Bachelor thesis
Artificial Intelligence

Radboud University

Improving the Tractable Lexicon Size for the Rational Speech Act Model

Author:
Jelle Philip Hilbrands
s4336348
jelle.hilbrands
@hotmail.com

First supervisor:
L.D. van de Braak, MSc
Donders Institute for
Brain, Cognition and
Behaviour
l.vandebraak@psych.ru.nl

Second supervisor:
dr. I.J.E.I van Rooij
Donders Institute for
Brain, Cognition and
Behaviour
i.vanrooij@donders.ru.nl

June 21, 2021

Abstract

The Rational Speech Act model adapted to resolving misunderstandings in conversation is conjectured to be intractable. A combinatorial explosion in searching through the space of possible lexicons is likely responsible for this intractability. Coming from a fixed-parameter tractability view, where intractable problems can become tractable for restricted input parameters, I have introduced and tested two constraints on possible lexicons in an attempt to improve upon the feasibly computable lexicon size. Stemming from an intuition that humans use these constraints as well, one constraint is the use of knowledge or facts for generating lexicons. Some words can have very specific and unambiguous meanings. The other constraint is about the use of similarity or neighborliness to generate new lexicons. The other person will likely have a similar meaning for the same word. I named these constraints factual lexicons and neighboring lexicons. The former is formalized as fixing a certain amount of mappings of signal and referent that are always existent. The latter is formalized as considering similar lexicons as compared to the current one available to the agent. Both constraints were formalized in the Human Language Constrained Rational Speech Act model. The results indicate that factual lexicons improve to a lesser extent the time and space performance of the model and result in degraded model accuracy when resolving misunderstandings. Neighboring lexicons improve to a much larger extent time and space performance and also improve upon accuracy. As of writing, both the intractability of the problem and the fixed-parameter tractability approach are unproven. Nevertheless, these results hint at the capability of constraints to improve upon the feasibly computable lexicon size for the RSA model. However, I am not convinced that the constraints are a fixed-parameter solution to an unproven problem.

Contents

1 Introduction

In this thesis I have taken a closer look at communication between two persons. A model for communication has been introduced, and was expanded upon to also reflect resolving misunderstanding in a dialogue by others. Some parts of this extension are conjectured to be intractable. By taking inspiration from the methods of fixed-parameter tractability, I have introduced two ideas that constrain the size of the input for the model. My research question and subsequent formalizations, model implementation and analyses shed some light on the conjectured intractability of this model.

1.1 Pragmatic inference in communication

Communication by spoken language is a tool at our disposal that sets us humans apart from all other animals on this planet. It is one of our defining traits (Seidenberg and Petitto, 1987). The effectiveness and speeds at which we both talk to each other and process and infer meaning from the words and sentences uttered are amazingly fast. Our brains have evolved in such a way that it is automatic. When I speak, and I am talking to some other person, and you happen to overhear the conversation, then you cannot help but process the words you hear and try to make sense of what you just heard. This process is an automatic system that cannot be suppressed (Posner and Snyder, 2004). The emergence of language evolved from, among other things, human interaction that needed such a tool in order to communicate effectively (Levinson, 2006).

Communication can be thought of as two people entering in a dialogue with several common rules, boundaries, and expectations in place. In this dialogue, the two interlocutors take up a specific role. These are most commonly referred to as the speaker and the listener. These roles can swap between the two as the dialogue moves forward. The speaker tries to convey a message to the listener such that the listener understands the message's intention. The listener, in turn, tries to make sense of the message and infers the intended meaning of the message. Already, several assumptions about good and clear communication become visible. These implicit assumptions in conversations have been researched and captured in several maxims by Grice (1975). Interlocutors that are considered to be rational adhere maximally to the Gricean maxims. These are considered to be the cooperative principles that guide almost all successful cooperative communication. The four maxims are:

1. Maxim of Quality (Try to make your contribution one that is true)
2. Maxim of Quantity (Make your contribution as informative as is required)
3. Maxim of Relation (Be relevant, relatable)
4. Maxim of Manner (Be perspicuous, clear, unambiguous)

These principles have been expanded upon by, for instance, Sperber and Wilson (1986) and Leech (2016). An emergent property of rational dialogue is that of pragmatic inference or scalar implicature. Pragmatic inference means understanding a message and the underlying meaning at a deeper level than just the literal meaning of the message. A step of inference is needed when decomposing the literal message and assigning meaning to the deeper level of the message. It is assumed that humans use pragmatic inference and that speakers and listeners alike make use of the fact that we all do so in order to craft the most appropriate utterance and to understand that utterance (Wilson and Sperber, 2006).

1.2 The Rational Speech Act model

Nearly a decade ago, a formal model was introduced by Frank and Goodman (2012) to capture the pragmatic reasoning in conversation. This framework, called the Rational Speech Act (RSA) model, proposes a specification at the *computational level*[1]. In this specification it is assumed that pragmatic speakers and listeners have available a lexicon consisting of signals (the words uttered) and referents (what they mean or try to convey). The agents in the model can reason recursively about each other's lexicon distribution using Bayesian inference. However, the distribution over lexicons between two agents can be different, which poses some problems. The symmetry difference and ambiguity difference between lexicons in agent pairs are a source for possible misunderstandings in conversation. Resolving those misunderstandings can be costly (Blokpoel et al., 2019), or resolving misunderstandings can become problematic if the referent can no longer be disambiguated (van de Braak et al., 2021). A further explanation of the RSA framework and the model is given in section 3.1. This framework has been expanded upon and proven successful in explaining various different language games (Frank and Goodman, 2012; Frank, 2016; Khani et al., 2018), implicatures (Goodman and Stuhlmüller, 2013; Bergen et al., 2016) and convention formation (Hawkins et al., 2017; Cohn-Gordon et al., 2018; Hawkins et al., 2020).

1.3 Extended version and a problem

In an extension of the RSA model introduced by van de Braak et al. (2021), the authors model how RSA can be used to specify a model of communicative repair. They model what happens when agents engaged in dialogue have difficulty understanding each other and have to come to some form of agreement on what they believe to be talking about. A more in-depth explanation is given in section 3.2. The model implementation that was used could feasibly run computations for a lexicon size of 4 signals by 3 referents. This is a rather small lexicon. The feasible size is in practice restricted by a combinatorial explosion that occurs when all possible lexicons are created. Assuming that lexicons are binary, meaning that a signal either maps to a referent or it does not, the number of possible lexicons is determined by $2^{referents \times signals}$. This exponential growth can quickly become enormous even for relatively small sizes of signals and referents[2].

In the specifications of the extended RSA model, intractability looms. Although there is no formal proof yet, (but see Woensdregt et al., 2021, for proof of a similar model), the model is presumed to be intractable in the lexical updating part of the model (Blokpoel, personal communication). This intractability assumption comes from the fact that an agent has to reason over all possible lexicons when trying to figure out what lexicon the other agent could have used that lead them to their misguided inference and subsequent misunderstanding. It is clear now that searching through that space to find and update the lexicon becomes increasingly more extensive because of the combinatorial explosion in the space of possible lexicons due to increasing lexicon sizes.

[1] As defined as one of the three levels of analysis by David Marr in his influential posthumously published book Vision. (Marr, 1982)

These small lexicons and their toy-like size are nothing compared to the real world that the RSA framework tries to model. Speakers and listeners as humans have a much larger vocabulary than just 4 signals and 3 referents as used by van de Braak et al.. There is no real consensus on how extensive an adult speaker's vocabulary is, but estimates range from 20.000 words upwards to 35.000 words for native English speakers (Test Your Vocabulary Project, 2013). That is a difference of 5 orders when comparing the toy world to the real world. If the model is as intractable as conjectured, then under the widely held assumption of $P \neq NP$ (Garey and Johnson, 1979), there exists no algorithm that can go through the entire lexicon search space in polynomial time. Intractability and its effects on the explanatory power of computational-level human cognition models have been researched time and again, Kwisthout et al. (2011); van Rooij and Wareham (2012); van Rooij et al. (2019) to name a few. In the RSA model, one can wonder how much explanatory power this model has if the language the agents use to reason in can only hold very small lexicons because of the intractability constraint.

1.4 Fixed-parameter tractability

Starting in the latter half of the 1990s and gaining traction in early the 2000s, so-called fixed-parameter (in)tractability was introduced (Downey et al., 1999) and expanded upon in cognitive science (van Rooij et al., 2008). This fixed-parameter approach states that computational level models can run tractably for arbitrary input size n and are only intractable for certain aspects of the input. These aspects are called parameters. If those parameters are small or do not change much, and the problem is fixed-parameter tractable, then the problem as a whole can be computed in polynomial time. This approach is not a solution to intractability in the sense that it would solve the famous $P \neq NP$. Instead, it is a tool for looking at which parts of the problem cause intractability and a way to circumvent or limit these causes. I provide further details on this FPT framework in section 2.2.

Using the fixed-parameter approach to guide my intuitions, as explained later, I propose and explore a possible way to tackle a source of the intractability in this extended RSA model. A formal proof is outside the scope of a Bachelor's thesis. Instead, by using computer simulations I aim to provide a proof of concept that can warrant further research into the matter of fixed-parameter tractability for this RSA model.

1.5 Research question

"*Can the extended RSA model be improved in terms of feasibly computable lexicon size by constraining the possible lexicons between agent pairs?*"

This question stems from the ideas and reasoning that the explanatory power held by computational level models is related to the (in)tractability of such models, as mentioned earlier by, among others, van Rooij. Computational level models of human cognition need to be tractable since we as humans presumably do not have a way of solving intractable problems with our brains. There is evidence that suggests that there are many ways of constraining such problem-solving algorithms that we both consciously and unconsciously apply to intractable problems (Lieder and Griffiths, 2019).

[2] A lexicon of 3 by 3 has $2^9 = 512$ possibilities, a lexicon of 8 by 8 has $2^{64} = 1.844\,674\,4 \times 10^{19}$ possibilities.

1.5.1 Introducing constraints

To increase the computable lexicon size that the RSA model can handle tractably, I propose two additional constraints on the lexical search space that the agents have to reason over. The first constraint is that only those possible lexicons with very little difference from the agent's current lexicon should be considered by them when generating the set of possible lexicons to search through. Assuming that the other agent they are trying to understand is rational and speaks the same language, it is reasonable that the two agents' lexicons are similar in their mappings, even if not identical.

The second constraint is that lexicons have some mappings that cannot be changed or changed very few times to resolve a misunderstanding. These signal-referent mappings are so fundamental to the language that the agents reason in. Without them, the language does not make sense. Both agents are thus grounded in their common language, and the use of that language poses limitations on a lexicon. Think of sentence structures like certain adjectives always going together with specific nouns or mappings that are considered facts. For instance, the word red will refer to the color red and not blue or green. Therefore, all possible lexicons that would change the mapping for one or more of these elements can be ignored in the search.

By introducing these two constraints, we can pose an upper bound on the size of the search space. Knowing that the presumed intractability lies in searching through this space, making it sufficiently smaller could mean an increase in the size of lexicons that the agents can reason over in a feasibly computable time.

1.6 Overview

In order to be able to answer the research question, I will first explain in section 2 several notions in related work in order to further position this thesis in the relevant literature. Then in section 3, I will explain and show the RSA models by Frank and Goodman, van de Braak et al. and then introduce my extension on a computational level. Next, in section 4, I use this specification in order to implement my model, explaining further implementation assumptions made and show the results of the simulations. Finally, in section 5 the results are discussed, and I answer the research question.

2 Related Work

2.1 Computational models

In the ongoing effort to make sense of the world around us, we try to find an explanation for how and why things like humans, animals, computer systems or asteroids do what we observe them do. One such explanatory approach is to make a model of a complex system, trying to capture the significant elements of that system simplified in a model and see if we can understand its parts. Then, by understanding the individual parts, we try to explain the system as a whole. Countless models exist; think of the solar system, weather predictions, or a model for tidal behavior. In computational modeling, in which this thesis is embedded, the approach comes from a formal and mathematical background. Trying to define formulas, mappings of input to output, and their interplay to understand, explain, or predict a complex system. We are mainly interested in the behavior of humans, and thus we try to model that behavior to understand it and maybe even recreate it.

