with a pragmatic listener) and the simple one (only with a literal listener) - Scontras *et al.* (2019) made the observation that the performance of the simple one was even better. This simple RSA model is presented in the following.
The handled variables are u the utterance, s the state of world referred to by u and f the listener's preferences for relevant instances.
The simple listener updates their beliefs about s depending on u and their preferences f for instances of s.

$$P_{L_0}(s|u,f) \propto [[u]](s) \cdot P(s|f) \tag{2.5}$$

The simple pragmatic speaker chooses the utterance u to describe s maximizing the utility of u inferring about how the literal listener (L_0) would interpret s depending on u. This pragmatic speaker considers also the cost of u with $C(u)$, which represents possible social costs uttering certain things. The softmax scaling parameter α controls how much the speaker wants to be understood i.e. how much they maximize utility when choosing utterances (default $\alpha = 1$).

$$P_{S_1}(u|s) \propto exp(\alpha \cdot [log(P_{L_0}(s|u)) - C(u)]) \tag{2.6}$$

The observer observes both the utterance choice (by the speaker) and the object choice (by the listener) and can then infer the listeners preferences based on this current information $P_{L_0}(s|u,f)$ and previously gathered knowledge about the listener $P(f)$.

$$P_{S_2}(f|s,u) \propto P_{L_0}(s|u,f) \cdot P(f) \tag{2.7}$$

The utterance choice can also be optimized similar to $P_{S_2}(f|s,u)$. Considering previously gathered information about the listeners preferences the utterance choice can be optimized to maximize the information gain while also considering utterance costs $C(u)$.

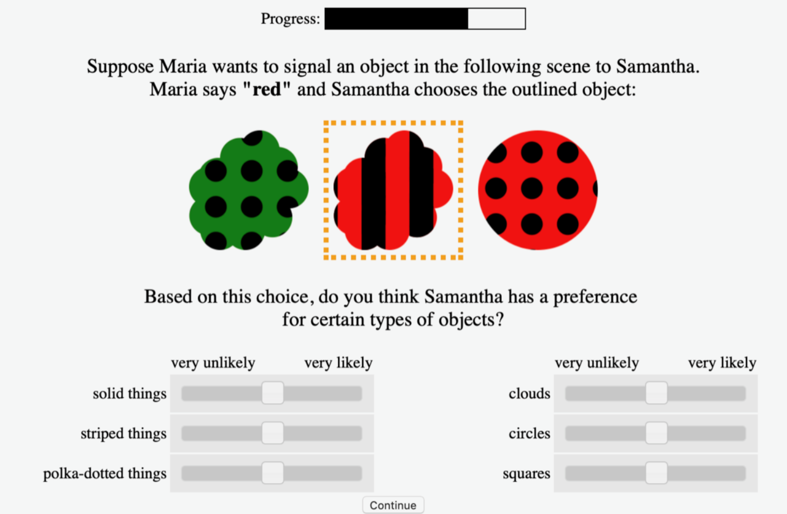
$$P_b(u) \propto \sum_{s:[[u]](s)} \lambda \cdot \mathrm{KL}(P(f), P_{S_2}(f|s,u)) - C(u) \tag{2.8}$$

Chapter 3

Research Question

The main questions which were tried to answer in this thesis are whether pragmatic reasoning about ambiguity and ambiguity resolving is used by humans in dialogues. Another question is whether this reasoning is predictable with an RSA model. This could lead towards a further understanding of human use of ambiguity and towards a broader understanding of the power of RSA models and their limits.

Conducting an experiment parallel to a modeling attempt with an RSA-variation - as done similarly by Scontras *et al.* (2019) - permitted a direct comparison and a possibility to address those questions. Although this topic is similar to the one examined by Scontras *et al.* (2019) there are some main differences. An example trial from one of their experiments is depicted in figure 3.1. The differences between this experiment and the experiment of the present thesis will be explained in the following and in Chapter 5 *Experiment*.



Progress:

Suppose Maria wants to signal an object in the following scene to Samantha.
Maria says "**red**" and Samantha chooses the outlined object:

Based on this choice, do you think Samantha has a preference
for certain types of objects?

Figure 3.1: This is an example trial taken from one the two experiments by Scontras *et al.* (2019).

First of all, a possible weakness of the experiments by Scontras *et al.* (2019) is that the scenarios were rather unrealistic. The story within the trials was not put into context and the task was only observational. The participants did not play an active role in the interaction and did not get any feedback. As language is a complex phenomenon, an approach to study it under more realistic circumstances might hold new insights. Therefore, this disadvantage was tried to be tackled by the approach of including an utterance choice and an observation task into one experiment, having the participant doing both tasks. Also, a feedback for the participant was included. This should create a task which might be more relatable to the lives of the participants.

The second change that was made was the concentration on only one feature and not all three of them (color, shape and texture). This helps the scenario to be more realistic, too, because it happens more often that people interact with the idea of finding out *something specific* about another person, instead of simply finding out *something*.

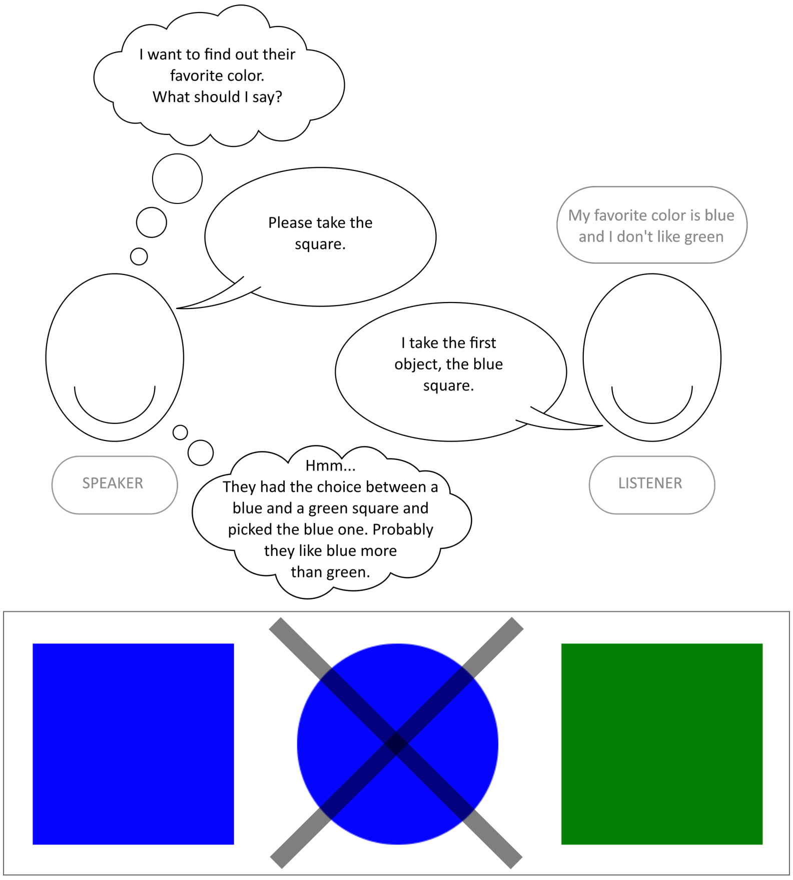

I want to find out their favorite color. What should I say?
Please take the square.
My favorite color is blue and I don't like green
I take the first object, the blue square.
SPEAKER
LISTENER
Hmm... They had the choice between a blue and a green square and picked the blue one. Probably they like blue more than green.

Figure 3.2: Example of the references game used in this study with the schematic reasoning chain to pick ambiguous utterances for information gain.

The third change was to put participants in the position of the speaker and observer who interacts in dialogues with a listener instead of simple utterance-response-tasks. The dialogues were created by having multiple utterance-response-tasks in a row, interacting with a consistent interlocutor.

The hypothetical reasoning chain in the dialogue depicted in figure 3.2 involves a pragmatic speaker/observer and a literal listener with preferences. The speaker reasons about which utterance they can make, in order to provoke the listener to reveal information about preferences with the previously gathered information in mind. If the speaker then uses an utterance which refers to more than one object - an ambiguous utterance - the literal listener has to choose. In this scenario the listener does not ponder on the speakers intentions, a possible context or anything else, but only on the objects, the utterance and they themselves. The two or more indicated objects differ in at least one feature (shape, texture or color). Because the listener has to choose between these objects, they have to choose depending on the features that differ. Therefore, the choice is based on the preferences for the feature values of the indicated objects. After observing the object pick of the listener the speaker/observer can deduce the preferences of the listener. To do that, the speaker/observer goes through a similar reasoning process as the listener thinking about which preferences would lead to this specific object choice under the present circumstances. The dialogue is created by repeating this situation with different objects. This gives the speaker the possibility to adjust their prior beliefs about the listeners preferences over and over again, always approaching the real preferences.