2.1.1 Intractability

Certain problems and optimal solutions are considered intractable, assuming the famous $P \neq NP$ conjecture. Finding a(n optimal) solution takes exponential time[3]. This exponential growth is a problem for computational modeling since some of the formulas used are intractable. They can only be considered applicable to practice for relatively small input sizes. Modeling human behavior with intractable models is problematic at least (van Rooij et al., 2019). I argue that intractable models are not a good explanation for human behavior. Our brains are evolved for problem-solving. We solve many problems daily and do so in a fast and efficient manner, definitely not on a scale of universe-time. How can a model be a good explanation if it does not scale to the stage it tries to model? (Woensdregt et al., 2021). As stated before, the alleged intractability of the extended RSA mdoel lies in the immense lexical space and the incapability of any known algorithm to search through that space in polynomial time. The search space is defined by, assuming binary lexicons, $2^{\text{signals}\times\text{referents}}$. In the case of a human-level vocabulary, which has about 20.000 words and maybe even more than 20.000 referents, the size of the lexicon that represents that vocabulary and the number of possible lexicons that follows is enormous. Our brains cannot even comprehend such a large number. I cannot even begin to explain how large a number with 100 million digits is, which is the result of $2^{20.000\times20.000}$.

2.1.2 Pitfalls of handling intractability

After a problem has been proven intractable (again, assuming $P \neq NP$), solving it becomes prohibitively difficult for all but small instances. Let us say that a computational model provides a good explanation and prediction for some aspects of human behavior, but it is considered intractable. The model only runs tractably for relatively small sizes of input. This does not scale to the larger real-world stage. A common approach is to use approximation. Finding the optimal answer is not necessary to run the model

[3] A small example. Say that finding a solution to a problem is dependent on the input size n of that problem, and it takes 2^n operations to find that solution. The input size is 100; thus 2^{100} operations are needed. If a computer could do 10^{12} computations per second, it would still take 4×10^{10} years in order to find that solution. This number is on the same timescale as the age of the universe.

but instead use some sub-optimal answer. This answer satisfies some, but maybe not all parts of the model. Using this approach, a model can still be a valid prediction and explanation to some parts of human behavior. However, there is a pitfall by using an approximation to combat intractability. The approximation is not guaranteed to be good in terms of accuracy, or maybe not in terms of structure (Kwisthout et al., 2011; Kwisthout, 2013). Finding an approximated local optimum is not necessarily the global optimum. Furthermore, there is no guarantee that a suitable approximation method is tractable in either form. Regarding this approach to *solving* the intractability in the RSA model, RSA models how humans use pragmatic inference to communicate effectively. Using approximation techniques to arrive at a pragmatic inference, we would inherently be unsure about what the other is saying. This is true to some extent, and Bayesian inference is rarely about being 100% sure. However, I argue that using approximation techniques will be detrimental to pure, effective communications.

Another argument is that computational power will increase, and in time we can reduce the resources needed to solve intractable problems. According to Moore's law, the amount of transistors on a computer chip roughly doubles every two years. However, this doubling has been slowing down in the past decade. Taking the example of footnote 3, doubling the computers speed every two years, in about 6.6 years, the time it takes to find the solution is *only* about the age of the earth, one scale lower than the age of the universe, still a very, very large number. This argument does not attempt to solve intractability by finding better solutions but by finding solutions faster. This reasoning is not what we need for solving intractability because it inherently does not tackle the critical issue that finding a solution takes exponential time, not polynomial time. In the case of the RSA model, if it is indeed the model of how humans use pragmatic inference in communication, it would mean that RSA is somehow implemented in our brain. Our brains do not double their computing speed every two years. That computing speed has been set by way of evolution over an expansive time period. The increasing speed argument thus is not the way to prove the viability and explanatory power of the RSA model.

Other approaches of dealing with intractable models explore the idea that the theoretical grounds from which these models stem need to be addressed. The field of computational cognitive modeling should be more focused on building models that are better embedded in the real-world constraints that the cognitive phenomena are in. Their theoretical grounds may have lost touch with the practical field that they are modeling (van Rooij and Baggio, 2021).

2.1.3 Explanatory power

Models such as the RSA model can help us in gaining insight into human communication. In the toy world, the model can effectively explain how interlocutors arrive at mutual understanding. The framework was based on results gathered by Frank and Goodman (2012). In their experiment, they had only used 3 referents and 2 signals. Nevertheless, that size is just it. Their experimental model does not scale to the real world, where lexicons are enormous. Such a model then can not be a reasonable explanation for how we humans communicate. It might be so in the basis, the idea might be correct, but it can not be the whole picture, the real explanation. Humans react almost instantly, or at least in mere seconds, to a signal received from another agent. The computing time necessary to go through the entirety of the lexical space is on an entirely different level, planet, or even universe. I, therefore, argue that the RSA model in its current form is not in good shape to explain how humans use pragmatic inference with such great precision and effect.

2.1.4 Human approach

If humans indeed use the RSA model in their communication, there must be some way that the inherent intractability is circumvented. There has been much research done into the effects of intractability and 'as-if' approaches (van Rooij et al., 2018), cognition and approximation techniques (van Rooij and Wareham, 2012) and understanding the limited resources available to human cognition (Lieder and Griffiths, 2019). While most models have intractable functions in pure theory and unbounded input domains, their intractable functions might prove to be tractable for a subset of inputs when these models meet practical world-bound domains. A framework was provided and introduced as fixed-parameter tractability (Downey et al., 1999). I intend to guide my approach by this framework for improving the tractable lexicon size for the RSA model.

2.2 Fixed-parameter tractability

Downey et al. introduced a general framework for making problems Fixed-Parameter Tractable (FPT). The authors detail the framework and how FPT works for several different problems and known algorithms in their publication. They also discuss what is not FPT and how this framework does not apply to all known NP-complete problems.

2.2.1 General idea

The idea is that a solution to an NP-complete problem can still be found in tractable (polynomial) time. Only a small part of the problem is intractable. This intractable part is then 'parameterized.' Take for example the VERTEX COVER problem, see figure 1 for an illustration (Weisstein, 2009). Given a graph G of vertices V and edges E, a vertex cover is a subset S of V such that every edge in E is incident on a vertex in S. In other words, every node in the graph is chosen such that all edges have at least one of their endpoints touching a vertex in S. All red-colored vertices are in S. Finding a set S such that S is minimal (as few vertices as possible) is considered to be NP-hard. The parameters in the FPT version of the VERTEX COVER problem are the size n of the graph G and the upper bound on the number of vertices k allowed to cover the graph. For other problems parameters for FPT may come from hardware-specific limitations like the number of transistors on a computer chip or the degrees of freedom when a robot is planning its next move. We look at known, informative, and finite distributions over that parameter. These smaller ranges of values limit the amount of complexity stemming from one or multiple parameters. Usually, these parameters are relatively small, not much bigger than $k = 60$, as given by Downey et al. (1999).

FPT uses practical information about the problem domain and applies that to make informative and good choices in designing an algorithm. This similar to the use of heuristics. FPT-algorithms are a more explicit form of the already used and implicit distributional bounds of heuristic parameters. The key difference between the two is that heuristics are inexact approximations and generalizations of a problem, without provable performance. Whereas FPT-algorithms are exact and proven to be tractable for the limitations that they pose.

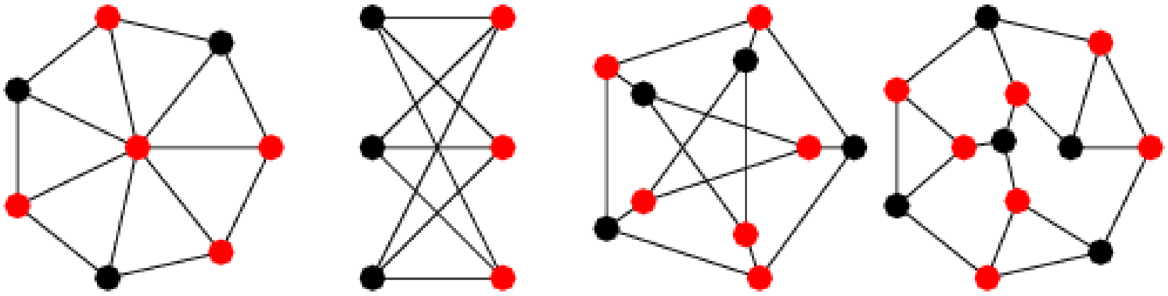


Figure 1: Red vertices are in a vertex cover S.

A subset of NP-complete problems is considered to be in the class of FPT problems for one or more parameters. First, one needs to find a viable translation of the problem to a parameterized version of the problem and then prove it is either in $W[1]$ (the parameterized analog of NP) or not $W[1]$-hard. If it is neither, then the problem is considered fixed-parameter *intractable*.

2.2.2 What FPT is not

It is not the silver bullet that silences all intractability problems. Shielding the part of the input that is tractable from the intractable part but feasibly computable using the parameterization mentioned above still does not make the problem suddenly tractable. The problem still is intractable. Moreover, only a subset of problems that are NP-hard can be reduced to be FPT. Thus, there is still a set of problems that is provably unsolvable by applying the FPT framework.

3 Computational Models

3.1 Frank and Goodman

3.1.1 Background

According to Frank and Goodman and their Rational Speech Act model framework, agents adhering to this model produce signals S and infer referents R in compliance with Gricean communication (Grice, 1975). These signals and referents are collected as a mapping in a so-called lexicon $\mathcal{L}$. Using Bayesian inference, speaking agents choose the most probable signal they think the listener will infer to mean the intended referent. Listening agents will choose the most probable referent given the signal from the speaker, assuming the speaker could not have chosen a better signal. Thus, both speaker and listener take the perspective of the other to infer the intentions of the other. This perspective happens in a recursive fashion that ends in a literal listener who accepts the lexicon without question.

3.1.2 Computational level theory

Although Frank and Goodman's RSA model is simple in its Bayesian inference definitions, these definitions are recursively defined. Speaking and listening can happen at higher orders, denoted by n. A speaker of order n forms a distribution over possible signals given the referent r and lexicon $\mathcal{L}$ that is proportionate to the presumed inference made by a listener of one order lower, and some cost function. A listener of order n forms a distribution over possible referents given the signal s and lexicon $\mathcal{L}$, that is proportionate to a prior over all referents $Pr(r)$ and the presumed inference made by a speaker of the same order. This recursion ends as a literal listener who has a distribution over referents given the signal and lexicon proportionate to a prior over all referents and a lexicon. See figure 2, reprinted from Blokpoel et al. (2019) with permission.

$$S_n\,(s|r,\mathcal{L}) \propto exp(\alpha\, log\, L_{n-1}(r|s,\mathcal{L}) - cost(s)) \tag{1}$$

$$L_n\,(r|s,\mathcal{L}) \propto Pr(r)\; S_n(s|r,\mathcal{L}) \tag{2}$$

$$L_0\,(r|s,\mathcal{L}) \propto Pr(r)\; \mathcal{L}(s,r) \tag{3}$$

3.1.3 Remarks

Although this framework is simple in its definition, many parameters play a role here. It has been found that creating agents in higher orders do not really improve or are more easily differentiated from first-order agents when going beyond the second-order of inference. Whether this phenomenon is a limit of human cognition and its capability for pragmatic recursive reasoning or something intrinsic in the RSA framework has been studied before (Frank, 2016; van Rooij et al., 2019; Blokpoel et al., 2019). The effects of different priors over referents have been researched, but only in single interaction games. The author concluded that the effect of differing priors was not that strong, so having uninformative or uniform priors is generally a good idea (Frank, 2016).

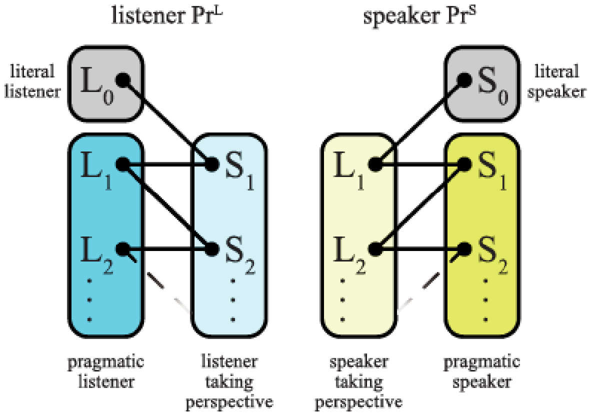


Figure 2: Recursive speaker and listener.

Reprinted from Blokpoel et al. (2019) with permission.

How a mapping of signals and referents is made, particularly the effects of ambiguity and asymmetry, is important in effective communication. Ambiguity means that one signal could have multiple referents, and one referent could have multiple signals. Asymmetry comes from a difference in the lexicons agents use to reason over. Having some degree of ambiguity and asymmetry between agents who communicate tends to facilitate the communication (van Rooij et al., 2019). Two parameters that do not affect the accuracy of RSA are the number of referents and the number of signals. However, these meta parameters are critical in the size of lexicons that agents can reason over tractably, as stated before.

3.2 Van de Braak et al.

3.2.1 Background

Stepping beyond single reference games and building towards a richer model of human conversation, van de Braak et al. have built upon an RSA specification in a paper by Hawkins et al. (2017). Hawkins' model takes into consideration a conversation history and attempts to model convention-formation in iterated reference games. A listener uses Bayesian inference to arrive at a distribution over signal-referent mappings a speaker uses, given the previously used signals. Van de Braak et al. extend this model by making a distinction between so-called ostensive and non-ostensive conversation. Ostensive communication is defined there as being able to use other means than verbal communication to indicate a referent, e.g. pointing to it. Non-ostensive communication is defined as the lack thereof. Arriving at mutual understanding through words and the uncertainty inherently accompanied by this is more difficult when the referent in question can no longer be clearly disambiguated by pointing to it in the physical world. They have tried to model this communicative repair by introducing initiators and responders. Both agents take on the role of speaking and listening, but they differ now in who has asked what signal or referent they are uncertain about. The theory below explains at a formal level the roles of initiator and responder.

3.2.2 Computational level theory

The equations below are an extension upon equations (1, 2, 3) in section 3.1.2. I refer the reader to the paper by van de Braak et al. for a full specification of both models. I will only include the non-ostensive model, because this model is a more relaxed version of modeling communicative repair.

$$\Pr_{L_n}(\mathcal{L}|h) \propto \Pr(\mathcal{L}) \prod_{(s,r)\in h} L_n(r|s,\mathcal{L}) \tag{4}$$

$$\Pr_{L_n}(r|s,h) \propto \sum_{\mathcal{L}} L_n(r|s,\mathcal{L}) \Pr_{L_n}(\mathcal{L}|h) \tag{5}$$

$$\Pr_{S_n}(s|r,h) \propto exp\Big(\alpha \ln\Big(\sum_{\mathcal{L}} \Pr(\mathcal{L}|h)\, L_{n-1}(r|s,\mathcal{L})\Big) - cost(s)\Big) \tag{6}$$

$$\Pr_{L_n}(\mathcal{L}|h) \propto \Pr(\mathcal{L}) \prod_{(s_{\text{initial}},s_{\text{request}})\in h} \sum_{r} S_n(s_{\text{initial}}|r,\mathcal{L}) L_n(r|s_{\text{request}},\mathcal{L}) \tag{7}$$

NON-OSTENSIVE RESPONDER

Input

A set of signals S, a set of referents R, and the conversation history $h = ((s_{\text{initial}}, s_{\text{request}}), \dots)$. An observed signal $s_{\text{observed}} \in S$ and an order of reasoning n. A linguistic bias representing the agent's preferred lexicons $\Pr(L)$. A rationality parameter α, a certainty threshold η and a cost function $cost : S \to \mathbb{N}$.

Output

A clarification request s_{request} if $H(\Pr_{L_n}(r|s_{\text{observed}}, h)) > \eta$ (Eq. 5, Eq. 7); or otherwise a the inferred referent r_{inferred} and a special confirmation signal *aha!* Here, s_{request} is sampled from $\Pr_{S_n}(s|r_{\text{inferred}}, h)$ (Eq. 6, Eq. 7) and r_{inferred} is sampled from $\Pr_{L_n}(r|s_{\text{observed}}, h)$ (Eq. 5, Eq. 7).

NON-OSTENSIVE INITIATOR

Input

A set of signals S, a set of referents R, and the conversation history $h = ((s_{\text{initial}}, s_{\text{request}}), \dots)$. A referential intention $r_{\text{intended}} \in R$ and an order of reasoning n. A linguistic bias representing the agent's preferred lexicons $\Pr(L)$. A rationality parameter α, a certainty threshold η and a cost function $cost : S \to \mathbb{N}$. Optionally, an observed clarification request $s_{\text{request}} \in S$.

Output

The initial signal s_{initial} sampled from $\Pr_{S_n}(s|r_{\text{intended}}, h)$ (Eq. 7, Eq. 7) if no clarification request was made, or else a clarification response s_{response} sampled from the same distribution if $H(\Pr_{L_n}(r|s_{\text{request}}, h))$ is above η (Eq. 5, Eq. 7), or a special confirmation signal *indeed!* if entropy is below η and r_{intended} matches r_{inferred} sampled from $\Pr_{L_n}(r|s_{\text{request}}, h)$ (Eq. 5, Eq. 7).

3.2.3 Remarks

The authors clarify that their extended RSA model applied to this part of human conversation still has pitfalls. Breaking free from the huge amount of underdetermination is difficult, even more so in non-ostensive repair. Underdetermination here means that the agents in the model are unsure about what exactly was meant, even after communicative repair took place. While the authors focus on the comparison of both the ostensive and the non-ostensive repair of miscommunication and reaching mutual understanding, I have only included the non-ostensive model. I choose to expand upon this model because the inherent conjectured problem of intractability is present for both models. However, if the referent can be clearly disambiguated by means other than verbal communication as is the case in the ostensive model, the constraints that I intend to implement lose their communicative meaning for modeling resolving misunderstanding. Being able to point to the sun, and say sun would make it hard for the other agent to attain to the fact that 'sun' - 'big bright orb glowing in the daytime sky' means anything other than what is being pointed at. There would be no use for the agent to use a fact to generate possible other lexicons to resolve a miscommunication.

3.3 Human language constrained RSA

3.3.1 Background

Having touched upon both RSA models, I can now introduce the proposed extensions on a formal level and provide a computational level specification. First, I will address the size of the lexicon space and the so-called consistent lexicons in that space. Second, I introduce my two constraints that are parameterized as ν for skepticism and γ for neighborliness, and their resulting lexical sets are expanded upon. Having done that, I then provide a computational level specification.

3.3.2 Size of the lexicon space

As stated before, the size of the the set $\mathcal{P}$ containing all possible lexicons, assuming binary lexicons, is $2^{\text{signals}\times\text{referents}}$. This space can be huge for even small signals s and referents r. However, this set is constrained by the mere virtue of being lexicons embedded in human language. Theoretically, this set is all lexicons that are possible, but the actual number of *allowed* lexicons is a smaller set. In the following text, I will assume a small lexicon of 3 signals and 3 referents. The number of possible lexicons then becomes $2^{s\times r} = 2^{3\times 3} = 2^9 = 512$. A small number, manageable for the explanations below. Do not forget about the exponential growth; an also arguably small setup of 5 signals and 5 referents would have a total of 33.554.432 possible lexicons.

3.3.3 Consistent lexicons

All possible lexicons are 512 in total. This number is reduced by the implicit assumption in the RSA framework that all lexicons the agent reasons over are considered consistent lexicons. This assumption translates to each signal having at least one referent it refers to and each referent having at least one signal that signals to it. Imagine a lexicon as a two-dimensional matrix. Then there can not be a row or column consisting of only 0's. All these lexicons are thus discarded before any actual search is done through the lexical space. This means that the consistent lexical search space is reduced from all possible lexicons by excluding those lexicons that have only 0's on either any row or any column. The formula to calculate this number is somewhat complex (Jovovic, 2008). Using the inclusion-exclusion principle applied to $m \times n$ binary matrices where m is the number of signals and n is the number of referents, we get a definition for the set of consistent lexicons $\mathcal{C}$:

$$\mathcal{C} = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \left(2^{n-k} - 1\right)^m \qquad \text{(consistent set)}$$

In the running example of 3×3 lexicons, the total number of consistent lexicons is now only 265, about half of the total 512 possible lexicons. This halving effect does not hold, however. For a 5×5 lexicon, the number of consistent lexicons is 24.997.921, which is not around half of 33.554.432, but more around three-quarters.

3.3.4 Factual lexicons

The number of consistent lexicons can be reduced even further by introducing my two constraints, as discussed before. First, I will discuss the effect of introducing facts to the lexicons. A fact is a fixed entry where a mapping of signal and referent has a value of 1. This value has to be 1 in order to denote an existing mapping. A value of 0 means that no mapping between that signal and that referent is present. This mapping is a design choice and an assumption on my specification. The factual set is derived from the set of consistent lexicons by placing a constraint on the lexicons. Only those lexicons that have the fact in them are included. Let ν be a fact or a list of facts, each fact meaning that a certain signal has a mapping to a certain referent, and l is a lexicon. The factual subset $\mathcal{F}$ is then defined as:

$$\mathcal{F} = \forall\, l \in \mathcal{C} : \mathit{contains}(l,\, \nu) \qquad \text{(factual set)}$$

where *contains* is a simple operation that tests whether the specific signal-referent mapping(s) ν is (are) present in the lexicon l. The testing of this factual presence is an operation done in constant time. Computing this factual subset is thus dependent on the size n of set $\mathcal{C}$ and the number of facts in ν. In terms of Big O notation, computing the factual set has a time complexity of $\mathcal{O}(\#\nu \times n)$. Unfortunately, there is no nice closed-form formula to calculate the number of lexicons in the factual set as derived from the consistent set. However, there is a formula to compute the number of factual lexicons derived from the entire possible set. The number of facts in ν limits the degrees of freedom in the power of 2. Where $\mathcal{P} = 2^{s \times r}$, the amount of factual, but also including inconsistent lexicons, becomes $\mathcal{F}' = 2^{s \times r - \nu}$. The following numbers were obtained through the implementation of the formulas and definitions above. In our running example, the number of factual lexicons for a single fact is reduced from 265 to 161. This is a decrease in size of about 39%. Introducing a single fact to a 4×4 lexicon gives a set of 23.045 factual lexicons out of 41.503 consistent lexicons. This is a decrease of about 44%.

3.3.5 Neighboring lexicons

The other constraint that I have discussed is introducing agents using neighboring lexicons when generating their possible lexicons. I postulate that humans make use of this when resolving misunderstandings in communication. One is more likely to consider different lexicons close to the current lexicon than to consider lexicons far removed from the current lexicon. The terms close and far are used as a measurable distance in two-dimensional signal-referent space where two lexicons are close to each other if they only differ on very few mappings and far away if they differ on many such mappings. One implementation of this difference or neighborliness is by using the Hamming distance. This distance is defined as the absolute number of differences in each signal-referent mapping between two lexicons. This distance is then normalized over the total lexicon size such that regardless of the lexicon size, neighborliness is the same for small and large lexicons. In terms of a formal definition, where γ is the neighborliness parameter between 0 and 1 inclusive that defines a lower bound on how close another lexicon must be in order to be in the neighboring set, and $\mathit{Hamming}(l_1, l_2)$ is the function that defines the Hamming distance between two lexicons. The set of neighboring lexicons $\mathcal{N}$ is then defined as:

$$\mathcal{N} = \exists\, l_{\text{current}} \in \mathcal{C}, \forall\, l \in \mathcal{C} : \frac{\mathit{Hamming}(l_{\text{current}},\, l)}{s \times r} \geq \gamma \qquad \text{(neighboring set)}$$

The time complexity for finding the Hamming distance is done in $\mathcal{O}(s \times r)$, the size of the lexicon. This means that finding the neighboring set can be done in $\mathcal{O}(s \times r \times n)$, because the remaining operations are a constant factor and are repeated n times, where n is the size of set $\mathcal{C}$. The size of the set $\mathcal{N}$ is dependent on the neighborliness parameter γ. As with the factual set, there is no closed-form formula to calculate the number of neighboring lexicons as derived from the consistent set. The problem of choosing the members of set $\mathcal{N}$ can be thought of as an n *choose* k problem. A lexicon can be flattened to a 1-dimensional binary string. The length of that string becomes n. Flipping k elements of that string can be done in $\binom{n}{k}$ possible ways. Of course, k has an upper bound defined by its relation to the neighborliness parameter γ. Parameter k is the highest positive non-zero integer that still satisfies $\gamma \leq \frac{k}{s \times r}$. However, the maximum k is not an indicative number of the size of $\mathcal{N}$, as this number does not include the lexicons that can be generated using fewer than k flips. This would be a sum of all values up to k, as in $\sum_{i=1\ldots k}^{n} \binom{n}{i}$. Even still, this number would count possible lexicons several times since it can be the case that a flip can occur twice in the same place, making the resulting lexicon identical to the starting point. Finally, the resulting number, even if accounted for counting lexicons multiple times, would still be a derivative from the total set $\mathcal{P}$ and not the consistent set $\mathcal{C}$.

In our running example, by choosing a γ of 0.70, the set of neighboring lexicons is of size 36. Only the 30% closest lexicons are considered in order to resolve a misunderstanding in communication between agents. The chosen value for gamma is arbitrary for this example.