The reasoning in this scenario bases on certain hypotheses and provokes others. Therefore, the following hypothesis have been deriven: (1) Humans pick strategically ambiguous utterances to perform the task. (2) Only in cases where ambiguous utterances are picked (from now on also called *informative* utterances), the preferences can be learned. (3) Preference detection accuracy rises over the trials with ambiguous utterances chosen (from now on also called *informative* trials). (4) The RSA model is capable of predicting how humans perform in the observation task in dialogue-like reference games.

Chapter 4

Model

4.1 Computational cognitive modeling using Bayesian inference

Computational cognitive models allow simulating human reasoning if the model presumptions are accurate. Following the principles of Bayesian probabilistic inference, computational cognitive models can become powerful tools to study cognitive processes like learning, vision, memory, language and much more. These models have the ability to handle uncertainties about the state of the world. In inference processes under these uncertainties, prior beliefs about the object of interest are updated in the light of evidence. The priors are derived from experience and from differently learned knowledge (see Griffiths *et al.*, 2008).

These prerequisites offer a promising approach to study the strategic use of ambiguity in dialogues. Following the Gricean approach that language is a form of rational behavior, rational models should have the power to enlighten the phenomenon of ambiguity use and resolution (see Goodman and Frank, 2016).

The priors in this model are the knowledge of the speaker about what preferences the listener has. These are then updated by observing the choice by Bayesian inference. The uncertainties in this model are about the listeners preferences.

4.2 RSA model

The model used to predict the experimental data is a variation of the RSA model used by Scontras *et al.* (2019). Their variation is explained in Box 2. In their research, they introduced two models: one simple one and one fully pragmatic one. The differences are the implementations of the listener: a literal listener vs. a listener that reasons about how the speaker might reason. In accordance with Sikos *et al.* (2019), Scontras *et al.* (2019) came to the conclusion that the simpler model of RSA is better suited to predict this kind of human behavior and reasoning.

Although the prediction of the behavior and reasoning of a speaker which chooses utterances pragmatically, could be modeled using RSA, this thesis focuses - due to time constraints - on the behavior and reasoning in the observation task. The observer monitors the listeners reaction (object choice) to the speakers action (utterance referring to a set of objects). With this information and the utterance being ambiguous the observer can infer the listeners preferences for object feature values. The observer can gather information from multiple action-reaction observations with the same interlocutor. With each and every observation they form a more precise presumption about the listeners preferences.

The model can be formalized in the following way:
The simple listener updates their beliefs about the state of the world s depending on the utterance u and their preferences f for instances of s.

$$P_{L_0}(s|u, f) \propto [[u]](s) \cdot P(s|f) \tag{4.1}$$

The observer observes both the utterance choice (by the speaker) and the object choice (by the listener) and can then infer the listeners preferences based on this current information $P_{L_0}(s|u, f)$ and previously gathered knowledge about the listener $P(f)$.

$$P_{S_2}(f|s, u) \propto P_{L_0}(s|u, f) \cdot P(f) \tag{4.2}$$

The outcome of this process can then be taken again as $P(f)$ for the next observation and prediction, while the listener and their preferences remain the same. In this way iterating over sets of objects leads to more and more accurate predictions for the listeners preferences. A rising number of observations leads to rising accuracy and and if iterating infinitely, the model should be able to always exactly predict the hierarchy. This is a typical Bayesian learning process with incremental improvement of the results.

The model predictions can be optimized by changing the parameter for preference softness γ. This parameter determines how much preferred a feature value is when it is chosen. When set to $\gamma = 0$ preferences will be interpreted as absolute, whereas $\gamma \to \infty$ leads to a uniform preference-distribution.

There is also the possibility to adjust the obedience of the listener. $\Xi = 0$ leads to the literal listener completely following the instructions whereas $\Xi \to \infty$ leads to the listener only following their preferences and mind, while ignoring the instructions completely.

In the presented model γ was set to $\gamma = 1$ by sight control and $\Xi = 0$ due to the compatibility of the model to the experiment where the listener obeyed completely. Because of time constraints an optimization of these parameters was not executed but should surely be done in the future.

The model was implemented in R and tested with the data used in the experiment. The exact implementation can be viewed in Chapter 9 *Appendices: Model and Simulation*.

4.3 Predictions

The predictions for human reasoning in the observation task, which resulted from the simulation with the model, will be presented in the following.

The model predicted better performances in the observation task in blocks where more ambiguous utterances were chosen. Also performance trajectories within blocks improved further with the more extensive use of ambiguous utterances. This is congruent with the hypotheses (2) and (3).

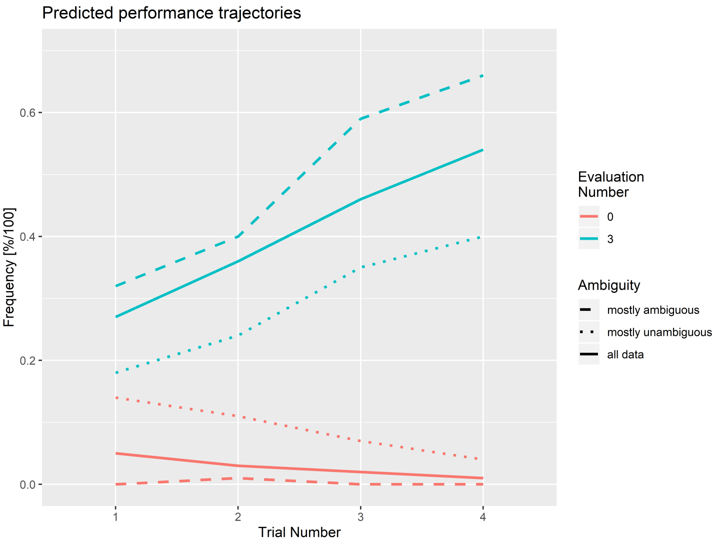


Figure 4.1: Predictions for human behavior and reasoning based on model simulation with experimental data.

In Figure 4.1 these predictions are shown. A very good performance would be a high rate of the evaluation number = 3 (blue) and a low rate of evaluation number = 0 (red). It can thus be deducted that most of the preferences are detected correctly. Running the simulation with data from participants, who picked mostly ambiguous utterances (dashed lines), the performance is better than in a simulation with the data from all participants (solid lines). The performance in the simulation with data from participants, who did not pick many ambiguous utterances (dotted lines) is worse than in the all-data-simulation. Also the trajectory of evaluation number = 3 is steeper in simulations with mostly ambiguous blocks (dashed blue line) than in both other simulations.

Chapter 5

Experiment

There are different possibilities of conducting an experiment to answer the named questions. Online experiments hold the possibility to easily manipulate a high number of variables and to easily present different kinds of stimuli. The recruitment of participants is easy as well. With these advantages in mind, the choice for this study fell onto an online experiment.

5.1 Setup

The experiment was carried out on Amazon.com's Mechanical Turk crowdsourcing service, with 100 persons with US-American IP addresses. The participants were compensated with around 2$ for their participation[1].

The participants completed the experiment on their private computers online.

The experiment was coded in HTML & CSS & JAVASCRIPT. The stimuli sets were created using PYTHON.

5.2 Participants

Of the 100 participants in the experiment 98 indicated to be English native speakers. The data of other two participants was excluded from the analysis to avoid problems related to language barriers. The vast majority claimed that they read the instructions carefully and did the task accordingly and right (97). Only 3 participants were confused by the task or did not do the task as it was intended. The data of those 3 people was excluded, because it can be surmised that they did not do the task correctly.

Of the included 95 participants, 43 identified as women, 48 identified as men, 3 identified as having another gender and 1 person did not answer. The average age was 37.91 years (median 35.00). The youngest person was 21 years old and the oldest was 68 years old. Around half of the people had graduated college (48 participants) followed by 28 who only had gone to some college and 12 who only

[1] The experiment was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 198647426.

had graduated high school (5 had a higher degree, 1 person only went to some high school and 1 person did not answer).

It took participants in average 9.3 minutes to complete the experiment (Mean 9.3, Median 8.8, Min 2.6, Max. 24.1).