3.3.6 Applying all constraints

When we apply all three constraints to the lexical search space, these being consistent, factual and neighboring, we get a much smaller set of *allowed* lexicons. The set of allowed lexicons $\mathcal{M}$ in my model is then defined as the intersection of the four sets:

$$\mathcal{M} = \mathcal{P} \cap \mathcal{C} \cap \mathcal{F} \cap \mathcal{N} \qquad \text{(allowed set)}$$

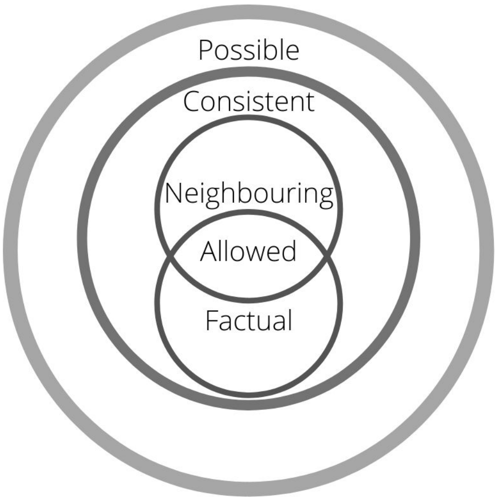


Figure 3: Illustration of the relations between lexical sets

An illustration of this relation is given in figure 3. In our running example of 3 × 3 lexicons, fixing ν at 1 fact and γ at 0.70, the resulting set $\mathcal{M}$ has a size of 22. This is an almost 95% size reduction of the superset $\mathcal{P}$, which holds 512 possible lexicons.

3.3.7 Computational level theory

HUMAN LANGUAGE CONSTRAINED RSA MODEL

Input

Everything from the NON-OSTENSIVE MODEL for both the RESPONDER and the INITIATOR with a change in the distribution over prior lexicons $\Pr(\mathcal{L})$. The neighborness parameter $\gamma = [0, 1]$ and the list of facts ν where each fact is a non-zero mapping between a signal s and a referent r in the lexicons from set $\mathcal{L}$. The set $\mathcal{M}$ (Eq. allowed set, 3.3.6) replaces set $\mathcal{L}$ in the distribution over linguistic prior, and it thus becomes $\Pr(\mathcal{M})$.

NON-OSTENSIVE RESPONDER

A set of signals S, a set of referents R, and the conversation history $h = ((s_{\text{initial}}, s_{\text{request}}), \dots)$. An observed signal $s_{\text{observed}} \in S$ and an order of reasoning n. A linguistic bias representing the agent's preferred lexicons $\Pr(\mathcal{M})$. A rationality parameter α, a certainty threshold η and a cost function $cost : S \rightarrow \mathbb{R}^{+}$.

NON-OSTENSIVE INITIATOR

A set of signals S, a set of referents R, and the conversation history $h = ((s_{\text{initial}}, s_{\text{request}}), \dots)$. A referential intention $r_{\text{intended}} \in R$ and an order of reasoning n. A linguistic bias representing the agent's preferred lexicons $\Pr(\mathcal{M})$. A rationality parameter α, a certainty threshold η and a cost function $cost : S \rightarrow \mathbb{R}^{+}$. Optionally, an observed clarification request $s_{\text{request}} \in S$.

Output

Everything from the NON-OSTENSIVE MODEL for both the RESPONDER and the INITIATOR.

NON-OSTENSIVE RESPONDER

A clarification request s_{request} if $H(\Pr_{L_n}(r|s_{\text{observed}}, h)) > \eta$ (Eq. 5, Eq. 7); or otherwise the inferred referent r_{inferred} and a special confirmation signal *aha!* Here, s_{request} is sampled from $\text{Pr}_{S_n}(s|r_{\text{inferred}}, h)$ (Eq. 6, Eq. 7) and r_{inferred} is sampled from $\text{Pr}_{L_n}(r|s_{\text{observed}}, h)$ (Eq. 5, Eq. 7).

NON-OSTENSIVE INITIATOR
The initial signal s_{initial} sampled from $\text{Pr}_{S_n}(s|r_{\text{intended}}, h)$ (Eq. 7, Eq. 7) if no clarification request was made, or else a clarification response s_{response} sampled from the same distribution if $H(\text{Pr}_{L_n}(r|s_{\text{request}}, h))$ is above η (Eq. 5, Eq. 7), or a special confirmation signal *indeed!* if entropy is below η and r_{intended} matches r_{inferred} sampled from $\text{Pr}_{L_n}(r|s_{\text{request}}, h)$ (Eq. 5, Eq. 7).

3.3.8 Remarks

The HUMAN LANGUAGE CONSTRAINED RSA MODEL has several assumptions. First and foremost is the assumption that a fact ν is not mutually exclusive with other referents bound to the same signal. The input just states that a certain referent must always be referred to by that signal. I chose to leave it like this to account for utterances where the speaker could have another intended referent than the commonly accepted fact. Think of cases like lying about the intended referent or using the knowledge that a signal *always* means one referent in one context, but not necessarily so in another context. The word sun always refers to the big bright glowing orb in the daytime sky here in our solar system. It could be entirely different for another alien system. Granted, the latter is an extreme example, but it drives my point home, I believe.

The second assumption is that both agents modeled in this framework share the same list of known facts ν. In turn, this would mean that what I model is to be considered world knowledge or conventions, the things that are known to *all* agents (that speak the same language), everywhere. I chose this assumption because otherwise, agents could have vastly different facts consisting of fixed signal-referent mappings and thus different resulting lexicons. This would mean that agents could never reach an understanding when using the RSA framework because they do not consider the lexicon of the other to be possible. However, relaxing this assumption and formalizing the idea that everybody has different ‘facts’ could be a good idea for future research.

My third assumption comes into view when we consider where the list of facts comes from. As stated before, the list of facts is world knowledge by every agent in the model. Who or what determines world knowledge, and how does it emerge the same for all agents? I have used arbitrary facts for my model implementation since the used lexicons themselves do not represent real-world mappings of words and meanings.

My final assumption concerning facts is about the value a fact takes. A 1 denotes an existent mapping in the binary lexicon, and a 0 a non-existent mapping between that signal and that referent. I choose to have facts take a value of 1. A fact is thus always an existing mapping. However, it could also be possible for facts to have a 0 as mapping. When this is the case, it follows that a signal-referent mapping cannot possibly exist for an agent adhering to that 0-fact. An example of this would be that the word sun could never refer to the (most of the times) glowing orb in the nighttime sky, more commonly referred to as the moon. Applying this extra constraining factor for non-existent facts could be an intriguing research entry point.

4 Simulations

4.1 Implementation

For the implementation of the HUMAN LANGUAGE CONSTRAINED RSA MODEL (HLC-RSA), I built upon the work done by van de Braak et al.. Their code is available in appendix B. In order to be able to compare my extension with that of the authors, I ran the model with the same parameters. There are 500 agent pairs of the first order, each reasoning over a lexicon of 4 signals and 3 referents. The rationality parameter α is set to 5, and since it is irrelevant for the answer to my research question, the cost function was omitted. The distribution over the lexical prior is a binomial distribution with $X = 0.5$. A dialogue between two agents started with the initiator agent communicating an intended referent to the responder agent. Each dialogue consisted of 6 turns. All agent pairs engaged in 6 such dialogues, with each dialogue having a random intended referent at the start. I have implemented the factual and neighborliness constraints as specified in my model. Both agents share the same list of facts, and both agents have the same neighborliness parameter value. My code, images, and data interpretation scripts are available in appendix B.

4.1.1 Assumptions

Most of these parameters were left the same across both models to ensure valid comparison. However, I did have to change the value for the binomial distribution. The change was made because of limitations in the underlying software that was used. This change, in turn, means that the agents in the model are no longer asymmetrical. The effects of asymmetry on the RSA model are explained further in Blokpoel et al. (2019). In short, breaking this asymmetry, or instead adhering to symmetry, no longer allows agents to enjoy some of the benefits of asymmetry.

4.1.2 Limitations

Even in a constrained version, the implementation still has its limitations. The most important constraint is that the model could not run for lexicons greater than 4 by 3. A more extensive lexicon would cause a StackOverflow error due to the high number of all possible lexicons. These are still generated before any constraints are applied. Increasing the allowed space to the limits of my machine (8 GB of RAM) did not resolve this issue. The problem lies in the Java Virtual Machine that the model runs on, something I have no influence over. However, I remain optimistic because the runs that were computable did so in a relatively fast time.

Another limit I had to keep into account was the implementation of the neighborliness parameter. The neighborliness value had to be the same for both initiator and responder. The code that I built upon expects that distributions over possible lexicons are the same for both agents. This expectation meant that both agents had to consider the same set of neighboring lexicons. What that means for the change in specification accompanied by this is the following. Instead of choosing the most neighboring lexicons at the start of each turn, the agents choose the same set at the start of each dialogue. This is then becoming a model where agents have a prior over lexicons that remains uninfluenced by the other agent. Several lexicons are available for an agent to consider, but that amount will not change throughout the dialogue due to the neighboring lexicons. It will still change due to other parameters and updating in the model.

4.2 Testing

In order to test for my research question, I vary the two parameters to see if they have any effect on the accuracy of the model. Furthermore, I count the time it took for the model to run, benchmarking it on my machine for a fair comparison. Because of the limitations mentioned above, I cannot answer any implementational questions regarding lexicon sizes larger than 4 by 3. The math as specified in sections 3.3.2 through 3.3.6 will have to provide for this. I will provide a theoretical answer in the conclusion section.

4.2.1 Varying the facts parameter

I have varied the list of facts for a 4 by 3 lexicon. There are four possibilities that I tested. These are either no facts, one fact, two facts for two different signals and referents, and two facts for the same signal, but different referents. I chose the last variation for a complete analysis. As stated before, having a fact that is multiple referents and one signal could be meaningless in particular conversational contexts. The first variation is an unconstrained model to test the effect of neighborliness on its own. At the same time, the middle two values were chosen to see if the model's accuracy takes a hit when these facts constrain the search space.

4.2.2 Varying the neighborliness parameter

The neighborliness parameter was varied between four levels for the same lexicon of 4 by 3. These values were 0.0, 0.50, 0.70 and 0.90 respectively. The first value is to determine if there is an effect of facts regardless of neighborliness and to run an unconstrained version of the model for time and space comparison. The other values are to show a possible effect of the neighborliness parameter. I did not go higher than 0.90 because that would constrain the model so that it could not be any real-world model. A value of 1.0 means that the agents only have one lexicon in their possible search space. That value is unrealistic and is not insightful or a good model of resolving misunderstandings.

4.2.3 Analyses performed

Three different analyses were done to understand and explain the effects of constraining the lexical search space in HLC-RSA. The first is of clarification sequence length, the second a dialogue analysis, and finally an exposition of the actual numbers of reaching factual understanding and the success rate. Some of the results are compounded in a single image in appendix A. For the full resolution individual images, I refer the reader to appendix B.

4.2.3.1 Clarification sequence length

These results are an analysis of the clarification sequence length as van de Braak et al. performed. It investigates how many turns in successive dialogues were needed for the agents to come to a believed understanding. Figure 12 in appendix A shows the full results of this analysis.

4.2.3.2 Dialogue

This analysis focuses on the factual (mis)understanding the agent-pairs reach, in what turn of the dialogue, and how often they are correct in their understanding. This analysis is shown in full in appendix A figures 13, 14, 15 and 16. The image is split up over four pages to make it readable yet still informative.

4.2.3.3 Factual and perceived understanding

The third analysis is done to gain some insight into the understanding that the agents reach. Note that because of the number of referents, the chance level for a correct guess is 33.3%. The data in table 1 is split up between when the agents had perceived understanding when they had reached factual understanding and when they gave up when they had reached factual understanding. The latter means that the agents will guess and have guessed correctly, given all the information they have gathered over the 6 turns in a dialogue.

4.2.3.4 Time and space analysis

My final analysis is about the time it took to run a specific model and the model's lexical search space. To this end, several scatter plots granted enough insight into the relation between ν, γ, and time and space used.

4.3 Results

4.3.1 Clarification sequence length

The orange line is the baseline that the model was compared against (figure 4). This can be seen as an HLC-RSA model with $\nu = 0$ facts and $\gamma = 0$ neighborliness. The blue line is the HLC-RSA model for which the parameters were varied, plotted over the orange baseline for varying parameter levels. For some combinations of ν and γ, the decrease in clarification sequence length is very similar to that of the unconstrained version (figure 5). This length is already low for other levels from the start of the first dialogue, or there is only a small decrease (figure 6). Increasing the number of facts makes the HLC-RSA model behave more like the non-ostensive baseline, whereas increasing the neighborliness makes the HLC-RSA model behave more like needing a constant number of turns regarding the clarification sequence length. There is an interaction effect between the two parameters. For the full image of all parameter variatons, I refer the reader to figure 12 in appendix A, or to the repository for the high resolution image.