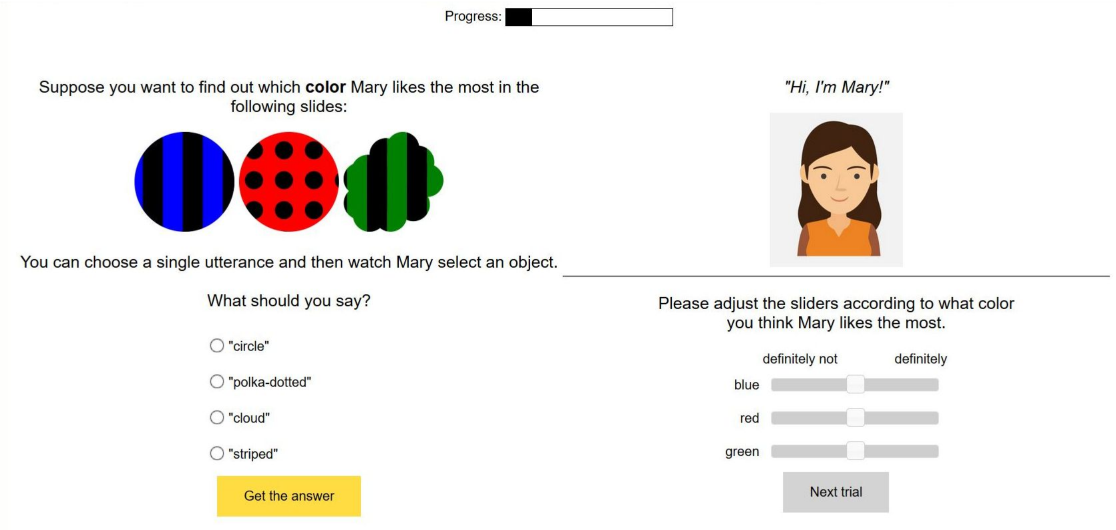
5.3 Design, Methodology and Procedure

The challenges, which the research questions posed upon the experiment, were firstly to create a dialogue-like setting in the tasks and secondly to determine only one feature the participants have to find out about. The previously conducted experiments by Scontras *et al.* (2019) served as the basis for the experimental structure and coding.

The experiment used so called *reference games*, previously suggested by Frank and Goodman (2012) (see Figure 2.1). This method of choice and interpretation of referential expressions allows to find out about how humans behave, when being confronted with ambiguity. These small games simulate situations containing a speaker, a listener and several objects. In the present experiment these objects were drawings of geometrical figures differing in the features *shape* (cloud, square or circle), *texture* (solid, polka-dotted or striped) and *color* (blue, red or green). The goal for the speaker was to find out about the listener's preferences for one of these features (e.g. color) - the target feature. The listener's preferences were implemented in a hierarchical manner. For example a set of hierarchical preferences could be "*a* is liked more than *b* which is liked more than *c*". If then an object with the target feature value *a* was present it was always chosen. If *a* wasn't present, objects with *b* were always chosen over objects with *c*.

In the experiment the participants acted as the speaker and observer. The task for the speaker was to select one utterance (e.g. "circle") to communicate to the listener and to observe the simulated listener's object choice. The number of utterances per trial depended on the feature values present in the three current objects. After that, the listener picked one of the objects following the indication, the participant was asked to indicate with three sliders, which preferences for the three target feature values the listener had, in their opinion, depending on the previously seen object choice as shown in the example trial in Figure 5.2. The sliders were encoded with values between 0 (liked the least) and 1 (liked the most).

The objects were simple pictures of clouds, squares and circles in the three colors red, green and blue with different textures; solid, striped and polka-dotted. With these three features *shape, color* and *texture* having each three feature values resulted in 27 different objects. Three of them formed an object set. These sets did not contain identical objects. One target feature was selected for every set. This feature would be the one for which the speaker should find out the preferences of the listener. To avoid confusion and to guarantee possible informativeness of each



Progress:

Suppose you want to find out which **color** Mary likes the most in the following slides:

You can choose a single utterance and then watch Mary select an object.

What should you say?

○ "circle"

○ "polka-dotted"

○ "cloud"

○ "striped"

"Hi, I'm Mary!"

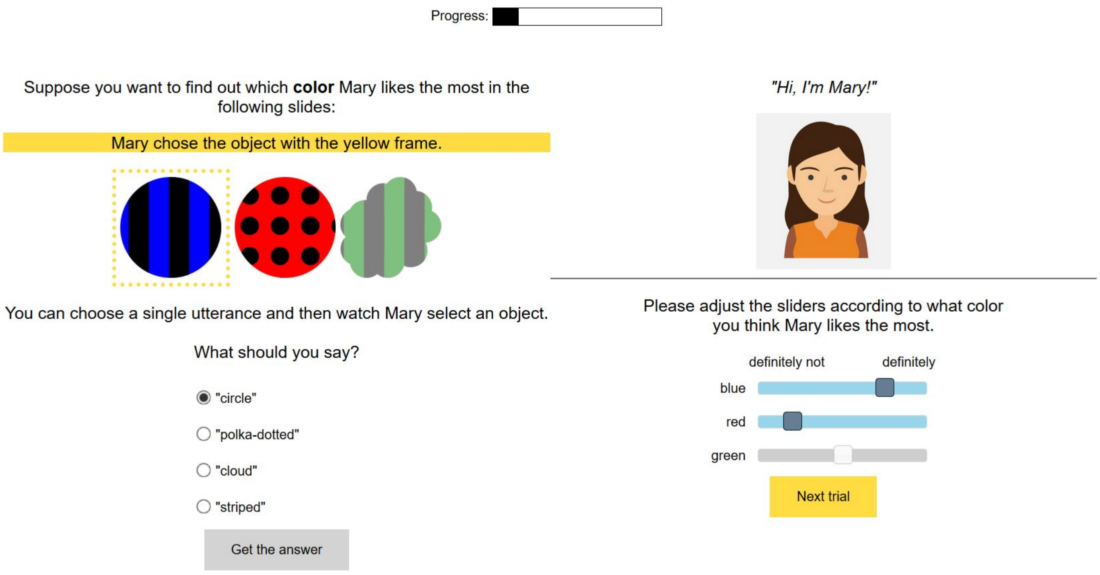
Please adjust the sliders according to what color you think Mary likes the most.

Figure 5.1: Example of one trial i.e. reference game before completing the tasks.

trial, only object sets which allowed ambiguous utterances for an informative positioning to the target feature values were used. The number of available utterances per trial ranged from two to five. The feature values of the target feature were not selectable as utterances. This guaranteed that, relative to every object set, at least one ambiguous utterance could be stated. Thus every object set was potentially informative. Nevertheless, the sets could differ based on their grade and type of informativeness. For example, one set could allow multiple ambiguous utterances referring to two objects, whereas a different set could allow only one ambiguous utterance referring to all three objects of the set.

The assumption is that the only way for the speaker to find out about the listener's preferences is to make ambiguous utterances, which means utterances that point to more than only one object. In that way, if these pointed-at objects differ in the target feature, letting the listener choose one and only one of the objects forces them to reveal preferences for feature values. Of which kind these preferences are - intrinsic, socially-induced or polite - does not matter for the game.

In order to create the dialogue-like setting, the experiment consisted of four reference games (trials) with the same listener talking about the same preferences for the same target feature. This leaded to the speaker having the opportunity to make use of the previously gathered information about the listener, to then choose utterances accordingly. Likewise, the sliders i.e. the knowledge about the listener's preferences could be adjusted repeatedly, to come to the final guess in the fourth trial. This was aimed at tracing a continuous learning process. These four trials formed one block. The amount of four was chosen, because it kept the balance between detection of the real preferences and reproducing the knowledge already gathered about the listener's preferences.



Suppose you want to find out which **color** Mary likes the most in the following slides:

Mary chose the object with the yellow frame.

You can choose a single utterance and then watch Mary select an object.

What should you say?

● "circle"
○ "polka-dotted"
○ "cloud"
○ "striped"

"Hi, I'm Mary!"

Please adjust the sliders according to what color you think Mary likes the most.

Figure 5.2: Example of one trial i.e. reference game after completing the tasks.

One whole experiment consisted out of four blocks. Likewise this number was chosen, because it kept the balance between learning a strategy for doing the task and getting bored. This resulted in each participant doing 16 reference games. Since the task may be confusing and difficult in the beginning, the first block served as a possible practice-block. The idea to declare the first block as a practice block was discarded, in order to not make the participants feel as if they did unnecessary work.

To get the participants to put effort in the task and to motivate them even more, each block was embedded in a little story. The participant i.e. the speaker should find out about the listener's preferences, to give them a present for their birthday. For example the instructions could be: “Tomorrow is Mary's birthday and you don't know her well but you want to give her a birthday present. But what should you choose? You have the choice between a number of objects and now you want to find out which one she will like the most”. The assumption underlying this method to introduce a story was that people are more motivated to make out someone's preferences in the setting of an online study if the context is relatable and fun. Also, creating an environment simulating a real interaction with another human being could strengthen the effect of the behavior in dialogue. It is assumed that it is more likely that participants use ambiguous utterances with another human than with a computer.

After each block of four trials, participants were asked to adjust a slider, according to how certain they were that the preferences they entered in the previous last trial were the real preferences of the listener.

Subsequently an evaluation was made, with the aim of revealing how good the participant had detected the listener's preferences. The grading was defined in four levels (evaluation number 0 to 3) comparing the hierarchy of the three target feature values given by the participants with the predefined preferences hierarchy of the listener. Concrete values were not compared, as the scaling was not explained to the participants. More details on the calculation of the evaluation number can be found in Box 3.

After this series of "four tasks - certainty - evaluation - small pause" the series began anew. Following four runs of this series (blocks) participants were asked to give some demographic data about themselves, such as gender, age, level of education, native language, if they thought they did the tasks correctly and if they enjoyed the experiment. Additionally, there was the possibility to leave comments.

Box 3: Calculation of the *Evaluation Number*

The evaluation number is a measure for how good a guess of the preferences-hierarchy is. It is a number between 0 and 3 with 3 being the best and 0 the worst. This method was chosen for simplicity and comparability. One can also imagine different methods.
The hierarchy of the real preferences is defined in the order {a, b, c} with "a" being the most and "c" the least liked one.
The categorization is oriented on how many right pairs are in the set. The pairs are: "a over b", "a over c" and "b over c".

- When the guessed order is identical(the first value is "liked", the second is "okay" and the third is "least liked"), then the evaluation number **3** is assigned.
  All three pairs are present.

- The evaluation number **2** is assigned, when the guessed order is {a, c, b} or {b, a, c}.
  Only two pairs are present.

- Guessing the order {b, c, a} or {c, a, b} leads to an evaluation number of **1**.
  Only one right pair is present.

- Having the order reversed completely leads to a **0** {c, b, a}.
  There is no pair in the right order.

5.4 Experiment Results

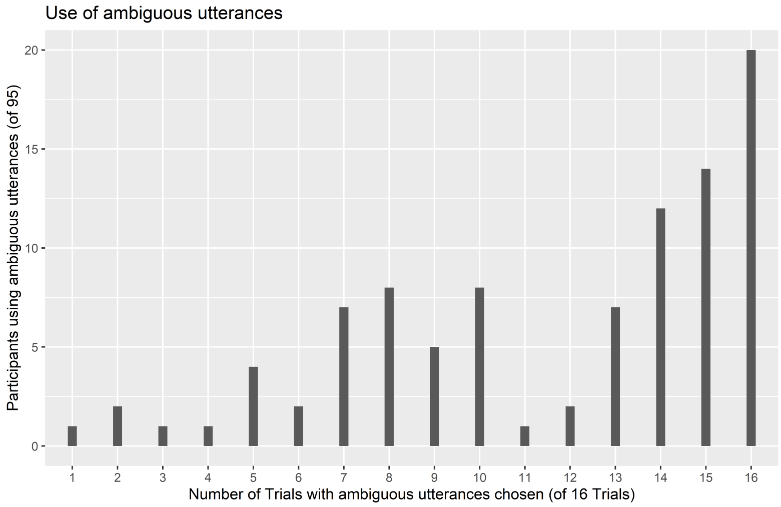


Figure 5.3: Choice of utterances whose indications were ambiguous relative to the present objects. Out of 95 participants 20 chose solely ambiguous utterances.

Number of possible utterance per trial	**2**	3	**4**	5
Occurrence [%]	**5.3**	15.7	**63.6**	15.5

Table 5.1: Distribution of number of available utterances between whose the participant had to choose per trial. The majority (63,6%) of trials offered the choice between four utterances.

In the experiment, participants were confronted with different object sets. Due to these differences also the set of available utterances and its length varied. The size of these sets ranges from two to five utterances. Most trials offered four utterances to choose from (63.6%). In Table 5.1 these differences are described in detail. From these sets, participants were to choose between either an utterance, which could be interpreted as ambiguous relative to the current objects, and an utterance, which would point straightforwardly to only one object. In Table 5.2 the amount of ambiguous utterances in the set of available utterances is broken down. 35.8% of the trials had a set of available utterances from which 50% were ambiguous. There were no trials without having at least one selectable ambiguous utterance. In only 5.3% of the trials all utterances were ambiguous. Half of the participants chose more than 81.20% of the utterances to be ambiguous (Min. = 6.20, 1st Qu. = 50.00, Median = 81.20, Mean = 72.77, 3rd Qu. = 93.80, Max. = 100.00). In the results depicted in Figure 5.3, one can observe that the amount of ambiguous utterances chosen by participants varies over nearly the whole range.

Percentage of ambiguous utterances among the set of utterances to choose from [%]	20	25	**50**	67	**100**
Occurrence [%]	15.5	27.8	**35.8**	15.7	**5.3**

Table 5.2: Distribution of percentage of how many ambiguous utterances were in the set of available utterances from whose the participant had to choose each trial. For the size of the set of available utterances see Table 5.1.

There was a positive correlation between (1) how certain the participants were that they detected the preferences correctly, (2) the actual performance, and (3) the number of ambiguous utterances chosen in the block (all p < 0,001; tested with clmm [cumulative link mixed model fitted with the laplace approximation] from the ordinal-package in R).

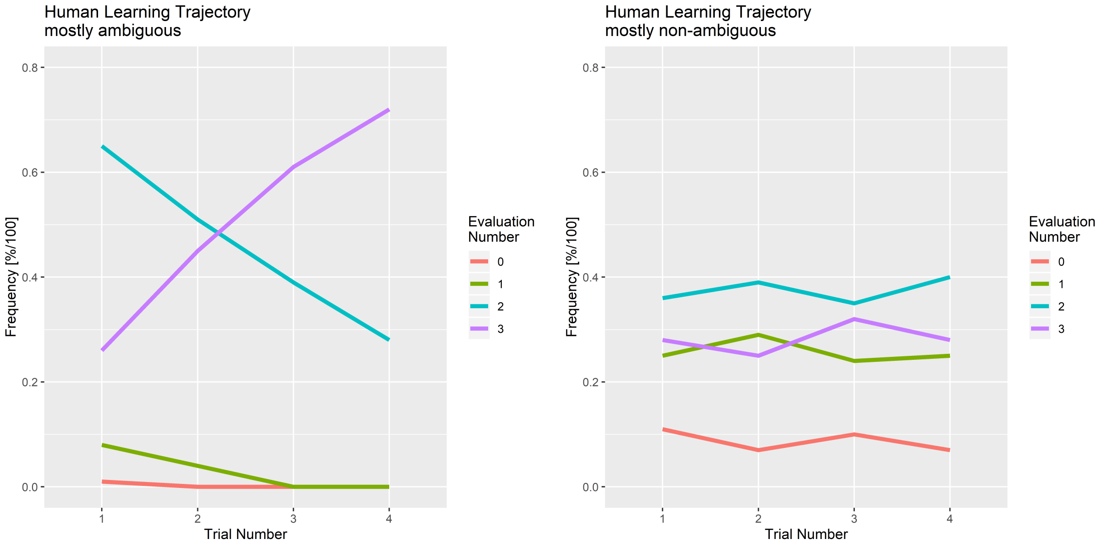


Figure 5.4: These two plots depict the performance over the four trials. In the left plot, only the data from participants in the 3rd quartile of using ambiguous utterances (15-16/16 utterances chosen ambiguously) is displayed. In the right plot, the data from participants in the 1st quartile of using ambiguous utterances (0-8/16 utterances chosen ambiguously) was taken. The learning trajectories with ambiguous utterances in the left plot show a clear improvement of the performance in contrast to the right plot where an improvement can rarely be seen. Evaluation number equals 3 is the best performance possible and it is clear to see that the frequency of a high evaluation number rises with each trial.