4.3.2 Dialogue

In the baseline model, understanding comes mostly after the 3rd dialogue has taken place (figure 7). After the second half of the conversation, the clarification sequence length is short, < 3 on average. However, this understanding is not always correct, as can be seen by the orange-colored bars that represent factual misunderstanding. The agents believe they understand each other but are mistaken in their assumptions. When the agents reach turn 7, it means they have given up on trying to understand each other and will guess at this point. What can be seen in this analysis are the different effects of ν and γ. For lower neighborliness, the HLC-RSA model behaves very similarly to the baseline

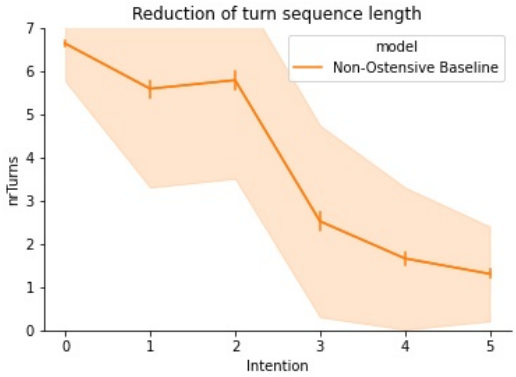


Figure 4: Non-ostensive baseline. The number of turns (y-axis) needed to resolve a misunderstanding decrease after each successive dialogue (x-axis).

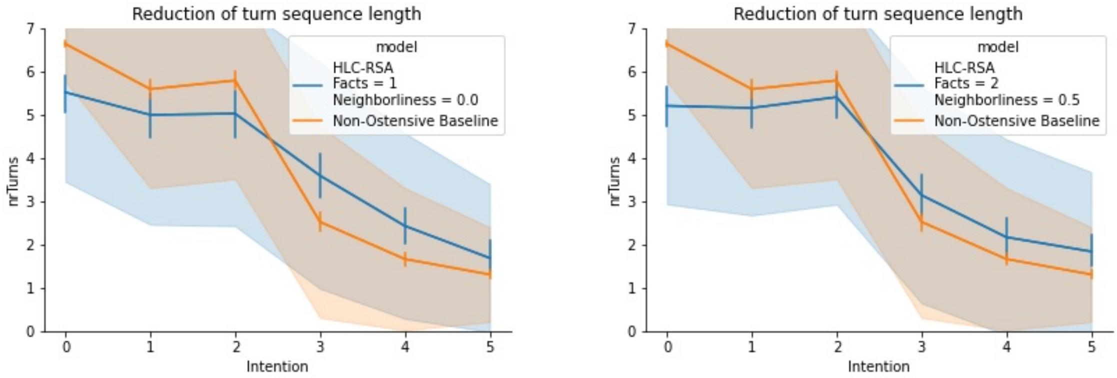


Figure 5: HLC-RSA behaves similar to the baseline for some parameter levels.

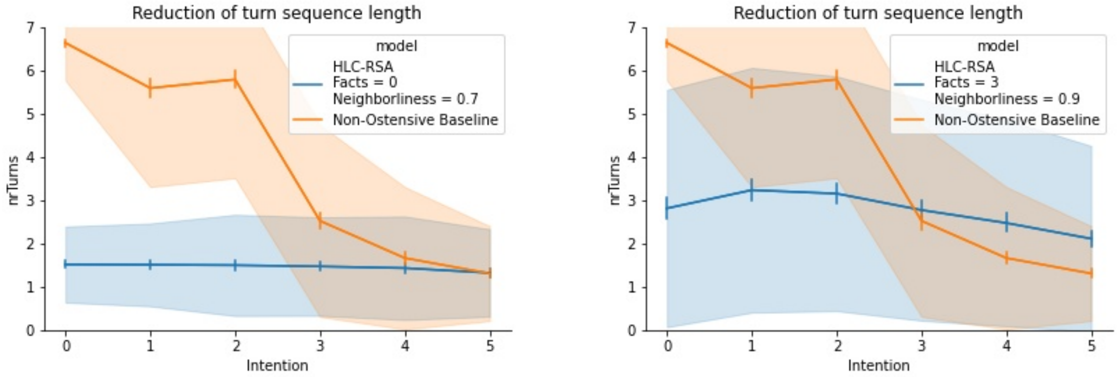


Figure 6: HLC-RSA starts low or decreases very little for high γ.

regardless of ν (figure 8). When γ increases, the agents do not need as many turns in their first dialogues to reach understanding. It is interesting to note here that agents reach a higher factual understanding than the baseline model. The variance in levels for the facts parameter introduces more factual misunderstanding for the agents, as seen by the increase of orange-colored bars when the facts increase.

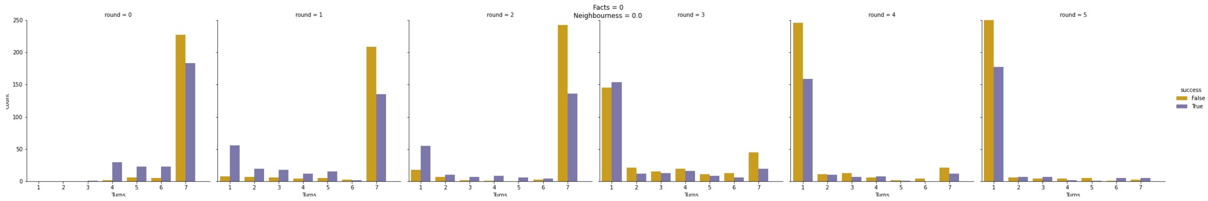
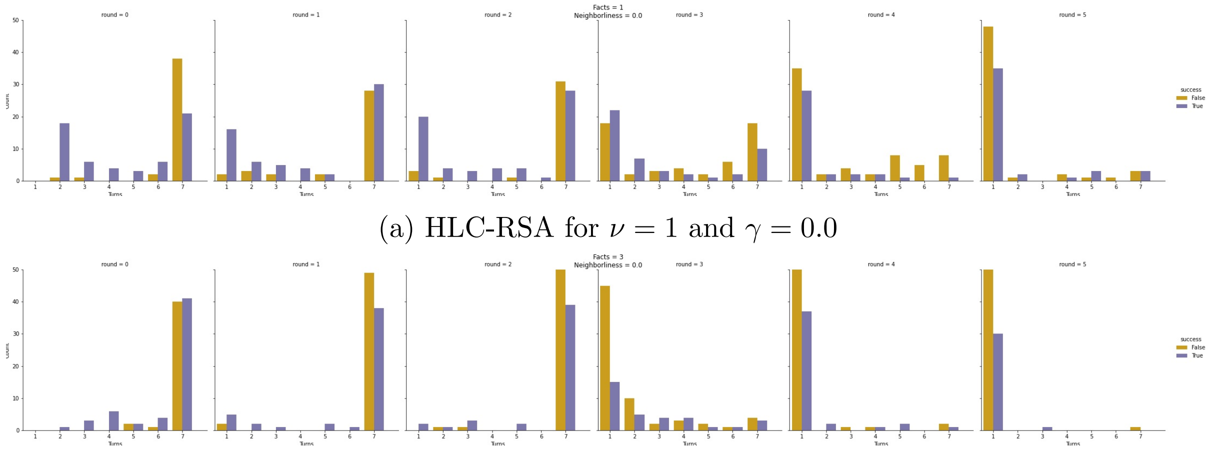


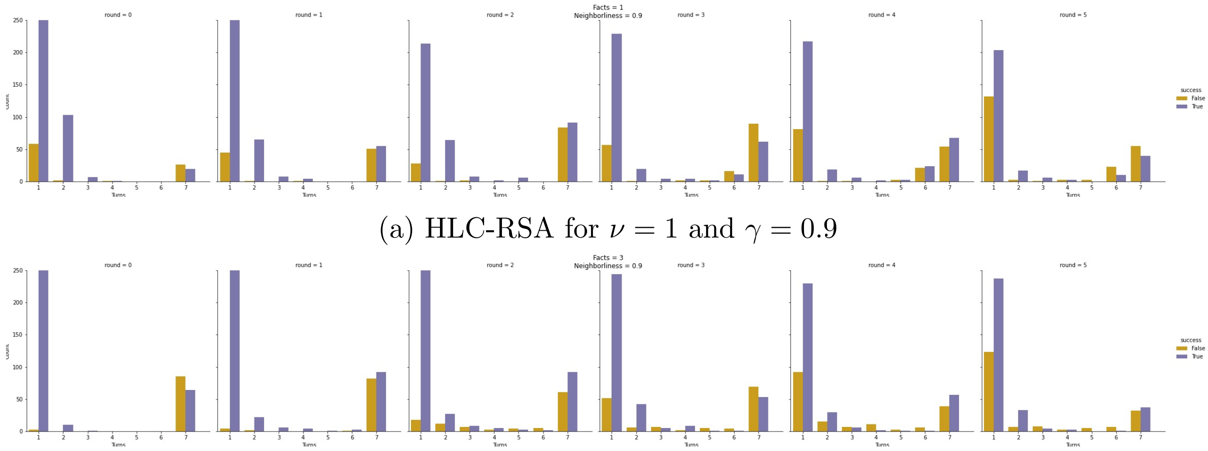
Figure 7: Non-ostensive baseline. Dialogue rounds progress from left to right, each panel is split on the number of turns the agent-pairs needed to reach understanding. That understanding is split further on factual understanding (purple bars) and misunderstanding (orange bars).



(a) HLC-RSA for $\nu = 1$ and $\gamma = 0.0$

(b) HLC-RSA for $\nu = 3$ and $\gamma = 0.0$

Figure 8: HLC-RSA performs similar to the baseline for low neighborliness.



(a) HLC-RSA for $\nu = 1$ and $\gamma = 0.9$

(b) HLC-RSA for $\nu = 3$ and $\gamma = 0.9$

Figure 9: HLC-RSA outperforms the baseline for high neighborliness.

4.3.3 Factual and perceived understanding

In the non-ostensive model agents reached perceived understanding in about half of the dialogues, and gave up in about 40% of the cases (table 1, see table 2 in appendix A for data sorted on facts first). The effects of the parameters can be seen here as well. As the neighborliness increases, so does the percentage of agents that have perceived factual understanding. The agents that give up and guess take an informative guess; they are above chance level. Increasing the number of facts dampens the effect of the neighborliness for the percentage of agents that reach perceived factual understanding. The agents that give up do not vary that much across both ν and γ. It remains a bit above chance level but varies within 10%.

Neighborliness	Facts	Perceived understanding	Give up
Non-Ostensive		50.23 %	39.66 %
0.0	1	57.48 %	42.47 %
0.0	2	65.57 %	40.91 %
0.0	3	41.69 %	45.35 %
0.5	0	66.81 %	39.86 %
0.5	1	72.21 %	40.99 %
0.5	2	54.57 %	46.93 %
0.5	3	47.47 %	43.57 %
0.7	0	91.04 %	46.88 %
0.7	1	69.96 %	41.96 %
0.7	2	68.55 %	50.14 %
0.7	3	56.22 %	45.35 %
0.9	0	99.57 %	0 %
0.9	1	78.78 %	48.28 %
0.9	2	67.49 %	51.3 %
0.9	3	81.14 %	51.77 %

Table 1: The cases where the agents reached factual understanding split over when they had perceived understanding or when they had given up. Sorted on neighborliness first.

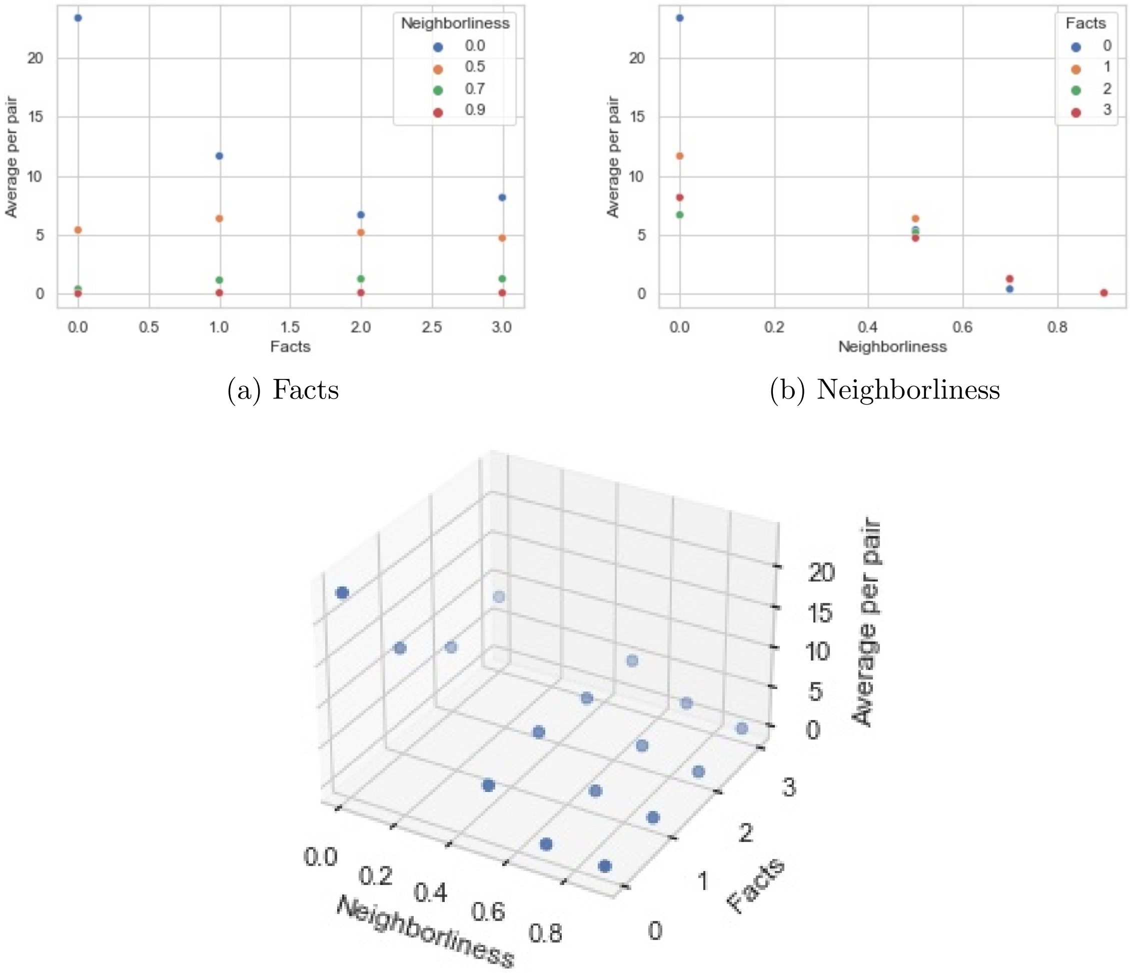
4.3.4 Time and space performance

Besides checking the performance of the HLC-RSA model when compared to the non-ostensive version by van de Braak et al. using the first three analyses, there is also the time and space performance of the model to consider. This last analysis, combined with the former three, will provide the insight needed to answer my research question. If the model accuracy for some levels of facts and neighborliness stays similar to that of the non-ostensive version, and the performance along the lines of time needed for computation and lexical search space to go through is better than the non-ostensive version, then an improvement in the tractable lexicon size follows. To this extent, I kept track of the average time in minutes that each agent pair needed to complete their dialogues and the lexical search space these agents needed to go through in their dialogues.

4.3.4.1 Time

The timing was done by running the model implementation for all levels of parameters on the same machine. This ensured a fair comparison. The results of the timing data can be seen in figure 10. Here, three scatterplots are shown. Scatterplot 10a shows the relation between the number of facts and the average time per pair, 10b shows the relation between the amount of neighborliness and the average time per pair. These two results are combined in 3D in 10c.

The effect of neighborliness alone (10a, left-most column) is a huge decrease in average time needed (98% for $\gamma = 0.0$ and 0.7). When introducing facts to this plot, the decrease in time is about constant. There is thus a limited effect of facts on time needed in computation. This effect is also visible in 10b, where the dots that represent the different levels of facts are close together. There is an effect of facts alone, as seen in the left-most column, but this effect is smaller than neighborliness.



(a) Facts

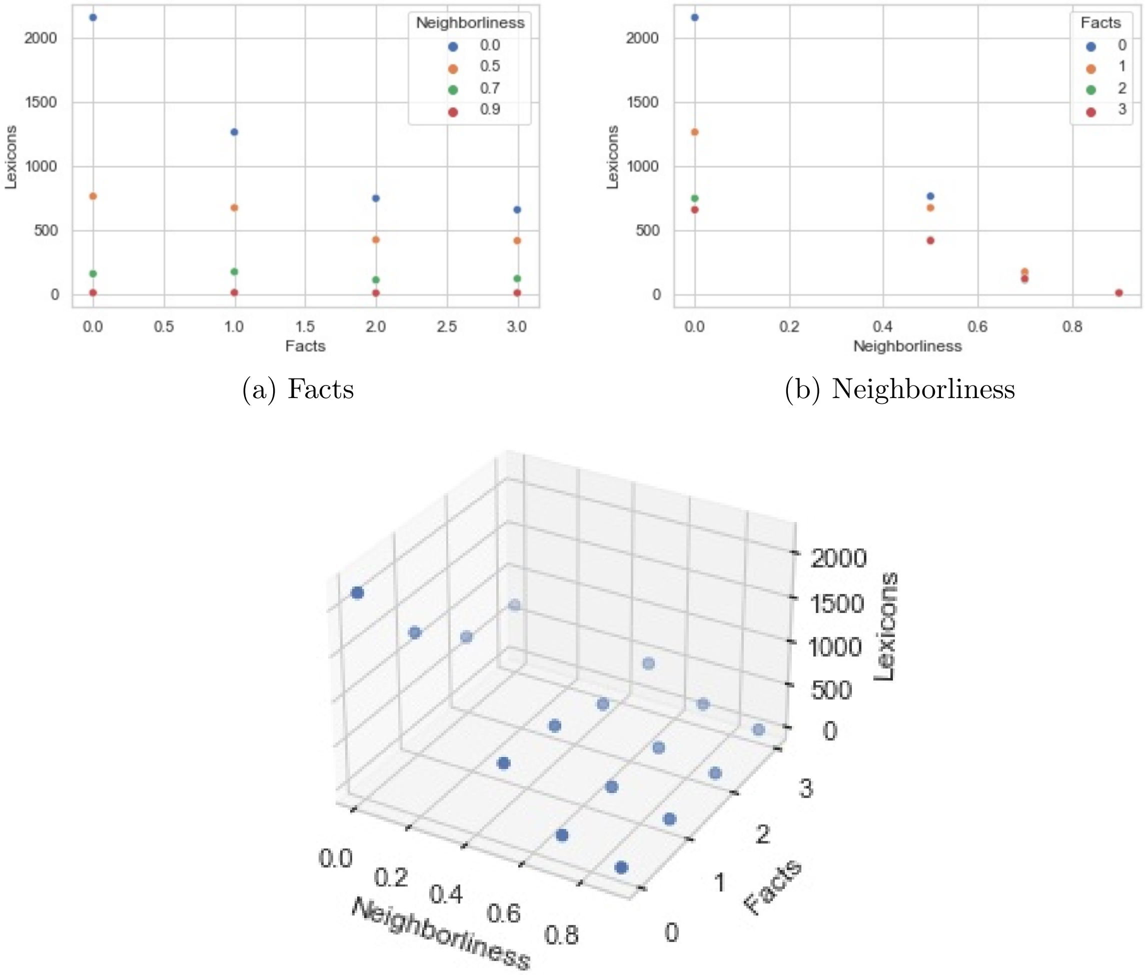
(b) Neighborliness

(c) Combined

Figure 10: Scatterplots of facts and neighborliness against average dialogue time in minutes.

4.3.4.2 Space

I compared the lexical search space by tracking the effects of the parameters on the number of allowed lexicons through the implementation of the constraints as specified in sections 3.3.2 through 3.3.6. These results are gathered in figure 11 and are organized in the same manner as the time analysis. Again, the effect of neighborliness is more present than that of facts. The search space decreases more when γ goes to 1, but there is an effect of ν as can be seen in the left-most column of plot 11b. The facts are not all in almost the same place as was the case in the time analysis, but their individual appearance can be seen.



(a) Facts

(b) Neighborliness

(c) Combined

Figure 11: Scatterplots of facts and neighborliness against size of lexical search space.

5 Discussion

5.1 Analyses

My results have shown that it is sensible to constrain the input size of possible lexicons. Summarizing the results of the performed analyses, the HLC-RSA model was very similar in behavior to the non-ostensive version for $\nu = 2$ facts and $\gamma = 0.5$ neighborliness for all tests. These parameter settings show that the factual and neighboring constraints do not necessarily impede the model's accuracy and positively improve both time and space performance. For these settings, the time reduction was 78%, and the search space reduction was 85%. These results prove hopeful for improving the feasibly computable lexicon size. It is noteworthy that even though these reductions seem prominent, the actual average time needed for the agents was 5 minutes, searching through 324 lexicons.

Interestingly enough, HLC-RSA even outperformed the baseline model for higher ($\gamma \geq 0.7$) neighborliness. Higher neighborliness resulted in shorter clarification sequence length, more eminent correct perceived understanding, time performance on a scale of seconds, and a search space of only double digits. Unfortunately, this positive effect was dampened for an increase in the number of facts.

It could be that the positive result of neighborliness is not a direct consequence of it. When the neighboring constraint has such a significant effect on the lexical search space, as seen from analysis 4b, it could be the case that the agents reason over so few lexicons that there is an a priori effect of this. Already resolving the misunderstanding in the first turn, as seen from analysis 2 and partly from analysis 1 and 3, could result from the underlying prior distribution over lexicons that the agents have. Agent pairs have a greater chance of random success because of the small distribution. Randomly sampling an initial lexicon from the prior is more likely to be the same lexicon for both agents in a pair. Because simulations of this model were made with equal binomial distributions for both initiator and responder, I can not provide any further insight into this matter.

Ultimately, these results show that the introduction of neighboring lexicons is an idea worth further scientific pursuit, but factual lexicons in their current form are not. I still believe that the idea behind the factual lexicons, namely the application of world or domain knowledge to the generation and search through the lexical search space, is worthy of further scientific pursuit.

5.2 Theoretical search space

Looking back at the definitions of facts and neighborliness (sections 3.3.4 and 3.3.5), these can not be considered to be a fixed-parameter tractability argument to the RSA problem. Although a correct reduction from the set of consistent lexicons to factual lexicons is lacking, the reduction from the possible set still is exponential in its characteristic. Facts reduce the degrees of freedom in this exponential formula ($2^{s \times r - \nu}$). The resulting number in the exponent could be reduced to smaller numbers as ν approaches the size of a lexicon, but as we have seen from the analyses, a bigger ν is not helpful for the model. The set of consistent lexicons is also exponential in its form, so even if an equation for the correct reduction could be found, this would still likely result in an exponential form.

When examining the formalization and resulting approaching equations that govern the neighborliness concept, it becomes clear that this set is likely not fixed-parameter tractable. The parameter γ determines the amount of k flips that are maximally allowed to generate neighboring lexicons by $\gamma \leq \frac{k}{s \times r}$. Then summing over all lexicons that can be generated using a maximum number of flips is a formula that is similar to $\sum_{i=1...k}^{n} \binom{n}{i}$. This equation needs to be rewritten into an inclusion-exclusion principle to account for the lexicons generated by flipping the identical mapping twice, resulting in an equal lexicon, and it needs to be incorporated into the equation for consistent lexicons. The latter has an exponential property, and the former is dependent on k, which can grow to the total size of a lexicon. However, following the results, HLC-RSA performs better for higher neighborliness, thus lower k. Arguing that the former might hold still leaves the latter exponential form. Conclusively stating that the neighborliness approach is not fixed-parameter tractable is outside of this Bachelor's thesis scope. I believe that it is not, looking at the equations given.

5.3 Conclusion

In conclusion, the results have shown that constraining the lexical prior does not necessarily impede model performance. Agents in the HLC-RSA model were able to resolve misunderstandings similarly to the non-ostensive model by van de Braak et al., but their time needed and the size of the search space was substantially smaller. An increase in the number of facts made for poorer resolving of misunderstanding, although there is an improvement in time and space performance in these specific agents.

Increasing the neighboring lexicon constraint reduced the time needed by a lot, 98% in one case. For low neighborliness, agent pairs performed similarly to the unconstrained version. For high neighborliness, the agents could even outperform the non-ostensive model both in resolving misunderstandings and in time and space performance. The positive effects of this parameter cannot be directly attributed to it, as it needs to be taken into account that the a priori reduction of the lexical prior results in more spontaneous success the agent-pairs have. The prior is based on a binomial distribution. When the prior is reduced substantially, agents have a higher chance of sampling the same lexicon, coming to an understanding in the first turn of a dialogue.

The formalizations introduced in this thesis cannot be proven to be a fixed-parameter tractable solution. Combining the interplay between the different equations and, as of writing, unproven but highly likely intractable nature of searching through the lexical space to prove or disprove fixed-parameter tractability of this problem is outside the limitations of a bachelor's thesis. The results as shown might incentivize readers to pursue this problem and either prove or disprove my notions.