An improvement was able to happen both within trials in a block or over all blocks in an experiment. An adjustment of the strategy (to choose ambiguous utterances) over the blocks was found (p < 0,001; tested with a “Cumulative Link Mixed Model fitted with the Laplace approximation”-method in R), but the performance of preference-detection did not improve (tested with the same method).

The data shows shows a clear improvement in the course of the trials, in those cases where participants chose mostly ambiguous utterances (see Figure 5.4).

More results will be presented and discussed in Chapter 6 *Results* and Chapter 7 *Discussion*.

5.5 Experiment Discussion

Consistent with Scontras *et al.* (2019), the results reveal that some humans choose ambiguous utterances more often than an average speaker (see Figure 5.3). This supports the hypothesis that humans pick strategically ambiguous utterances to perform the task (1). The ability to strategically choose ambiguity over linguistic clarity has already been suggested in existing literature (Piantadosi *et al.*, 2012; Erev *et al.*, 1991; Wasow, 2015).

Participants had a good grasp of how good they did in the experiment, which is shown by the positive correlation between certainty and both performance and amount of ambiguous utterances chosen. This supports the hypothesis that participants choose strategically and not just by reflex ambiguous utterances to perform this task. The positive correlation between certainty about that the participant did a good job (self-reported) and the actual performance indicates that certainty could function as an indication of how much participants "understood" how to solve the task.

The fact that not all participants used only or mostly ambiguous utterances offers room for interpretation whether there might be different groups of participants: the ones who prefer to choose ambiguous utterances (a); the ones who do not realize that there are ambiguous and non-ambiguous utterances or do not think that this difference would help them doing the task (b); and the ones that prefer to choose non-ambiguous utterances (c).

One of the hypotheses is that detection performance rises over informative trials (3). This can be confirmed with the learning trajectory-results. Participants learned more about the listeners preferences with each informative trial in a block. In blocks which are mostly non-informative no improvement was observed. These results lead to the conclusion that indeed humans are capable of updating their beliefs about preferences of an interlocutor with every piece of evidence. This evidence can only be acquired through informative trials. Further, this conclusion indicates that the modeling approach using Bayesian inference, which in this case uses the same process as, according to the results, humans do, is promising.

Participants significantly increased the number of ambiguous utterances per block over the experiment. This may imply that they learned some kind of strategy to gain more information. With the premise that the strategy is pursued with the intent to derive information from the ambiguity resolution, it would then follow that participants also improved their performance over the blocks. This, however, was not the case. Possible reasons might be that there are two separate mechanisms to pick ambiguous utterances and to derive information from the ambiguity resolution and that some participants could only learn the first one. This hypothesis would have to be tested in another study where the two tasks would have to be separated. A different explanation is that there might be another reason than information retrieval for participants to choose ambiguous utterances in this task. Generally, this is an interesting observation and should be considered as a starting point for more research on strategy learning and information deduction.

Chapter 6

General Results

In the following chapter, the results of the experiment and the results of the model analysis will be presented. The evaluation was done in R. The exact implementation can be viewed in Chapter 9 *Appendices: Model and Simulation.*

Learning and detection success was measured in four levels. The assumed hierarchy of preferences was extracted from the slider values and compared to the real (implemented) preferences. Depending on how similar this hierarchies were, an evaluation number between 0 and 3 was assigned. More information about this evaluation process has been described in Chapter 5 *Experiment* and in Box 3.

In over half of all blocks in the experiment, the hierarchies of the preferences were correctly detected (see Figure 6.1).

Still around 14% of the participants were either guessing the hierarchy completely or nearly wrong. The model presented this effect, as well. A comparison of the data with and without the first block (which could be a "training block") it did not show any difference in effect.

Participants were able to choose between an ambiguous and a non-ambiguous utterance four times per block. Looking at the distribution in over 46% of the blocks, all utterances were chosen ambiguously (see Table 6.1). In Figure 6.2 the learning success of model and participants was evaluated, based on how much ambiguity they used. In those cases where participants chose zero ambiguous utterance, no clear learning trend can be seen in the data. This observation can explain the 14% wrong or nearly wrong guesses, which were mentioned before. In the blocks, where participants chose only ambiguous utterances, over 98% of the preferences-sets were detected correctly or nearly correctly. Similarly the model predicts in over 95% of the blocks an evaluation number of 2 or 3.

Also under this aspect, the predictions of the model are accurate.

Number of ambiguous utterances per block	**0**	1	2	3	**4**
Number of blocks	**20**	41	65	81	**173**

Table 6.1: Distribution of use of ambiguous utterances per block, not grouped by subject.

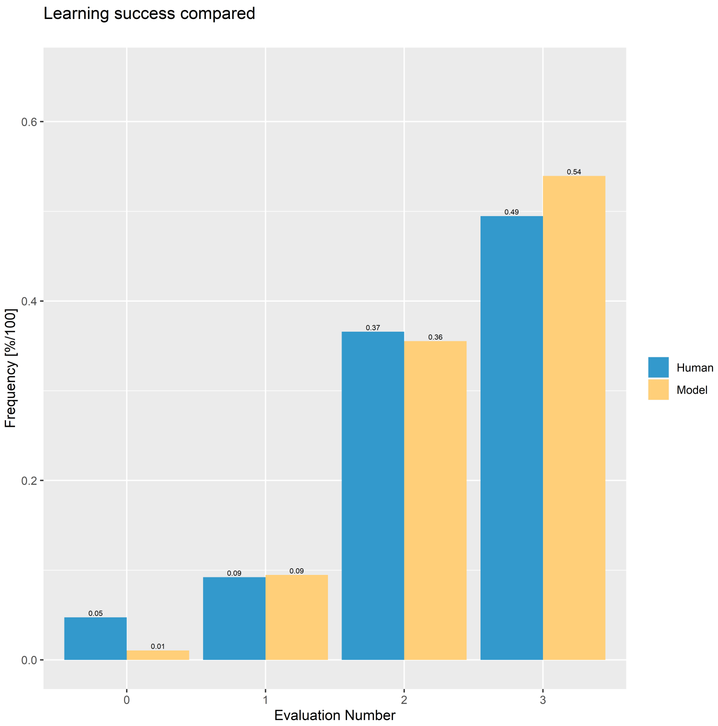


Figure 6.1: In around 50% of all blocks, the hierarchies of the preferences were detected correctly (evaluation number equals 3). Model predictions (yellow) are accurate.

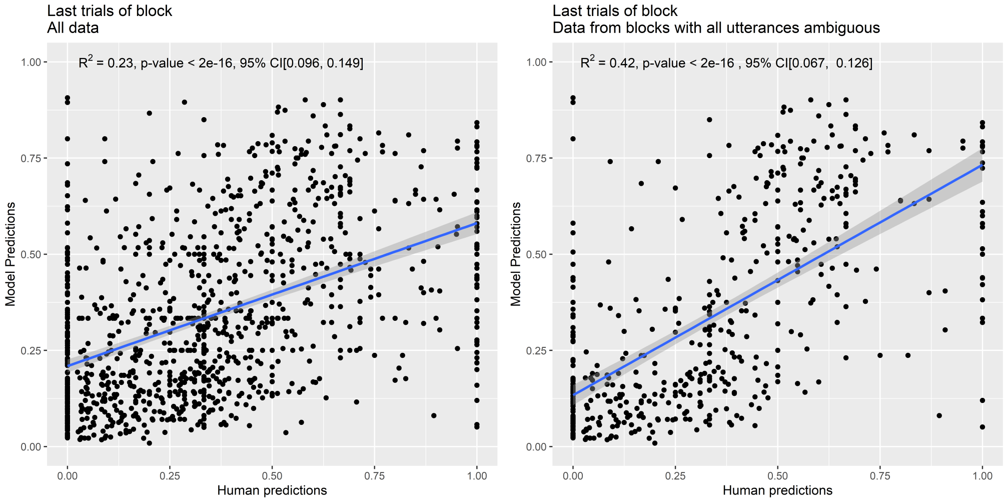


Figure 6.3: The raw preference values predicted by the model (y-axis) are plotted with the human slider values for the preferences in the experiment (x-axis). The blue line describes a simple linear regression. In the left graph, all data from the experiment was taken with the corresponding model predictions. The model fit is even better in those cases where exclusively data from blocks with informative trials has been used (right graph).

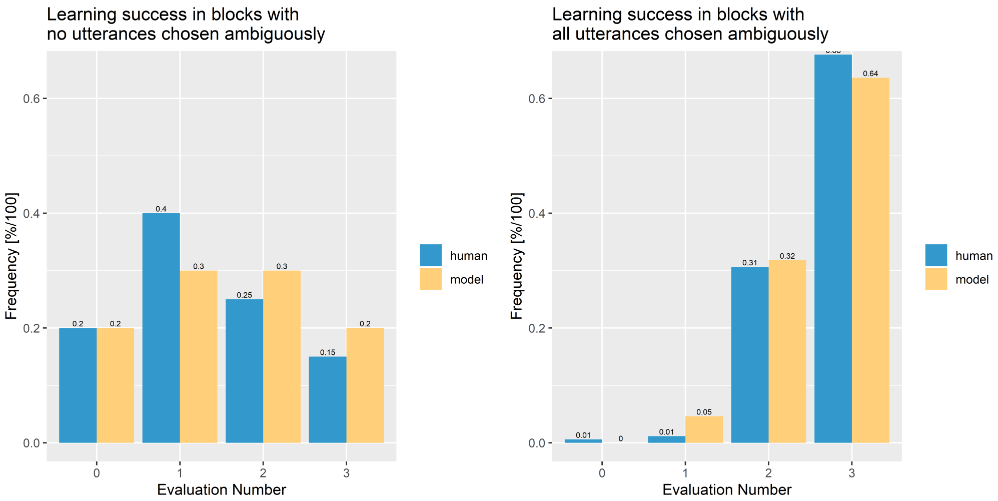


Figure 6.2: Learning success varies strongly between non-ambiguous (left graph) and ambiguous blocks (right graph). Only in the ambiguous ones a clear success is visible (for size of data sets see Table 6.1). Also the model predicts this observation.

The raw human results, without taking the step towards evaluation numbers, are the slider values varying between 0 and 1. In Figure 6.3 (left), the human data is compared with the corresponding model predictions. A significant positive correlation between human data and corresponding model predictions in a linear regression analysis is observed ($R^2 = 0.23$, p-value $< 2e - 16$). The relatively low $R^2 = 0.23$ contains a high uncertainty. In Figure 6.3 (right), the subset of blocks which were ambiguous was taken and the just described comparison was made, too. The resulting correlation was even stronger than in the all-data-case ($R^2 = 0.42$, p-value $< 2e - 16$).

Figure 6.4 depicts the performance trajectories of the four trials of each block. A general improvement (increase in evaluation number 3 and decrease in all the others) can be observed. The model approximates the human data well. A learning of the preferences can be observed. The learning effect is even stronger, when only the data from participants which chose mostly ambiguous utterances is used for the evaluation. In this case, the curve of the best performance rises even more drastically. In the opposite case (mostly unambiguous utterances), no improvement can be observed.

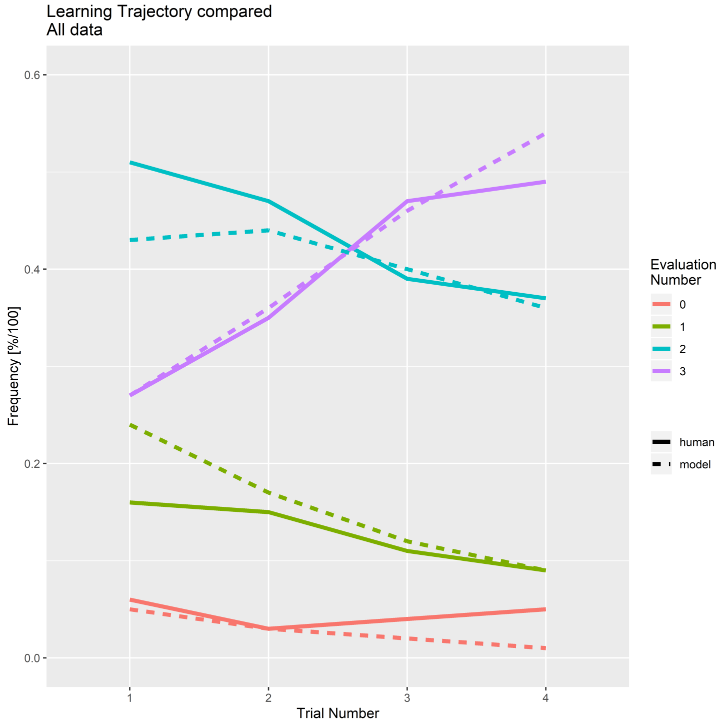


Figure 6.4: The model shows a very similar learning trajectory to humans over the course of the four trials ($R^2 = 0.96$, p < 0.001). The violet line depicts the number of correct preference-guesses. Whereas the number of bad performances goes down during one block, the number of correct guesses goes up.

Chapter 7

General Discussion

The experiment was designed on the basis of the hypothesis that the preferences of the interlocutor could be learned only in the trials where ambiguous utterances had been chosen. The results show this clearly and confirm therefore that the experimental task - the reference games - are suitable to study whether humans strategically use ambiguous utterances to find evidence for the interlocutor's preferences. The modeling resulted in the same observation. The non-uniform distribution of the model predictions in the case of only data from unambiguous blocks (see Figure 6.2, left) can be explained with the calculation of the evaluation number (see Box 3). This resulting distribution reflects an untilt and neutral outcome, as predicted.

Only this conclusion allows further interpretation of the results. It falls in line with and confirms the applicability of previous research that was conducted with reference games (Scontras *et al.*, 2019). It further offers a stable ground for more research using reference games to study ambiguity in communication and maybe even phenomena beyond.

Both experiment and simulation with the RSA model allowed to test the hypothesis that improvement trajectories in blocks with informative trials can be observed. The results confirm this hypothesis. Both model and humans improve their performance with every informative trial. This implicates that humans update prior knowledge with every piece of new evidence. The mechanism of Bayesian inference in the model enables the model to do the same. The observation that the model and humans behave very similar when encountered with new evidence enables the conclusion that humans use a process which might be similar to Bayesian inference.

Overall, this thesis aimed also at studying the power of RSA models to simulate and explain human reasoning and behavior. For this thesis, some existing RSA models were examined, developed and enhanced, towards a version of RSA which should be able to predict human reasoning about disambiguating behavior of an interlocutor in dialogues. The results indicate that this was successful. There were not found any major differences between model predictions and human data. The reasoning in uninformative trials, as well as in informative trials was predicted correctly. Also, the learning trajectories within blocks of humans were predicted accurately. Extending the findings of Scontras *et al.* (2019), it was successful to model human reasoning in dialogic reference games about ambiguity with a development of the RSA framework. The novelty is the capability of the model to have a similar learning capability as humans in dialogues, who can acquire new evidence with every new utterance and response. These findings coincide with the assertion of Goodman and Frank (2016) that in fact RSA models "provide a computational framework for integrating linguistic structure, world knowledge, and context in pragmatic language understanding" (p. 1).

Even though the model is already in good accordance with the human data, an optimization of the parameters γ and Ξ has a big potential to make the predictions more accurate. At the same time, optimization must be handled with care, in order not to overfit the model. To validate an optimized model (and generally speaking also the unoptimized model) this model has to be tested on data on which the model was not fitted (see Wilson and Collins, 2019). The validation of this proposed model and a possible optimization is therefore strongly required.

It needs to be noted that only one part of the experiment was simulated by the RSA model: the observation task. The utterance choice task (the first task in the experiment) instead was not modeled. Therefore, the incorporation of this second part in the model should be one of the first steps in a future continuation of this thesis.

An analysis based on different degrees of ambiguity which are possible with different object sets in combination with target features as done similarly by Scontras *et al.* (2019) (there called "ambiguity classes") was not possible in this study. The multi-trial dialogue design did not permit an analysis with a reasonable amount of classes. Future work should incorporate some form of ambiguity classes, in order to observe the reasoning and behavior, when confronted with different grades of ambiguity.

Even though the study was designed with the intention to simulate human behavior in realistic situations, the setting was still quite artificial. Important factors, such as a broader context of the situation, possible costs of utterances and preferences, the possibility of questions, general social rules, the relationship between speaker and listener, more relevant objects, or else, were not incorporated. Hence, the setting is not complex enough to display human behavior and reasoning in daily life. This holds an enormous potential for future work to find out how far cognitive modeling can get to simulate human behavior and reasoning with ambiguity and generally in communication. Additionally, a common weakness of behavioral experiments is the experimental lab-setting. The question of how applicable the obtained results are to behavior in real life cannot be answered definitively.