Finally, to answer my research question. "*Can the extended RSA model be improved in terms of feasibly computable lexicon size by constraining the possible lexicons between agent pairs?*" The answer to this is: *yes*. The results, specifically those of analysis 4, indicate that HLC-RSA can increase the feasibly computable lexicon size without compromising accuracy in modeling resolving misunderstanding, as seen in analysis 1-3.

By being agents embedded in human language and the rules and regulations that govern it, it becomes clear that there is something more happening in our brains than just what the formal RSA model suggests. My thesis has shed some light on some of the possibilities that we could be using, and I hope that other researchers will delve deeper into this matter.

References

Bergen, L., Levy, R., and Goodman, N. (2016). Pragmatic Reasoning through Semantic Inference. *Semantics and Pragmatics*, 9.

Blokpoel, M. (2021). Personal Communication.

Blokpoel, M., Dingemanse, M., Woensdregt, M., Kachergis, G., Bögels, S., Toni, I., and van Rooij, I. (2019). Pragmatic Communicators can Overcome Asymmetry by Exploiting Ambiguity. *OSF Preprints*.

Cohn-Gordon, R., Goodman, N., and Potts, C. (2018). An Incremental Iterated Response Model of Pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 81–90. Linguistic Society of America.

Downey, R., Fellows, M., and Stege, U. (1999). Parameterized Complexity: A Framework for Systematically Confronting Computational Intractability. In *Contemporary trends in discrete mathematics: From DIMACS and DIMATIA to the future*, volume 49, pages 49–99.

Frank, M. (2016). Rational Speech Act Models of Pragmatic Reasoning in Reference Games. *PsyArXiv Preprints*, pages 1–62.

Frank, M. and Goodman, N. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability.* Freeman.

Goodman, N. and Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, 5(1):173–184.

Grice, H. P. (1975). Logic and Conversation. In *Speech Acts*, pages 41–58. Brill.

Hawkins, R., Frank, M., and Goodman, N. (2017). Convention-Formation in Iterated Reference Games. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 482–487.

Hawkins, R., Frank, M., and Goodman, N. (2020). Characterizing the Dynamics of Learning in Repeated Reference Games. *Cognitive Science*, 44(6):e12845.

Jovovic, V. (2008). Number of {0,1} n X m matrices with no zero rows or columns.

Khani, F., Goodman, N., and Liang, P. (2018). Planning, Inference and Pragmatics in Sequential Language Games. *Transactions of the Association for Computational Linguistics*, 6:543–555.

Kwisthout, J. (2013). Structure Approximation of Most Probable Explanations in Bayesian Networks. In *Proceedings of the 12th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 7958 LNAI, pages 340–351.

Kwisthout, J., Wareham, T., and van Rooij, I. (2011). Bayesion Intractability is not an Ailment that Approximation can Cure. *Cognitive Science*, 35(5):779–784.

Leech, G. (2016). *Principles of Pragmatics*. Routledge.

Levinson, S. C. (2006). On The Human Interaction Engine. In *Roots of Human Sociality*, chapter On The Hum, pages 39–69. Routledge, 1st edition.

Lieder, F. and Griffiths, T. (2019). Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources. *Behavioral and Brain Sciences*.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, New York.

Posner, M. and Snyder, C. (2004). Attention and Cognitive Control. *Cognitive Psychology: Key Readings*, pages 205–223.

Seidenberg, M. S. and Petitto, L. A. (1987). Communication, Symbolic Communication, and Language: Comment on Savage-Rumbaugh, McDonald, Sevcik, Hopkins, and Rupert (1986). *Journal of Experimental Psychology: General*, 116(3):279–287.

Sperber, D. and Wilson, D. (1986). *Relevance: Communication and Cognition*. Harvard University Press Cambridge.

Test Your Vocabulary Project (2013). TestYourVocab.com.

van de Braak, L., Dingemanse, M., Toni, I., van Rooij, I., and Blokpoel, M. (2021). Computational Challenges in Explaining Communication: How Deep the Rabbit Hole Goes. *PsyArXiv Preprints*.

van Rooij, I. and Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*, 16(1).

van Rooij, I., Blokpoel, M., Kwisthout, J., and Wareham, T. (2019). *Cognition and Intractability*. Cambridge University Press.

van Rooij, I., Evans, P., Muller, M., Gedge, J., and Wareham, T. (2008). Identifying Sources of Intractability in Cognitive Models: An Illustration Using Analogical Structure Mapping. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30, pages 915–920.

van Rooij, I. and Wareham, T. (2012). Intractability and Approximation of Optimization Theories of Cognition. *Journal of Mathematical Psychology*, 56(4):232–247.

van Rooij, I., Wright, C. D., Kwisthout, J., and Wareham, T. (2018). Rational Analysis, Intractability, and the Prospects of 'as if'-Explanations. *Synthese*, 195(2):491–510.

Weisstein, E. W. (2009). Vertex Cover.

Wilson, D. and Sperber, D. (2006). Relevance Theory. In Horn, L. and Ward, G., editors, *The Handbook of Pragmatics*, chapter 27, pages 606–632. Blackwell.

Woensdregt, M., Spike, M., de Haan, R., Wareham, T., van Rooij, I., and Blokpoel, M. (2021). Why is Scaling Up Models of Language Evolution Hard? *PsyArXiv Preprints*.

A Images

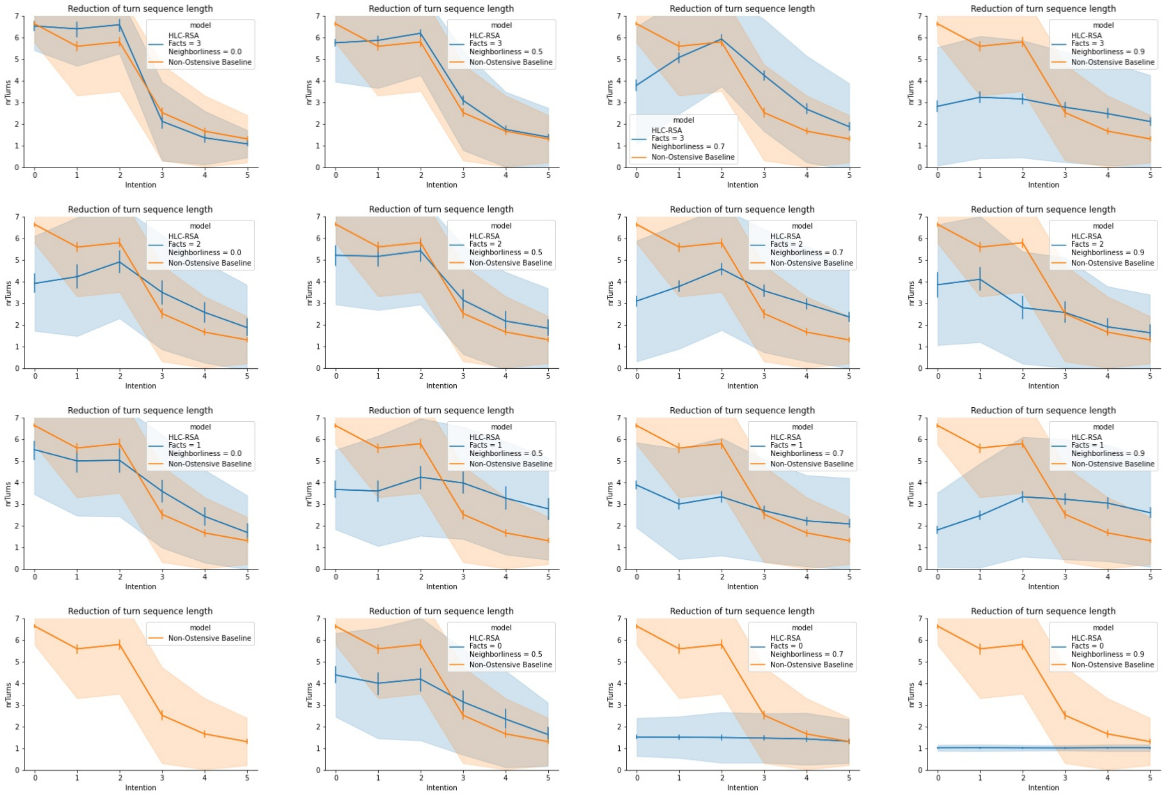


Figure 12: Analysis 1: Clarification length for different parameter values ν (y-axis 0, 1, 2, 3) and γ (x-axis 0.0, 0.5, 0.7, 0.9).