The results on ambiguity use give hints that there might be different groups of participants (as mentioned in Chapter 5.6 *Experiment Discussion*). That, of course, might be the case, due to the fact that this was set as an online experiment, where some participants possibly were not doing the task with the utmost effort. An other reason might be different general strategies. The evidence that there was no improvement over the course of the blocks implies that participants held on their strategies from the beginning until the end. An opportunity for future work would therefore be the further investigation of possible groups based on different strategies in this task. Implications from these different groups might be that there are different types of social behavior and this type of task could then be used to discriminate them. Even in clinical psychology there might be an opportunity of application.

During the development and evaluation of the study, the topic of mutual understanding in connection to ambiguity as discussed here came up. One interesting question for future work on mutual understanding and ambiguity in dialogues might be, whether mutual understanding can be measurable by observing the degree of successful disambiguation. In other words, how much two persons deduct the same meaning when exchanging ambiguous utterances with the intentions that the other person should understand. Following this idea in a situation when both persons always understand what state of the world the other person is referring to with their utterance, a feeling of understanding and being understood - mutual understanding - would be reached.

The model in this thesis, as well as possible future research and model development with the idea of mutual understanding, might lead to insights for the development of digital assistants. One obstacle of digital assistants is to understand the meaning of expressions and sentences uttered by humans. I has been stated earlier that natural language is highly ambiguous and therefore the integration of a tool for disambiguation, as presented here, can have a big impact on the performance of digital assistants.

In summary, it can be said that human language is complex and so must be the models which try to describe it. But it remains clear that every model is always at least a little bit wrong and the task is to find the model that is least wrong. Wilson and Collins (2019, p.27) remark that “A modeler’s work is never done. To paraphrase George Box, there are no correct models, there are only useful models (Box, 1979).”

Many questions remain: Which mechanisms do ambiguity and disambiguation underly in natural language? How do these cognitive processes work? Cognitive modeling can be one way to answer some of these questions. This thesis is now based on evidence and in turn encourages further research on how far cognitive modeling can get to simulate human behavior, as well as how complex the models must be in order to reach this point. The study of this thesis also gives answers on how ambiguity is strategically used and why it is purposeful. The underlying question inquires after the real nature of rational and pragmatic human reasoning and behavior.

Chapter 8

Conclusion

In research about rational and pragmatic reasoning in communication many questions are unanswered. This thesis therefore uses a cognitive modeling approach to study rational and pragmatic reasoning about ambiguity and disambiguation. An RSA model, which successfully simulated the human behavior in the conducted experiment, was designed. Results show that humans have the ability to use ambiguity strategically, with the intention to gather information about the interlocutor. It might be derived that humans use a similar process to Bayesian inference in order to update their prior beliefs about the preferences of the interlocutor, when confronted with new information. Also, it could be found that reasoning about disambiguating behavior of the interlocutor can reveal their preferences. This reasoning is modelable using the computational RSA framework. This model was able to formalize inferences about behavior in context.

The attempt to study reasoning about and behavior with ambiguity in a realistic situation did not fully succeed, due to the still artificial setting. Despite the fact that all hypotheses were confirmed, more general research on this topic is necessary. This holds the possibility to understand better, how and why humans reason and behave the way they do.

Chapter 9

Appendices: Model and Simulation

Structure

First, the model will be shown, then the model testing and then the statistical test.

RSA-Model

The developed RSA-Modelling Framework is implemented in the following pages.

The simple listener function determines the hypothetical listener's object choice given the objects to choose from and its preferences, determining P(obj | utt, listener's object preferences).

```
    knitr::opts_chunk$set(fig.width=4, fig.height = 3)

simpleListener <-
  function(utterance,
           mapUttToObjProbs,
           listenerObjectPreferences) {
    objPosterior <-
      mapUttToObjProbs[utterance, ] * (listenerObjectPreferences + 1e-100)
    if (sum(objPosterior) == 0) {
      return(objPosterior)
    }
    return(objPosterior / sum(objPosterior))
  }
```

The simple pragmatic speaker considers all "imaginable" (i.e. implemented) preference distributions over objects of the listener. Starting with a prior assumption over the possible listener's preferences ("preferencesPrior"). This prior has one value for each feature value possible. Only the priors for the target feature values > 0 because only they are relevant for the task. It then infers the posterior over these preferences given the listener makes a particular object choice. The priors in the beginning have a uniform distribution but with each trial the priors are the previous posterior preference presumptions. This leads to a Bayesian learning process. "utterance" is an index referring to one of the relevant utterances ("relevantUtterances) which are all present feature values in the current objects i.e. P(listener's feature value preferences | utterance, object choice by the listener, prior over preferences).

```
simplePragmaticSpeaker <-
  function(utterance,
           obj,
           preferencesPriorAll,
           relevantUtterances,
           currentObjects,
           mapUttToObjProbs,
           objectPreferenceSoftPriors) {
    preferencesPrior <- preferencesPriorAll[relevantUtterances]
    prefPost <- rep(0, length(relevantUtterances) + 1)
    for (pref in c(1:length(preferencesPrior))) {
      # prior over the preferences the speaker is interested in
      if (preferencesPrior[pref] > 0) {
        pp <-
          simpleListener(utterance,
                         mapUttToObjProbs,
                         objectPreferenceSoftPriors[[pref]])
        prefPost[pref] <- pp[obj] * preferencesPrior[pref]
      }
    }
    for (pos in c(1:length(relevantUtterances))) {
      preferencesPriorAll[relevantUtterances[pos]] <- prefPost[pos]
    }
    if (sum(preferencesPriorAll) == 0) {
      # no evidence for any preferences... -> no inference
      return(preferencesPriorAll)
    }
    return(preferencesPriorAll / sum(preferencesPriorAll))
  }
```

To know how well the model did, the evaluation number similar to the one used in the experiment is calculated. The best is 3 and the worst is 0. This number is calculated by comparing the simulated preferences the choosing listener has with the model or human predictions. To not compare absolute or relative numbers, as they might vary in quite big ranges, the hierarchy is compared. This results in having truthfully predicted which preferences the listener has (evaluation number = 3) in contrast to not having a clue about them (evaluation number = 0).

```
evaluate <-
  function(allUtterancePref,
           preferencesPrior,
           targetFeature) {
    index <- targetFeature * 3
    indices <- c(index - 2, index - 1, index)
    tarFeaPref <- allUtterancePref[indices,]
    if (length(preferencesPrior) > 3) {
      tarFeaPrefPrior <- preferencesPrior[indices]
    } else {
      tarFeaPrefPrior <- preferencesPrior
    }
    prefRank <-
      order(as.numeric(tarFeaPref[, 3]))
    prefPriorRank <-
      order(tarFeaPrefPrior)
    if (identical(prefRank, prefPriorRank)) {
      evalNum <- 3
    } else if (identical(prefPriorRank, c(prefRank[1], prefRank[3], prefRank[2])) ||
               identical(prefPriorRank, c(prefRank[2], prefRank[1], prefRank[3]))) {
      evalNum <- 2
    } else if (identical(prefPriorRank, c(prefRank[2], prefRank[3], prefRank[1])) ||
               identical(prefPriorRank, c(prefRank[3], prefRank[1], prefRank[2]))) {
      evalNum <- 1
    } else if (identical(prefPriorRank, c(prefRank[3], prefRank[2], prefRank[1]))) {
      evalNum <- 0
    }
    return(evalNum)
  }
```

Simulation

The model has some determining factors for how it performs. These are the not-obey-instant and the soft-preferences-value. The not-obey-instant is a value between 0 and 1, with 0 the listener follows always it's preferences and 1 not at all. The soft-preferences-value manipulates how strong the listener's preferences are. With the value being 0 the listener has absolute preferences and absolute non-preferences. Augmenting this value leads to a uniform-prior model.

```
notObeyInst <- 0
softPrefValue <- 1
```

Here the model gets tested with the data from the experiment. It allows direct comparison of the performance of the model with the human behavior. The data to feed the model is the data which was also fed to the participants in the experiment. For each combination of objects, utterance and previous trials the model adjusts its posterior presumptions which then become the next prior presumptions.

```
for (worker in c(0:totalWorker)) {
  for (block in c(1:totalBlock)) {
    blockdata <-
      subset(inputData,
             blockNr == block - 1 &
               workerid == unique(inputData$workerid)[worker + 1])
    targetFeatureNum <- blockdata$targetFeatureNum[1]
    preferencesPrior <- getPreferencesPrior(targetFeatureNum)
    preferencesPriorIndices <- which(preferencesPrior != 0)
    allUtterancePref <-
      getAllUtterancePref(
        c(
          blockdata$simPreference0[1],
          blockdata$simPreference1[1],
          blockdata$simPreference2[1]
        )
      )
    ambiguousUtteranceCount <- 0
    for (trial in c(1:maxTrialNum)) {
      row <- row + 1
      currentObjects <-
        c(blockdata$orderObjNum1[trial],
          blockdata$orderObjNum2[trial],
          blockdata$orderObjNum3[trial])
      allPresentFeaValues <- determineAllFeaValues(currentObjects)
      inputData$allPresentFeaValues[row] <-
        toString(allPresentFeaValues)
      relevantUtterances <- determineValidUtterances(currentObjects)
      utteranceGeneral <- as.integer(blockdata$utteranceNum[trial])
      utterance <- which(relevantUtterances == utteranceGeneral)
      ambiguous <- isAmbiguous(allPresentFeaValues,
                               utteranceGeneral,
                               currentObjects,
                               targetFeatureNum)
      inputData$ambiguous[row] <-
        ambiguous
      ambigRatio <- countAmbigUttRatio(allPresentFeaValues,
                                       currentObjects,
                                       targetFeatureNum)
      inputData$ambigRatio[row] <- ambigRatio
      if (ambiguous) {
        ambiguousUtteranceCount <- ambiguousUtteranceCount + 1
      }
      inputData$ambiguousUtteranceCount[row] <-
        ambiguousUtteranceCount
      mapObjToUtt <-
        determineObjectToUtterancesMapping(currentObjects)
      mapUttToObjProbs <-
        determineUtteranceToObjectProbabilities(relevantUtterances,
                                                currentObjects,
                                                mapObjToUtt,
                                                notObeyInst)
      mapUttToPref <-
        getMapUttToPref(relevantUtterances, allObjects, allUtterancePref)
      objectPreferenceSoftPriors <-
        getObjectPreferencePriors(
          relevantUtterances,
          currentObjects,
          softPrefValue,
          mapUttToObjProbs,
          mapUttToPref
        )
      mapUttToObjToPref <-
        getMapUttToObjToPref(
          currentObjects,
          targetFeatureNum,
          relevantUtterances,
          allUtterancePref,
          allObjects,
          mapUttToPref
        )
      obj <-
        which(currentObjects == blockdata$simulatedAnswerObjNum[trial])
      preferencesPrior <-
        simplePragmaticSpeaker(
          utterance,
          obj,
          preferencesPrior,
          relevantUtterances,
          currentObjects,
          mapUttToObjProbs,
          objectPreferenceSoftPriors
        )
      inputData$preferencesPrior1[row] <-
        preferencesPrior[preferencesPriorIndices[1]]
      inputData$preferencesPrior2[row] <-
        preferencesPrior[preferencesPriorIndices[2]]
      inputData$preferencesPrior3[row] <-
        preferencesPrior[preferencesPriorIndices[3]]
      evalNumModel <-
        evaluate(allUtterancePref, preferencesPrior, targetFeatureNum)
      inputData$evalNumModel[row] <- evalNumModel
      humanResponse <-
```

```
        c(
          blockdata$normResponse0[trial],
          blockdata$normResponse1[trial],
          blockdata$normResponse2[trial]
        )
      evalNum <-
        evaluate(allUtterancePref, humanResponse, targetFeatureNum)
      inputData$evalNum[row] <- evalNum
    }
  }
}
```

For each trial the utterance chosen by the participant is evaluated if it is ambiguous for the target feature or not. This shows which participants use ambiguity to detect information.

```
inputData$ambiguousUtteranceCount <- as.factor(inputData$ambiguousUtteranceCount)
ambiguityUsed <- matrix(nrow = totalWorker + 1, ncol = 3)
for (worker in c(0:totalWorker)) {
  ambiguityUsed[worker + 1, 1] <-

unique(inputData$workerid)[worker + 1]
  ambiguityUsed[worker + 1, 2] <-
    round(sum(inputData$ambiguous[which(inputData$workerid ==

unique(inputData$workerid)[worker + 1])]) / 16 * 100, digits = 1)
  ambiguityUsed[worker + 1, 3] <-
    sum(inputData$ambiguous[which(inputData$workerid == unique(inputData$workerid)[worker + 1])])
}
ambiguousWorker <-
  subset(ambiguityUsed, ambiguityUsed[, 2] > quantile(ambiguityUsed[, 2], 0.75))[, 1]
inputDataAmbiguous <-
  subset(inputData, workerid %in% ambiguousWorker)
nonAmbiguousWorker <-
  subset(ambiguityUsed, ambiguityUsed[, 2] < quantile(ambiguityUsed[, 2], 0.25))[, 1]
inputDataNonAmbiguous <-
  subset(inputData, workerid %in% nonAmbiguousWorker)
```

The experimental human data was cleaned for language and assurance that the task was done accurately. Only participants which specified that their native language was English were kept. Two participants with different native languages (Italian and Urdu) were excluded. This happened to avoid problems and hassles related to language barriers. Also, as the study scrutinized a psycho-linguistic phenomenon, possible interference with other native languages as English were tried to minimize. Three participants who answered to the question “Did you read the instructions and do you think you did the HIT correctly? - Yes, No or Confused” with “No” or “Confused” were excluded too. In these cases the data cannot be considered in the evaluation. (HIT is an abbreviation for Human Intelligence Task and is to be used here synonymously with “experiment”.)

Thanks for reading until here. Have a good day!

Bibliography

Brown, P. (2015). Politeness and language. In J. D. Wright, editor, *International encyclopedia of the social & behavioral sciences*, pages 326–330. Elsevier, Amsterdam.

Chomsky, N., Belletti, A., and Rizzi, L., editors (2002). *On nature and language: With an essay on The secular priesthood and the perils of democracy*. Cambridge Univ. Press, Cambridge.

Crystal, D. (2015). *A dictionary of linguistics and phonetics*. Language library. Blackwell Publlishing and Credo Reference, Malden, Massachusetts and Oxford [England] and Boston, Massachusetts, 6th edition edition.

Erev, I., Wallsten, T. S., and Neal, M. M. (1991). Vagueness, ambiguity, and the cost of mutual understanding. *Psychological Science*, **2**(5), 321–324.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science (New York, N.Y.)*, **336**(6084), 998.

Gao, C. and Ren, X. (2013). A pragmatic study of ambiguity and puns in english humor. In *Proceedings of the 2nd International Conference on Management Science and Industrial Engineering (MSIE 2013)*, Paris, France. Atlantis Press.

Gärdenfors, P. (2014). *Geometry of meaning: Semantics based on conceptual spaces*. The MIT Press, Cambridge, Massachusetts.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, **20**(11), 818–829.

Grice, H. P. (1975). Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 26–40. Academic Press, New York.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun, editor, *The Cambridge handbook of computational psychology*, Cambridge handbooks in psychology, pages 59–100. Cambridge University Press, Cambridge.

Hirst, G. (1987). *Semantic interpretation and the resolution of ambiguity*. Studies in natural language processing. Cambridge University Press, Cambridge.

Macagno, F. and Bigi, S. (2018). Types of dialogue and pragmatic ambiguity. In S. Oswald, T. Herman, and J. Jacquin, editors, *Argumentation and Language - Linguistic, Cognitive and Discursive Explorations*, volume 32 of *Argumentation Library*, pages 191–218. Springer International Publishing, Cham.

Nichols, K. (n.d.). Dialogic rsa: A bayesian model of pragmatic reasoning in dialogue.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, **122**, 280–291.

Scontras, G., Achimova, A., Stegemann, C., and Butz, M. (27.07.2019). On the purpose of ambiguous utterances: Poster presented at the 41st Annual Meeting of the Cognitive Science Society.

Sikos, L., Venhuizen, N., Drenhaus, H., and Crocker, M. (2019). Reevaluating pragmatic reasoning in web-based language games.

Sun, R. (2008). Introduction to computational cognitive modeling. In R. Sun, editor, *The Cambridge handbook of computational psychology*, Cambridge handbooks in psychology, pages 3–20. Cambridge University Press, Cambridge.

Wasow, T. (2015). Ambiguity avoidance is overrated. In Susanne Winkler, editor, *Ambiguity: Language and Communication*, pages 29–47. DE GRUYTER.

Wilson, R. C. and Collins, A. (2019). *Ten simple rules for the computational modeling of behavioral data*.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Tübingen, 28.11.2019

Ort, Datum

Unterschrift