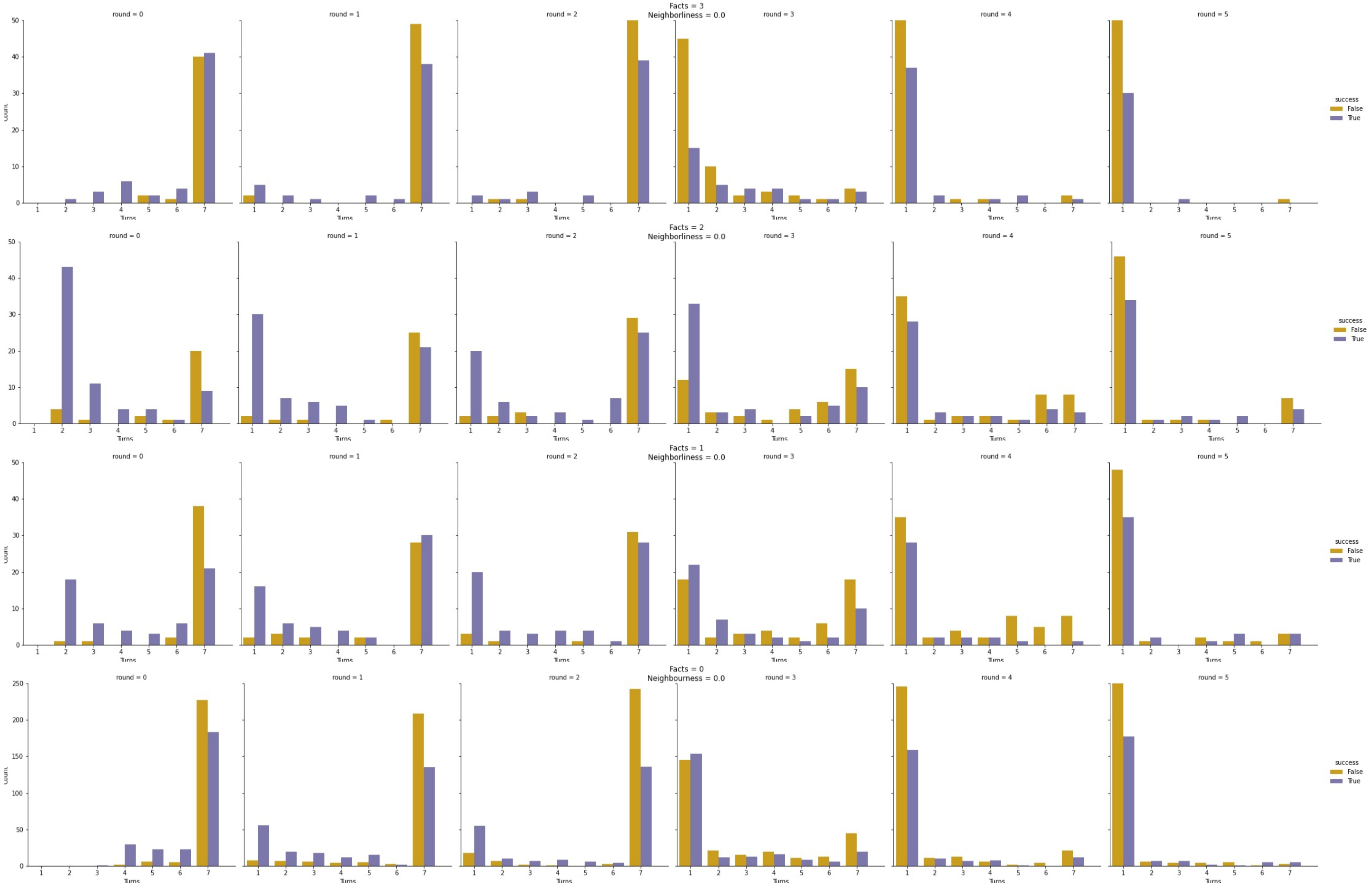


Figure 13: Analysis 2: $\gamma = 0.0$

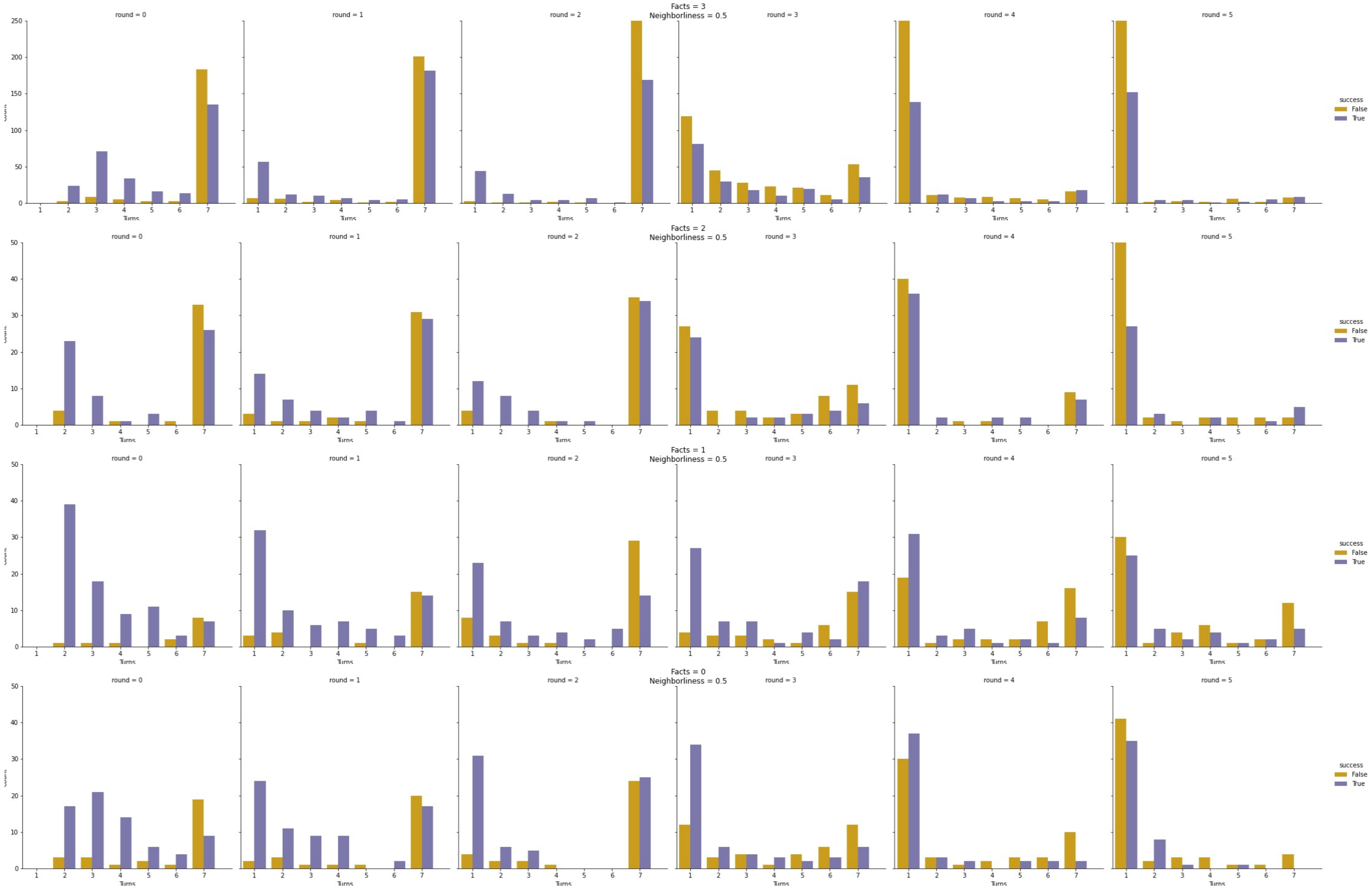


Figure 14: Analysis 2: $\gamma = 0.5$

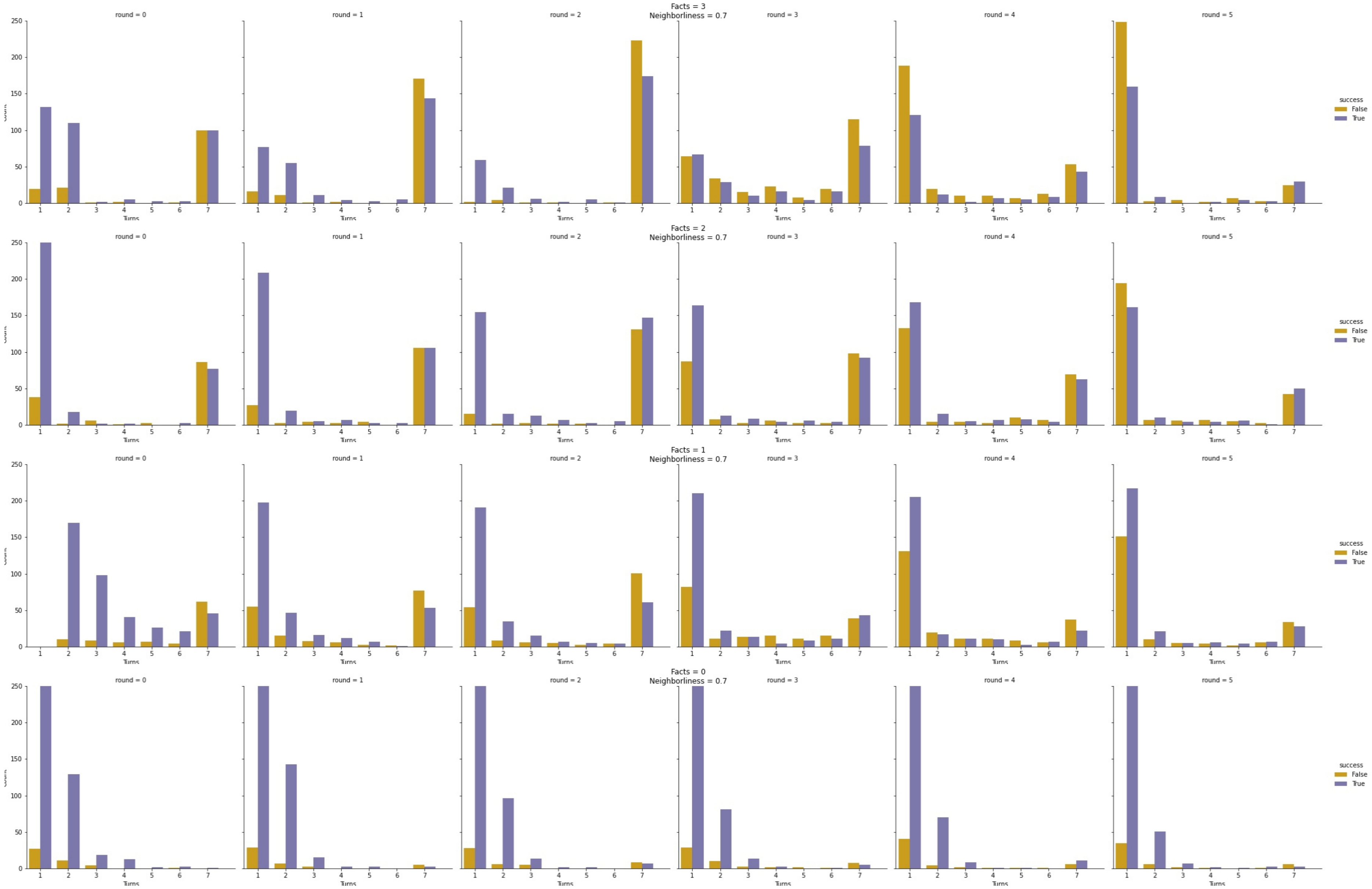


Figure 15: Analysis 2: $\gamma = 0.7$

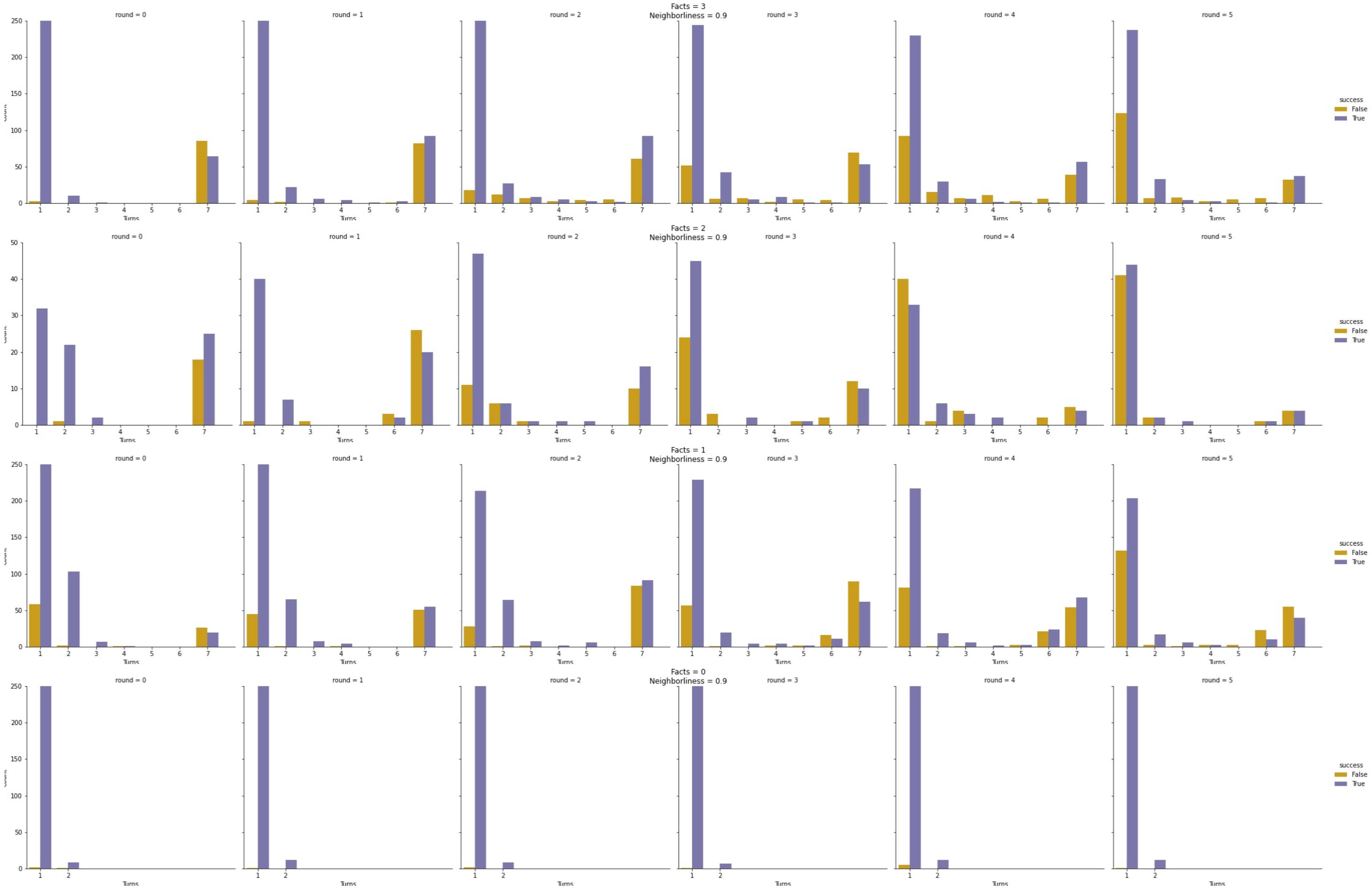


Figure 16: Analysis 2: $\gamma = 0.9$

Facts	Neighborliness	Perceived understanding	Give up
Non-Ostensive		50.23 %	39.66 %
0	0.5	66.81 %	39.86 %
0	0.7	91.04 %	46.88 %
0	0.9	99.57 %	0 %
1	0.0	57.48 %	42.47 %
1	0.5	72.21 %	40.99 %
1	0.7	69.96 %	41.96 %
1	0.9	78.78 %	48.28 %
2	0.0	65.57 %	40.91 %
2	0.5	54.57 %	46.93 %
2	0.7	68.55 %	50.14 %
2	0.9	67.49 %	51.3 %
3	0.0	41.69 %	45.35 %
3	0.5	47.47 %	43.57 %
3	0.7	56.22 %	45.35 %
3	0.9	81.14 %	51.77 %

Table 2: Analysis 3: The cases where the agents reached factual understanding split over when they had perceived understanding or when they had given up.

B Code

See the repository located at:
`https://github.com/jelio94/bsc-ai-thesis-jelle-hilbrands.git`
for the full resolution images from analysis 1 and 2, the Jupyter Notebook scripts used to generate them, the data from my simulations and the full model by van de Braak et al.